ABSTRACT

Title of dissertation: CHANGE DETECTION: THEORETICAL AND APPLIED APPROACHES FOR PROVIDING UPDATES RELATED TO A TOPIC OF INTEREST

Kristine M. Rogers, Doctor of Philosophy, 2024

Dissertation directed by: Professor Douglas Oard
University of Maryland, College of Information Studies

The type of user studied in this dissertation has built up expertise on a topic of interest to them, and regularly invests time to find updates on that topic. This research area—referred to within this dissertation as "change detection"—includes the user's process of identifying what has changed as well as internalizing the changes into their mental model. For these users who follow a specific topic over time, how might a system organize information to enable them to update their mental model quickly? Current information retrieval systems are largely not optimized for addressing the long-term change detection needs of users. This dissertation focuses on approaches for enhancing the change detection process, including for short documents (e.g., social media) as well as longer documents (e.g., news articles).

This mixed methods exploration of change detection consists of four sections. First, this dissertation introduces a new theory: the Group-Pile-Arrange (GPA) Change Detection Theory. This theory is about organizing documents relevant to a topic of interest in order to accelerate an individual's ability to identify changes and update their mental model. The three components of this theory include: 1. Group the documents by theme; 2. Pile the

grouped documents into an order; and 3. Arrange the piles in a meaningful way for the user. These steps could be applied in a range of ways, including using approaches driven by people (e.g., a research librarian providing information), computers (e.g., an information retrieval system), or a hybrid of the two.

The second section of this dissertation includes the results of a survey on users' sort order preferences in social media. For this study, change detection was compared with three other use cases: following an event while it happens (experiential), running a search within social media, and browsing social media posts. Respondents recognized the change detection use case, with 66% of the respondents indicating that they perform change detection tasks on social media sites. When engaged in change detection tasks, these respondents showed a strong preference for posts to be clustered and presented in reverse chronological order, in alignment with the "group" and "pile" components of the GPA Change Detection Theory. These organization preferences were distinct from the other studied use cases.

To further understand users' goals and preferences related to change detection, the third section of this dissertation includes the design and prototype implementation of a change detection system called Daybreak. The Daybreak system presents news articles relevant to a user's topic of interest and allows them to tag articles and apply tag labels. Based on these tags and tag labels, the system retrieves new results, groups them into subtopic clusters based on the user's tags, enables generation of chronological or relevance-based piles of documents, and arranges the piles by subtopic importance; for this study, rarity was used as a proxy for subtopic importance. The Daybreak system was used for a qualitative user study, using the framework method for analyzing and interpreting results. In this study, fifteen participants engaged in a change detection scenario across five simulated "days." The participants heavily leveraged the Daybreak system's clustering function when viewing results; there was a weak preference for chronological sorting of documents, compared to relevance ranking. The participants did not view rarity as an effective proxy

for subtopic importance; instead, they preferred approaches that enabled them to indicate which subtopics were of greatest interest, such as pinning certain subtopics.

The fourth and final component of this dissertation research describes an evaluation approach for comparing arrangements of subtopic clusters (piles). This evaluation approach uses Spearman's rank correlation coefficient to compare a user's ideal subtopic ordering with a variety of system-generated orderings. This includes a sample evaluation using data from the Daybreak user study to demonstrate how a formal evaluation would work.

Based on the results of these four dissertation research components, it appears that the GPA Change Detection Theory provides a useful framework for organizing information for individuals engaged in change detection tasks. This research provides insights into users' change detection needs and behaviors that could be helpful for building or extending systems attempting to address this use case.

CHANGE DETECTION: THEORETICAL AND APPLIED APPROACHES FOR PROVIDING UPDATES RELATED TO A TOPIC OF INTEREST

by

Kristine M. Rogers

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Professor Douglas Oard, Chair/Advisor
Professor David Grossman
Professor Bill Kules
Professor Philip Resnik
Professor Beth St. Jean

Dedication

"Life can only be understood backwards, but it must be lived forwards." - Søren Kierkegaard

Acknowledgments

Looking back at my educational journey, my studies at the University of Maryland may have been foreshadowed. Years ago, as an elementary school student in Kaneohe, Hawaii, my class was assigned homework where we had to learn about a state and draw their flag. I pored through my family's encyclopedia set (this was well before Google), finally settling on the state with the most beautiful flag I had ever seen: Maryland!

An early inspiration for delving into the world of information happened in the late 1990s. I was listening to NPR during a commute, when they aired a feature about David Shenk's new (at the time) book, *Data Smog* [243]. I found it so intriguing that I emailed him, and sent me a complimentary copy of his book. I continued thinking about ways to organize information to understand what is happening in the world. Eventually I made my way to change detection, the topic of this dissertation. I never thought I would gain such an interest in chronological ordering along the way!

There are too many people to thank for their direct and indirect roles leading to my dissertation completion, but I will attempt to thank as many as I can. I could not have completed this research without the support of my organization, which provided time and funding for pieces of this journey. I would also like to thank the iSchool for its support of my research, including funding key parts of my research and supporting my studies. I am grateful for my research participants (including the ones who tried to sneak in without meeting the study criteria). I learned much through the studies, and appreciate people's willingness to spend time participating in my fairly complicated studies. I am thankful for the Graduate School Writing Center, and for its director, Dr. Linda Macri; one of the dissertation retreats she organized proved critical for development of the "Pile" component of the GPA Change Detection Theory. I am also grateful for Michael Cole of Lexis-Nexis, who aided in identifying options for leveraging news articles for the Daybreak study.

I would like to thank my advisor, Dr. Doug Oard, for his patience and wisdom throughout the process. He helped me think through my research topic at a deep level, which led to uncovering intriguing connections and use cases. I would also like to thank my dissertation committee for their interest in and support of my research, and for their feedback along the way. In particular, I would like to thank Dr. Beth St. Jean for her detailed feedback, Dr. Philip Resnik and Dr. Bill Kules for their insightful comments, and Dr. David Grossman for his ideas and enthusiasm related to potential applications of this research.

On the personal side, I would like to express my thanks for my communities focused on birds, technology, music, and baseball. These groups and the friends I've made through them have helped to keep my spirits up through the dissertation research and writing process. I also send my thanks to my friends (especially Andi, Paul, Jean, and Annette) who attempted to help me remember to have fun, even in the final stages of the dissertation; I look forward to being more adventuresome after I finish this research. I would also like to thank my friends who provided thoughts and input related to my dissertation content—in particular, thank you to Steve for always being willing to serve as a sounding board for my ideas, and to Marianna for commiserating along the way.

And finally, I would like to thank my family for their support of my dissertation research, starting with my little feathered dragons—especially Drake, Kyrie, and Archie, who were a bright spot throughout the process. While my whole family has provided encouragement for my research, I want to thank my parents for both serving as examples, and discussing their own dissertation research with me [216, 219]. I also am grateful for my mom for taking time to review some of my writing along the way. Finally, I am extremely grateful to my sister, Emily, for being willing to assist me in conducting the Daybreak study; I appreciate that she took the time to aid in every session, no matter the outcome.

I wish I could personally thank all who aided me along the way, and sincerely appreciate all of the support I received during my dissertation research and writing process.

Table of Contents

List of Figures

List of Tables

Chapter 1: Overview of Dissertation Research

For people who follow a topic over time, there is a lot of friction in the current processes they use to get updates. It can be a bit like a treasure hunt, picking through various sources of information, trying to find something that they care about. They might disentangle stories and information provided by a wide range of sources, pulling out pieces that cover what matters to them; in some cases, the pieces that matter might be only a small fraction of the information conveyed by the source. This dissertation focuses on approaches for organizing information to enable users to detect what has changed related to their topic—which we refer to as "change detection"—to enable them to get updates to their mental model more quickly. How might a system organize information in a meaningful way that enables the user to pull in the pieces they need to understand what is happening? Throughout this dissertation, we focus on change detection as an individual process for obtaining new information on their topic of interest.

The idea of change detection can be thought of in the context of the 1993 film *Groundhog Day*, in which a man relives the same day over and over again. Whereas other search approaches (e.g., ad hoc search) focus on running a different query every day, change detection focuses on running the same query over and over again. We define change detection as the processes through which a user identifies and characterizes changes related to a topic that they follow over time.

For this use case, we assume that individuals are not searching for absolute truth, nor do they seek information on all possible topics in the universe. Rather, they want a detailed

ongoing understanding of what is happening on a topic they care about. Danish philosopher Søren Kierkegaard stated that "truth is subjectivity" and "subjectivity is truth" [143]. Things are meaningful in the context of the experience of the individual; we cannot easily separate our understanding from our personal experience. Because of that, this dissertation addresses change detection (getting updates on a topic over time) as a task focused on subjective truth. We look at this as an individual effort—an individual person has a mental model, which is updated through the process of reviewing information from a variety of sources. These individuals have specific interests, generally for personal or professional reasons; they intend to learn, and are willing to spend time going through information—these are not users who are simply dabbling in a topic or looking for a summary or overview. Their learning can come from a variety of sources; in this dissertation, we primarily focus on learning from social media and news articles.

Over time, new information becomes available and stories change. We refer to this as change detection because it can take effort to detect and understand what has changed related to the topic. Organizing information in some sort of flow—orderings that are meaningful to the user—can aid in identifying changes or interesting information that may otherwise have been missed. For example, an individual might review information about their favorite baseball team from a variety of sources. Thinking of the user as a bundle of interests, what they want to see about baseball may differ from someone else's interests. Different information sources might only cover certain pieces of what they want to know. The user takes what is of interest from each source, then reassembles this information into something that is meaningful to them (mental model updates).

How do users perform change detection tasks today? While systems exist to answer specific ad hoc questions, or provide general overviews of events that are happening in the world, search systems tend not to be optimized for the change detection use case. It would be beneficial to have a system that focuses on understanding and meeting the specific

standing information needs of users who have followed a topic over time. When users have to force fit their use case into an interface that was not designed for that need, they may end up frustrated, spending extra time trying to meet the need—if they are able to find their desired updates at all. Future research could explore applications of this idea beyond the two studied data types, to include multimedia files.

The goal of this dissertation is to build a framework for understanding users' change detection needs, and how information can be organized to aid in updating their mental model quickly. For this purpose, we have devised a theory that we call the Group-Pile-Arrange (GPA) Change Detection Theory. It focuses on three segments of the process for organizing information: grouping documents by theme, piling them into some order within those groups, and then arranging the piles in order of interest to the user. This research informs the creation of systems that can help the user find updates on a topic that they follow over time, to reduce the amount of friction they face in the process. Additionally, we design an evaluation approach to measure the effectiveness of one aspect of change detection systems.

1.1 Components of Dissertation

This dissertation includes research on five specific areas related to change detection. To understand the change detection process and improve systems' ability to meet the needs of users, we focused on the following:

1. **Theory:** This area consists of a theoretical framework for designing change detection systems, using three components: group, pile, and arrange. This theory—the GPA Change Detection Theory, is meant to be generalizable to include a variety of types of systems that could support change detection tasks.

2. **Prevalence:** To understand the prevalence of the change detection task, we con-

ducted a survey about sort order preferences in social media.[1] This survey gave us insights into the prevalence of change detection among survey respondents, and how it compares to various other use cases in social media: experiential uses (e.g., following a live event), browsing, and ad hoc searches in social media; these were used to compare sort preferences.

3. **Understanding user preferences:** In addition to aiding us in understanding the prevalence of change detection tasks in social media, the sort order survey also gathered information on what organizational approaches the respondents prefer for their social media feeds—specifically related to clustering (group) and sort orders (pile) for change detection as well as the other studied use cases.

4. **Testing user preferences:** We designed and implemented a user study for Daybreak, a prototype change detection system, to determine the extent to which the participants were able to complete change detection tasks successfully using their preferred approaches for organizing documents and subtopics. We then conducted the Daybreak user study to test the preferences expressed in the sort order survey and gain a deeper understanding of approaches that participants applied when they need to convey information they have learned from the topic they follow.

5. **Automated evaluation approach:** We designed an automated evaluation approach for change detection tasks, focused on comparing system orderings for subtopic clusters. We demonstrated how it might be used in a system evaluation, and ran a sample evaluation on the Daybreak study results. The primary focus of this evaluation design is on arrangements of piles of documents, highlighting cases where the system prominently displays piles of greater importance to a user.

[1] Social media includes sites such as Facebook, Instagram, and Twitter (renamed to "X" in July 2023). In this dissertation, we refer to this social media site as Twitter, which was the name of the system at the time the data was collected.

We now introduce these research areas in more depth, and indicate where in the dissertation we cover each concept.

1.1.1 Chapter 3: The GPA Theory of Change Detection

Chapter 3 introduces a theory to characterize the change detection needs of a user. The user provides a topic of interest and an externalization of their model. The system—which could be implemented through people, a computer system, or a hybrid of the two—follows a three step process for organizing results about the topic. First, the system groups the documents into subtopic clusters. Next, these groups are organized into piles by organizing documents into some order. Finally, the piles are arranged to present piles of greatest interest to the user first. The Group-Pile-Arrange (GPA) Change Detection Theory serves as the foundation for the remainder of the dissertation research.

While the theory is presented early in the dissertation, chronologically, it was formalized later in the research process—specificially, during the analysis of the Daybreak user study results in Chapter 6. During the analysis process, it became clear that it would be beneficial to separate the design story into the theoretical motivation for the design (Chapter 3) and the actual Daybreak study design in Chapter 5. We determined that this would be clearer than presenting the theory and design together, in part because elements of the same theoretical motivation helped to shape the study in Chapter 4.

1.1.2 Chapter 4: Understanding Sort Order Preferences in Social Media

To start down the path of understanding user needs and change detection, Chapter 4 focuses on understanding the prevalence of the change detection use case relative to other use cases. This section focuses on understanding users' preferences for organizing social media posts in relation to change detection tasks, which we generally refer to as sort order

preferences. Intuitively, it is clear that people follow specific topics over time; this research helps to determine how prevalent the change detection need is. Is this simply an edge case, or an activity that users perform?

Chapter 4 contains the findings from a survey of social media users to understand their sort order preferences. The survey asks questions about four use cases that users perform in social media, to contrast users' information needs between change detection and other use cases. Two of these use cases are update tasks; in other words, these two focus on information retrieval activities designed to understand a topic over time. The change detection task addresses the information needs of people who follow a topic over a long period of time. In contrast, the experiential tasks relate to the information needs of respondents who are following a specific event with social media often serving as a "second screen"—for instance, a conference, speech, or sporting event. For further contrast, we compared social media usage for browsing (reviewing posts without a specific goal) and for ad hoc searches.

The survey on social media sort order preferences was offered online to US-based Internet users aged 18 or older. The survey was administered through the Qualtrics system. In particular, the survey covered the group and pile components of the GPA Change Detection Theory. Insights from the sort order survey informed the clustering and sort order decisions used in the Daybreak system design and user study that are covered in Chapters 5 and 6, and also feeds into the evaluation design discussed in Chapter 7.

1.1.3 Chapter 5: Design of Daybreak System and User Study

Based on the GPA change detection theory and the sort order survey results, we created Daybreak, a prototype computer system designed to meet users' change detection needs. This chapter includes the design and implementation of the Daybreak system, as well as the design of an online user study to test the system in real-world use cases.

The prototype system enabled participants to leverage core GPA capabilities: within the Daybreak system, participants could view documents grouped in personalized subtopic clusters, and had the ability to select between relevance and chronological sort orders. The Daybreak system organizes retrieved documents into subtopic clusters, which are arranged by importance. For this study, subtopic importance was defined as rarity (least prevalent subtopic first), based on the idea that less prevalent information could be valuable in this type of scenario. Documents that are about the overall topic but not tied to any of the participants' subtopic categories were displayed in an uncategorized section at the bottom of the subtopic list.

The scenario for the participant was that they were a blogger who typically follows a specific topic of interest. One week, one of their colleagues—who follows a different topic, but a similar one—is going to be out of the office for a work week. They have asked the participant to fill in for them for that full week. During that time, they follow their colleague's topic in search of updates. At the end of the five-day simulation, the participant wrote an outline for a story about key developments related to the topic. The user study was designed to last approximately 90 minutes.

The study participants leveraged the Daybreak prototype in a simulated five-day change detection session covering one of five topics: Red Sox baseball, cryptocurrency, global health, space, or extreme weather. The user could choose their sort and clustering options, then read documents, and apply tags and tag labels. After each day in the scenario, the participants were asked to fill out a questionnaire on the system and their experience. After Day 5, the participants completed a storytelling task in which they outlined key events over the course of the five days. At the end of the session, we held a semi-structured interview with each participant.

This study assumed that users encounter a high volume of news articles, and does not have sufficient time to review each one. The study included a higher volume of document

results than we would expect an individual user to reasonably be able to read through in one sitting. This mimics real-world scenarios, in which a user might state a preference for seeing all posts in chronological order, but this might be infeasible in a high-volume stream of news articles. Additionally, the time limit per day enabled us to keep the total session length bounded.

1.1.4 Chapter 6: Results from the Daybreak User Study

Based on the selection process, we chose five topics for the Daybreak user study: Red Sox baseball, extreme weather, global health, space, and cryptocurrency. We ran three sessions for each topic, each with different participants, resulting in 15 complete sessions. This user study aided in understanding in a more specific way participants' preferences when they are performing a change detection task. This included understanding how participants preferred to apply clustering or sort orders, depending on their processes. The study revealed that participants heavily made use of subtopic clustering based on their tag labels, with only one user deciding not to use clustering due to their preference for seeing all documents in one list. The participants' sort order preference varied more, with a weak preference for reverse chronological sort overall. The participants supported the idea of ordering subtopics to include more important or interesting clusters first; however, they did not find rarity to be a useful proxy for subtopic importance. Insights from the study fed into the evaluation design discussed in Chapter 7.

1.1.5 Chapter 7: Evaluation Design for Change Detection Systems

The final substantive chapter of this dissertation focuses on an automated evaluation of one aspect of change detection systems using Spearman's rank correlation coefficient. This evaluation focuses primarily on comparing approaches for arranging subtopics. We

devised a correlation-based evaluation approach to compare participants' ideal subtopic ordering for a simulated day with the system-generated arrangement of subtopics. As an exemplar, we looked at two sets of ideal rankings: a sample session from the author, and the results from the user study. We ordered the subtopic clusters in a variety of ways, and applied our evaluation approach to determine which orderings were more closely aligned with the user's preferences. These sessions demonstrated the approach for an evaluation that could be implemented as a future study.

1.2 Methodology

Our interest in understanding change detection led us to ask a range of questions about change detection. Given the variety of our questions and the types of information needed to answer each one, this dissertation uses a mixed methods approach to understand users and their change detection needs. Here we detail the research methods and research questions used for the studies that comprise this dissertation.

1.2.1 Research Methods

We framed the research in this dissertation into a variety of studies to understand change detection concepts, and to view different aspects of users' needs and applications. Each study was designed to build on prior studies, using a different research method. We used the following methods within each chapter:

- **Chapter 3 - Theory:** We devised a theory to serve as the foundation for the change detection research.
- **Chapter 4 - Survey:** We applied survey methods to conduct a survey and interpret survey results related to user preferences in social media.

- **Chapters 5-6 - Mixed methods:** For the Daybreak user study, we included a selection survey, questionnaires, and a user study of a prototype system. To collect information and interpret results for the selection survey and questionnaires we used survey methods, including analyzing responses from a qualitative and a quantitative perspective. For the user study itself, we used the framework method—a qualitative approach for interpreting collected data from the context of a framework; in this case, we use the structure of the GPA Change Detection Theory to frame the research.
- **Chapter 7 - Evaluation design:** We designed an automated evaluation approach. This enables comparison of system ordering options without requiring manual user evaluation of multiple result sets.

1.2.2 Research Questions

To frame this research, we set up a series of research questions (RQs) focused on aspects of change detection. These research questions are introduced in Sections 4.3.2 and 5.1.4, and answered in Sections 4.4 and 6.3. The complete set of research questions for this dissertation are as follows:

Chapter 4:

- **RQ4.1:** How prevalent is change detection? (Section 4.3.2)
- **RQ4.2:** Would users accept clustering as an approach for organizing posts? (Section 4.3.2)
- **RQ4.3:** How do users prefer to have results sorted for a change detection? (Section 4.3.2)
- **RQ4.4:** How many posts do respondents feel they need to see when performing a change detection task? (Section 4.3.2)

Chapters 5 and 6:

- **RQ5.1:** Does tagging and tag label generation aid users in representing their mental model of a topic? (Section 5.1.4)

Questions related to "Group" Concept

- **RQ5.2:** Does organizing search results by subtopic clusters aid users in performing change detection tasks? (Section 5.1.4)
- **RQ5.2a:** What information retrieval approaches would be effective for transforming a user's tags into clusters of relevant documents? (Section 5.1.4)

Question related to "Pile" Concept

- **RQ5.3:** Does organizing search results within subtopic clusters in some sort order aid users in performing change detection tasks? (Section 5.1.4)

Questions related to "Arrange" Concept

- **RQ5.4:** Does arranging subtopic clusters in some order aid users in performing change detection tasks? (Section 5.1.4)
- **RQ5.4a:** How should the system handle documents that do not fit in any existing subtopic cluster? (Section 5.1.4)

Question related to GPA Change Detection Theory

- **RQ5.5:** Does the system help users develop and externalize mental models? (Section 5.1.4)

1.3 Contributions

This dissertation research is intended to address gaps observed in academic literature related to the practice of change detection. It also provides an approach for evaluating systems designed for change detection. We introduce the contribution areas here; the details and findings are included in Chapter 8. Specific areas of impact include:

- **Connections:** Cross-disciplinary research focusing on change detection as a specific use case, contrasting users' needs and preferences with other search use cases.
  - We studied change detection as an end-to-end arc within a single research program. Where other studies might focus on intellectual contributions related to one aspect of the research, this design includes a cross-disciplinary range of research: starting with a theory, performing studies both in social media and news articles, and into the evaluation design.
  - We looked at change detection across two content types: social media and news articles. This provides a starting point for understanding commonalities and differences in change detection needs across sources.
  - We connected concepts across a wide range of fields, including information retrieval, human-computer interaction, information studies, cognitive psychology, behavioral economics, and more. We identified and connected related concepts, theories, and ideas that intersected with change detection.
- **Theoretical:** We developed a change detection theory to provide a foundation for our study; this addressed a gap in research, and introduced a new method for thinking about the ways that users seek and receive updates on topics of interest.
  - We devised a theory to describe factors and considerations related to change

detection, in a way that could be applied to a wide range of scenarios—whether supported by people, computers, or some combination.

  - We used examples to illustrate possible applications of the GPA change detection theory.

- **Survey:** Our sort order survey provided details about change detection, how it is applied in social media, and users' organization preferences for posts. We used the survey to compare and contrast change detection with other use cases.

  - To aid in understanding the prevalence of change detection relative to other social media use cases, the sort order survey helped us understand how familiar this use case was to survey respondents.
  - The survey identified respondents' preferences for how they would like to see social media posts organized when they are performing change detection tasks.
  - Our research provided insights into differences between social media use cases, and how different user tasks can be identified and addressed in social media.

- **User Study with a Prototype System:** We designed and built a system that supported a user study for simulated change detection task on a variety of topics.

  - We designed and implemented Daybreak, a prototype system that implemented change detection functionality based on the GPA Change Detection Theory and input from survey participants who follow a topic over time.
  - We devised and conducted a user study that leveraged the Daybreak system in a simulated five-day session where a user followed changes in news related to a topic, then drafted an outline summarizing their learning.

- **Evaluation Design:** We applied a correlation measure in an example evaluation for the change detection use case.
  - We designed an automated evaluation process that leverages Spearman's correlation coefficient to compare system arrangements for subtopic clusters.
  - Using a sample session by the author as well as data from the Daybreak user study, we performed a sample evaluation using the automated evaluation approach.
- **Code:** For this research, we created a number of artifacts that can serve as exemplars for other researchers' research related to change detection.
  - We developed the prototype Daybreak system. This includes the HTML and Javascript-based user interface, Python-based back end, and implementation of indexing with the Indri search engine.
  - We generated Python code that operationalizes the subtopic cluster ordering comparisons between various systems.

1.4 Summary

This dissertation addresses the topic of change detection: organizing information so that a person who follows a topic over time can receive updates to their mental model quickly. The research discussed here includes a theory describing change detection, a study of sort order preferences for change detection and other use cases in social media, a system (Daybreak) for users focused on change detection, a user study leveraging the Daybreak system, and an evaluation design focused on ordering of subtopics.

Next we focus on the foundational concepts behind the change detection concept and the components of this dissertation. In Chapter 2 we start our with a review of literature

addressing concepts relating to users' interests, expertise-building, and information seeking behaviors. We then address the technical foundations for this research, drawing from literature from information retrieval, human-computer interaction, and related fields.

Chapter 2: Literature Review

How does a user move from large amounts of information to the specific subsets that they care about? Here we focus on the foundations for change detection, showing the world in which it takes place and the technologies that can support it. We define change detection as a process through which users identify and characterize what has changed in relation to a topic of interest to them. For this use case, the user is someone who needs or intends to spend time with the relevant information; this it is an information task related to expertise development. We note that this is not a prediction task; rather, the objective is to review information on recent events to understand the current situation. To understand the task of change detection in more detail, we step through the process through which users understand and follow topics, ultimately ending in gaining of expertise. We begin by looking at the learning process of the searcher, leading into their development of mental models of a topic, and ultimately to the phase where they become an expert on the topic. We include a discussion of information overload, biases, and other issues the user encounters along the path to expertise. From there we turn to the technology side, to identify capabilities and ideas from information retrieval that can be combined as we attempt to generate a more efficient and effective capability that can help users address their change detection needs.

This review of relevant research covers topic areas that set the stage for the broader dissertation. We draw these from literature on information studies, cognitive psychology, reading comprehension, computer science, information retrieval, human-computer interaction, and more, to build a foundation for the concept of change detection.

2.1 Foundations of Knowledge

Ackoff provided an often-used description of the differences between data, information, knowledge, and wisdom [1]. In his description, *data* represents properties of objects and events. Data that is processed becomes *information*—having been made more useful with descriptions that answer questions such as "who, what, when, where, and how many." *Knowledge* is information that can answer questions of "why." *Understanding* is about answers to the questions—especially the "why" questions. Finally, Ackoff defines *wisdom* as exercising judgment based on what is known.

A number of researchers have attempted to clarify terms such as information and knowledge and how they can be used by people as well as computer systems. Buckland looked at the ambiguity related to definitions of information [42]. To disambiguate some of the ways people use the term "information," he broke the concept of information into three different areas: 1. *Information-as-process*, which is about communicating facts or changes; 2. *information-as-knowledge*, in which information reduces uncertainty about something; and 3. *information-as-thing*, a term used to represent documents and other data. Minsky created a unified theory of information centered around the concept of a frame—a data structure that represents some state or situation [183]. His idea for representing knowledge in a structured way enabled descriptions that could be used in areas such as artificial intelligence.

Central to this dissertation is the idea of acquisition of knowledge focused on a topic—"the subject of a discourse or of a section of a discourse."[1] Here we explore some of the philosophical underpinnings for the concepts addressed throughout this dissertation. For centuries, philosophers have debated the nature of truth, knowledge, reality and related concepts. Greek philosopher Plato wrote about a cave where people experienced things

[1]Definition from Merriam-Webster dictionary, at https://www.merriam-webster.com/dictionary/topic

about the world by observing shadows on the wall, giving only an impression of what is real [203]. The shadows served as their reality, but how accurately do they represent the real world? A common description of truth includes objective truth (what actually exists) and subjective truth (an interpretation of what is real based on a person's lived experience). An adjacent concept to these is in current discussions of "post-truth"—dealing with issues related to the difficulty people have in discerning what is real or true in an increasingly complex information environment. The term has its origins in the early 1990s, but was popularized in the mid-2010s [107].

Different people may interpret the same event in very different ways—a phenomenon known as the Rashomon Effect, named after a 1950s movie in which multiple people describe very different and plausible interpretations of the same event. While the origin of the term is debated, one of the earliest uses described differences in interpretations by ethnographers [111]. The term has been applied in a wide range of situations, from eyewitness accounts in courtroom scenarios to political activity to interpretations of a speaker's remarks [14, 239].

This idea of multiple possible interpretations is important to the case of change detection. People interpret their topic of interest based on their personal framework and interests; two people may define or outline the same topic in very different ways based on background and experience. Bruner describes how human domains are constructed through narrative principles. "Reality," he stated, "is represented in the act of knowing." He notes that narrative not only represents reality, but also aids in defining reality [40]. As a toy example, take a set of scenes from the Disney movie *The Little Mermaid*. The main character, a mermaid named Ariel, collects a variety of human-made objects. She has no prior context for or understanding of these objects, and constructs meaning based on discussion with other non-human characters. This leads to a humorous scene in which Ariel—unaware of human societal conventions—uses a fork as a comb. Was her interpretation incorrect, or

simply an unpopular albeit logical approach? Throughout this dissertation we aim to support individually constructed knowledge. While one person's interpretation of a topic may seem unusual or even "wrong" to another, the approach to change detection supports the organizational scheme preferred by that individual.

2.1.1 Key Structures of Information

Throughout this dissertation we refer to a number of important structures used to convey information. We start with concepts that relate to organizing information. Most importantly, there is the idea of a "topic"—the subject of discourse we have previously introduced. Van Oosten explains a topic as a concept similar to Rosch's basic level categories for objects, from the field of psychology [221]. She described superordinate topics as generalizations, with basic level topics referring to major participants or parts of the superordinate topic. Subordinate topics are minor parts of the basic level topic. She refers to the prototypical topic as being the basic level topic in her hierarchy [279, 280]. Our research also refers to subtopics, which we consider as themes or aspects that are related to the topic, but at a lower representational level. These are similar to the subordinate topics described by Van Oosten.

We also introduce here the concepts of taxonomies and ontologies. These each refer to methods for categorizing topics, or information. While sometimes used interchangeably, these typically have different structures. A taxonomy is typically more structured and hierarchical, whereas an ontology is a broader representation of knowledge and interconnections, often represented as a network [105, 281].

There are a variety of ways through which we internalize and convey topics and subtopics. When people learn or know topics, we think of this as being represented in their brain as a mental model. This concept, from the field of psychology, describes our internal repre-

sentations of external reality, and serves as a framework for organizing our knowledge. We discuss the role of mental models in more depth in Chapter 3.

Documents are another core construct for this research. Documents serve both as sources of information for the individual with the change detection need, and they can be outputs from the change detection process—which may become sources for information for others with change detection needs. Levy described three characteristics of document objects: they are communicative artifacts, they are external and public (separate from their creators, and accessible to others), and they are stable or relatively permanent [158]. Buckland traced back the history and evolution of the term "document." He ultimately comments on the modern concept of the document as a representation of socially constructed knowledge, and having some relevance [43]. In this dissertation we typically use the term "document" to mean text, but some of the principles of change detection may also apply to multimedia files—video, image, audio, and other contents. We leave that extension to future research.

Claude Shannon, a researcher known as much for his quirky approaches as his innovative ideas [248], pioneered concepts such as digitization of information as 0s and 1s—leading to digital storage of information [242]. He is also considered the "father of information theory" for his 1948 paper that covers how a message flows through a channel between a sender and a receiver, storing and interpreting information mathematically. In this paper he introduces the concept of entropy (now called Shannon's entropy), which conveys the amount of information contained in a message [241].

2.1.2 Theories and Information

Chapter 3 of this dissertation features a theory developed for the change detection use case. Why does theory matter? We provide here some background on theories and theory

building to set the stage for the theory building presented there. In particular, we focus on the role and purpose of theory in the field of information studies.

What is theory? According to foundational works by Kaplan and Merton, the general concept is that theories are about answering the question of "why" [139, 182]. Whetten described the building blocks for theory, and a process for assessing theoretical contributions. An important component of this is what he refers to as the "theoretical feedback loop"—new applications of the theory should not only reinforce its utility, but strengthen it [293].

A number of writers have also attempted to define what a theory is not. Sutton and Staw listed five items as being commonly mistaken for theory: references, data, lists of variables and constructs, diagrams, and hypotheses or predictions. Theses would relate to the method, rather than the theory itself [258]. In response, Weick discussed the process of theorizing, describing theories as a continuum. While the five areas described by Sutton and Staw may not be theories themselves, Weick points out that the process of theorizing may generate these items as interim processes [258, 291].

In 2001, Pettigrew and McKechnie discussed how theory applies to the field of information science. They found that while information science draws heavily from theories in external fields, few of the theories in this domain were cited by fields outside of information science [195]. More recent research into the diffusion of ideas across fields shows that this has changed; for example, a 2020 study of publications about information behavior, a concept from information studies, show that this information is being cited and researched in a wide range of fields [297].

Theories can be tested through qualitative or quantitative studies. A variety of research methods can be applied to interpret and analyze the data, ranging from quantitative approaches such as statistical and survey methods to qualitative approaches involving interpreting unstructured data [68].

2.2 Dealing with Information: The Foundations of Change Detection

Now that we have addressed approaches that can be used to organize and study information-related topics, we cover other topics relevant to the concept of change detection. Here we discuss the strategies that individuals use to cope with large volumes of information, take interest in a topic, and learn about that topic as part of their long-term information seeking practices.

2.2.1 Managing Information Overload

In today's age of information, there is too much information available to know or follow everything. Writings about information overload go back for decades, including David Shenk's 1997 book, *Data Smog*. In this book, Shenk discusses the affect of the growth of information availability, effects of the information overload (to include both mental and physical effects), and various coping mechanisms people apply to deal with data—including reducing exposure to information [243]. While the Internet may provide opportunities to inquire on a broad range of topics, no individual has the capacity to know every topic at a detailed level. Even a desire to dive into depth on a large number of topics brings to mind the story of Faust, with his desire to "know everything that can be known," according to the retelling by Goethe [95]. In spite of possible Faustian hopes, it is impossible to know everything; people instead specialize in specific topics.

The idea that too much information is available is not a new one. In 1945, Vannevar Bush described the "...growing mountain of research." Due to specialization, thousands of people were producing new information on various topics, making it difficult for researchers to keep up with the latest information available. Bush foresaw future innovations that could aid in dealing with information [45]. Still, the volume of information available—including to experts for use in research—has continued to grow significantly, reaching lev-

els that Goth referred to in 2010 as "...a 'Can you top this?' exercise in numbers" [98]. This increase in information volumes has led to rethinking of how to deal with large volumes of information, as well as serendipitous uses for information. Information made available for one purpose sometimes turns out to be relevant to another purpose—for instance, data that Google made available for machine translation purposes turned out to be highly useful for predicting the next word in a sentence, given a specific term.

Operating in an information-saturated world, it is unlikely that the user has sufficient time to explore all possible angles of their topic. Eriksen describes the "falling marginal value of slowly acquired knowledge" [79]. Some of the same technologies that were intended to complete tasks and save people time now monopolize people's time. Instead of spending time understanding topics in depth, people hurry through their update processes. When completing change detection tasks, the user faces what researchers have termed an "explore-exploit tradeoff." This addresses the fact that people can determine whether to "explore" and try something new or "exploit" and go back to something they already know or like—they don't have sufficient time both to experience something tried-and-true as well as explore something new [296]. There are some things that many people know, and many things that only few people know (a sort of "long tail" map of knowledge); users engaged in change detection tasks do not want to miss what everyone knows ("exploiting" common themes), but also cares about the latter—knowing information that others do not know ("exploring" concepts that may give them an information advantage).

2.2.2 Topics of Interest

We have previously introduced the concept of a topic, or the subject of a discourse. We now turn to the idea of interest in a topic. What causes a user's interest in a specific topic? Why are people interested in certain things? For this section, we define "interest" as

a willingness to provide attention to the topic. As noted previously, an individual does not possess the time, energy, or bandwidth to pursue a Faustian desire to learn about everything. They must necessarily filter or specialize. Similarly there are many things that people encounter on a consistent basis, but do not stop to think or wonder about. In relation to the natural environment, author Helen McDonald refers to this concept as a "green blur" [169]; in other words, people often pass by things in nature without noticing them. Chachra adapted this idea to infrastructure, calling it the "gray blur" [50]. This concept applies to a wide range of topics: From afar, you might not notice details, but if you delve into a single topic, a detailed and complex story may emerge.

Kubovy viewed people's lives as a collection of strands, in which certain strands are at the foreground at different times [150]. Similarly, different topics may be at the forefront of an individual's interests at a given time. There are certain topic areas that individuals want to understand in detail. Why do they focus on these things and not others? We cannot know about or focus on everything; people tend to specialize on certain topics for a range of reasons. Schiefele, et al. developed a general theory of interest, where there is a person, an object, and a relationship. The person has some positive view of their object of interest, and also finds value in taking actions that relate to that interest [236]. As a follow-on to the theory of interest, Schiefele, et al. researched the role of interest in learning and education. They found that when an individual has an interest in a topic, it has positive effects on their comprehension of the topic [237]. Krapp and Prenzel looked at the concept of "interest-triggered learning." They found that interests in the topic led to better learning results, to include a more desirable "deep level learning" [149].

It is human nature to want to fit in or belong. This could explain some of the reasons for individuals' interests. Tajfel and Turner developed a theory of social identity. Each person has both a personal identity and social identity. As part of their social identity, they can be a member of ingroups and outgroups [259]. Asymmetric knowledge is one way that people

become insiders. It can be important for a person to keep up with what others know or are saying, and add their own flavor to the discussion. Knowing things enables people to do things that help them belong—for instance, by creating and conveying information that helps others.

When looking across a topic of interest, the subtopics vary in their level of importance. The level of importance may also vary depending on the user's level of interest, or the task they are hoping to accomplish with the data. Why are some more important than others? This may have to do with the way the information is perceived as fitting into the overall structure [303].

We do not mean to imply in this that a user's interests cannot change. While someone may be interested in a specific topic, the interest could evolve, or a new topic could arise that takes greater precedence in the person's mind. Jia, et al. researched this in the academic space. They found researchers' interests evolved over time. Even within a single field, the topics that one researcher covers may change as the field changes [129]. For this research into change detection, the user has a specific interest in a topic, which they study and seek information about over time. Even if there is some evolution around the specifics of the topic, the user continues a deliberate, repeated process of seeking out new information on the topic. Novelty can be a significant driver of users' interests, with novelty-seeking behaviors applying from children to adults [254]. Curiosity and novelty seeking applies to human behaviors ranging from technology use (e.g., information seeking behaviors) to human exploration and movements over geographic space [12, 278].

2.3 Cognitive Tasks

Before we focus on defining technology to support users' change detection needs, what is the user trying to accomplish? What are the goals that the change detection task sup-

ports? In this section we draw together the steps through which the user gains and maintains expertise; later, we look at ways that the task can be supported with systems. We note that change detection covers individual learning tasks that enable a user to build out their personal mental model as they gain expertise on a topic. It is not intended to be about collaborative aspects of learning; rather, we focus on learning and developing expertise as an individual constructive process, and leave collaborative information seeking and learning for separate research.

2.3.1 Information Seeking

Users engaging in change detection tasks have long-term information needs that need to be met. Ideally, this need can be transformed into a question or query that would enable a system to return the perfect information to fill this need; in reality, this transformation between need and query is difficult. Wilson focused on the range of information seeking behaviors applied by individuals to meet information needs, depending on the type of information sought. Wilson's work identified a shift from a systems-oriented view of information seeking to more of a human-centered approach for defining needs—a broadening of the range of ways the need could be addressed. Wilson synthesized definitions of human information behaviors; he distinguished between requirements (requests for new capabilities for a system) and what people are trying to get from their use of a system. He focused in particular on information behaviors such as information seeking, information searching, and information use [298].

For the change detection task, these information needs persist over time. However, before we delve into the repeated nature of these needs (also known as standing information needs), we explore the process through which an individual gains expertise on their topic. With the ongoing aspect of information needs specifically, we focus here on standing infor-

mation needs: cases where the information needs remain over time. For change detection, we assume that the user is following their specific topic of interest over a long period of time. Change detection does not take place once; rather, the user is going to be checking on their topic today, tomorrow, next week, and so forth. Getting updates is sometimes included as a category of information seeking behaviors, though "keeping up to date" has more often been viewed as a fairly casual task [13].

Within the field of information studies, there are a number of theories related to information discovery. This includes Bates' berrypicking model, which describes how user needs evolve as they encounter information; their process for obtaining information may not be straightforward or consistent, as the user modifies their approach based on information they encounter [27]. In another approach, Pirolli and Card addressed information seeking and the interactions as information foraging. Modeled after concepts from biology, information foraging addresses the fact that useful information may be spread throughout a document set. Users must "forage" through documents in search of relevant information. The user in this case follows the "scent" of information, tending to spend more time in document subsets that appear to be more productive [198, 199]. Based on this concept, Pirolli and Card devised the "Scatter/Gather" approach to cluster information in a way that enables a user to identify more productive sets more efficiently. They devised the ACT-IF cognitive model to judge the relevance of information presented to the user, and then maximize the rate at which relevant information is provided to them [200]. Savolainen compared berrypicking and information foraging, two common approaches for exploratory search. He found that berrypicking tended to focus on the search aspect of the exploration, whereas information foraging emphasized the browsing component [235].

Another information seeking activity that has similarities to change detection is environmental scanning. This use case tends to be tied to a business or other organization; employees research their area of operation, looking at competitors and others in their space

to identify new developments [55, 268]. The key idea in environmental scanning is to gain an understanding of what external factors might affect an organization [7]. The concept of horizon scanning is similar to environmental scanning; the difference is that it looks out further in time, attempting to detect future threats or technology disruptors. Bas looked at the interrelated concepts of foresight and innovation as critical components of detecting potential disruptions in the context of horizon scanning [26]. For change detection, a topic could include some long-term, over-the-horizon subtopics. There might not be documents relevant to those subtopics on a regular basis, but some of these less common topics might inherently be of higher interest to the user when they do arise.

2.3.2 How Individuals Learn

In order to identify how to address users' change detection needs, we explore how users learn about a topic, eventually reaching a point of expertise. Here we explore processes through which people comprehend a topic, process updates, and expand their understanding of the topic [25, 31, 192]. When an individual encounters information, they go through a process called sensemaking (or sense-making) to understand the content. They have existing knowledge, and some cognitive gap. When they encounter information, it helps them to bridge their knowledge gap [70, 71]. Zhang looked at sensemaking from an individual perspective, framing it from the perspectives of learning and cognition. The sensemaking process can be iterative in nature, as the user alternates between sensing and sensemaking to understand what is happening [304]. While not studied in this dissertation, on a broader level, sensemaking can also be an organizational construct—people working together to understand or make sense of a change or action [112].

Philosophers and academics have long debated concepts related to narrative text for understanding—going back at least to the Greek philosophers. Aristotle focused on narra-

tive structure for storytelling, indicating that you need a beginning, middle, and an end for a complete story [19]. His approach is focused on retelling of a complete story; it is somewhat different from change detection, where the reader may be in the midst of an ongoing story. In his book *The Culture Code*, Coyle observed, "When we hear a story... our brains light up like Las Vegas, tracing the chains of cause, effect, and meaning. Stories are not just stories; they are the best invention ever for delivering mental models that drive behavior" [65].

Pearson and Cervetti reviewed the changes in the reading comprehension field from the early 20th century thru the early 21st century—including research in this field, as well as the relationship between policy and research. They traced the use and impact of models of reading comprehension from the past few decades that were designed to understand how users represent the meaning of text [191].

In order for a reader to understand a text, the concept of coherence relations helps to tie the text together—meaning that the ideas are arranged in a way that is meaningful to the reader [102]. Graesser, et al. provided additional perspectives on the topic of coherence relations. They found that more coherent text makes it easier for the user to understand the meaning of the text; in contrast, text that is less coherent (e.g., having conceptual or structural gaps) makes it difficult for the reader to understand the meaning [102].

One such method for organizing information that is familiar to many readers is chronological ordering. Van der Meer, et al. looked at the chronological aspects of coherence relations. As a reader goes through a text, they create a situation model—a mental model of the of the situation being described. The temporal angle of the information can be important to the mental model. Their study shows the influence of temporal information on memory [277]. A study by Flower and Hayes revealed that sophisticated writers' early drafts are sometimes chronological as they are thinking through their eventual message organization. Experienced writers later reoroganize their material to have a more hierar-

chical focus [86]. As an example from the process of creating a fictional story, we look at the one applied by author V.E. Schwab, and how she distinguishes between chronological order (the order in which events happen) and narrative order (the order in which the story is told by the author). When creating a new book or series, she starts from the point where she wants the series to end; from there, she focuses on a single character's story at a time, outlines each scene, and writes chronologically from each character's position. After this process, she weaves the story lines together into narrative order [164].

Telling a story in chronological order might not be considered narratively interesting from a creative writing perspective. However, if the goal of the reader is to learn and not simply to enjoy a story, the use of chronological order can help a reader understand the situation [252, 253]. From the reader's perspective, coherence relations are important for understanding a story and learning how what is happening relates to the topic. Zoran observed that the narrative continuum can include concepts of both time and space [308].

Pintrich defined self-regulated learning as a case where the learner has control over when they learn and what they explore. He included four phases in the self-regulated learning process: Forethought, planning and activation; monitoring; control; and finally reaction and reflection [197]. While these were applied primarily to an academic setting, they have direct relevance to the change detection process—where a user is learning about information on a topic of interest to them.

2.3.3 Output of Change Detection

Once the user has received their updates, what are some potential outcomes from following a topic? While not part of change detection specifically, the information can be used in support of a variety of outcomes, ranging from sharing in conversations, to writing articles about the topic, to creating other types of products to express what they learned.

Information may assist in solving a problem, alerting the reader to potential issues, or identifying opportunities to share information further. Some people organize information for understanding and retelling stories [138, 225, 228, 301]. In the information age, individuals have many types of content that are competing for their attention. How do they decide what information to act upon or propagate? Hodas and Lerman looked at the processes that people apply in formulating their decision on whether to re-share information online [118]. Unlike prior studies, which had focused on novelty of information, they found that visibility was an important factor—users tend to focus on information that is easily seen (for instance, near the top of the screen). Another potential outcome of change detection tasks could be serendipity—the user could draw unexpected connections between multiple areas as a result of information that they encounter [87, 167, 179].

Information can be used in many ways, ranging from writing documents to more artistic outcomes. People might picture the writing of a news article as an outcome of the change detection process. This is just one of many ways to convey the stories gleaned through the process. Some have argued that journalism is about creating a story that can shape audiences [288]. Art can also involve portraying subjective versions of a topic, conveying information while adding an emotional charge to the message by mixing it with music, poetry, and other media. During an interview and discussion about the responsibility of an artist in conveying reality or influencing people, Thom Yorke, musician and lead singer for the band Radiohead, said, "I think no artist can claim to have any access to the truth, or an authentic version of an event" [44].

2.3.4 Overcoming Issues and Biases

As individuals add information to their mental models, they should also be cognizant of potential biases and other issues that can affect their understanding of their topic. Pariser

studied potential issues with filter bubbles—potential situations where search engines return information so closely tied to a user's interests that they miss information that would provide an understanding of other perspectives [190]. Filter bubbles can have an impact on what a user sees in the future; if a system interprets their interests too narrowly, information that would provide alternative viewpoints are omitted. Even factors as basic as which words are used in a query can have an impact on the type of search results that are provided.

Confirmation bias is a related issue that experts should almost always try to avoid. In this case, the user may (intentionally or unintentionally) seek out only evidence that confirms prior-held beliefs. In some cases, evidence that goes against their beliefs is actively rejected due to the strength of the belief [187, 205].

While it may seem that organizing information for quick reading or reviewing might be useful, Posner found that adding "cognitive speed bumps" (sometimes also referred to as "mental speed bumps" or "cognitive interventions") could be beneficial [206]. Adding a bit of friction can aid the reader by slowing them down enough to ensure that they pay attention to the topic. If the process is too easy, the reader risks rushing through a text with little or no comprehension. That said, there is a need to distinguish between cases where adding cognitive speed bumps could help (e.g., by encouraging the reader to slow down and concentrate a little bit more) compared to where they are problematic (e.g., making it difficult to fill in relationships between events). Podsakoff, et al. discussed inclusion of "speed bumps" into surveys to encourage respondents to slow down and focus on the questions [204]. Galante adapted this idea to design processes, including features that would make the user slow down and think. He drew upon ideas from Kahneman's book "Thinking, Fast and Slow," explaining how cognitive "speed bumps" can stimulate slower, more methodical thinking known as "System 2" processes [89, 134]. Given that users engaged in change detection may typically review large amounts of information, there may

be a need to incorporate these sorts of speed bumps into systems to ensure that they do not unintentionally lose focus and miss information.

2.4 Technology Supporting Change Detection

Now that we have described the user needs, information seeking behaviors, and actions that the user takes to become an expert, we move into a discussion of some of the technological foundations that support change detection. In this section, we discuss principles related to information retrieval, human-computer interaction (HCI), and other related system and search concepts.

2.4.1 Information Retrieval

What is an information retrieval system? An information retrieval system represents the user need or intent as a query, and matches user to the information they need. Taylor identified four levels of question formation for addressing a need with a search system. This includes: 1. the actual (visceral) need, which is unexpressed; 2. the conscious need; 3. the formal statement of the question; and 4. the statement presented to the information system (compromised need) [262]. The query for the search can be thought of as the compromised expression of the need: it is intended to match what the user believes the system can do.

Perceived gaps in knowledge on the part of the user could cause the user to take action, potentially triggering the situation described in Belkin's theory of Anomalous States of Knowledge (ASK). The user has some unmet need—in this case, perhaps an unclear understanding of what happened during a certain point in time—and this triggers a process through which they formulate a query to seek information to fill the gap [30].

Information retrieval systems use a variety of approaches for organizing search results, ranging from chronological ordering to relevance ranking. Ranking based on relevance

does not have a single definition or algorithm; rather, it is a method for organizing results based on a mathematical representation of the user's need, often based on concepts such as the number of times query terms appear in a retrieved document [61]. The concept behind relevance ranking was introduced and refined by Salton and his research teams in the mid- to late-20th century [229, 230, 231]; the concept of relevance ranking has been adapted and expanded in many ways since then. For example, two well-known relevance ranking models are Sparck-Jones' TF-IDF [250] and BM25 (Okapi Best Matching 25) [215]. Modern search systems apply a variety of algorithms intended to retrieve relevant documents for users [93]. These algorithms integrate such concepts as novelty, diversity, and recency in an attempt to provide better results to users [72, 283].

Depending upon the type of system, the product could be a result set that the user can review for the information they need, or even simply an answer. Information retrieval aids in matching users to the data that they need, narrowing down from the volumes of information available to the information relevant to them. There are a variety of types of information retrieval approaches available, including search and discovery concepts such as ad hoc search (often a one-time query based on a current information need), recommender systems, and more [104, 172]. Broder categorizes the types of information retrieval tasks into three categories: informational, navigational, and transactional [38]. An information retrieval system is one that is designed to retrieve information for the user to fill their need. In a search system, the user's need is represented by a query. The query is sent by the system to an index, which contains representations of the contents of documents. The system retrieves the documents that matched the query.

Depending upon their eventual goal, users of information retrieval systems may satisfice rather than satisfy their needs. They read some of the documents available at the time until the basic information need is met. They may not review all available relevant results, given that there are other commitments drawing the user's time [269].

2.4.2 Human-Computer Interaction

Human-computer interaction (HCI) deals with understanding and meeting users' needs with computer technology. This field is at the crossroads between computer science, information studies, and other fields, and aims to improve users' interactions with technologies for a wide range of use cases [170]. A core idea behind HCI is that it is important to get input from the users of the system—whether directly or indirectly. This can be accomplished through the application of research methods such as surveys, analysis of logs and other technical data, user studies, and through user activities such as discussions and journaling [127, 155]. Some of the steps involved in a HCI study might include discussions with the targeted user group, creation of wireframes or prototypes to convey potential interface designs, and user studies to test the effectiveness through which a system meets a user need—with the key principle being to put the user need first in a user-cented design process [289].

How can current technologies be assembled to test how well a system meets the user's change detection need? Fan, et al. show that people want updates, but available interfaces often lack personalization for their areas of interest [81]. This idea may be relevant to change detection, though its emphasis is on exploring a concept, whether or not there is prior knowledge about it.

While talking to users represents an obvious approach for gathering user needs, there are a variety of approaches that can be used to gather relevant information. It is possible to interpret artifacts generated by users to understand their needs in a tacit way. For example, Limam, et al. applied semantic analysis and clustering to users' search logs [162]. Through this technique, they were able to identify semantically similar search terms that users were leveraging in their searches.

2.4.3 Automated Evaluation of Information Retrieval Systems

HCI presents some methods for understanding user interests and providing thoughts about systems. However, it is not always sufficient or efficient to ask the user their opinion of individual systems. Instead of performing a user study for every search capability, evaluation measures are used as a proxy to represent how well the system is able to meet a generalized user need. In order to compare information systems more broadly, we can also leverage automated evaluation approaches. These enable comparison of systems in a standardized way that can expand beyond what a user might provide in other types of studies. It is useful to be able to compare information retrieval systems, to understand which approaches work best for meeting specific user needs.

Information retrieval evaluations provide a baseline for performance on a task, enabling longer-term work on improving capabilities. These evaluations typically leverage one of three approaches: pointwise, pairwise, or listwise comparisons. Pointwise comparisons look at an individual document's relevance to a query. Pairwise approaches look at the relative rankings of pairs of documents, and remains a popular method for search engine rankings. Listwise comparisons use an entire set of documents or other items in their comparison [47]—for instance, a comparison between subtopic clustering result sets using a correlation measure, as we do in Chapter 7.

For a standard information retrieval evaluation, we look at the results provided by a system (for instance, relevance ranked results), and determine the extent to which the results are in fact relevant to the user. The primary components that are needed for an evaluation include an information retrieval system, a collection (e.g., documents), topics (the information need) and queries, human relevance judgments (assessors' views as to whether individual documents are relevant to a specific query), result sets from a system, and an evaluation measure. The system uses the query (the compromised expression of the need,

referring back to Taylor's description noted earlier) to retrieve the result set. The evaluation measure is a mathematical approach for assessing how well the result set performed. The measure leverages the human relevance judgments (documents that have been annotated for relevance to the topic) to determine how effective the retrieved results were. The measure produces a score comparing the results set retrieved with the ground truth for that query; this approach enables comparisons between performance of different systems on a given topic [62, 172, 175, 181, 287].

Common evaluations in information retrieval include comparing factors such as precision (were the right results retrieved?) and recall (were all the relevant results retrieved?). There is a tradeoff between precision and recall, with one tending to decline as the other increases [41]. We tend to view change detection as a recall-focused task, where users are interested in seeing all the documents—or at least as many as they can view within available time—as opposed to expecting a single document to satisfy their information need.

Some search approaches are designed to improve diversity of results—including across a variety of subtopics. Traditional information retrieval measures, such as Mean Average Precision, do not account for query performance against aspects or subtopics of the overall topic. Researchers have adapted existing measures and created new ones to account for document and subtopic diversity, such as the aspect-focused Normalized Discounted Cumulative Gain (α-NDCG) [2, 168]. Axiomatic approaches are applied to the process of information retrieval evaluation as a rigorous, structured method for analyzing how well systems and measures exhibit specific properties [82]. Even though we do not include axiomatic analysis in this dissertation, we did structure the evaluation design in Chapter 7 to support potential future axiomatic analysis.

2.4.4 Topics in Information Discovery

Building upon the more conceptual descriptions of topics presented earlier, in studying various use cases for information retrieval, researchers have defined "topics" in different ways based on the user need. The definition used in this dissertation is most similar to the one applied in the annual Text REtrieval Conference (TREC). In that interpretation, a topic is a subject-based set of themes that an individual would recognize as being related.

As described by Hoyle, et al., a topic can be thought of as "a set of terms, [that] when viewed together, enable human recognition of an identifiable category" [121]. Another sense of the term "topic" is observed in the concept of topic modeling—in that research area, a topic is based on probability distributions: topics can be grouped together mathematically based on term usage [285]. While identifying clusters of like documents has been a feature of topic modeling for many years, identifying a suitable, representative label for those clusters (one that is meaningful for a human) has been difficult. Systems such as ALTO have been designed to combine machine-generated labels and human annotations to produce better label sets [208].

An additional application of the term "topic" comes from the Topic Detection & Tracking (TDT) program, which was created in the late 1990s to review multilingual broadcast news sources. As defined for TDT, "...topics consist of a seminal event plus any directly related events and activities" [163]. The goal of TDT was to indicate which stories first identify a new development, and then to find other related stories from the stream of broadcasts [10]. TDT included five research tasks: Topic Tracking, Link Detection, Topic Detection, First Story Detection, and Story Segmentation, with Topic Detection and Topic Tracking being the most closely linked to the change detection tasks studied in this dissertation [84].

2.4.5 Organizing Information of Interest

For the change detection use case, we have previously indicated that the user has a topic of interest. Tied to that topic are typically a variety of subtopics—themes or aspects that represent smaller, subordinate pieces of the topic. How are subtopics defined? How might a system treat these subtopics? Here we focus on areas related to subtopics in systems, including identification of subtopics and subtopic ordering. We also introduce ideas that influenced this dissertation research early on, including subtopic identification and importance.

As an outcome of the change detection process, users may be interested in taking things they have learned and applying them in various ways. To do so, they need a way to return to information of interest. Knowledge management functionality could be a useful way to highlight subtopics or specific items of interest. Personal information management capabilities not only include methods to find and organize information, they also include a capability to retrieve and synthesize relevant information that was previously found by a user [132].

Implementations of topics and subtopics in system can take a variety of forms, ranging from hashtags (tags often marked with the # character) to folders or tag labels containing terms or lists of phrases. While folders or tags may be more about knowledge management—marking information of interest for future review or retrieval—hashtags also serve another function, of marking information so that others interested in that idea can find it, or of providing commentary. For example, a user might add a hashtag such as #yeahright after their message as an indicator of sarcasm. Santos-Neto, et al. looked at a variety of tagging systems to understand individual and group behaviors related to tagging, tag reuse, and interest sharing [233]. Sets of tags generated by users result in folksonomies, a port-

manteau of "folk taxonomies"—collections of users' tags within a site. They can reveal more angles of understanding of content than other forms of taxonomy [9].

In addition to manual approaches for identifying topics and subtopics, automated approaches can be applied to associate or group text by related theme. Topic modeling includes a set of automated approaches for identifying content that contains common themes, using probabilistic approaches. Other techniques such as Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NNMF) are examples that use matrices of term counts from documents to find similar themes [142]. Word embeddings found using neural methods have also been used to identify topics—for instance, to identify themes within social media [60].

Some subtopics may be of greater interest than others. If we want to organize subtopics, we can identify potential ways that subtopics might be organized based on the user's interests. One way to approach this would be to think of it as a sorting process, except the items being sorted are the subtopics rather than documents or other information tied to those subtopics. Sorting of subtopics can be accomplished in a variety of ways. Existing work on organizing subtopics has focused on complex calculations of using word relatedness, part of speech, as well as the influence or importance of poster (in the case of microblogs). Shajalal, et al. used a subtopic importance calculation in their document ranking that was derived from the part of speech of the terms (noun, pronoun, verb, and adjective) and term popularity [240]. Others have attempted to identify the importance of the subtopic to the user. Takaki, et al. use entropy and the probability of a query term appearing in a subtopic to determine the importance of a subtopic for a use case involving patent search [260].

2.4.6 Systems Designed to Provide Updates

The concept of having systems provide updates is not a new one. Researchers have applied a broad range of approaches for organizing data in ways that are useful, depending upon the use case and collection being searched. In this section, we discuss search and discovery approaches for finding new information of interest within collections of news, email, and social media.

Kuhlen and Preston studied concentrations of topics in news sources over time, and how the flows changed based on events [151]. They found that political, geopolitical, financial, and natural disasters were most influential in the flow of news topics. In a study of news articles over a 5-year period in the 2010s, Hendrickx and Ranaivoson found a significant increase in the number of articles that were recycled between news outlets within a European market. They found that this corresponded with the decrease in numbers of journalists over the time period. Their research revealed potential issues with diversity of news sources and perspectives over time [113].

Technologies that attempt to understand and interpret political texts are also relevant to change detection. As noted by Grimmer and Stewart, automated interpretation can provide benefits in reviewing large amounts of information quickly. However, systems experience difficulties in accurately interpreting what is happening. New approaches—including validation approaches—are needed to make further advances in systems that assess text on political topics [103].

Many researchers have studied inefficiencies related to email. In one review of email from a behavioral economics perspective, some of the issues cited included interruptions to workdays and a lack of useful organization or flow [196]. Some of the suggestions for improved interfaces included grouping messages based on factors such as whether the

message is informational or requires a reply, and if a reply is required, what the urgency level is (when a response would be needed).

Various systems have attempted to improve users' experience finding relevant information within email messages, in particular to find important messages, events, and other details. In 2012, the Internet company AOL introduced a system that served as a personal email assistant [140]. One of the goals of the personal email assistant was to ensure that users did not miss incoming information in their inboxes. It integrated with email clients and other systems to identify information of interest to users—to include event and calendar information. This information was extracted and provided in a dashboard format to highlight key updates to the user. The service was discontinued in late 2017, after AOL was acquired by Verizon and integrated into their Oath subsidiary [83]. Email assistants have largely remained focused on dealing with incoming messages one at a time, including application of manually curated filters and features that identify or interpret information within individual messages. In recent years, Google's Gmail service has introduced an ability to filter messages using a recommender approach (finding new emails that are similar to the selected email message), and taking some action (e.g., applying a label or marking them as important) [299].

Kulshrestha, et al. characterized social media production, as well as a user's "information diet" [152]. They adapted their idea from Johnson's book *The Information Diet: A Case for Conscious Consumption* [130]. They defined an information diet as "the topical distribution of a given set of information items." This research revealed that users tend to consume information diets consisting of only one or two topics (as might be the case for a user engaged in change detection). While the authors' intent was to identify opportunities to broaden users' information diets, the research revealed that users tend to focus on specific topics of primary interest to them.

Even though this dissertation focuses on individual information needs, we note that sys-

tems that contain social features—either as the primary or secondary role—can sometimes lead to changes in the user's behavior. Lim and Datta looked at communities in social networks based on their topical interests. Over time they observed the evolution in the users' interests, and the effect on the community [161]. Sousa, et al. looked at interactions across user networks on Twitter, and found that users' interactions appeared to be primarily focused on social interactions; however, users who have broader interactions "had a slight tendency to separate their connections depending on the topics discussed" [249]—behavior that could be relevant to users in change detection tasks.

Search systems have applied a variety of tactics to understand users' interests within the system and refine results accordingly. In personalized information retrieval, search results are customized based on the system's interpretation of the user's interests or focus, with the goal of providing improved results [226]. Teevan, et al. have explored approaches for personalization via algorithms as well as through query term use [264]. Some of these approaches—for instance, the WhittleBit system from the early 2000s [246]—leverage explicit feedback from users. WhittleBit asked users to provide a thumbs up or thumbs down for a document, and used that to refine future search results. Other systems have leveraged tacit approaches instead of requesting active user feedback to gauge user interest; for example, recommender systems make use of user actions to understand what is of interest [157].

Thinking ahead to future approaches for accomplishing change detection tasks, Large Language Models (LLMs) provide new opportunities for providing updates to users. LLM-generated summaries of changes on a topic could provide value to the user, and point them to starting points in their periodic exploration of subtopics or documents. Researchers are already exploring the idea of summarizing new information in LLMs [305], and a change detection system like Daybreak could make a logical integration point for these kinds of capabilities.

2.4.7 Related Search and Discovery Approaches

The change detection task is a form of session search taking place over the course of days, weeks, or months. The user issues an initial query, selects and reads documents from the result set, then reformulates or re-runs the query in a series of "search iterations" [49]. In our use case, the user may keep the original query terms, and search for them on a later day to receive a different set of documents. Over time, the user may evolve the query to fit with observed developments. For now, we focus on the query and results it produces a day at a time. Foundational information is needed to understand this task; in the future, perhaps it will be possible to look at change detection from the more complex session search viewpoint, to include how the expert's query evolves over time as concepts grow and adapt. Research into users' long-term search history has been used to improve effectiveness of both ad hoc and recurring queries [261]. This type of research could further apply to change detection-related information retrieval tasks.

Exploratory search is designed to aid users as they discover new things and connections. White and Roth frame this as including two components: the problem context and the search process. The search process consists of exploratory browsing or focused searching [295]. Perer and Shneiderman discuss guiding domain experts through exploratory search in a "systematic yet flexible" way [193]. Even though we view change detection as distinct from exploratory data analysis (in particular, in cases where exploratory data analysis is aiding users who are new to a topic), there may be user behaviors that overlap between these use cases.

Even though we are not focused on creating an aggregated search interface in this research, we are drawing upon some of the ideas in designing an interface that bundles together content in some way—in our case, subtopic clusters. Aggregated search focuses on providing search "verticals" alongside core search results to provide additional context

to users. These verticals might not only contain text content, but also include multimedia contents such as images or videos. Within verticals, results can be organized or sorted in a specific way; additionally, the verticals themselves are sorted in relation to the users' query [17]. Aggregated search has its roots in federated search, expanding upon the original idea of providing search results gathered from a variety of different collections. The federated search literature has dealt with difficult issues such as truncating vs. interleaving results. Dealing with issues like interleaving results can be problematic due to federated search engines applying very different scoring mechanisms. This makes it difficult to decide which document to place next in a combined set of search results, a process referred to as collection fusion [120]. These concepts can provide inspiration for approaches for organizing and displaying change detection result sets. Even though they may all be of the same source type, each subtopic might need to be treated differently—with ones of greater interest to the user displayed more prominently.

2.5 Summary

In this chapter, we have provided a review of the literature relevant to change detection. We started with foundational philosophical concepts related to knowledge and truth, then focused on the user and their needs for relevant information. We looked at practices for reviewing information, and learning about a topic—while attempting to avoid negative issues. We transitioned from there into foundational concepts related to technologies supporting change detection. This includes background about the fields of information retrieval and human-computer interaction. In Chapter 3 we build upon concepts from this literature review with the GPA Change Detection Theory, which provides an organizing principle for this dissertation.

Chapter 3: Formulating a Theory about Change Detection Systems

3.1 Introduction

With the growing amount of information available, how can a user know what is happening related to their topic of interest? Change detection is defined as the process through which the user attempts to identify whether something has changed, and what has changed about a topic of interest that they follow. When engaging in this task, it can be difficult for a user to determine whether they have seen the range of new developments on their topic. To address this gap of aligning users with new information on a topic of interest, this chapter provides a theoretical framework for the change detection concept, which we call the GPA (Group-Pile-Arrange) Change Detection Theory. The goal of this theory is to provide the user with a personalized view of new developments related to their topic. While this user may later use their knowledge to contribute to collaborative discussions, the process we study here focuses on an individual's construction of knowledge.

What are the key tasks performed by an individual who is trying to determine what new information has emerged on their topic of interest since the last time they checked? We intentionally scope this as an individual information seeking task, in which an individual user seeks updates to their personal mental model. The user comes to the problem with a baseline familiarity about the topic, and they intend to learn what has changed by reviewing a set of documents that are filtered based on their relevance to the topic of interest—not a random (from the user's perspective) set of documents. The theory supports a system in a

general sense—one that can be used to describe or design systems supported by manual, computer-driven, or hybrid approaches for meeting the user's change detection need. To support this theory, we draw from literature across a variety of fields, to include reading comprehension, cognitive psychology, behavioral economics, information retrieval, and information studies.

Imagine that a researcher has a research librarian assigned to work with them. The librarian knows what topic the researcher cares about or follows closely, and actively seeks information of interest to that person, using a process traditionally known as selective dissemination of information (SDI) [115]. Given that they can generally focus on only one thing at a time, the researcher has to process the documents in a linear fashion, by reading documents individually, in some sequence. What if the librarian took this one step further, to make it easier for the researcher to understand what has changed? The librarian could aid the user in their updates by grouping the documents in a way that is meaningful to the researcher based on aspects of the topic of interest. Before this librarian hands the document set to the researcher, they could then organize the set in a way that emphasizes developments of greatest interest to the researcher first. Over time, the researcher works with this librarian to refine this process, both in how the researcher describes the topic of interest, and how to break the topic down into aspects. Such collaborative processes for finding and organizing information have existed a long time, and have been supported by research librarians and others as well as with computer systems.

3.2 Theoretical Foundations

The GPA Change Detection Theory serves as a foundation for understanding users' approaches for finding updates, with the goal of supporting design of manual-, computer-driven, or hybrid systems for meeting these needs. Given the capabilities of modern tech-

nology, a computer system dedicated to this task could provide information more efficiently that a human might, giving the user a broader understanding of the range of new developments on their topic of interest.

For many years, researchers and academics have refined the definition of theory, and how theories are refined, to include the role of data within the process. Based on a review of theoretical and conceptual frameworks, Imenda argues that theories tend to emerge from deductive reasoning, whereas conceptual frameworks come from inductive processes [124]. Varpio et al. further expand on this idea, describing the loop from data to theory and back. They refer to the flow between theory and data collection as the "objectivist deductive approach," and the flow from data to theory as the "subjectivist inductive approach" [284]. We apply a similar approach for the theory development here. While the data that informs theory could come from new experimentation, theory can also follow from examination of existing literature. Looking across a range of theories and prior experiments in a similar research space can provide insights that lead to new theory [122, 156, 176].

In Figure 3.1 we show the cyclical flow between data and theory, where perspectives gleaned from data inform the theory. The refined theory produces gaps that can be addressed by collecting additional data.

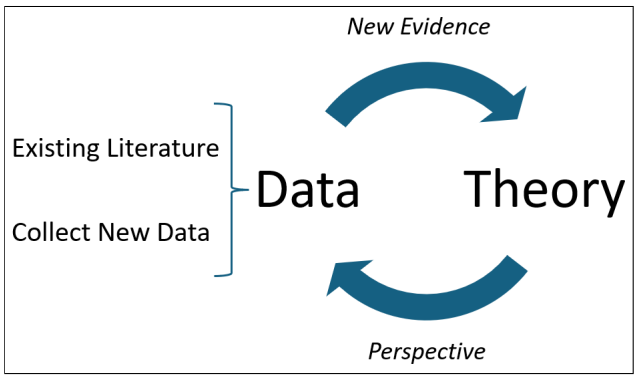


Figure 3.1: Interplay between data (new data or data from existing literature) and theory.

In this dissertation, the theory is presented ahead of the studies to provide a foundation for explaining observed behaviors and stated user needs. Components of the concept that ultimately became the theory were used to define the scope and structure for the studies covered in Chapters 4, 5, and 6. Chronologically, the theory was formalized after data collection for the Daybreak study.

The GPA Change Detection Theory aids in framing what capabilities and components need to be present to address the user's underlying need. Change detection is about learning and building expertise related to the topic of interest; this repeated practice of finding information can be looked at like an athlete practicing for a game by using a structured approach. As the athlete practices the game, their skills increase and they gain an ability to identify what is new or different on their topic more quickly. Similarly, the user and the system engage in a cycle where each learns from the other as the task progresses. There are potential benefits not only to learning more about the topic and associated subtopics, but also improving the user's information seeking skills related to change detection.

We formulate a foundational theory for change detection inspired by the approach described above. Recognizing that most people do not have a team of librarians aiding in their searches, we have designed this theory to be sufficiently broad as to incorporate a range of real-world use cases ranging from SDI to computer systems designed specifically for the task. At the core is this three-pronged concept: 1. Group: documents relevant to the topic are grouped by subtopic, 2. Pile: documents in the group are organized using a sorting or organizing process into piles, to form superdocuments (compilations of documents about the same subtopic), and 3. Arrange: the piles are arranged in some order before being presented to the user. In addition to framing the Group-Pile-Arrange (GPA) Change Detection Theory, this chapter discusses options for optimizing organization in each of the three steps, as well as the inputs into the processes that are framed by the theory. We first describe the steps, then discuss the literature related to each step. The GPA Change De-

tection Theory can apply to a broad range of topics, from financial analysis to sports and more. Throughout this chapter, we use the example of a person following news about their favorite baseball team as a consistent example.

As described in Figure 3.1, our theory emerged from review of literature and data from the field of knowledge management. This research was inspired in part by informal observations of user behaviors in a variety of systems supporting news articles, email, social media, and other collections. Users leverage tagging, hashtags, and related capabilities for organizing information—including for topics that they follow over time. Research shows how users apply personal information management techiques (including foldering and tagging) to keep track of information of interest [33, 56, 210]. If tags and folders are artifacts that demonstrate a user's interest in aspects of the topic, could they provide further benefit in other parts of the process, beyond serving as a bookmark for returning to documents later? Some research has leveraged users' tags in relevance ranking algorithms [32]. As a framing question, we expand on this by exploring whether tagging—which is beneficial for knowledge management after they find something useful—could also be helpful for driving the organization of results from the retrieval process.

3.3 Task Overview

Change detection is about getting updates on a topic over time. The user has a topic of interest and some amount of expertise on that topic. The user looks for new developments on the topic on a periodic basis. Our assumption here is that the user finds and reviews text—whether on paper, through a GUI, screen reader, or other device—in their workflow of getting updates. Throughout this chapter we refer to the smallest block of text viewed by the user as a document. The document is featured at the start of the change detection process, in which a user views documents and indicates which ones are of interest. We

discuss the process that happens when the user views additional documents, then loop back to how the system provides relevant documents.

3.3.1 Interactions with Documents

Behind this scenario is an ongoing interplay between document creators, documents, the system, and the user, depicted in Figure 3.2. A document creator could be a person (author), organization, or a computer system. The creator generates documents for a range of reasons, including personal interest, organizational interest, summarizing a broader topic, and more. In this chapter we do not delve into the reasons documents exist. In general, however, we note that there is no direct alignment between the author's reason for creating the document and a reader's interest in the topic. The creator may be producing documents that meet a range of needs, which are likely not customized to the interests of a single user. This means that only a portion of the document may be directly relevant to the user in the change detection scenario. Users want to know how new information aligns with their interests; thus, the change detection system plays an important matchmaking role to fill this gap. Ideally, the system not only identifies that the document is of interest to the user, but also indicates the segment of the document with the relevant information.

3.3.2 Supporting a System

The change detection theory focuses on support to a system in the broadest sense of the word; a system is defined as a regularly interacting or interdependent group of items forming a unified whole.[1] The purpose in this case is to provide an individual user with updates on their topic of interest. This can be accomplished through technologies such as computer systems, or via human networks such as selective dissemination of information.

[1] Definition from https://www.merriam-webster.com/dictionary/system

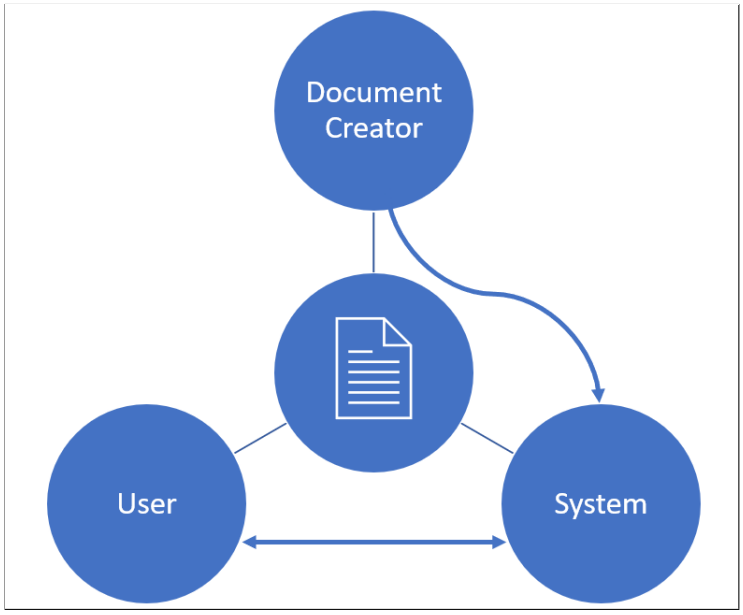


Figure 3.2: Interactions between the document creator, the document, the system, and the user.

The key to change detection is that the system is personalized to one user's need, rather than a group need. Everything should be framed according to the individual user's way of framing the topic. This personalized system needs to address the question of how the user views the topic and its aspects or subtopics, in order to frame new information in a way that addresses this worldview.

To accomplish the task of aligning documents with the user need, the system first retrieves new documents on the topic of interest. Many systems perform information retrieval tasks. What is unique about the change detection theory is that it takes the results, then performs three primary steps: group the documents by subtopic, organize the documents into piles, and arrange the piles, as depicted in Figure 3.3. We will refer to these three steps as the GPA function. In the first step, the documents are grouped into subtopics based on the user's way of framing the topic. Next, the documents within the group are organized into piles to form a superdocument, which is a set of related documents presented in some logical order. Finally, the piles themselves are arranged, starting with groups that are of

greater interest or importance to the user. While the three overarching components of the theory appear fairly simple, there is some complexity in how the system accomplishes these tasks—especially in personalizing the results to the user's interest.

3.4 Defining the Function of the GPA Change Detection Theory

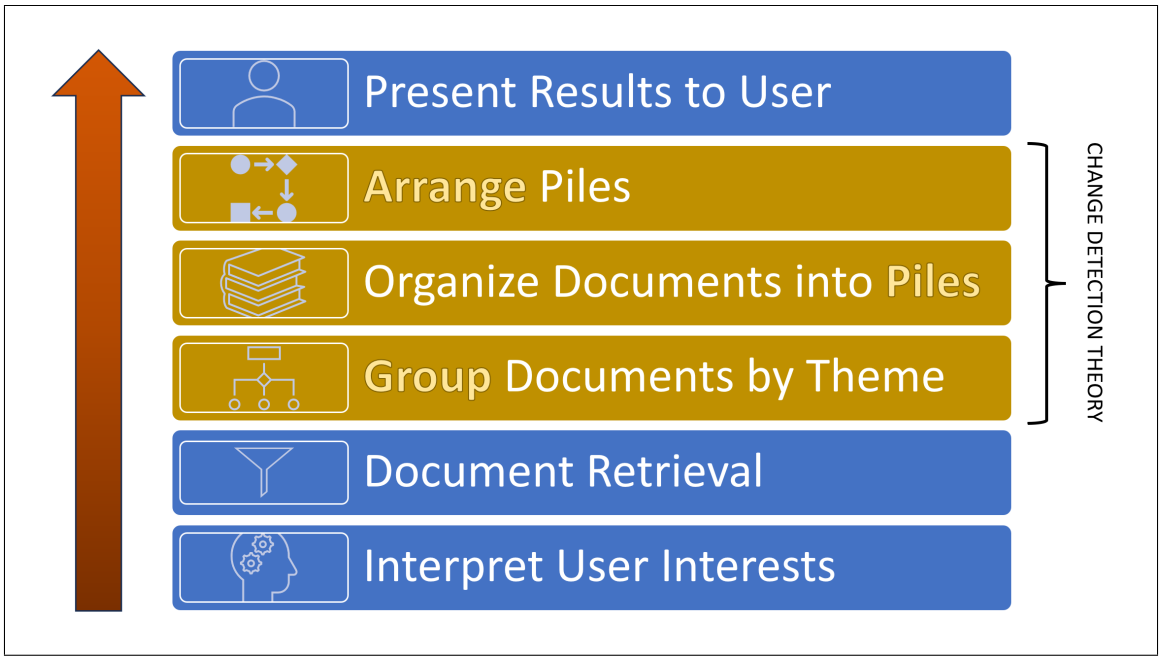


Figure 3.3: Overview of the steps within the change detection theory, organized from interpreting user interests (bottom) to presenting results to users (top).

To support the core of the GPA Change Detection Theory, information is needed from the user regarding their topic of interest. The GPA function organizes new documents and presents them to the user; the user interacts with the documents, indicating subtopics of interest. At times, information is externalized for use in various ways—presented in an article, discussed in a conversation, or used as a seed for further searches. Figure 3.4 expands on earlier graphics to highlight the role of input and output data relative to the change detection function.

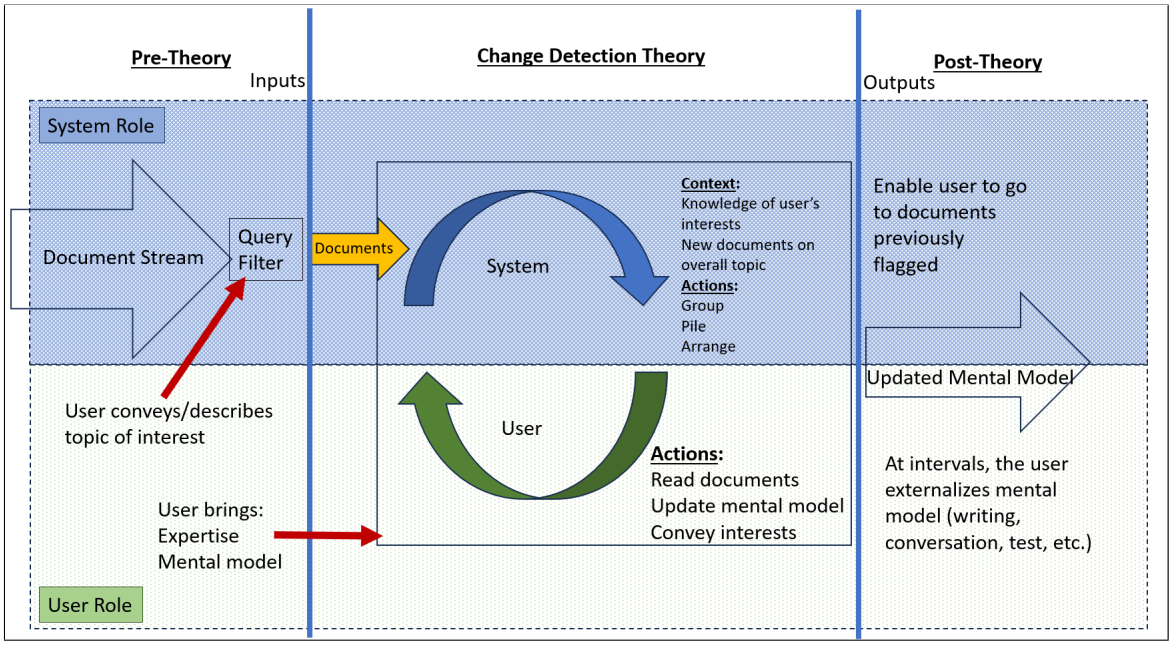


Figure 3.4: Role of the system and user before, during, and after the change detection process.

3.4.1 Inputs: What the User Provides to the GPA Function

For the GPA function to work, the user somehow must convey their interest in a topic to the system. Depending on the desired implementation approach, this can be accomplished by holding conversations (in cases where there is a human-driven system, such as SDI), by generating user-defined query terms (if it is a computer system), or (for either case) by providing exemplar documents that represent the user's topic. This information is passed into the function in the form of a query, which is used to filter the system's broader stream of documents into a subset that is related to the topic of interest.

3.4.2 The Steps of the GPA Function

We divide the change detection components into two parts: user actions and system actions. User actions include such things as reading a document, externalizing an inter-

est, or updating a mental model—"An internal representation having—in some abstract sense—the same structure as the aspect or portion of external reality that it represents."[2] The system actions include performing the GPA function on documents and maintaining queries.

We start our description of the core of the change detection theory by focusing on a user's interaction with a single document. What is that experience? How does the content affect the user's mental model? The user expresses interest in some of the documents. From there we step outward to understand how these particular documents made it to the user through the query, the Group-Pile-Arrange process, and the query refinement process.

Querying for Relevant Information

At the beginning of the change detection cycle, the system queries available documents to filter down to the ones that are related to the user's topic of interest. Where did the documents originate? The initial set of documents was identified based upon the user's topic, which may have been conveyed to a computer or to a person assisting with the research. Expanding to the system view, the system has an awareness of what is available across the latest set of documents within the collection, and must determine which documents to convey to the user. The first step is for the system to run a query based on the user's topic of interest in order to produce the relevant document set for that session. To create a common understanding of the topic between the user and the system, the system maintains a query that combines the initial topic plus information from user actions that further refine their topic of interest. This can be accomplished by providing the user with a mechanism for noting documents of interest, as well as for highlighting specific snippets within the docu-

[2]Definition from https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100150482

ment that were of interest. The user could also indicate general subtopics of interest within a document.

User Encounters a Document

One of the first actions performed by a user is the viewing of a document. The user has some sense that the document is related to their topic of interest. The user reads through the document, attempting to identify something new that can be aligned with or used to update their mental model. Some of the documents viewed by the user may only tangentially align with the user's goals, even if they are on the topic. Why is this the case? The document may have been created because it was required by some organization, due to personal interest, or even generated automatically by an algorithm. From the perspective of the user, one document might be full of relevant, useful information, whereas another document only contains a sentence that ties to the user's interests or expands upon information in their mental model. Two different users might look at the same document, and depending on their interests, may find completely different things interesting.

User Interacts with Multiple Documents

We have started by looking at a specific user interaction—beginning at the document level, and identifying documents and snippets of interest. We now expand outward to discuss the process that led to the documents presented to the user by the system. First, we note that the system provides information that arrived since the last time the user looked—whether daily, weekly, or some other periodicity.

3.4.3 GPA: Core Steps of the Change Detection Theory

What makes the change detection system unique is what happens after the query: the GPA function itself. In order to enable the user to more easily identify what has changed on their topic of interest, documents are grouped by subtopic, sorted into piles (superdocuments), and then the piles are arranged in some way. Here we detail these processes in more depth.

Step 1: Group - Documents Grouped to Align with Mental Models

The first step in the process involves organizing documents into groups based on the relationship of the documents to the subtopics externalized by the user while they reviewed individual documents. A single document could be assigned to multiple groups, if it contains information that is relevant to multiple subtopics.

Given that the user does not know everything about the topic—or what it might be in the future—the system should include the ability to identify the emergence of new subtopics. That is, there must be some way for the user to view documents on topic that do not fall into any of the user's existing or pre-identified subtopics. For instance, if a new baseball player joins a team, there should be a way for the user to encounter this information; it should not be ignored or eliminated simply because it does not align with subtopics from prior sessions. At the very least, the system should maintain a separate group for documents relevant to the topic, but which do not align with the other groups.

Step 2: Pile - Documents Organized into Superdocuments

When the pile phase begins, we have unordered sets of documents that cover specific subtopics. We make a simplifying assumption here that the documents are independent,

that this sorted set of documents is viewed in a linear manner, and that there is contiguity in the user's workflow of looking at document after document in the pile. Within this theory, we refer to the organized pile as a superdocument. The superdocument includes component documents that are arranged in a way designed to help the user focus on content and understand the changes related to the subtopic, while also providing the flexibility to move within a specific pile, as needed. With this arrangement of piles, the user experiences the set of documents about the subtopic as they might with a book, where individual documents (like chapters) contribute to some broader story. An author has organized the book with an intentional flow; if desired, a user generally follows the flow of the book, but could skip between chapters in an order other than the author intended.

Step 3: Arrange - Organize the Piles Based on Importance

In addition to having a pile, we must also have a label for the pile. This label should represent the externalization of the user's mental model through tagging or other approaches. The pile labels are helpful for scanning; looking over the range of available subtopic labels can indicate to the user about what areas have changed, and what ideas they may encounter as they review the documents. This is an important concept as the change detection system performs the third step, of arranging the piles.

Because not all groups matter equally to the user, there needs to be some organizing of the piles to align to the user's interests. As an example, in a linear approach the system could first provide the user with the piles that are likely of greater interest to the user, to provide a logical starting point. The ideal order for this approach would be one in which the superdocuments presented to the user first are about the subtopics that are most important to the user. This concept aligns with the idea of subtopic (aspect) importance—identifying the higher-interest piles. The goal is to arrange all of the piles, though the user's interest level in

some piles may be roughly equal. In these cases, order could be assigned arbitrarily based on alphabetizing the list of labels, chronological order based on most recent document, or other techniques.

3.4.4 Outputs from Change Detection Process

As the GPA function proceeds, the user may have a need to externalize parts of their mental model, whether formally or informally. This could include telling stories, writing articles, or other output. As the system keeps track of information needed to understand the user's interest and organize data, this same data could support the user's ability to revisit documents for use in writing an article or summary, remembering some key event, and other tasks.

As we have emphasized throughout this dissertation, change detection is an individual process through which a user updates their knowledge of a topic. This step is important for solidifying the user's understanding of changes related to the topic, which can then feed into the potential output areas. One logical follow-on to the user's learning process could be a collective process, through which a group or team looking at the same topic—or different angles of a topic or similar topics—collaborates to build a shared understanding.

3.5 Discussion: Ideas Supporting the GPA Change Detection Function

Now that we have introduced the broad concepts of the GPA change detection function, we will delve deeper into foundational literature and concepts that align with each step. This section explains the rationale for inclusion of components of the GPA function. We start with the inputs into the GPA function, then move into the core Group-Pile-Arrange steps, and end with the output from the GPA function.

3.5.1 Inputs: User Interests and Mental Models

In preparation for a deeper dive into the change detection theory components, we first explain the initial state of the user's knowledge, and the components that they externalize as inputs into the GPA function. The user has a preexisting interest in a topic to the point of wanting to identify changes related to the topic over time. Due to the user's enduring interest and prior expertise development on the topic, they bring to the change detection task an existing mental model.

Developing an Interest in a Topic

At the core of this change detection theory is the idea that the user focuses on one or more specific areas of interest, and that the user continues to seek updates on this topic over a long period of time. The process starts with an individual focusing on a topic. What is a topic? A dictionary definition of a topic is "a subject that is written about, discussed, or studied."[3] For the change detection system, the user will need to convey the topic, possibly using a short phrase or sentence. Aspects of the topic of interest may change, yet the core area of interest remains stable over time. For instance, a user who is interested in a baseball team (e.g., the Washington Nationals) may have interest in different things that are happening—summaries of recent games, the team's chances of entering the playoffs, or expected trades in the off-season.

In order for change detection to be relevant to a user, they must be interested in one or more topics to the point that they regularly seek updates. Outside of a general discussion of motivations and externalizations of mental models, this research does not attempt to address in detail the driving force behind users' interest in the topic—for instance, whether

[3]Definition from
https://dictionary.cambridge.org/us/dictionary/english/topic

individuals are performing change detection tasks for professional or personal purposes. We are addressing a more generalized view of this use case, looking at characteristics of a system that would meet a variety of needs ranging from work to hobbies. We leave the question of distinguishing work-focused from hobby use cases to future research. That said, as Sir Arthur C. Clarke quipped in his short story "The Man who Ploughed the Sea," "Harry had long ago discovered that a considerable number of Americans put quite as much effort into their hobbies as into their professions."[4]

How does a user get their initial spark of interest in a topic? In exploring this spark, Loewenstein provides a psychological definition of curiosity, known as the "gap theory" [166]. Under this definition, curiosity is the identification of a gap in knowledge, which causes an individual to feel a sense of deprivation; this leads to a need to fill that knowledge gap. Once the curiosity gap is filled, why does the interest persist? Harackiewicz looks at the combination of individual interest (someone finds a topic or activity interesting) and situational interest (the individual is taking a class or performing a task related to the topic). In some cases, the user takes these interests with them to other contexts for the longer-term; these persistent interests are referred to as enduring interests [109].

Hidi described a 4-stage model of interest development: first, the individual's situational interest is triggered through an exposure to, for example, something surprising, new, or with some personal connection. Second, the individual maintains that situational interest. Third, an individual interest in the topic begins to emerge, meaning that the individual begins to seek out information on the topic. Here the topic is starting to give signs of being an enduring interest for the individual. Fourth, the user has a well-developed individual interest, and actively engages with information—choosing to learn more about the topic. For example, this process is particularly relevant to the field of education, as teachers work to inspire interest in their students [116, 117].

[4]Clarke, Tales from the White Hart, 1940

The user's enduring interest is an "information need"—defined by Taylor as the the "conscious and unconscious need for information not existing in the remembered existence of the investigator" [263]. In this case, the user is regularly devoting a certain amount of time to focus on their topic. They have a willingness to keep looking at the developments related to the topic, and each time they do so, they are practicing their skill. This is not only about learning the topic, but also crafting techniques for finding information, including operating more efficiently and quickly recognizing important changes.

Motivations to Continue Following a Topic

A user engaged in change detection may have extrinsic or intrinsic motivations for keeping up their interest; this likely is about more than simply learning and knowing things. The user may have some specific application in mind for their acquired knowledge. Some of these motivations may relate to what they want to do with the information later. Users may operationalize these change detection concepts in different ways. For example, outcomes from their information discoveries could include identifying opportunities for financial investments, conveying information in an article, leveraging knowledge in a game of trivia, or demonstrating comparable knowledge during a conversation with other fans of a sports team. While we do not distinguish between professional and casual use cases, we do look at potential motivators for their change detection activities.

We can view the user's motivations from the perspective of "carrot" (positive incentives) vs. "stick" (negative incentives). The user wants to be up to speed on key developments, and not missing out on aspects that may matter in some way. Carrot motivations for change detection could include:

- **Seeking an information advantage.** By obtaining information before someone else, a user could profit from the information—a financial incentive for finding informa-

tion. Akerlof famously described this information asymmetry in terms of people willing to purchase a car. Someone who knows the actual condition of the car has an advantage over people who do not know the current state, whose monetary offers may be over or under the actual value of the car [3]. Similarly, someone who has more information about the current situation of a company might give them an advantage for stock purchases or sales.

- **Identifying rare information.** In physical and informational situations, people have an interest in things that are scarce. Even if they are not profiting from it, knowing information that is not known by others can have an advantage [307]. In research about the application of information theory and Shannon's entropy to economics, Chen noted, "the value of information is inversely related to the number of people who understand it" [53].

- **For the joy of knowing something.** For example, curiosity about learning interesting new facts can stimulate the reward center in the brain [137]. Some users may want to learn more unusual information because they enjoy knowing trivia, or want to use it in some sort of competitive setting [128].

- **Identify opportunities to combine knowledge.** Experts with different pools of knowledge are able to bring them together for an information advantage, such as the creation of a new, innovative capability [238].

In contrast, individuals may engage in change detection because they wish to avoid negative situations. Drawing from ideas in behavioral economics, an individual may disproportionately avoid a costly risk, even when an opportunity may exist in that space [212]. "Stick" (negative) motivations for change detection include:

- **Missing information:** Some users feel a concern about being aware of information

that other people have seen or are discussing; this concept is sometimes referred to as a Fear of Missing Out (FOMO) in a social media context. This could include missing out on knowing something (e.g., what your friends are doing on social media) as well as more substantial issues, such as missing an opportunity to attend an event. Many people indicate that they are motivated by a fear of missing out for a variety of reasons, including missing opportunities that are better than the ones they currently are taking [11, 73]. Additionally, there may be an impact if they are not aware of some piece of information on their topic that others already know. For instance, a common form of risk aversion is that people do not want to miss out on financial opportunities [16]. In the field of behavioral economics, researchers have found that risk avoidance or loss aversion can be more of a motivator than reward seeking [267]. This could extend to avoiding situations where someone has an information disadvantage, caused by knowing less than someone else.

- **Avoiding making incorrect decisions.** Businesses and other organizations need to understand competitors and the current state of their field in order to make good decisions. As Kahaner described, having information can enable an organization to make informed decisions, which can contribute to success [133]. Without competitive information, an organization may make suboptimal decisions, leading to potential failure. Maungwa found that root causes of business failures included issues such as an inability to articulate information needs effectively and poor information seeking practices—especially in cases where searches were done by proxy, on behalf of users [177].

Expertise Development

Once the user has an interest and motivation to keep up with the topic, they begin to develop their expertise. One of the most familiar models of expertise revolves around time: many have heard that it takes "10,000 hours of practice" to become an expert on a specific topic. This originates from research by Ericsson et al., in which they estimated that an expert had experience consisting of 20 hours a week, 50 weeks a year for 10 years (10,000 hours) [78]. Others have estimated that expertise is based less on number of hours and more on the quality of the practice—specifically, the level of focus applied to the task [97]. An expert's deep knowledge about their topic enables them to perceive meaningful patterns in their domain more quickly than novices. Analysis of short and long-term memory shows that experts represent problems at a deeper level than novices. The time they have spent analyzing their topic makes them aware of issues and gaps when they come across new relevant material [54].

Collins looked at various models of expertise, finding that two primary types of expertise are in ways of thinking (conceptual tasks) and ways of doing (physical tasks). Expertise in a domain comes from, among other things, building up knowledge and, as Collins stated, "becoming embedded in the social life of the domain" [59]. A topic is one example of such a domain—in this case, a domain focused on knowing things, as opposed to doing things. This depth of knowledge allows the expert to recognize divergence from existing patterns. Drabenstott found distinctions between the ways that domain experts and non-experts applied information seeking strategies. Domain users were able to apply more sophisticated search strategies and applied scaffolding that was not available to the novice users [74]. A variety of strategies can help to construct and explain knowledge from multiple documents, including identifying corroborating information, interpreting sources, and comparing information [24]

Expertise is not only about knowing things, it also relates to what people do with that information. Tetlock examined expertise and expert behaviors from a number of angles. He found that labeling someone an "expert" can be misleading; experts frequently make incorrect judgments and predictions [265]. Still, some experts are disproportionately effective at making predictions or forecasts based on the information they have gathered. Tetlock refers to these individuals as "superforecasters," and describes approaches that individuals and groups can apply to improve their use of information [266].

The User's Mental Model

The user has existing expertise on their topic of interest, and that expertise is growing. When the user encounters new information, it is stored in the brain as part of a mental model. This concept, foundational to the change detection theory, is derived from the field of psychology. An early forerunner of the mental model concept was first introduced in 1943 by Craik, who theorized that the brain uses analogies and neural models to describe the world [66]. Philip Johnson-Laird published a foundational description of mental models in his 1983 book on the topic. He described two distinct levels of mental models: the first level includes the working models that people construct to represent the world. The second level is the one that cognitive researchers construct to attempt to explain mental activity [131, 171]. For the change detection use case we are focused on the first level: the internal models that people construct to understand aspects of the world.

Since the initial idea was documented, the research around mental models has expanded significantly, to include research into the processes through which mental models are created and stored in the brain. In a survey of literature around mental models, Treur describes studies that aligned mental model components with specific brain regions. Not only did

prior studies show brain mappings for conceptual and spatial concepts, they also addressed mappings of feelings and emotions [272].

Externalizing Mental Models

Rather than using a complex depiction of brain mappings and interactions for the change detection theory, we simplify the concept of mental models by using the analogy of a concept map as an externalization of a user's mental model. A concept map is similar to a mind map; mind mapping is an example of a technique that people use to break a topic down into subtopics. Introduced on a British television program in 1976, mind mapping provides a visual method for displaying aspects of the topic [46]. Initially described in the 1980s, a concept map builds upon the idea of the mental model and adds the ability to depict additional interconnections within and across subtopics [274]. In particular, here we are focusing on depicting a segment of the mental model that is specific to the change detection topic of interest, as opposed to the person's entire mental model.

How do people describe the components of their mental model? The fields of cognitive psychology and linguistics include a concept called the "basic level" that refers to the subcategory level that users find the most understandable for a specific scenario. For example, when we point to a rolling desk chair and ask someone to assign a label to it, many choose "chair." Why is this? How do humans have an ability to select a label that is generally understandable? Rosch describes "basic level" categories through which humans understand and generate labels for perceived objects. At this level, the description term "carr[ies] the most information... and are, thus, the most differentiated from one another." Categories one level higher in abstraction (more general, e.g., "furniture") are known as "superordinate categories," and categories lower in abstraction (more specific, e.g., "desk chair") are called "subordinate categories" [174, 221].

For topics and subtopics relating to change detection tasks, the system needs to know what the user considers to be the right level of detail for the subtopics associated with the topic. That is, while subtopics may fade in or out or otherwise morph over time, they need to be able to search for information using the right level of specificity to provide relevant information. For example, the user's topic may fall at the subordinate level—not just "sports" (superordinate) but "baseball" (basic level), and not just "baseball" but "the Washington Nationals" (subordinate). Within the topic we would anticipate further subtopics based on events and entities relating to the broader topic. The subtopics are not the full story, but they can provide pointers to interesting or useful items, allowing the user to review specific documents to get updates.

Initializing Use of the System: The Cold Start Problem

When the user first comes to the system, we have a cold start problem. At the beginning, the user does not know much about the system, and the system does not know the user, their topic of interest, or their mental models. The cold start problem is related to what Rashid, et al. described in recommender systems: a system designer has a number of options to gain insights into the user's needs, including random, popularity-based, and other strategies. Each approach represents a guess about what the user might find interesting, and is confirmed or refined as the user encounters and reacts to the data [213].

3.5.2 User Interaction with Documents

In order to initiate the group-pile-arrange steps, we include a discussion of the interactions that a user has with documents. We initiate this with a description of the query and query refinement processes, before covering the user's interactions with the retrieved documents.

To better meet the user needs over time, the user and system must exchange information about system functionality and user interests. This initiates the process of understanding the user need. Over time, the system continues to refine its ability to represent the user's needs, and adapt the document filtering approach to meet those needs.

Document Retrieval Cycle Supporting Change Detection

The change detection topic is represented as a standing query; we expect that it is fairly stable, though the query may need to be refined over time to ensure coverage of the current state of the topic of interest. Ideally, the system can refine the query based on the users' actions, and minimize the need for the user to adjust the query manually. From there, the system uses information from the user to determine how best to organize the query results for presentation to the user. There may be too many documents to view one by one, so there is a need for the the system to provide some ordering to give the user an indication of where to start.

To what extent should a system be able to convey the results, as opposed to the explanation for why those results were selected? Tied into this idea of a system supporting a change detection need is the concept of uncertainty absorption: when summaries or subsets of information are provided, there is a reliance on the judgment of whoever created the subset, and the information used to generate it is likely not conveyed [173]. For the change detection system, we expect that the system would provide relevant results, but a complete explanation as to why certain documents were included within or excluded from a superdocument may not be easily conveyed. Users would likely seek transparency and explainability of the results set, to the extent possible.

Learning and Evolving a Topic of Interest

We explore how users initially learn about a topic and continue to study it, eventually reaching a point of expertise. Learning and expertise development play a key role in the change detection process.

Bates discusses the transformation of raw information (Information 1: "a pattern of organization of matter and energy") into knowledge (Information 2) through the assignment of meaning. It is through this process of categorizing and understanding that people become informed [28]. People follow steps to comprehend a topic, track updates, and expand their understanding of the topic [25, 31, 192]. Zhang looked at sensemaking from an individual perspective, framing it from the perspectives of learning and cognition. The process is iterative in nature, as the user alternates between sensing and sensemaking [304].

Learning from a new document is not a standalone process; as an individual encounters new information, the process of reading activates information learned from prior texts read by the individual. The end result could be an integration of the new information into a broader representation—the mental model [29]. When learning about a topic, an individual identifies patterns (subtopics) within a topic. When new information appears, it allows them to compare the new information to prior information and make inferences and associations to update their mental model [37]. This could include new information about an existing subtopic or new subtopics.

As the user becomes more familiar with their topic, there may be a decrease in the amount of time required to recognize new developments and add them to their mental model. VanLehn provides a framework through which people learn and practice cognitive tasks. In the framework's final phase of learning, individuals expand beyond accuracy; as practice increases, their performance of the cognitive task also increases in speed [282].

The fact that the change detection task is repeated over time can be thought of not only

as a way to build knowledge of the topic, but to build strength relative to the information seeking process. The user is regularly practicing their approach for spotting new potential information. The user and system both improve their abilities over time, as the user improves their understanding of the topic. The system also learns from the user's behavior. Cockburn, et al. adapted this idea to computer system interfaces, discussing how deliberate practice by users in consistent interfaces—for instance, to begin to transition the novice into an expert, they can perform the task in the exact way an expert would (e.g., not a simplified version of the task), even if they initially perform the task slowly [58].

User Expresses Interest in Document Contents

As the user reviews documents, they identify items of interest within the set. The user can indicate their interest through implicit or explicit approaches. Implicit indicators include spending a long time reading a specific document or returning to the same document multiple times. Explicit indicators are ones through which the user actively expresses their interest: tagging a document with a label, highlighting a useful paragraph, and similar actions.

As the user reveals their interest in specific documents or snippets to the system, this could eventually be leveraged in multiple ways. The user could be highlighting information due to an interest in returning to it later as reference material for some externalized output. Additionally, the highlighted or flagged information can serve as a source for the system itself to learn the user's interests and refine future result sets.

Conversely, the user could note items that are not of interest—for example, something that has similar terms to their topic of interest, but is in actuality unrelated. This could serve as a set of exclusions for tuning document results in future change detection sessions.

While query reformulation might be a more appropriate method for refining the search, this could serve as another signal as to which documents are not relevant to the user's interests.

3.5.3 Concepts Supporting the GPA Function Components

In this section we explain workings of the GPA function, and underlying concepts. This includes grouping the documents in alignment with the user's mental model, organizing the documents within a subtopic into piles (superdocuments), and arranging the piles by subtopic importance.

Step 1: Group - Aligning Documents with the User's Mental Model

Why do people group things together? This can be a simplifying approach to help identify aspects or subtopics. We differentiate this concept from identifying exact duplicates. In the GPA function, we are grouping like content based on their relationship to the concepts within the user's mental models, and do not want multiple copies of the same exact document to be present. What we do want are a grouping of documents that each represent some unique view of the subtopic, based on the perspective of the document's originator or author. Grouping can also aid in understanding and interpreting the utility of results, as described by Oard and Resnik in [188]. In a cross-language search setting, they applied clustering to support document selection as well as to aid in refining queries.

It may be appropriate to account for creation of groups of documents that are not of interest. Ultimately, query refinement may serve as a better approach for removing irrelevant or uninteresting information; however, some users may set up subtopics to obscure content that they do not want to see into defined groups, at least on a temporary basis. We note that this could include subtopics that are relevant to the query, but of little to no interest to the user. For instance, a user may want to see information about trades related to a baseball

team, but wants to ignore human interest stories related to specific players on the baseball team.

Step 2: Pile - Creating Superdocuments with a Logical Flow

Research into user focus has shown that users' ability to perform tasks declines as users switch between tasks [223]. Studies have shown that users are able to retain more information when they focus on one document at a time [23]. Thus, to promote focus and reduce distraction, within each group we want to achieve a state where individual documents are presented to the user one at a time in some order. The goal of this phase is to organize the documents within the groups into piles.

It is especially important that there be some logical flow for the documents within the superdocument. The specific organization applied across these related documents should enable the user to understand what has changed with respect to that subtopic. In order for a reader to extract the underlying meaning from text, it is helpful to have some sort of underlying coherence to tie the text together—the ideas are arranged in a way that is meaningful to the reader [102]. Zwaan, et al. discussed temporal, spatial, and causal organizational structures for organizing texts [310]. In a later paper, Zwaan, et al. dove deeper into chronologically organized texts, and how this structure serves as a sort of default order. While other ordering is possible, a chronological organizational approach is easily understood, and aligns with the order in which we experience reality [309]. Kelter, et al. found that readers spent more time attempting to process new information when the gap between the time of the two represented events increased. By reducing the temporal difference, users were able to process the information more rapidly [141]. We recommend chronological ordering as a potential default approach for organizing superdocuments, recognizing that there should be flexibility to apply other approaches, such as relevance-based sorting.

The concept of a flow also applies in another way: The concept of flow as in getting into a state of focus or concentration. Hoffman and Novak studied this concept in an online setting, focusing on identifying when individuals achieve a state of flow in online systems and virtual communities [119]. For the change detection task a state of flow could aid the user in focusing on information and identifying changes.

Step 3: Arrange - Present the Most Important Superdocuments First

What are the ways that a system could define or operationalize subtopic importance? The user could convey what aspects are of greatest interest to the user. Alternatively, other approaches could be implemented. As an example, rarity could be used as a simplifying factor representing subtopic importance in a linear implementations. That is, in cases where a user wants information advantage, they may benefit from knowing something about a subtopic that is not commonly known, or not found by other people reviewing the topic.

Tying back to the concept from behavioral economics about loss aversion, we want to make sure that the user sees "enough" (as they define it) documents. At the very least, there is an importance of seeing the labels for the piles, which would provide an understanding of how the current set of superdocuments map to their preexisting mental models. This becomes a sort of terrain mapping exercise for the session, where the user sees the range of subtopics covered across the currently available set of piles. This allows them to gain some understanding of the current state of the topic, even if they do not view documents related to all aspects—for instance, if they only review superdocuments on subtopics that they consider more important.

We now explore the horizon scanning behavior of the user in more detail, through which the user reviews subtopic labels to get a sense for the range of labels available for that day. This step helps to satisfy the user's need to understand the range of changes within the topic.

In contrast, when considering the superdocuments for each subtopic, the user sometimes "satisfice" their need rather than satisfying it. Simon described the economic concept of "satisficing" a need as being when an individual or organization seeks only to reach some satisfactory point before halting, rather than taking the time to satisfy the need by finding the best possible answer or option [244].

Within groups, there may be information that reinforces information found elsewhere, as well as information that duplicates content from other documents. For instance, if there are six documents and all say that a particular baseball team won (and not much else), it is unlikely that the user needs to read them all. There is a law of diminishing returns for these reinforcing documents—seeing one more document that only discusses the score of last night's baseball game may not be useful, unless it introduces a new angle. That case could be addressed by having documents appear in multiple subtopics, or highlighting a specific area within the document that is new or of interest.

Spink evaluated qualitative and quantitative criteria that researchers use to decide when to cease information discovery activities. Factors included the amount of time available for the discovery process and belief that the user has sufficient information to complete the task [209]. We see these as key factors that drive users engaged in change detection to decide how much information to review—they halt when they believe they have identified as many relevant changes in subtopics as they can, in the time available for this activity.

Based on the user's interaction with the organized piles, the system refines its understanding of the user's interests. The topic may shift or evolve over time, resulting in the need to refine the query producing the result set. As Bothma observed, the "world," or "broad context in which an information need occurs," may remain constant until a natural or human intervention alters the need [34]. At that point the user or the system might need to adjust the query or filter to ensure that the documents produced are relevant to the user.

3.6 Expanding on the GPA Change Detection Theory

The GPA Change Detection Theory in its current format was designed as a foundation for the research covered within this dissertation. For broader application, the theory could be expanded to address a broader range of system designs. In this section we address the current scope of the GPA Change Detection theory, to indicate potential areas for enhancement of the theory, external concepts that could build upon the theory, and research spaces that are adjacent to the theory.

3.6.1 Enhancement of GPA Change Detection Theory

To enable the theory to apply to a wider range of system implementations, the following adjustments are recommended:

- **Additional data structures:** The theory presented here largely focuses on data structures involving lists, which is the organization approach used throughout this dissertation. We recognize that other presentation options may also be helpful for presenting results to users; this includes trees, maps, and visual representations of clusters of documents. A future enhancement to this theory could more generally include concepts such as these.

- **Topic flow:** At present, the theory treats incoming documents as a steady stream from day to day. Further research could address whether variations are needed in cases when topics develop more slowly, and represent more of a trickle than a stream.

3.6.2 Building Upon the Change Detection Theory

Here we introduce a number of areas that we consider as external to the GPA Change Detection theory, that could provide future branches of research:

- **Multi-language content:** The theory currently assumes that information flows to the user in a format that they can comprehend. While not part of the theory, additional ties could be made to the upstream task of getting content into one or more languages that make it more easily understandable by the user.

- **Professional vs. personal use:** The theory has been generalized to support both professional and hobby or other personal change detection needs. Follow-on research could look at distinctions between these cases, perhaps resulting in specialization to support either case more specifically.

- **Individual vs. collective change detection tasks:** This theory intentionally focuses on an individual's mental model development process. Change detection focuses on an individual's construction of knowledge, and systems that can provide highly personalized results. This is a different process than the one a group would undertake to align mental models and construct a collective approach for understanding a topic. While not part of this theory, we see collective change detection tasks as an important (albeit separate) potential research area. Connected research could look at how groups work together to discover information. We acknowledge that sensemaking can also be an organizational construct—people collaborating to understand or make sense of a change or action [112]. However, the goal of the change detection theory remains the support of individual rather than group information seeking activities. We leave the research on how groups or teams create and adapt their collective information seeking approaches for future research that can be connected to the individual processes described by this theory.

3.6.3 Adjacent Use Cases

We now distinguish the change detection process from other information discovery and information retrieval concepts that may appear to be similar. Some of these activities can be initiated while the user is performing change detection tasks, or come after change detection tasks; regardless, they are separate use cases requiring different information retrieval approaches.

- **Exploratory Search:** The intent behind exploratory search is to investigate a topic that is new to the user, to learn about it and reduce uncertainties and knowledge gaps [294]. Exploratory searches tend to start out with general search terms, and involve a combination of searching and browsing behaviors. Unlike exploratory search, change detection is focused on a specific topic of interest. In contrast to users who are exploring a topic, users engaged in change detection have existing expertise, and the emphasis is on filling the recent knowledge gap related to an already familiar topic.

- **General news search:** This use case focuses on getting a general understanding of local or global news, or following current events in a broad way. The focus is on browsing the latest news without a specific intent or topic in mind. Systems designed to address this use case tend to focus on features such as novelty, emergence of trending news, content and source diversity, and influence or popularity [88, 106, 136, 273]. While there may be some personalization to ensure that news feeds show at least some topics of probable interest to the user, the emphasis is on breadth of coverage rather than depth on a single topic, which is the purpose of change detection.

- **Serendipitous discovery:** While someone is reviewing documents, they may iden-

tify new, serendipitous connections to another topic area. For instance, while performing change detection tasks related to a baseball team, the user might identify something of interest about a player who may be traded to the team. This could cause the user to initiate separate research into that baseball player's background and skills. In a survey of research on serendipitous information seeking, Foster and Ford found that browsing behaviors could result in serendipitous discoveries [87]. The information seeking activities of users engaged in change detection could lead to accidental or serendipitous discoveries and explorations into tangential topics; however, this is a supplement to change detection and not the activity that is being studied here. The user may discover and explore adjacent concepts based on what they see during their change detection session. They might perform interim research into the area that they observed during the change detection task, either in parallel with change detection tasks or as a separate exploratory session later.

3.7 GPA Function in Practice

Now that we have a framework for the GPA function, and have distinguished the concepts from other, potentially similar-sounding use cases, we introduce the three examples of change detection systems, and how each one aligns with the change detection theory. The change detection theory is not focused on any specific type of system; as a result, in this section we look at three examples of systems that can be supported under the change detection theory. The systems discussed are: first, Selective Dissemination of Information, a manual system of discovering and sharing information; second, retrieval approaches leveraging technologies such as Google Alerts or Really Simple Syndication (RSS) and search engines to compile updates; and third, a hypothetical future system supporting human-

machine teaming. We demonstrate how the theory can be applied to each of the three systems.

We provide in Figure 3.5 an example of the GPA function in use for a topic related to a local baseball team. In this example, the system retrieves new documents that are grouped into five subtopics: Roster, Injuries, Merch (short for "Merchandise"), Owners, and Recap. The groups are then piled into an order—in this case, they are sorted in reverse chronological order. Finally, the piles are arranged to display the ones of greatest interest to the user more prominently. The game recap pile is displayed first, and the Merch subtopic—of least interest to the user—is displayed last.

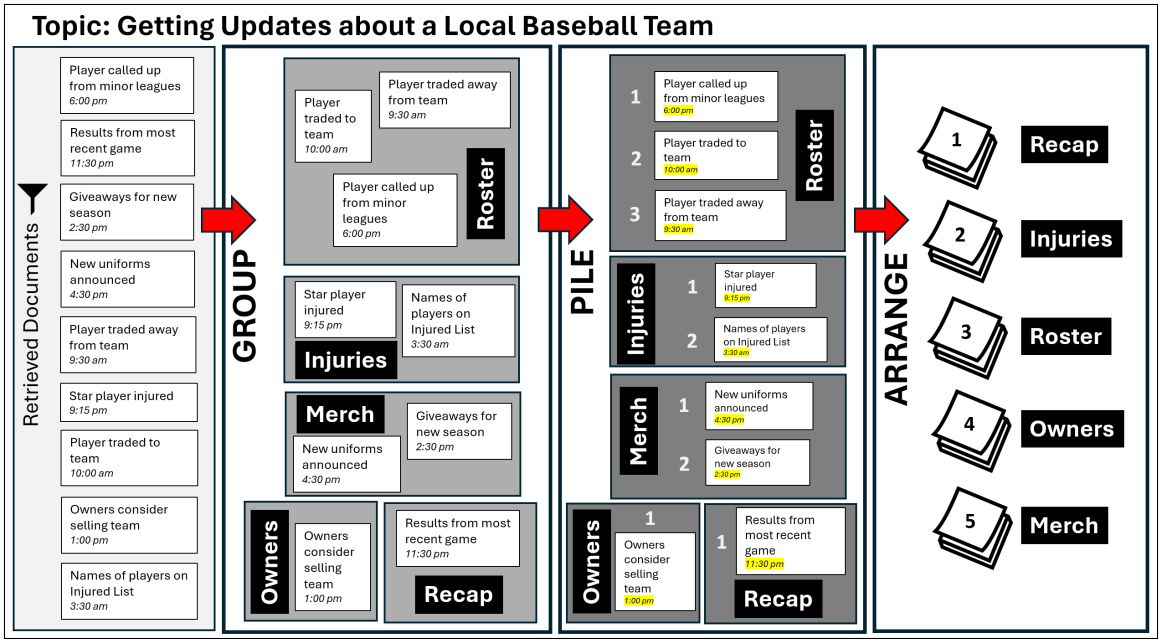


Figure 3.5: Example of the GPA function in use for a topic about a local baseball team. In this example, the system retrieves documents, groups them by themes (e.g., "Roster"), creates sorted piles in reverse chronological order, then arranges the piles to lead with the one of greatest interest to the user ("Recap").

3.7.1 Selective Dissemination of Information

The idea of selective dissemination of information (SDI) is an implementation of the change detection use case that predates modern personal computers, though some pieces of the use case may persist to some degree—for instance, specialized research librarians for Library of Congress. The individual following the topic reads through journal articles, news articles, and other documents related to their topic. They also have conveyed their interest to librarians, colleagues, and others who know their interests. They may document the aspects of the topic that they care about on index cards, as in the example described by Richards, et al. [214].

How does the user improve what information they are receiving? In this case, the process happens through conversations with people, in which they clarify their interests and refine what information is provided to them. By having effective conversations with the people supporting this approach, the documents can be filtered effectively to what matters most to the user. However, a human-driven process may not scale effectively.

Note that this team approach for addressing the user need does not make this a collaborative information seeking process; this is still about one person's change detection-related information seeking behavior, with a change detection process that is supported by other people. The supporting team in this case is aiding with uncertainty absorption, as described earlier—these individuals use whatever processes they need to apply to find the relevant documents that they will provide to the user. However, they likely will not provide full context about their search methods; it is the results themselves that the user wants to see—consistent with the uncertainty absorption concept described earlier.

3.7.2 Current Approach: Bundles of Automated Searches

In this example, users accomplish their change detection tasks with a variety of computer technologies that they assemble manually to provide some overarching view of the topic. Historically, this approach has been implemented with a mix of applications, to include leveraging Google search alerts, manual Web searches, social media feeds, and RSS feeds that funnel new information into a central place for the user. Still, even if it is clunky, users try to recreate a capability to pull information together into a central location. They often curate these items manually, as various developments and changes happen in relation to the topic. As an example, a TechCrunch reporter described his former process of manually assembling hundreds of RSS feeds, with results sorted chronologically, as part of his journalistic update process [69].

Even though this approach can be beneficial in getting topic and subtopic results to the user, it is a manually intensive way to maintain information. As new subtopics emerge, the user may need to create new queries, or remove old ones. Additionally, some of the technology may cease to be supported—for instance, RSS is no longer supported at the same the level it was in the past. As these technologies fall out of favor, it is left to the user to find a new way to find the information they need.

3.7.3 Specialized Computer System

Here we introduce the Daybreak system concept in the context of the GPA Change Detection Theory, a computer system optimized for the change detection use case. Given the capabilities of modern computing, we envision a system that can address the needs of users in a more automated way. This improves upon the manual selective dissemination of information approach in scaling queries. Unlike the bundled search approach, a computer system could reduce the need to manage a large set of queries representing subtopics.

Ideally, the system could emphasize human-machine teaming angles through which the user is learning more about the system, and the system is learning about the user based on actions the user takes to label and organize documents of interest. The technology would serve as more of a partner to the user, endeavoring to understand the user's interests and filter information to them. How might this system improve what information is shared? In addition to meeting the basic group-pile-arrange component of the use case, there could be an ongoing dialog of sorts between the system and the user.

Figure 3.6 depicts components that could be present in a computer system designed to support the change detection use case.

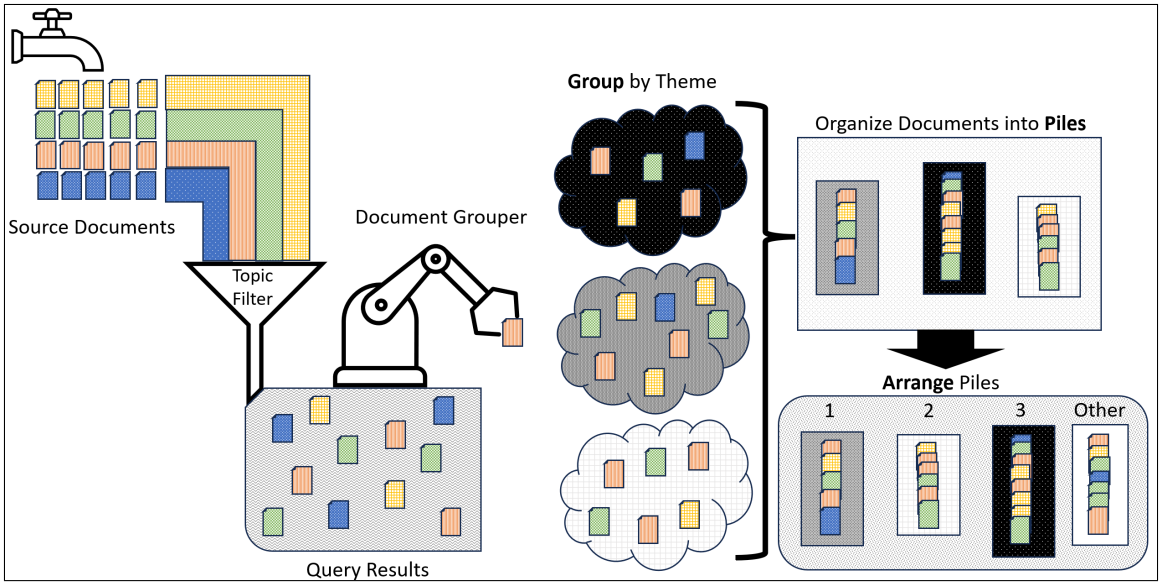


Figure 3.6: Overview of tasks to be performed by a computer system in support of the change detection task.

3.8 Summary

The change detection theory addresses use cases where a user is interested in getting updates on a topic that they follow over time. The components of the change detection

theory include interpreting the user's interests, retrieving documents, then grouping the documents by subtopic, organizing the grouped documents into piles, and arranging the piles for presentation to the user. This theory is intended to address creation of systems—whether manual or computer-based—to enable the user to accomplish the task of getting updates on their topic more quickly. Following this, it is necessary to understand the scope and range of change detection tasks (e.g., is it recognized by users as a task that they perform?) and to test a system design based on the features of the change detection theory to determine whether the system does, in fact, meet users' needs. To start down this path, the next chapter focuses on understanding the change detection task among users of social media, through a survey.

Chapter 4: Sort Order Preferences in Social Media

In this chapter, we start to test the ideas laid out in the GPA Change Detection Theory defined in Chapter 3. One of the initial questions we wanted to address was about the prevalence of change detection tasks. Is this concept something that seemed familiar to many users? How often do they perform these tasks? How do users prefer to have social media posts organized? To answer these questions, we conducted a survey in which we compared respondents' organization preferences for change detection tasks with other social media use cases: change detection (following a topic over time), experiential (following an event as it happens), browsing, and searching.[1] The hypothesis studied was that respondents' views on the most appropriate ordering would vary based on task, and that the respondents focused on change detection would express preferences that align with the GPA Change Detection Theory. In the survey, we focused on the group and pile aspects of the theory.

Survey responses from 188 social media users were analyzed to identify display preferences for social media posts. All four use cases were found to be common, although responses regarding browsing (the most common category) indicated that the respondents engaged in two recognizably distinct types of browsing activity: time-oriented browsing for updates (often focused on friends, family, and news) and general browsing (e.g., for entertaining or humorous content). Respondents who engaged in change detection, experiential, and time-oriented browsing activities expressed strong preferences for chronological sort-

[1]The results of this sort order survey were presented in the 2022 proceedings of the HCI International Conference and published under Springer's Lecture Notes in Computer Science [217].

ing. Those who engaged in general browsing and searching preferred relevant documents first. This chapter discusses the implications of these findings for task-based design of information retrieval systems for these use cases.

4.1 Overview

"Listened to Agatha Christie audio book today but all the chapters played in random order so it turned into more of a mystery than intended."[2] In January 2012, cross-country skier Felicity Aston posted this tweet while chronicling a groundbreaking solo cross-country ski trip across Antarctica. This comment seems humorous because we expect stories to be presented in the "right" order, as laid out by the author. When presented in an order different than what the reader expects, this can cause confusion or angst. And yet this is the experience that many social media users have come to expect when visiting sites such as Twitter and Facebook in search of updates. Instead of following individual narratives, users see posts in an order that an unknown algorithm provides. While the algorithm may be "right" on a granular level (people do respond to the posts that are ranked as highly relevant), the overall interaction can feel disjointed. Social media users often consume content in the order presented, but some users post frustrated comments on social media sites, including asking for the return of additional sort options, such as chronological sort.

We approach this case of users seeking updates on people and topics as being like the act of following a story over time. To the user, a social media feed represents a complex mix of multiple queries; a typical feed includes a combination of people, organizations, groups, and topics, each of which a user follows for a different reason. A user might follow US football teams and players to get information on the current game, friends and family members to get an ongoing sense of what is happening in their lives, or a photography

[2]Tweet available at `https://x.com/felicity_aston/status/156917318293794816`

hashtag to get inspiration for an art project. Which aspect they wish to see during any given social media session will depend on their intent at that moment. For instance, if a football game is taking place, during the event the user might wish to focus primarily on commentary from fellow fans that they follow.

We studied the impact of the type of activity in which a user is engaged on their preferences for organizing posts, focusing on four use cases: change detection (following a topic over a long period of time), experiential (following an event as it happens), browsing, and searching. We were particularly interested in determining whether time-bounded social media behaviors are insufficiently addressed by social media sites. The hypothesis studied is that users' views on the most appropriate ordering would vary based on task, with users preferring clustering and chronological sort ordering for update-related activities such as change detection and experiential use. To understand when one organization might be preferred over another, we surveyed 188 social media users to understand their sort order preferences for specific use cases. The results indicate that users had distinct organization preferences for different use cases.

All four types of activities were found to be fairly common among the set of social media users who responded to the survey, with browsing (both time-oriented and general) being the most popular category; 93% of respondents indicated that they have performed some browsing tasks. Searching was the next most common, being performed by 63% of surveyed users. Change detection followed close behind, with 61% of the respondents following a specific topic over time on social media. Experiential use was the least common, though nearly half of respondents indicated that they have followed a specific event on social media.

The survey results identified some of the characteristics of social media use for each of the analyzed use cases; these insights can help to guide the design of new system capabilities that improve the experience for specific use cases, and to anticipate some of the

challenges that may arise when introducing new capabilities. Further analysis revealed that respondents engaged in two recognizably distinct types of browsing: time-oriented browsing for updates (often on friends, family, and news) and general browsing (e.g., for entertaining or humorous content). This is noteworthy because feeds offered by many social media sites are better suited for general browsing, which was less common among respondents. We also noted that time-oriented browsing shares some characteristics with change detection.

Delving into the grouping aspect of the GPA Change Detection Theory, we asked about both presentation order and about clustering preferences. We included questions about clustering to adjust for users' optimism about their ability to keep up with masses of streaming information. In Present Shock, Rushkoff describes the "quest for digital omniscience," in which users believe they can catch up on what they missed; they actually won't (or more precisely, can't) because of overwhelming volumes [224]. Clustering by account, event, or other alternatives are ways to help users get a sense for what is happening, even if not every individual post is present. Users expressed an interest in applying clustering to update tasks in particular, including for change detection.

Regarding the pile aspect of the GPA Change Detection Theory, respondents preferred chronological ordering for change detection. They also expressed a preference for this sort order for experiential use cases and for time-oriented browsing. Many of the respondents who perform searches prefer relevance-ranked results, though a large percentage also indicated interest in newer posts. For general browsing, respondents wanted relevant or popular posts. For most use cases, respondents appeared to support the idea of clustering; only users in time-oriented browsing preferred no grouping. There was a preference for recall-oriented feeds for the change detection, experiential, and time-oriented browsing use cases; users indicated a fear of missing out on something interesting—they wanted to see all posts related to the topic or new theme.

4.2 Background

According to a 2021 Pew Research Center poll, roughly 70% of the adult population of the United States uses social media [20]. The U.S. Census Bureau estimates that, as of July 2018, there were approximately 332 million people in the United States, 77.7% of whom are adults [276]. This means that as many as 181 million adults are current social media users. Understanding what users want to accomplish in social media can help to identify and meet their needs.

Many social media sites' initial designs used chronological sort orders for displaying users' posts. Over the course of time, many of the popular sites transitioned from chronological sort orders to relevance rankings. Figure 4.1 provides the dates that four of the major social media sites—Facebook, Twitter, Instagram, and Snapchat—made this shift. The result has included negative reactions over time from parts of the user base, directed at sites like Instagram, Facebook, and Twitter. New short-post social media sites have emerged recently, to include Bluesky and Threads; notably, Bluesky has largely maintained reverse chronological ordering for posts [194].

Figure 4.1: Timing of social media moves to proprietary sort orders.

In late 2018, Twitter responded to user comments by broadly reintroducing a capability to resort posts chronologically—though the default remains relevance ranking. Notably, users tend not to switch away from default settings. This phenomenon is related to the concept of "status quo bias" from the field of behavioral economics, which was identified by Samuelson and Zeckhauser. Even when offered an opportunity later to change to an alternative, users often remained with the option they previously selected (or that was selected

for them—for instance, as a default) [232]. Kahneman, et al. expanded on this concept, looking at a variety of fields in which users continue to maintain their status quo—to include in electrical billing and insurance policies, even when switching to a new plan would save them money [135]. Cranor and Wright discussed default settings as they apply to human-computer interaction: "the default settings that inevitably accompany most pieces of software often have great influence on how that software is used." Even in cases where configuration is available, starting with configurable features turned off by default reduces the likelihood that they will ever be used [67].

Highlighting specific content can be useful for maximizing interactions such as click-through rates for individual social media posts, when the goal is to drive users to profitable content. For example, Dujeancourt and Garz studied the impact of Twitter's sort order change on tweets published by German newspapers. Within 30 days of the change, they found significant increase in engagement—with a 20% increase in likes and a 15% increase in retweets. They further found what they described as a "rich-get-richer" dynamic, in which the organizations that received high levels of engagement prior to the algorithm change received even more engagement after the change [75]. Still, this push to highlight fragments of individual content comes at a cost to understanding a broader story.

Researchers have looked at the fragmentation of messages in modern media for many years. In *Amusing Ourselves to Death*, Postman observed that the advent of the telegraph ushered in an era of discontinuous, fragmented messages—a "decontextualized information environment" lacking social or intellectual context [207]. Sadler extended this concept into the social media world, examining the implications of fragmented stories across sites such as Twitter [227].

People use social media to learn, share, or be entertained. Others visit to participate in a "mass shared experience," such as live tweeting a sporting event. In these cases, a device becomes a "second screen," in which the user simultaneously watches a television

show (e.g., the British TV show "The X-Factor") and participates in a worldwide online conversation [165]. Some people visit social media because they are bored or alone. One study demonstrated that pictures on social media sites can help reduce feelings of loneliness [202].

To understand some of the user behaviors in social media, along with associated sort implications, one study looked at learning to rank in social media to predict which posts were most memorable, so these items could be preserved for future reminiscence [185]. Kim, et al. performed a survey to understand students and their use of social media; they found that students' social media activities ranged from academic research to entertainment [146].

Many studies in the field of information retrieval have compared approaches for relevance ranking, focusing on augmenting relevance with such features as recency, document diversity, sentiment, and popularity. We were unable to find studies that address the underlying assumption that relevance ranking is preferable to the user over non-relevance sort approaches, such as chronological orderings of posts. There has also been work on reviewing historical posts to aid in storytelling—for instance, gathering content about a user from their friends' pages to produce a more robust story of their life [228]. However, there seems to be less of a focus on how to meet users' needs as they process pieces of the many stories they encounter in social media. Other research has looked at searching of social media data in crisis situations to verify what is happening and support crisis decision making, using natural language processing techniques to discover and track events and subevents [125].

Research has characterized the cognitive impact of the large volumes of information provided to users as they review social media. Users' divided attention can reduce the likelihood that a user will find, much less act upon, an individual post by a friend [118]. Within only a few seconds, users of sites such as Twitter make decisions about whether a post is of interest. Some of the indicators provided on these sites—for instance, likes and

retweets—may distract the user rather than aiding in focusing their attention on the content [64]. Kim & Sin highlighted the importance of understanding the task that the user is trying to perform so as to offer "purpose-based personalization" [145].

The behavior of users when performing an update task may be repetitive. In the case of change detection, they may run a daily search for the same topic; when following an event, the user may run a search for a term or a hashtag over and over while the event is happening. These repeated queries are like a session search that runs over an extended period of time—days, months, or years. Session search considers queries in an evolving way; users update and refine their searches based on what they find [49]. Similarly, users who are getting updates on social media may adjust their searches for hashtags, people, and other things based on specifics at a given moment.

One might think of the user's process for following updates on social media as a reading comprehension task. Reading is a "constructive act" in which the reader creates mental representations of the material [108]. Given the active, participatory nature of the medium, social media users can be thought of as "reading-to-write." In other words, users often are poised to react to something they view or read (liking, commenting, and more). Flower, et al. note, "the reader as writer is expected to manipulate information and transform it to his or her own purposes" [85]. Put into this context, users' expressions of a desire for chronological ordering may not be surprising. "To achieve a proper understanding of the situation described by a text, the reader needs to know when the described events took place both relative to each other and relative to the time at which they were narrated" [311].

The reverse chronological sort ordering originally used by many social media sites was understandable and explainable. Additionally, temporal organization may be connected to the way that people store information internally. For many years, researchers have thought that the brain uses some temporal mechanisms to store information. A wide range of research supports the idea of temporal organization of memory. For example, Clark and

Bruno studied episodic (long-term) memory by using a variety of information recall tasks; their results demonstrated likely temporal organization of memory, as well as providing some evidence of possible spatial organization [57].

No matter the method used by the brain for interpreting and remembering information, current social media feeds do not appear to be optimized for broad comprehension or recall; instead, the interface appears to favor reactions to individual posts. The burden currently is on the user to determine how a new post connects to an overarching story about a person, group, or topic. Social media adds a further cognitive burden, discussed by Barzillai, et al. Users often weigh credibility when putting the post into a broader context [25].

The rising use of artificial intelligence (AI) has led to a push for explainable algorithms, which has extended into the domain of social media. While the primary focus of document organization in this dissertation is on user preference and understandability, potential negative side effects from other sort or recommendation approaches could include issues related to a lack of transparency. For example, there are concerns that relevance-based recommender systems that are tuned to users' past behavior could create filter bubbles or amplify certain topics; in contrast, a reverse chronological ordering shows all relevant posts from subscribed users or topics. In 2022, the Global Partnership on Artificial Intelligence published a study about potential impacts from recommender systems on the domain of terrorism and violent extremist content [100]. This study highlighted research by Huszár, et al. which revealed the results of a long-term Twitter randomized experiment comparing users who saw relevance ranked tweets with users who remained in reverse chronological sort mode. They found that political amplification occurred under the relevance-based algorithms [123].

Not all social media sites have adopted proprietary sort ordering. Bluesky is adopting an approach that incorporates an "open and diverse market of algorithms," according to Bluesky CEO Jay Graber in a March 2023 blog post [101]. Rather than offering feeds

with proprietary orderings controlled by the company, Bluesky has encouraged their user community to contribute algorithms to aid with the organization of posts.

In an attempt to better understand how social media sites determine what posts to show users, some researchers have attempted to reverse engineer algorithms by analyzing their output to users. Bouchaud, et al. focused on tweets from a set of several hundred users who self-selected for their study, analyzing their Twitter feeds as pulled through a browser plugin. Their study focused on characterizing users (e.g., connections and political orientation) and observing the tweets presented to them. They also analyzed a portion of Twitter's algorithm that was made public in early 2023. Their study was observational in nature, and did not introduce any interventions to test the algorithm. Through their study, they found that Twitter favored new, popular content, and tended to show users content from users with similar viewpoints [35].

Lewandowsky, et al. studied the way that exposure to certain algorithms shapes people, through exposure to certain information. They also raised issues related to gaming of algorithms by "bad-faith actors" to provide problematic content in spite of moderation strategies. They noted the difficulty in adopting gaming-resistant approaches. Even an "exclusively recency-based" (e.g., reverse chronological ordering) order can be gamed; posters could ensure that their messages regularly show up in timelines simply by sending a large number of posts over time [159]. In an attempt to understand how to make recommender algorithms reflect societal values, Stray, et al. collected a set of values related to cross-discipline recommender systems from ethics, ethical system design, and well-being literature. They aimed to provide a framework that maintains emphasis on system utility, while also raising awareness of issues related to recommendation algorithms—to include the complexity of auditing personalized recommendations [256].

Even though concepts like click-thru rate maximization—approaches focused on getting a user to interact with a specific post—have been a driver for many social media ac-

tions, not all systems focus on interactions based on the next user click. For example, companies such as Yahoo! have incorporated long-term measures, to include looking at absence time (the time between users' visits) [76]. Additionally, music and audio site Spotify not only looks at day-to-day or post-by-post interactions ("instantaneous rewards"), but it also explores ways to increase user satisfaction over the long-term. This includes encouraging users to engage with new and different types and genres of content over time. In addition to measuring short-term rewards, they also focus on lifetime value (LTV)—"the sum of cumulative rewards" and long-term survival models (in particular, looking at metrics related to a user's time-to-inactivity) across the user base [51, 52]. While a broader perspective than maximizing interaction with individual pieces of content, this approach does raise additional questions about the extent to which an application should strive to change a user.

4.3 Research Methods

To understand the prevalence of change detection relative to organization preferences for other tasks in social media, we surveyed adult (18 and older) social media users in the United States. People engage in many behaviors on social media; we scoped our study to focus on update tasks—specifically change detection and experiential uses—to help us learn the extent to which respondents are satisfied with how sites enable them to follow a story over time. Even though our primary interest is in the change detection use case, we wanted to be able to contrast change detection respondents' preferences with those of people following a live event. We also asked questions about more general social media use, such as browsing and searching, to allow us to see how similar or different the actions and expectations of respondents are across those scenarios. This is not intended to be a comprehensive study of all user behaviors in social media; our main interest is on tasks

that relate to following a story over time. This helps us understand respondents' social media sort preferences and distinguish whether posts airing frustrations are isolated events (e.g., immediately after an interface change), or if they represent large-scale or long-term concerns.

Surveys are a widely used research method that can be useful for gaining a qualitative understanding of users' attitudes, intent, and information about their experiences across a population [189]. An online survey is a low-cost method to get input from a broad set of users. This approach can also provide insight into which social media systems users associate with specific tasks. Additionally, a survey allows us to gain some insight into why the users prefer a certain approach, and what users perceive as lacking in other social media sites. This is an initial step toward understanding the scope of story-driven update tasks across social media, which can be further investigated in complementary ways, such as with user studies.

4.3.1 Survey Design

We designed a study both to understand use case prevalence and to determine organization preferences for each use case. The survey questions (provided in full in Appendix B) included a mix of closed- and open-ended questions. Our survey on sort orders in social media included five sections. First, respondents viewed a consent page containing an overview of the purpose of the study and information about Institutional Review Board approval. They were then asked to consent to participate. Second, we asked three screening questions to validate that participants were adult social media users located in the United States. Third, we asked demographic questions focusing on location, gender, ethnicity, education, and employment. Fourth, we asked for background on which social media sites the respondents used, how frequently, and which of the four tasks they performed. The

fifth section consisted of sets of questions for each use case, including questions related to components of the GPA Change Detection Theory. Respondents were only required to complete one set of use case questions; if they indicated that they performed change detection tasks, they were preferentially sent to that block of detailed questions first. After answering for one, they could optionally answer questions for additional use cases. Upon completion of the survey, interested respondents could submit their email address for an optional raffle for one of five $25 gift cards. A copy of the IRB approval document is included as Appendix A.

While most of our questions were consistent across the use case sections, we included additional questions specific to the update tasks (change detection and experiential). For example, in the change detection section we asked respondents about their level of expertise on the topic, and the amount of time they have followed it. The questions related to the GPA Change Detection Theory about sort order preference and clustering preference were in all use case sections, to enable later assessment of whether these factors are also desirable in other use cases.

We distributed the survey online from December 2018-February 2019, using a convenience sampling approach. We reached out to potential participants via email and social media sites—specifically, on Twitter, Facebook, and Reddit. This included advertising to users during events with significant online followings, such as the Super Bowl. In addition to social media posts about our survey, we sent targeted posts to specific Twitter and Facebook users who had made comments about social media sort orders. Our email campaign focused on US universities, in particular on programs with an emphasis on communications, journalism, and library and information science.

4.3.2 Research Questions

We had four research questions (RQ) for this study. We believed that the use cases were all prevalent across our respondents, and that there would be a distinct pattern of sort and clustering preferences for each use case. We note that our research questions cover two of the three components of the GPA Change Detection Theory: group, which is noted here as clustering; and pile, which is indicated as sorting. We did not ask the respondents any questions about how they would arrange the clusters.

RQ4.1: How prevalent is change detection? We hypothesized that these use cases are common among social media users. We anticipated that browsing would be the most common, followed by searching. Of the update tasks, we believed that change detection would be more common than experiential use of social media.

RQ4.2: Would users accept clustering as an approach for organizing posts? This gets into the "group" concept from the GPA Change Detection Theory. We thought that respondents would be more interested in clustering for some tasks than others. For change detection and experiential use, we believed respondents would like to have new posts on the topic or event clustered by theme or development. We did not expect that respondents focused on browsing or searching would be as interested in clustering posts by theme, user, etc.

RQ4.3: How do users prefer to have results sorted for a change detection? Related to the "pile" step of the GPA Change Detection Theory, our hypothesis here was that the respondents focused on update tasks would prefer chronological ordering. We thought that respondents engaging in browsing were there for entertainment value and would have no clear sort preference. We hypothesized that respondents engaged in searching would prefer relevance ranking.

RQ4.4: How many posts do respondents feel they need to see when performing

a change detection task? Our hypothesis was that respondents engaged in update tasks would only want to see a few posts per event or development. We anticipated that respondents engaged in browsing would want to see more, and searchers would only want a few posts.

4.4 Survey Results

We received 193 valid responses from the survey participants. Five of these respondents indicated that they did not perform any of the four tasks being studied; we provide the results for the remaining 188 responses in this chapter. Despite the potential sampling bias that can be a factor in surveys, our findings provide a starting point for future studies to build upon. Given that there are approximately 181 million social media users in the US, at a confidence level of 95%, our results would have a margin of error (confidence interval) of 7% if our sampling was random.[3]

4.4.1 Overview of Responses

All four of our use cases were prevalent among the respondents, as detailed in the *Frequency* column in Table 4.1; this is the percentage of the 188 respondents who said they perform the task. Browsing was the most common task, with approximately 96% of respondents indicating that they browse social media without a specific task or goal. RQ4.1 involved understanding how prevalent change detection is relative to each use case within social media. Even accounting for the margin of error for the question of how frequently each use case is performed, it appears that many adult US users perform each

[3]Due to our convenience sampling approach there may be sample bias in which we oversampled for people from certain fields, or individuals who had preexisting strong opinions about sort orders. This caveat applies to all confidence level statements throughout this chapter.

of the tasks studied in the survey. Table 4.1 displays the percentage of the 188 respondents who indicated that they perform each of the four studied tasks.

Use Case	Frequency
Browsing	96%
Change Detection	66%
Experiential	49%
Searching	65%

Table 4.1: Percentage of respondents who perform each task.

Note that not all respondents completed all four of the more detailed use case sections. Participants were required to provide detailed responses for at least one use case section. They were sent to a section containing a task that they indicated they perform, and we preferentially assigned respondents to the change detection section when possible because of our interest in that use case. After completing one section, they were asked whether they wished to answer another section for another task they perform (if they performed multiple tasks).

As we had hoped, we have the tightest confidence intervals for change detection because that use case was disproportionately sampled. Use case sections answered by fewer participants have broader confidence intervals. For the change detection section we received a total of 116 complete responses (Conf. Interval $\pm$9.1%). There were 55 responses for the experiential section (Conf. Interval $\pm$13.2%). 93 respondents completed the Browsing section (Conf. Interval $\pm$10.2%). The searching section received the smallest number of responses, with 31 completing that section (Conf. Interval $\pm$17.6%).

4.4.2 Demographics of Respondents

The survey asked questions about the respondents' age, gender, race and ethnicity, location, highest level of education, current employment, level of computer use, and social media site use.

Age: Most respondents were under age 45. The most common group was 25-34 (32%), followed by 28% for respondents aged 35-44. Another 19% were in the 18-24 age group. 11% of the respondents were in the 45-54 age group, 5% in the 55-64 range, 4% were aged 65-75, and 1% between 75-84.

Gender: We received more responses from women than men. 67% of the respondents were female, while 29% were male. The remaining 4% identified as "other," or chose not to respond.

Race and Ethnicity: Most of the respondents indicated that they were white or Caucasian (77%). Asian respondents (12%) were the next most frequent, followed by black or African American (5%) and Hispanic or Latinx (3%). The remaining 4% selected other or prefer not to answer.

Location: We asked respondents to self-report their location by providing their U.S. zip code. More than half of the respondents were located east of the Mississippi river. The general locations of respondents are displayed geographically in Figure 4.2.

Education: Survey respondents were relatively well-educated. On their highest level of education completed, 39% said they possess Master's Degrees, and 31% indicated that they have Bachelor's Degrees. 9% of respondents have Doctorates, 7% have attended some college but have no degree, 5% possess a professional degree such as a J.D. or M.D., 5% have a high school degree or equivalent, 3% have an Associate's Degree, 1% attended trade or vocational training, and 1% have less than a high school diploma.

Employment: 54% of respondents had full-time employment, and 16% were students. Another 14% had part-time jobs. 6% were self-employed, 4% were unemployed, 2% were retired, and 2% identified as homemakers. The remaining 2% selected "other." The industries employing the largest numbers of respondents were education and libraries, technology, government, and healthcare.

Computer Use: Many respondents indicated that they are online for significant time

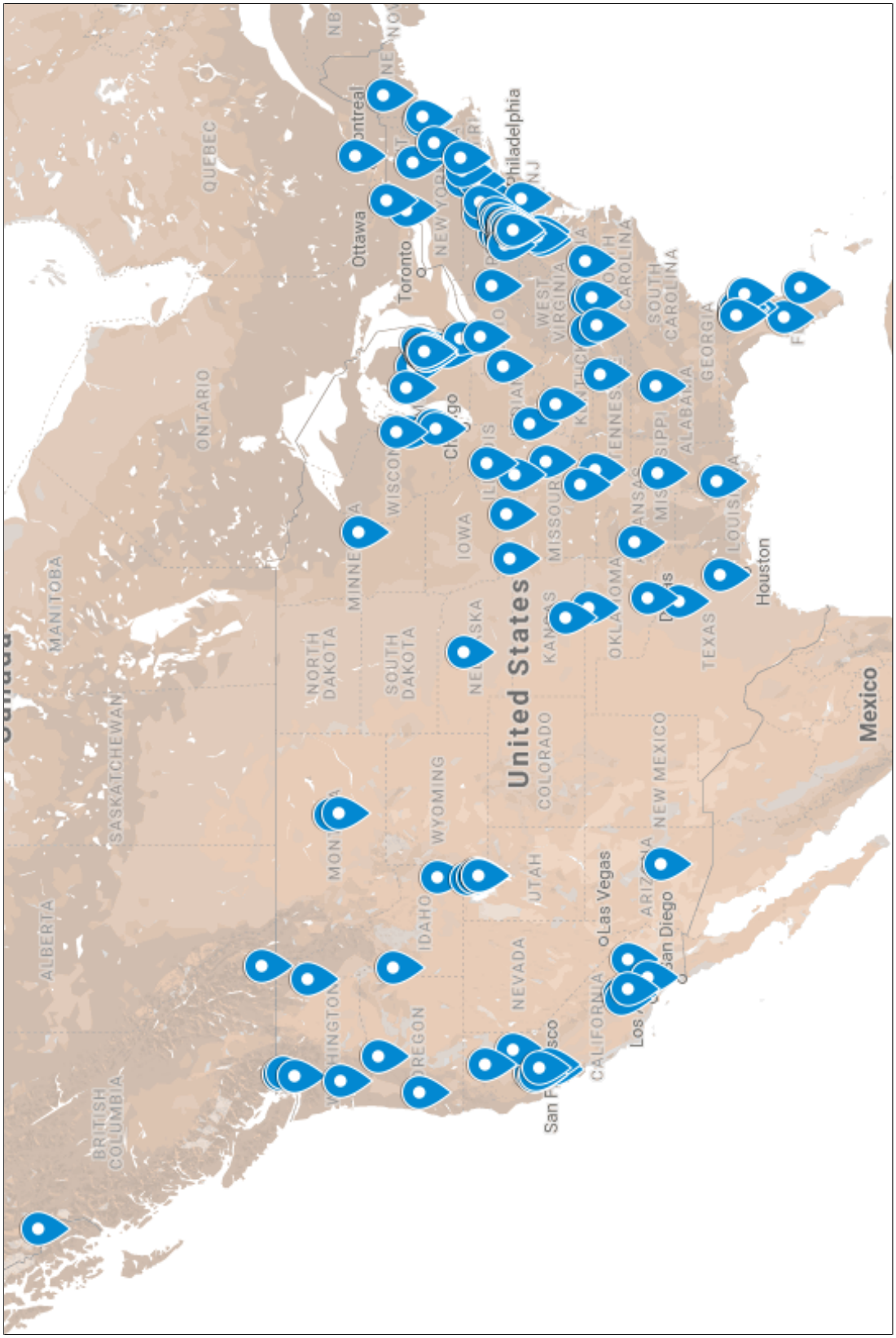


Figure 4.2: Locations of survey respondents.

outside of work or school. 37% are online 2-4 hours per day. 28% spend 1-2 hours per day online, and 20% are online 4-8 hours per day. 9% spend 8-12 hours per day online, 6% spend only 10 minutes to 1 hour online, and 1% spend more than 12 hours online.

Looking at the way respondents access social media sites, 44% access these systems via their phones. Approximately 31% of respondents use laptop computers, 14% use desktop computers, and 12% visit social media sites from tablets.

Social Media Site Use: The most commonly used social media sites among respondents were Facebook, Instagram, and Twitter. Social media sites that many of the respondents said they use rarely or never were Snapchat, Slack, and Reddit. Table 4.2 lists the frequency with which respondents make use of these social media sites. A total of 63 respondents reported using social media sites that were not mentioned in the survey questions. The sites most frequently mentioned included Tumblr, YouTube, Discord, WhatsApp, message boards, NextDoor, GoodReads, and Mastodon.

Site	Hourly	Daily	Weekly	Monthly	Rarely/Never
Facebook	15%	57%	12%	4%	11%
Instagram	13%	35%	15%	6%	32%
LinkedIn	2%	11%	21%	27%	40%
Pinterest	1%	6%	15%	21%	57%
Reddit	12%	17%	9%	7%	55%
Slack	10%	10%	7%	5%	68%
Snapchat	7%	12%	9%	3%	68%
Twitter	24%	23%	12%	9%	32%

Table 4.2: Frequency of social media use.

One question that could be asked about our analysis is whether respondents' opinions regarding organization might be systematically influenced by the organization of the social media sites that they currently use. To test for this, we performed chi-squared tests looking at the relationship between social media sites and respondents' level of satisfaction regarding current organization of posts, preferred sort order for the use case, clustering preference, and views on the number of posts that should be displayed. There were no

statistically significant relationships between social media site use and any of these characteristics for any of our use cases. Thus, we have no reason to believe that respondents' social media ordering preference is derived from their use of a specific existing site.

4.4.3 Change Detection

When asked whether they use social media to follow a specific topic or theme over a long period of time, approximately 66% of respondents indicated that they perform change detection tasks. Common topics that change detection-focused respondents follow over time include news and politics (19%), television shows (12%), movies (10%), the music industry (9%), sports (8%), and art (8%). Other frequently mentioned topics included science and technology, research on their area of expertise, and travel. One respondent explained that they look for "TV shows and what people are talking about them. I search for them using hashtags."

When asked why respondents perform change detection tasks, the primary motivators were curiosity (26%), entertainment (24%), and a desire not to miss out on something interesting (22%). The vast majority were driven by personal reasons; few indicated that they follow topics over time because they are required by work or school (7%). That said, in text responses several respondents noted that, while they are not specifically required to stay up to date on their topic, they do so for professional learning.

The most common social media sites that respondents use for change detection tasks are Twitter (29%), Facebook (21%), Instagram (16%), and Reddit (15%). They typically get updates late in the day, with 28% of respondents focusing on change detection tasks in the evening, and 18% at night. Early morning (16%) and around lunchtime (15%) were less common.

Our survey included several questions specific to change detection, designed to gauge

respondents' expertise and time spent focusing on their topic. Most respondents indicated at least an intermediate level of knowledge, with 22% at the expert level, 40% self-assessing that they have an advanced understanding, and 35% at the intermediate level. No respondents rated themselves as having only a basic understanding of their topic. A substantial majority have been following their topic over a long period of time, and 42% have been keeping up with their topic for more than 5 years. 27% of respondents have followed their topic for 2-5 years, and an additional 16% for 1-2 years. 42% of respondents spend 10 minutes to 1 hour getting updates each day, 25% less than 10 minutes a day, and 23% 1-2 hours a day. Most respondents had checked on their topic recently before taking the survey; 30% within the past hour, another 35% within the past day, and 26% within the past week.

4.4.4 Experiential

Experiential use cases were the least common, though still prevalent within the respondent groups. 49.5% of respondents indicated that they have used social media to follow or interact with a live event.

The kinds of topics respondents follow live are conferences and talks (20%), sporting events (19%), political events (17%), television shows (10%), and live streaming video (10%). Respondents also mentioned award shows, emergencies such as natural disasters, and professional events. Even though the primary experience took place during the event itself, some respondents also looked for context surrounding the event. As one respondent said about watching college football bowl games, "...I wanted see what others were thinking leading up to the game and after the football game ended."

The main actions that respondents take when interacting with live events are to browse posts of friends and people they follow (27%) and browse other users' posts (27%). 21% of

respondents also interact with their friends and people they follow, and 20% interact with other respondents' posts.

Twitter was the most common social media site that respondents use for engaging with a live event, with 44% indicating usage of that site. The next most prevalent sites for experiential use were Facebook (30%) and Instagram (13%). 29% of respondents for this section indicated that they follow events that take place in the evening (31%), at night (20%), or in the afternoon (17%). 40% of respondents said they spend 10 minutes to 1 hour focused on a live event; 24% spend 1-2 hours, 22% spend less than 10 minutes following the event, and another 11% spend 2-4 hours on the event.

4.4.5 Browsing

The most common category, 96.3% of responses indicated some type of browsing activity. Respondents' answers to an open-ended question about what they expected to see when browsing led us to the decision to split the category into two parts. Even after the split in browsing results, time-oriented browsing remained a very common task across respondents; 69 of the use case section responses provided details about the goals of this update task. Only 24 of the participants appeared to focus on general browsing.

Time-Oriented Browsing

The most frequent word that appeared within the category was "update." Even though the original question framed browsing as an activity without a specific goal, many respondents' responses indicated that there actually was a goal: to get updates on friends, family, and news. Further research is needed to determine why respondents responded to this as a distinct activity from change detection, since there are similarities between the two. These respondents might not know from session to session which items would be updated, but

they had a sense that it would fall into a finite set of categories. These respondents were also interested in topics such as art, health, movies, television shows, weather, financial information, music, and the outdoors. One respondent said, "It's everything. I follow specific bloggers, friends, family, local news, national news, international news, hobbies, and interests."

The main reasons why respondents said they perform time-oriented browsing were out of curiosity (31%), for entertainment (26%), out of a desire not to miss something interesting (15%), because their friends are talking about something (14%), and because their family is talking about something (7%). A common response supplied in the "other" section was that these respondents are motivated by boredom.

The main social media sites that people performing time-oriented browsing use are Facebook (30%), Instagram (18%), Twitter (16%), Reddit (12%), and Pinterest (9%). People perform this task in the evening (33%), at night (22%), in the early morning (17%), around lunchtime (12%), or in the afternoon (9%).

General Browsing

After splitting the browsing category, this became the least common, with only 24 of the participants providing detailed responses about general browsing. These respondents mentioned a variety of non-time-oriented topics of interest. For example, common topics included fashion, photography, memes, articles, and interesting ideas. News was still a popular topic within this set of respondents. One respondent stated, "I browse fashion, home decor, recipes, memes, and more on Pinterest on a daily basis." Another referred to searches for odd or overdone humorous content with their comment, "Unironically, dank memes."

Respondents engaged in general browsing undertake this activity out of curiosity (35%),

for entertainment (30%), out of a desire not to miss something interesting (14%), because their friends are talking about something (10%), and because their family is talking about something (6%).

The social media sites that respondents indicated were most used for general browsing were Facebook (24%), Twitter (21%), Reddit (19%), Instagram (16%), and Pinterest (9%). This browsing activity tends to happen later in the day, with 27% of respondents doing general browsing at night, 26% in the evening, or 15% around lunchtime.

4.4.6 Searching

Searching was a common use case, with 65.4% of respondents indicating that they have run searches on social media sites. Respondents run searches in social media for a wide variety of reasons. The most common explanations included learning more about a recent event (18%), looking up organizations and places (17%), and looking for specific users' profiles (15%). Respondents also use social media search capabilities to research events, news, trending topics, and to find articles. These searches are often driven by other content being posted to social media. For instance, as one respondent noted, they "... search for trending hashtags. Or look at someone's profile." Another looks "... for more information about someone who has commented."

Social media searchers run queries out of curiosity (30%), for entertainment (19%), because their friends are talking about something (17%), because they don't want to miss out on something (16%), because their work requires it (9%), or because their family is talking about it (7%).

The most common social media sites on which respondents ran searches were Facebook (26%), Twitter (21%), Pinterest (15%), Reddit (13%), and Instagram (10%). Respondents

typically run searches in the evening (33%), at night (23%), in the afternoon (14%), or in the early morning (12%).

4.5 Organization Preferences

Respondents expressed the largest amount of dissatisfaction with the way posts are organized for browsing, something we did not expect given that we initially thought of browsing as a casual, entertainment-focused task. That said, when asked specifically what orders they prefer for organizing social media posts, respondents preferred chronological order for update and browsing tasks. Clustering preferences were more varied.

4.5.1 Clustering

To address RQ4.2, we asked respondents whether they would like to have results grouped in some way. Figure 4.3 shows the most popular options, which for most use cases was a grouping by theme or development. None of the differences for this section were statistically significant based on a paired T-test, but they provide a general sense for possible preferences that could be explored further. For change detection tasks, respondents preferred to have results grouped by themes or developments relating to the overall topic (43%), with the next most popular being grouped by followed accounts (23%) and no grouping (20%).

Respondents focused on experiential use of social media also preferred results grouped by theme or topic (45%), with 26% wanting no grouping at all. Respondents engaged in time-oriented browsing tended to prefer no grouping (42%) or grouped by accounts they follow (28%). Respondents performing general browsing wanted results grouped by theme or development (67%). 37% of respondents performing searches wanted results grouped by theme or topic (37%), by verified or credible accounts (27%), or with no grouping (27%). Note that we did not ask whether participants had experience with these clustering

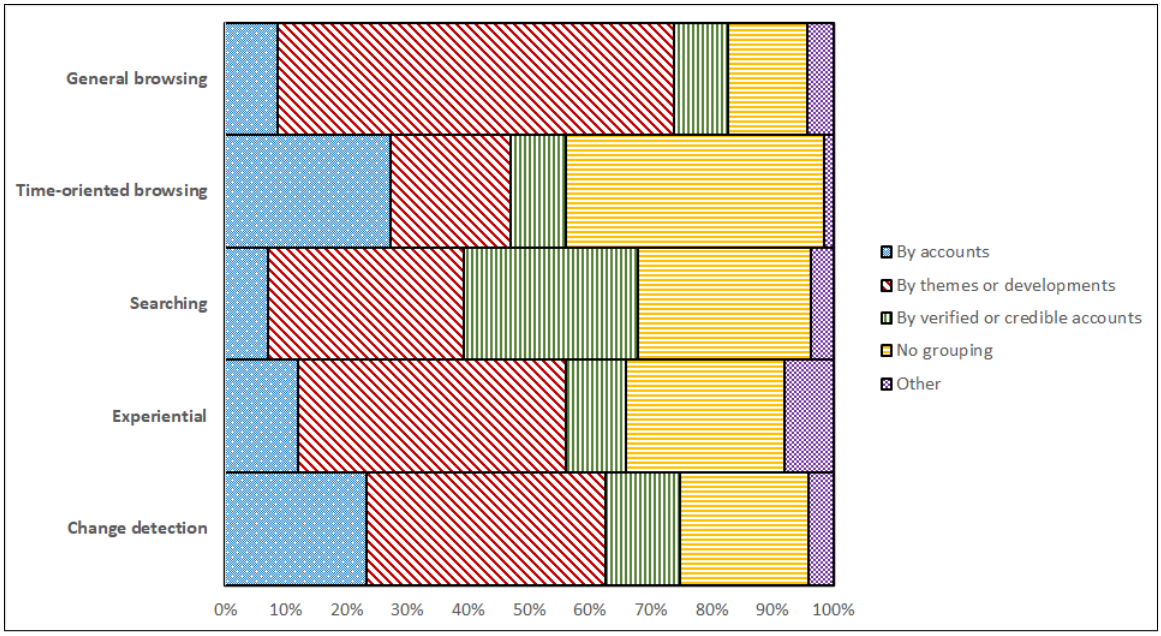


Figure 4.3: Respondents' preferred clustering approaches.

methods. These results may conflate actual clustering experience with preferences based on individual respondents' interpretation of clustering.

4.5.2 Preferred Sort Orders

To answer RQ4.3, we asked respondents about their sort order preferences in two ways: first, an open-ended question, followed by a closed-ended set of selections. After coding text responses, we found that respondents provided consistent responses to both questions. For change detection, experiential, and time-oriented browsing, respondents prefer chronological orders by a wide margin. Even accounting for margin of error, respondents engaged in these activities were significantly more likely to prefer chronological sort. Respondents who are running searches tend to prefer relevance ranked results, though these responses were not statistically significantly more frequent than those who wanted search results to be chronological. Those respondents who performed general browsing wanted results to

be either relevance ranked or ranked by popularity. Figure 4.4 displays the responses to the closed-ended questions about sort order preferences.

In their text responses, a number of respondents expressed their preference in terms of what they did not want. For example, 11 respondents responding to change detection noted that they do not want sorting to be based upon proprietary algorithms, or approaches using popularity as the driving factor. A common trend was expression of a desire not to miss out. "The algorithm-based feeds are harder to follow over long periods of time because you are guaranteed to miss something," according to one respondent. Another respondent was more blunt about preferences: "Literally in order of when they were posted. No freaking algorithms." A respondent describing organization preferences for browsing said, "...I just want a straight up timeline style feed. I want to see things as they happen, not two days later."

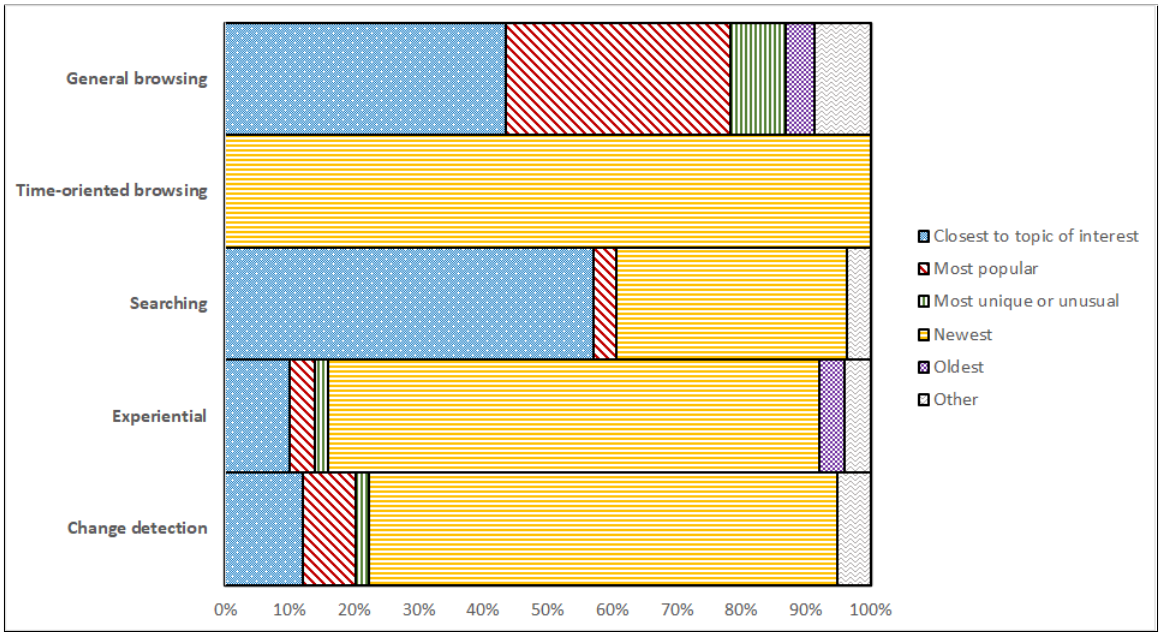


Figure 4.4: Respondents' sort order preferences.

4.5.3 Frustration with Current Sort Orders

While not a formal research question, we wanted a sense of the prevalence of respondents' frustrations with social media sort orders, to put social media posts asking for different sorting into broader perspective. For all use cases except browsing, the most prevalent response to this question was "neutral." Figure 4.5 details the levels of satisfaction for each sort order. Respondents expressed the highest amount of dissatisfaction with sort orders for time-oriented browsing. 52% of respondents indicated that they were unsatisfied or very unsatisfied, while only 17% said they were satisfied or very satisfied. In contrast, only 33% of respondents who perform general browsing were unsatisfied with current sort orders, equal to the number who expressed that they were satisfied or very satisfied.

Respondents who perform change detection tasks were statistically significantly more unsatisfied with current social media sort orders (t-value: 2.92, DF: 44.31, p-value: 0.01).[4] 37% of respondents indicated either being unsatisfied or very unsatisfied, compared to 19% selecting satisfied or very satisfied. Respondents engaged in experiential tasks had more neutral to positive feelings about current social media sort orders when following live events, with 55% of respondents indicating that they are neutral. 26% described themselves as satisfied or very satisfied, and only 20% said they are unsatisfied or very unsatisfied.

Interestingly, for the two update tasks there was a statistically significant difference between responses of different age groups. Respondents aged 35-54 were statistically significantly more likely to be negative than those aged 18-34 about current sort orders for performing change detection tasks (t-value: 3.51, DF 102.16, p-value <0.01). For experiential tasks, respondents from 35-54 were more likely to be unsatisfied than respondents from 18-34, though when a Bonferroni correction for multiple tests was applied, this result was not statistically significant (t-value: 2.39, DF: 44.05, p-value: 0.02).

[4]For statistical significance tests in this section, two-tailed one sample t-tests were applied.

Respondents engaged in searching had more positive than negative responses to questions about the status quo. 29% of these respondents were satisfied with their experience, and 25% were unsatisfied or very unsatisfied.

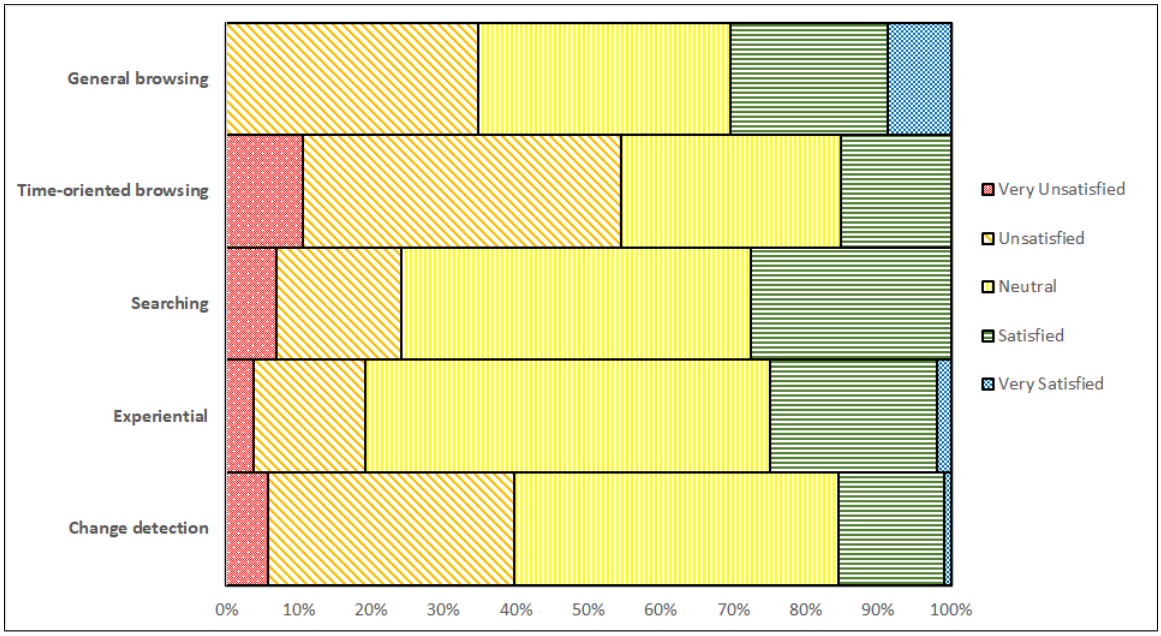


Figure 4.5: Respondents' satisfaction with current organization of posts.

4.5.4 Numbers of Posts

We asked respondents to answer two questions to help us understand RQ4.4: would they prefer to see all posts vs. only the best posts for certain categories, and what number of posts would they like to see for new developments. Differences in responses for these questions were not statistically significant, but provide general insights into preferences that can be tested further with more specific experiments. Respondents engaged in change detection prefer to see all posts by people they follow (49%), with "only the best posts related to the topic I follow" as a distant second (19%). As for the number of posts, they prefer 5-10 posts per theme (31%) or all posts on the theme (31%). For experiential use cases, respondents want to see all posts related to the topic (43%) or all posts by people

they follow (21%). Time-oriented browsing respondents were the most focused on seeing everything—they would like to see all posts by the people they follow (70%), and all posts on new themes that come up (43%), or 2-4 posts per theme (a distant second, at 24%). Respondents engaged in general browsing were split between wanting only the best posts by people they follow (21%) and only the best posts for each theme or development (21%). These respondents wanted to see 5-10 posts per theme (42%). Search-focused respondents wanted to see either the best posts related to the topic (28%) or all posts about themes or developments (28%). 34% want to see all posts on the theme.

4.5.5 Ending a Session

We asked how respondents decide they are done with their session for each use case. The results for this section provide general insights, but differences are not statistically significant. Change detection-focused respondents were split between "when I have found the relevant themes and developments" (25%), "when I have read all of the relevant posts" (25%), and "when I have answered my questions" (23%). For experiential use cases, respondents were generally done when the event was over (26%) or when they have read all the relevant posts (23%). Time-oriented browsers were done when they ran out of time (33%) or when they have read all the relevant posts (32%). General browsers were done when they ran out of time (42%) or met some other criteria (e.g., got bored, got an idea, or found something interesting to follow up on) (17%). Respondents engaged in searching were done when they answered their questions (48%), read all relevant posts (24%), or ran out of time (21%).

4.6 Implications for System Design

In many of the use cases studied here, the respondents intend to have a deep interaction with social media content. Responses about update tasks indicated that respondents engaging in these activities at least claim to favor recall over precision. That is, generally, respondents wanted to see "all" of something (all posts on the topic, all by people they follow, all on the theme, etc.). They may favor an interface that provides some amount of clustering with at least the ability to get to all relevant posts in the appropriate order (likely chronological).

Thinking of each user as a bundle of interests, a single feed can contain a large amount of unrelated information. For users with a specific intent—for instance, following posts about a live event—it can be difficult to find relevant commentary when the user sees all posts (relevant to the event or not) flowing into the same stream, organized by relevance to the user's overall interests, popularity, and other factors. Popular search engines have made a transition from a single stream of ranked results to blocks of related content, each organized by relevance to user needs (which can be thought of as a kind of aggregated search) [153]. We suggest taking a cue from aggregated search, in case such presentation approaches can be integrated into social media feeds as well. Aggregated search builds upon concepts from federated search, and groups different content together like sets of "information nuggets." Relevant content within these blocks (referred to as "verticals") are sorted, and then the blocks themselves are also sorted [148]. Disentangling a user's social media feed into related groups is a more complex undertaking than providing responses to a query. That said, based on our survey results, we believe there may be tacit indicators to help identify and adjust the organization of the feed for a specific activity, in a way that connects an individual post to an overarching story. For instance, a social media site could incorporate the ideas from the GPA Change Detection Theory to aid users performing

change detection tasks. Sites like Twitter provide topic and event-centric views in addition to user-curated lists. Twitter's Explore tab includes content customized based on user interests, as well as topics—to include some feeds related to live events [275]. Such features could be further adapted to meet specific user needs within the system.

The large numbers of "neutral" responses about current ordering of social media feeds may indicate some flexibility in accepting new arrangements, especially considering the mismatch between the current views and the preferred sort for update tasks (chronological). Users may not feel tied to a specific organization approach. Bron, et al. looked into design of aggregated search for multi-session search tasks; this revealed that some users tend to prefer tabbed interfaces over blended interfaces [39]. Arguello, et al. found that more complex tasks resulted in higher search and content interaction. Unlike Bron, et al. Arguello, et al. found a preference for blended aggregated interfaces, which could be attributed to users' technical sophistication levels [18]. This kind of comparison could serve as an initial starting point for update-focused interface design, with distinct views for specific topics, themes, or events, while making it easy for the user to return to a general feed for serendipitous discovery.

4.7 Limitations

We recognize that there are limitations inherent in using surveys. First, a survey represents only a snapshot in time. Behaviors change over time, and approaches shift as capabilities of systems change or users adopt new technologies. Only a limited number of people are willing to take surveys, and—based on our convenience sampling approach—our respondents could over-represent users who have strong feelings about sort orders within social media (sample bias). Given the length and complexity of our survey, we may have

further limited the number of respondents; we made a tradeoff decision to opt for a greater level of detail in the questions we asked of the social media user base.

We attempted to overcome these limitations in several ways. We shared the survey through a variety of channels to gather responses outside our local demographic. We also asked questions multiple ways to compare responses—for instance, respondents were asked to provide a textual description of their sort preferences, then later were able to select from a predefined set of categories. We also ran a small pilot of the survey involving read-aloud and online versions of the survey, which we used to clarify wording and adjust the overall survey flow.

4.8 Summary

We ran a survey to understand the prevalence of the change detection task among social media users, as well as to understand user behavior and user preferences relating to four distinct use cases within social media. The four use cases included two update tasks, which were divided into change detection and experiential use cases, as well as browsing and searching. The survey results revealed that respondents have distinct behaviors and data organization preferences for each of the use cases; the preferences expressed by change detection-focused respondents aligned with the group and pile aspects of the GPA Change Detection Theory. Additionally, analysis of results indicated that browsing preferences could be split into two subcategories—general browsing and time-oriented browsing. The results from this study can be used to inform the design for social media systems where users perform these activities, to improve users' experience with these systems.

Some of these findings were incorporated into the design of the Daybreak system, discussed in Chapters 5 and 6. Taking inspiration from the findings from this survey, we organized and ran a hands-on user study focusing on the change detection needs of users

who follow the same topic over a long period of time—pivoting to news articles rather than social media posts. The follow-on study tested whether the preferred organizational approaches identified in this study can lead to successful task completion when applied to news articles.

Chapter 5: Designing a Change Detection System and User Study

In the sort order survey described in Chapter 4, we asked respondents for input regarding the components of the GPA Change Detection Theory (Chapter 3). We learned that respondents recognize the change detection concept, and that it is a prevalent use case within social media sites. This study follows up on these concepts to determine the extent to which information organized in this way results in task completion and user satisfaction. The sort order survey provided details about user preferences for change detection tasks in social media. In this chapter, we connect the theory and the survey results to change detection tasks in practice to see how a system designed based on the GPA Change Detection Theory aids the users in a real-life scenario focused on news articles.

We developed the Daybreak system prototype based on the GPA Change Detection Theory (Chapter 3), to further test and validate what we learned from the sort order survey (Chapter 4). This chapter includes the design of the system, the study, and the process used to select participants for the study; the results of the study are discussed in Chapter 6.

5.1 Research Objectives

The Daybreak user study was designed to identify ties between the theory and users' practice of change detection tasks. Earlier in this dissertation we focused on theoretical aspects of change detection, then moved to an initial study to understand users' perceptions and preferences related to change detection through a survey. The next step was to build

a prototype system for use in a user study, to determine whether what we learned in the survey aligns with actual user needs in practice. In order to understand how a system might support change detection tasks, our research focused on seven questions that were derived from the GPA Change Detection Theory. These questions addresses the three core areas of the GPA Change Detection Theory: Grouping documents into subtopic sets, organizing the documents within the groups into sorted piles, and then arranging the piles by subtopic importance. We also look at the overall ability of the system to meet the change detection needs of the user study participants. Finally, we look at the alignment between the topics used for the study, participants' interests, and the change detection scenario.

5.1.1 Terminology Note

This chapter includes a number of terms that tie to the GPA Change Detection Theory. While we used different terms within the user study, these terms map to concepts from the theory. The key concepts from the theory are as follows:

- **Group:** The first grouping-related concept covered in the study is *tagging*. Users *tag* a document to indicate their interest in the contents or for other purposes. Users can apply *tag labels* to tagged documents. Tag labels are the user's representation of some concept from the document. While tags and tag labels are tied directly to user actions, a *subtopic* is an abstraction of the concept in the tag label. Finally, the user study refers to *subtopic clusters* (sometimes shortened to *clusters*) to represent groupings of retrieved documents; these are thematically-related sets of documents within the result set that align with tag labels and documents tagged by the user.

- **Pile:** The GPA Change Detection Theory notes user interest in organizing documents in some order, to produce a superdocument focused on a single theme. In the Daybreak prototype, this concept is represented in a cluster's *sort order*. Users have the

ability to change document sort orders across the result set, including within subtopic clusters.

- **Arrange:** In the theory, this concept focuses on arranging the piles in some order to aid the user. For the Daybreak system, we refer to this concept in a few ways. We describe the *system ordering* when discussing how we have arranged subtopic clusters. In the Daybreak study we also include the concept of *subtopic importance*—a system ordering that assumes that the user wants to see "important" subtopics first. We operationalize this concept of subtopic importance in the Daybreak system with subtopic *rarity*.

Here we define additional terms used in the user study design. We use the term *user* to represent a notional user—an individual who might use a change detection system. For specific individuals who we interacted with during the study, we applied more specific terms: *respondents* are individuals who completed our selection survey, with *qualified respondents* being individuals who met the criteria for participation in the study. We use the term *selected participant* to refer to qualified respondents who we contacted about participating in the Daybreak user study. We refer to the individuals who completed the study as *participants*, and we describe to their time spent participating in the Daybreak user study as a *session*. Each of the participants' five interactions with the Daybreak system is called a *day*.

5.1.2 Research Methods

As previously noted, our research objective was to build a system to test the components of the GPA Change Detection Theory in practice. Based on our research objective, we envisioned the Daybreak user study as a mixed-methods study consisting of a qualifying survey and a hands-on user test of a system. Through the qualification survey we gathered

some quantitative data about the respondents and their change detection practices. The user study was a qualitative user study, where we observed participants' interactions with the Daybreak system and asked for contextual information via questionnaires and a semi-structured interview. This chapter provides additional detail about the specific approaches used throughout. As noted by Creswell, a mixed-methods approach is based on "the assumption that collecting diverse types of data best provides an understanding of a research problem," starting with a survey and expanding to a variety of qualitative approaches [68].

While the primary purpose of our survey was for selection of participants, we also gained additional insights into respondents' change detection practices—including triangulation of findings related to the sort order survey discussed in Chapter 4. Surveys are effective options for deepening an understanding of trends and preferences [186]. Even though we structured our qualifying survey to focus on areas that would enable us to select users for the study, these questions revealed useful information about users' change detection activities and topics of interest. Our assessment of the survey results focused on broad themes present in the data.

5.1.3 Framework Method

For the Daybreak user study, we explored two potential research methods for analysis and interpretation of our results: grounded theory and the framework method (also referred to as framework analysis). A key difference between these two research methods is in the degree to which there is an attempt to align with an external framework. For studies that leverage grounded theory, the theory emerges from the coding and analysis of the collected data [63]. In contrast, the framework method is an approach for gathering and analyzing qualitative data for trends and relationship to some organizing framework. It is more deductive in nature, and more rooted in the existing research questions. The collected data is

tagged based on principles from the overarching framework using methods similar to the coding process in grounded theory [90]. The framework method has its roots in the policy and health research fields, and has been applied to a wide range of qualitative studies, from psychology to technical research [251]—in our case, the GPA Change Detection Theory. The framework method has similarities to grounded theory, which is also a process for analyzing qualitative data such as user sessions and interviews. We determined that the framework method would better suit this research, due to our research questions and their focus on the GPA Change Detection Theory.

Given that the aim of this study was to test the ideas of the GPA Change Detection Theory, we determined that the framework method would be a suitable research method to apply. The advantage of this approach was that it enabled us to review the collected data and devise codes that addressed specific components of the theory and related research questions. While more rooted in deductive approaches, the framework method does include inductive components to aid in capturing other emergent themes that had not been anticipated in the research questions. While our focus was on understanding change detection as framed in the GPA Change Detection Theory, we maintained some flexibility to add codes not specifically tied to the theory that would be applied as we observed unexpected concepts in the data.

5.1.4 Research Questions for the Daybreak User Study

In the sort order survey (Chapter 4), we learned that respondents are open to the idea of grouping documents into related clusters; we did not define the clustering concept within the survey, but attempt to do so in this study. As for organizing documents, the survey respondents who said that they engage in change detection expressed a preference for chronologically-ordered information, divided into subtopics. To address this further, we

designed and built the Daybreak system, recruited participants, and performed a user study in which participants leveraged Daybreak to perform a change detection task.

To understand users' behavior in a change detection-focused system, we produced seven research questions for this study. These questions were designed to help to understand the extent to which the GPA Change Detection Theory is adequate as a framework for supporting a user's change detection-related information seeking behaviors.

RQ5.1: Does tagging and tag label generation aid users in representing their mental model of a topic?

A foundational concept in the GPA Change Detection Theory relates to the role of a change detection system in updating an individual's mental model. The user compares new information to old, and determines whether it should be added to the mental model. Subtopics seek to represent a user's personal organization of knowledge on a topic; in the Daybreak system, participants indicate subtopics by applying tag labels. This research question evaluates the effectiveness of tags and tag labels as a representation of a portion of the participant's mental model.

GPA Change Detection Component: Group

RQ5.2: Does organizing search results by subtopic clusters aid users in performing change detection tasks?

Clusters operationalize subtopics for information discovery by organizing aspects of the topic in some way. The Daybreak system lets participants tag documents and apply tag labels, and also organizes future days' search results in alignment with the participant's personal tag labels. This research question focuses on understanding the utility of aligning search results with participants' personalized subtopic clusters.

RQ5.2a: What information retrieval approaches would be effective for transforming a user's tags into clusters of relevant documents?

In addition to understanding how participants respond to personalized clusters to organize search results for change detection tasks, we also want to understand how documents should be placed in those clusters. This question explores the information retrieval approaches used in Daybreak, and their perceived effectiveness.

GPA Change Detection Component: Pile

RQ5.3: Does organizing search results within subtopic clusters in some sort order aid users in performing change detection tasks?

Within the GPA theory, grouped documents are turned into piles by applying some sort order. For this research question, we look at two sort options within the clusters to determine whether participants prefer one over another for the change detection task.

GPA Change Detection Component: Arrange

RQ5.4: Does arranging subtopic clusters in some order aid users in performing change detection tasks?

The GPA theory discusses organizing the sorted piles of documents in some way to highlight important subtopics. For this research question, we test a specific approach for organizing subtopic clusters to determine if it aligns with participants' views of subtopic importance.

RQ5.4a: How should the system handle documents that do not fit in any existing subtopic cluster?

Given that change detection is about identifying new information related to a topic, it is reasonable to assume that there is some information that does not align with the subtopics that they have personally externalized as tag labels. For this research question, we test a concept for organizing the documents relevant to the topic but that do not align with existing subtopic clusters.

GPA Change Detection Component: Combining Group-Pile-Arrange Components

RQ5.5: Does the system help users develop and externalize mental models?

While there may be individual benefits in combining the concepts of grouping, piling, and arranging documents, the GPA theory focuses on meeting change detection needs by combining all three concepts into a single system. This research question addresses the extent to which the Daybreak system—as an implementation of the GPA theory—meets change detection-focused user needs. This is addressed through determining the degree to which the Daybreak system helped the participants reflect their knowledge gained by externalization of some newly updated aspects of their mental model.

5.2 User Study Design

Building on the sort order survey described in Chapter 4, we designed a study in which a group of users would complete change detection tasks in a prototype system, which we named Daybreak, based on the idea that someone might start their day by checking for updates on a topic of interest. The user study was designed to build on what we learned from the sort order survey, in order to address our research questions. Our goal was to create a scenario that mimics a real-world change detection use case in order to collect data on user behavior. In this case, we would have participants leverage the Daybreak system

to review news articles; at the end of the scenario, the user would be asked to externalize some of what they learned.

Between the sort order survey and the Daybreak user study we made one change to the type of data studied: while the sort order survey focused on user needs in social media, this study looks at usage of a system focused on news articles. The primary reason for the pivot from social media to news articles relates to the fact that it is generally easier to set up effective searches against long text (e.g., news articles) compared to short text (e.g., social media). Research into searching short text includes abstracts from papers, discussion forums, as well as social media. In the case of discussion forums and social media, there are additional issues that make the these types of collections more difficult to search. For example, some of the issues affecting searchability include the smaller number of terms per post (less signal), the lack of context in a short post, and the use of casual or slang language [211, 290]. Even though some researchers have identified ways to more effectively search social media text [77, 126], we determined that longer-form news articles would be more appropriate for the scenario, with more formal language and context in the documents. An additional reason for the selection of news documents was based on an observation from the sort order survey in Chapter 4. Respondents to the change detection section of the survey indicated that news was one of the most common topics they followed. While finding news through social media differs from searching for news articles directly, we expected that there would be commonalities. We recognize this shift as a limitation of our study, and note that future research could compare users' performance of change detection tasks across different collection types.

As an example, if the participant was a baseball fan who follows the Miami Marlins, they might be shown historical articles for the Boston Red Sox. The documents would be selected far enough in the past to reduce the likelihood that the participant remembered specifics from that date range, while still knowing enough about the overall topic (base-

ball) to recognize and follow the story. This aided in tracking the participant's process of learning and retelling a story gleaned from their document review.

5.2.1 Scenario

We wanted to select a scenario that would mimic a potential real-world change detection use case. The scenario that we chose was based on a professional change detection use case. We devised a scenario that would set the stage for a participant to look for daily updates on a topic that they follow over time. Given the complexity of customizing collections based on all participants' individual interests, we opted for a scenario that would enable us to assign multiple users to the same topic. To do so, we asked participants to pretend that they are a full-time blogger on the topic they said they follow on a regular basis. However, their fellow blogger was going on vacation for a week. This colleague follows a similar topic, but not the exact topic that the participant is interested in. For this simulation, the participants were asked to follow their colleague's topic instead of their own over the course of five simulated work days; during that time, they would review news articles relevant to the colleague's topic. At the end of the five days they would be responsible for a blog post about what happened on their colleague's topic during the time that they were away. To simplify the final task, instead of asking participants to prepare a full blog article, our scenario stopped with the generation of a summary outline only.

By focusing on a topic that was similar to their own topic of interest, we expected that the terms and ideas within the news articles should be familiar to the participant. We also needed to ensure that the scenario included a learning component—the participant learns the latest developments on the specific topic. We determined that selecting documents from an adjacent topic was not sufficient on its own; we did not want the change detection scenario to turn into a memory recall test because we had selected articles about which the

participant already had recent knowledge. To minimize cases like these, we added another dimension to the scenario: a sort of time machine. For each topic, we selected news articles from a few years in the past. The reasoning behind this was that even the users who were assigned a topic very close to their actual topic of interest likely would not be not aware of all historical developments on the assigned topic. That would ensure sufficient "new" learning to be consistent with a change detection task.

The sort order survey in Chapter 4 revealed that many respondents perform change detection at different points in the day, though many saw it as more of an end-of-day task. For purposes of this study, we did not specify what time of day that the participant's search process was taking place.

5.2.2 User Study Structure and Sequence

The primary goal for this study is to understand how well the system met the change detection needs of the participants, with a secondary goal of understanding the participants' experience in the Daybreak system more broadly. The system was designed to incorporate features that we believed would meet the participants' needs, including the capabilities detailed in the GPA Change Detection Theory and the sort order survey. The key features applied based on the change detection theory include supporting expansion of a participant's mental model, operationalizing the mental model by applying tags and tag labels, leveraging a sort order that supports following a story, and clustering to provide documents organized based on the participant's mental model.

Each user study session consisted of five eight-minute simulated days where a participant would review news articles in the Daybreak system. The session also contained additional segments designed to get additional contextual information. In all, each session included the following components:

Startup and Task Familiarization: At selection time, the participant was provided with their assigned topic, which was one of the following: Red Sox baseball, cryptocurrency, global health, space, and extreme weather. Before using the Daybreak system, the participant completed a pre-study questionnaire to provide additional context about their background and interests related to change detection, and to their assigned topic. After that step, the participant watched a six-minute training video about the Daybreak system. They were given an opportunity to ask questions about the task and the system prior to starting Day 1.

Day 1: The first simulated day was a "cold start," meaning that the system had no prior model of the participant's specific interests at that point. Daybreak ran a query for news articles from a single day that related to the participant's assigned topic. All documents appeared in one long list, which could be sorted in reverse chronological order or by relevance. The participant reviewed the results, tagged documents, and applied tag labels. Not all results needed to be tagged; the participant could use these tags to keep track of interesting documents. A single document could have multiple subtopic labels. The system displayed a list of previously assigned tag labels and associated document numbers to aid the participant.

Post-Day Activities: After each day's document review was completed, in between days the participant was asked to complete a brief questionnaire. This post-day questionnaire asked about what they thought the most important tag labels were that day, how many news articles the participant believed they viewed, and other details. We included some additional questions in the post-Day 1 questionnaire that weren't repeated on later days; this included questions such as how they define "enough." (e.g., to enable them to answer a daily question about whether they viewed "enough" articles that day).

Practice Storytelling Task: After Day 1, we sent the participant to a practice version of the storytelling task. The intent was to give the participant an idea of what they would be

asked to do at the end of Day 5. This was to familiarize the participant with the task, and to encourage them to focus their document review on the end goal. The story outlines created after Day 1 were not evaluated.

Days 2-5: For Days 2-5, the Daybreak system generated subtopic clusters based on the participant's prior tag labels. The system used tagged documents and tag labels in recommending new documents for the next day. The system ran a query against the next day's documents using the main query term, then clustered the documents into subtopic categories. Participants had the ability to choose whether or not to display their results organized by clusters. They also selected either relevance or reverse chronological sort options. Documents that were relevant to the main topic but not specific to any of the participant's tag labels appeared in an uncategorized section labeled as "other."

Final Storytelling Task: After all five days were complete, the participants completed the brief storytelling task in which they created an outline for a blog post, detailing the key events from the five days—a summary of key things that happened on the topic that they followed on behalf of their colleague. The storytelling task interface provided the participant with access to the documents they had tagged across all five days, categorized by tag label, to aid in preparing the summary.

Semi-Structured Interview: Following the storytelling task, we held a semi-structured interview through which the participant provided more detailed information on their overall experience, thoughts about the Daybreak system's interface, and various aspects of change detection. This was an opportunity to understand the reasoning of the participant as they worked on a change detection task, to include gaining insight into how well the Daybreak system met their needs.

5.2.3 Time Constraints

The participant in this scenario wishes to satisfy the need for a thorough review of information on a given topic in order to identify new developments related to their topic of interests. Rather than set up a scenario that could be infinitely long, we applied time limits on each day's review of news articles. We added the time limit for a number of reasons. First, we wanted to keep the overall session length around 90 minutes; minimizing the session length was in part to reduce the likelihood of participant burnout. Additionally, we expected that participants would focus their review of articles, and to place less emphasis on trying to reading all documents. and more on knowing the key developments. Over the course of the five days, we expected that the participant would become increasingly comfortable with their information seeking approach, and optimize their use of the time by prioritizing their approach for reviewing news articles related to their topic.

5.2.4 Conducting an Online User Study

Although the study was originally conceived as an in-person event on the University of Maryland, College Park campus, due to COVID-19, we pivoted to an online format and held the user study via the Zoom system. While online studies have been conducted for many years [99, 184], the COVID-19 pandemic led researchers to pivot to online studies in cases that may otherwise have been performed in-person [80, 154, 271]. This was the case for our study as well. It resulted in some benefits, to include flexibility of selecting participants outside of the local area, and in timing for the sessions, which did not have to be organized around availability of physical space.

Each online user study session was designed to last approximately 90 minutes. Sessions were recorded for later transcription and analysis. After completing the study, participants

received an Amazon gift card for the amount of $25. If a session terminated early, the participant received a $10 Amazon gift card.

We had developed the Daybreak system to run on a single laptop, and were able to mimic the approach for the planned in-person study by sharing our screen and providing each participant with keyboard and mouse control via Zoom. Given that the system was optimized for a computer rather than a phone or tablet, we needed participants who could connect into the call and use the system from a laptop or desktop. To aid in selecting participants with these devices, we included questions in the selection questionnaire about computer access.

We considered cyber security in our design; we only provided access to specific browser windows to perform the user study. If we thought that a participant was intentionally trying to access something else in the interface, we could terminate the system sharing.

The pivot to an online study provided some flexibility in the participant selection, since it no longer required that we select individuals who could go to the University of Maryland, College Park campus in person.

Each call included three individuals: the researcher (the author), the observer (who served as a notetaker and backup Zoom host), and the participant. Figure 5.1 depicts the setup of our Zoom sessions.

5.3 Overview of the Daybreak System

We developed and applied the Daybreak system to gain a deeper understanding of users as they perform a change detection task. The system and associated data were assembled to test whether the proposed design enables a user to complete change detection tasks successfully. As discussed earlier, we note that the sort order survey covered in Chapter 4 was focused on social media, while this study leverages news articles. The system and

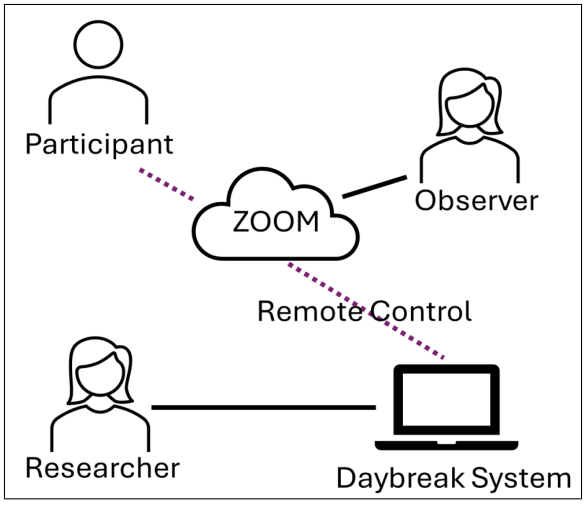


Figure 5.1: Composition of Daybreak user study sessions, including the researcher, observer, participant, and how they interacted with the Daybreak system.

associated data enabled us to test whether the proposed design enables a user to complete change detection tasks successfully.

5.3.1 Core Change Detection Functionality

Daybreak was designed to test how well specific functionality supports the participant's change detection needs. The system presents documents on a topic that the participant follows, without requiring that the participant run separate queries each day. The version of Daybreak used for the user study was a limited-functionality prototype system designed to study how well a system of that type meets the change detection needs of participants. The basic system functionality includes the following:

- Simulate the running of a (predetermined) query on a topic
- Display search results to the participant
- Allow the participant to switch between the clustered and unclustered view of document results, customized based on the participant's tag labels
- Allow the participant to change the sort order of the documents between relevance ranking and reverse chronological order

- Allow the participant to open and view individual documents
- Allow the participant to apply tags and tag labels to individual documents
- The following day, retrieve new documents and organize them into clusters based on relevance to prior tag labels
- Display additional documents that were not relevant to any of the clusters (e.g., potential new subtopics)

At the start of each day, the participant saw a blank Daybreak window, and had to take specific action on two items to see the day's news articles. First, they had to decide whether to use clustering. This was presented in the form of a switch: the participant initiated clustering by clicking on the switch. Once clustering was enabled through the switch, if they close to turn clustering back off, that required an additional click. We anticipated based on the sort order survey results (Chapter 4) that the participant would prefer to leverage clustering when performing a change detection task. As a result, we used that as the default switch option; participants needed to take additional action to move away from clustering.

Second, after selecting their clustering option, the participant had to select a sort order for the documents. The system provided two options as radio buttons: reverse chronological or relevance ranked (based on the document's relevance to the topic). Unlike with clustering, where the first click brought up a default, the participants could select either sort option with a single click. The options were presented to the participant in that order in the interface due to the strong preference for time-based sorting that was expressed in the sort order survey (Chapter 4).

At the end of each day, while the participant was completing the post-day questionnaire, we manually ran Python scripts that generated the next day's subtopic clusters. The output of this process was an updated version of the interface, with the addition of clusters that

were customized based on the participant's prior tag labels and new documents that were relevant to those tag labels. This process is further described in Section 5.3.3.

5.3.2 Document Collection

For the Daybreak user study, we selected five topics that are related to those followed by the selected participants. These specific search topics were based on topics that respondents to the selection questionnaire said that they follow on a consistent basis. Given our desire to compare individual preferences for change detection within and across topics, we wanted a set of static queries and documents that we could use to compare participant behaviors on similar documents. We kept the topics general enough to allow for reuse of topics across participant sessions.

As noted in the scenario description, we used older data to ensure that participants were learning the details of what took place as they completed the task. While we did give some thought to time frames when more documents might be prevalent, we also took inspiration from journalist Gene Weingarten's book *One Day*. In this book, the author chose a historical date at random, then researched events from that day to demonstrate "...that in life, there's no such thing as 'ordinary.' ...In the events of a single day—in that telltale grain of sand—you would find embedded in microcosm all of the grand themes in what hacks and academics call The Human Experience" [292]. We selected dates for the documents from several years in the past, old enough that participants would be unlikely to remember specifics—for example, a five-day block of baseball games in the late summer—in case some participants had familiarity with their assigned topic.

The Daybreak system uses historical news articles obtained through academic collections provided by Nexis Uni, a Lexis-Nexis news search capability provided for academic and public library use [160]. Based upon document retrieval limits for Nexis Uni queries,

we limited our retrieval to 100 documents per query, with each query representing the results from one day per topic, drawn from existing historical document collections. After some trial and error, we adjusted our Nexis Uni queries to produce at least 100 documents for each day, on each topic. For queries that produced more than 100 documents, we limited our document selection to the top ranked 100 documents provided by Nexis Uni.

We finalized the Daybreak system development and document ingest after completing the participant selection process described in 5.6.1. Our initial document collection—used for prototyping and testing—related to the Boston Red Sox. The remaining four topics were selected because they were deemed similar to the topics of interest of multiple qualifying survey respondents. We assigned 3-character codes for each topic.

The following five topics were the areas of focus for the user study. Figure 5.2 shows a timeline of the dates covered for each topic.

- BBL: Boston Red Sox baseball, during a road trip (July 21-25, 2014)
- FIN: Cryptocurrency, showing the emergence of new coins (November 20-24, 2017)
- HLT: Global health trends, with an emphasis on flu trends (March 19-23, 2012)
- SPC: Space, with an emphasis on satellites (April 29-May 3, 2019)
- WEA: Extreme weather, with a European heat wave (August 10-14, 2015)

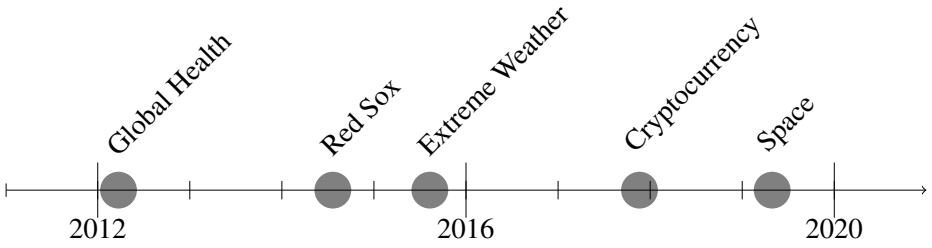


Figure 5.2: Date ranges for the documents selected for each topic for use in the Daybreak user study.

5.3.3 Information Retrieval Approach

Rather than providing streaming content as it arrives, the Daybreak system runs a query and provides relevant results to the participant once a day. To bound the participants' actions, some of the capabilities were controlled. For example, the top-level query terms for the user study were pre-assigned, and could not be modified by the participants. The purpose of this control was for pre-selection of documents used in this study, to ensure that we had sufficient documents that include a range of subtopics. If we had let the participant select their own terms, they might have ended up with insufficient documents. This also allows us to verify that we have some commonality in documents viewed and tagged across groups of participants assigned to the same topic, so we can observe how different participants tag the same documents. Other parts of the system included components that were customized based on participant actions during the study, such as applying tags and tag labels, generating clusters based on prior days' tag labels, and the ability to turn on clustering and select a sort order for the documents.

Before indexing the documents, we first normalized documents (e.g., removed non-alphanumeric characters and turned text to lower case). We then assigned a unique user study document ID for each document. For each day, the documents used within the Daybreak system were indexed with the Indri academic search engine, which uses a combination of language modeling and inference network approaches to find relevant documents [257].

To generate each day's subtopic clusters, we applied a combination of filtering and reranking. These processes were designed to produce a cluster containing a set of documents relevant to the participant's tagged documents and tag labels. Since this is a controlled scenario, the main task of filtering by topic—using the participant's subtopic labels—was performed ahead each of the days in the session. To generate these clusters,

we had initially tested a machine learning-based capability using Word2Vec, but the document collection for each topic (approximately 500 documents total, across all five days) was too small to generate sufficient signal to produce a useful model.

Instead of applying machine learning, we used a bag-of-words approach and an extraction-based methodology for identifying representative terms in a document. This was completed in advance of any sessions on that topic. For each document within a topic, we tokenized text, converted it to lowercase, and removed stopwords, numbers, non-alphanumeric characters, as well as the topic terms from the body text (for example, instances of the exact phrase "Red Sox" were removed from the Red Sox-related documents). The latter was done in order to identify key document terms other than the one used to select documents. We then used Rapid Automatic Keyword Extraction (RAKE) to identify the top ten phrases from the document. The RAKE tool splits documents into sets of words (tokens), removes stopwords, then uses word co-occurrence to extract phrases where words commonly appear together [222]. From these phrases, we selected 5-25 of the most frequently occurring words, depending upon the topic; these were stored as the key terms representing the documents.

As an example, we applied this process to an Associated Press article titled "De La Rosa hit hard, Blue Jays blank Red Sox 8-0" from 7/24/2014, which received document ID SOX-382 (Day 4) in our Red Sox collection. RAKE extracted 100 phrases from this document. Following are the top ten phrases that RAKE extracted from this document:

- 'rob rasmussen finished red sox right hander rubby de la rosa allowed runs earned'
- 'boston lh jon lester facing rays lh david price'
- 'red sox manager john farrell boston set season highs'
- 'season slugger david ortiz felt back spasms'
- 'games blue jays outfielder melky cabrera went'
- 'successful blue jays manager john gibbons'

- ‘red sox rookie brock holt returned’
- ‘big day de la rosa’
- ‘left handed hitters farrell francisco’
- ‘great competitor todd redmond worked’

From these ten phrases, the following 25 words were selected to represent this document:

- stroman
- red
- sox
- toronto
- right
- gibbons
- cabrera
- double
- runs
- time
- blue
- jays
- season
- francisco
- home
- hitter
- game
- games
- lh
- ortiz
- boston
- rookie
- innings
- made
- got

When we ran the session on a topic, we leveraged the terms from the extraction process to build the queries that we used to select documents for subtopic clusters. These queries were applied to the 100 documents retrieved for the following day for that topic. To generate query terms for each new day, we leveraged each of the tag labels created by the participant up to that point in the session (for instance, to generate the queries for Day 5, we leveraged all of the document tagging from the participant on Days 1-4). Each tag label became a separate query, and the results comprised the subtopic cluster. To generate the subtopic cluster queries, we combined the tag label with the most frequently occurring six RAKE-extracted words across all of the documents that were tagged with that tag label, from any day up to the point.

After some trial and error, the the Indri query generation approach we used was the combination of the tag label (in any order, within 3 words of each other in the case of multi-word tags) and the six RAKE words (in any order in the retrieved document) formed the query. Further research could investigate options for improving upon this bag-of-words approach for generating subtopic clustering queries. Each Indri query was run against the new day's document set within the topic-relevant corpus, which reranked documents relevant to each cluster. The query gave a higher weight (2.0) to the terms from the participant's tag label than to the six RAKE-extracted words extracted from the documents that participants had labeled with that tag (1.0).

This result reranking process was run manually by the researcher prior to each day. Results for the next day were presented as clusters based on the participant's tag labels, and sorted based on relevance to the documents that were previously clustered into that participant's defined subtopic.

As an example of our query generation process, we look at the query generated from subtopic "FENWAY" for Day 5 of a sample session by the researcher. Fenway Park is the home baseball stadium for the Boston Red Sox. We note that during the dates selected for the Daybreak document search, the Red Sox were away, playing games hosted by other teams. While they were away from the Boston area, Fenway Park hosted a soccer (a.k.a. football) match between Liverpool and Roma. In a week when the Red Sox are in Boston, we would expect that the query related to a Fenway Park would likely include terms relevant to the Red Sox games, to include information about team members, Red Sox organist Josh Kantor responding to Twitter requests for songs, the Fenway ground crew responding to a weather delay, and other concepts. Instead, for this Fenway example we can see that the terms were customised to the events that happened during the days of the study; terms relate to soccer and the Liverpool visit rather than the Red Sox.

In this case, these documents had been tagged with the tag label "Fenway." For those

documents, the top six RAKE-extracted terms from the tagged documents were: “liverpool boston red sox rodgers fenway.” These were combined to form the following Indri query:

```
#weight( 2.0 #uw3(FENWAY) 1.0 #uw( liverpool boston red sox rodgers fenway))
```

During the post-day processing, we organized the subtopic clusters by what we referred to in the study as “subtopic importance.” For the user study, the proxy we used to represent subtopic importance was rarity—in other words, the subtopics with the fewest associated documents were displayed first. This was a simplifying assumption to minimize the number of configurable items in the interface and reduce post-day processing steps. When participants selected clustering, the system displayed the clusters data using a single, straightforward algorithm that ran after each day: counting the number of documents retrieved for each subtopic cluster query, then sorting the results to present clusters with the fewest documents at the top and the most documents at the bottom.

The logic in leveraging rarity related to the idea of information advantage—we applied the idea that rare information can have high value. For the blogging scenario in the study, the idea was that a blogger might get attention by producing an article on a topic that others have not observed. This use of rarity as a proxy for subtopic importance sets a baseline for the subtopic orderings that we later discussed with the participant, and that we leveraged in the evaluation design in Chapter 7.

5.3.4 Daybreak User Interface and Back End Design

We created the Daybreak system as a bare-bones prototype, focusing on implementation of the key segments of the change detection theory—group (clusters), pile (sort orders within clusters), and arrange (organizing the clusters themselves). Figure 5.3 contains a screenshot of the Daybreak system. The design approach was inspired by the idea of aggregated search, in which the results are divided into subtopic categories based on data

source; instead of combining results from multiple data sources in one interface, the Daybreak interface shows subtopic categories separately.

Before building the Daybreak system, we used Balsamiq to create low-fidelity wireframes to represent the interface and desired functionality. We adapted the design to work using our desired technology. The Daybreak system was not intended to be a production-grade capability; we implemented it in a way that required manual steps between days. We implemented the Daybreak prototype using the following front-end and back-end technologies:

Front End: To implement the interface, we used HTML and Javascript. We leveraged iFrames, which provides inline frames within a webpage. These frames were used to present a banner at the top, and the main workspace at the bottom. The main workspace was divided into three frames—the results metadata, the document contents, and the prior tags. These were implemented as follows:

- **Banner:** Presented at the top of the page, this largely static frame contained the system name, topic, participant ID, the current day, and the countdown timer for the current day.

- **Results metadata:** On the left side of the main workspace we presented the clustering and sorting selection options at the top. Beneath that was a section containing the subtopic clusters (if selected) and document metadata. The subtopic clusters were implemented as an accordion that was minimized by default; at the beginning of a new day, the participant initially could only see the cluster labels. To view the list of documents within that cluster, the participant had to click on the name of the subtopic cluster—which also logged the fact that the user had clicked on the subtopic. Upon clicking, the system presented a table containing a list of documents within that cluster—metadata presented included the document title, publisher, and publication

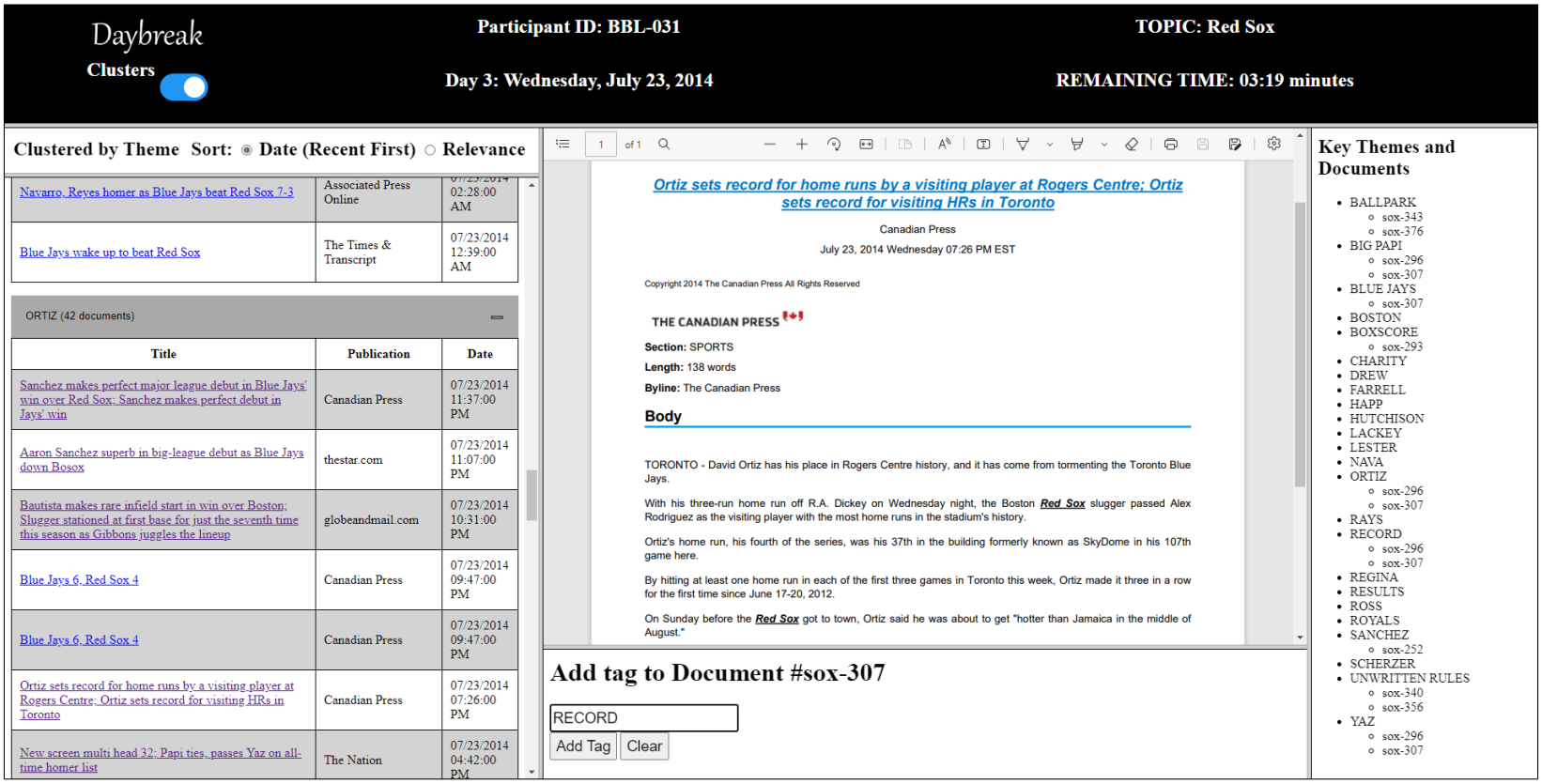


Figure 5.3: Final implemented interface for the Daybreak system. The "group" functionality is displayed in the subtopic clusters along the left side. The "pile" functionality is in the sort order used for the documents—in this case, documents within each subtopic cluster are sorted in reverse chronological order. The "arrange" functionality is in the order of the clusters, with the least populated one at the top of the pane.

date and time. The "Other" category was always presented at the bottom; this cluster contained any remaining documents relevant to the subtopic that were not retrieved for any subtopic clusters.

- **Document Contents:** The middle frame in the main workspace was used to present document contents to the participants. Documents were displayed as embedded Adobe Portable Document Format (PDF) files, which were generated by Nexis Uni. We renamed these PDF files with the assigned document number to enable display of the documents. We included a component at the bottom of the frame to support the participant's tagging process. This section listed the document number, and contained a text box for entering a tag label, and a submit button to add a tag. The participant could tag a document without adding a tag label.

- **Prior Tags:** The right-side section of the main workspace contained a tree of previously used tag labels and associated document numbers. The tree included tags from past days as well as the current day. For prior days, the tag label only (not the documents tagged with that tag label) were listed. For the current day, the tree displayed both the tag labels and the document numbers for documents tagged that day. When a new document was tagged, the system would add the document number to the block beneath the appropriate tag label; document numbers were sorted from lowest to highest for each tag label block. If the tag label had not previously been used, it was added to the tree, along with the associated document number. The prior tag tree could be be expanded or contracted using a plus (+) or minus (-) button.

Back End: Instead of implementing a database, the Daybreak system maintained logs and other data in the browser cache of the browser on which the Daybreak interface was displayed. We manually copied and saved these results for later analysis.

Separately, we created a set of Jupyter notebooks containing Python code. We used these to manually customize and set up each day's results. This included running the subtopic clustering searches and reranking in Indri and generating web pages.

While running a session, we separated the front-end and back-end process across two browsers to reduce confusion about which windows were intended to be presented to the participant, as opposed to windows which were used by the researcher to during a session. We leveraged Microsoft's Edge browser to display the interface to the participant, and used the Google Chrome browser to run the Jupyter notebooks that customized the result set for the participant.

5.3.5 User Study Artifacts for Analysis

To enable our eventual analysis of the Daybreak user study results and the context in which they were generated, we collected a variety of artifacts. This included the following:

- **Zoom Session Recording:** These recordings contain everything in the session after the showing of the training video, and ended after the semi-structured interview.
- **Surveys and Questionnaires:** This includes the participants' responses to the qualifying survey, the pre-study questionnaire, and all five post-day questionnaire responses; all of these artifacts were collected using Qualtrics.
- **Storytelling Task:** This was the final blog article outline crafted by the participant after Day 5. It was collected using Qualtrics. The initial outline created after Day 1 was a practice attempt at completing the task (for familiarization purposes), and was not considered one of the artifacts.
- **System Logs:** The logs contain information about participants' usage of the system. For Daybreak use, this includes documents viewed, tags and tag labels applied (by

document), subtopic clusters viewed. For the storytelling task, we collected logs about tag labels and associated documents viewed while the participant drafted their outline.

- **Researcher's and Observer's Notes:** These were documented during and immediately after the session, and were intended to capture specific issues or thoughts during the session.

5.4 Initiating the Daybreak User Study

In this this section we detail our interactions related to the Daybreak user study selection process. This includes the process we applied to test the initial prototype with pilot users, as well as the approach for finding and selecting participants through our selection survey. To run this qualitative user study, we aimed for recruitment of at least 12 participants [22]. We also used the information provided to aid in selection of specific topics to test within the Daybreak system, so that the participants would have experiences resembling their personal search for updates on a topic of interest. We planned to add additional participants until we reached conceptual saturation, a point where for each additional participant we are not learning as much new information [114]. Our Institutional Research Board (IRB) approval was for a maximum of 30 participants (including partial sessions). Additionally, our funding would not support more than 30 complete sessions. A copy of the IRB approval document is included as Appendix C.

5.4.1 Pilot Sessions

Prior to the formal study, we ran six pilot sessions to verify the timing, to ensure that the system worked as expected via the Zoom interface, artifacts were generated correctly, and to determine that the results were interpretable. During these tests, Zoom proved to

be sufficiently capable as a platform for remote use of software for a study that had not originally been designed with remote use in mind.

The pilot sessions were productive in optimizing the study as well as improving and hardening the system. All pilot sessions used the Red Sox topic, which was the earliest compiled collection, which was also used for prototyping and initial testing. Based on the pilot session, we addressed issues and made improvements. We simplified the overall structure of the user study by reducing the number of questions in the questionnaires, reduced the length of each day (initially ten minutes, but decreased to eight minutes after the initial pilot), and switched from a full storytelling task to an outline only. As for Daybreak system improvements, we developed bug fixes and ensured that unexpected input was handled. One such case involved introducing code to prevent system crashes in cases where participants added numbers as tags.

5.4.2 Qualifying Survey

Enrolling a sufficient number of participants is a difficult challenge for any study. The primary goal of the qualifying survey was to be able to select at least 12 participants for the user study. Because of the topics selected for the study, we were seeking adults located in the United States. The participant location was selected because information seeking behaviors may vary in other countries. We created an initial qualification survey that we used to find potential participants age 18+ located in the United States. This survey was created in the University of Maryland Qualtrics site. Advertising for the qualification study was performed through email and social media, targeting individuals who perform change detection tasks, and who were located in the United States. We emailed our announcement to professors and researchers across the U.S. in the fields of journalism, information and

library sciences, and related disciplines. To expand our reach, we posted about the study to research-related groups and hashtags on Facebook and Twitter.

In support of our participant recruitment process, we included questions such as the following in our qualification survey:

- Do they follow a topic over time?
- What data sources do they use to follow their topic of interest?
- How often do they seek updates? (daily, etc.)
- What topic(s) do they follow?
- What (if any) terms (hashtags, search terms, etc.) do they search to get updates?
- Basic demographic information, so we could attempt to recruit a diverse mix of participants

The complete qualifying questionnaire is included in Appendix D.

5.5 Characterizing Qualified Respondents

We received more than 200 initial responses to the Qualtrics selection questionnaire. All except one of the respondents indicated that they perform change detection tasks; the respondent who indicated that they did not complete change detection tasks was deemed not qualified to participate in the user study. Initially, we believed that 143 of the respondents were qualified. Many of the responses that were ruled out were non-US responses or clearly fraudulent responses. We identified some of the fraudulent responses to the qualifying survey by observing multiple submissions with similar details that were all sent 2-5 minutes apart, at fairly regular intervals—some of which originated from the same IP address. When IP addresses differed in such sequences, we suspect the use of a Virtual Private Network (VPN) to mask the true IP address. Some such respondents attempted to appear legitimate by creating multiple email addresses and completing the survey from different IP address.

Responses with these characteristics originated from locations that appeared to be in the U.S., China, India, or Kenya, based on Qualtrics metadata.

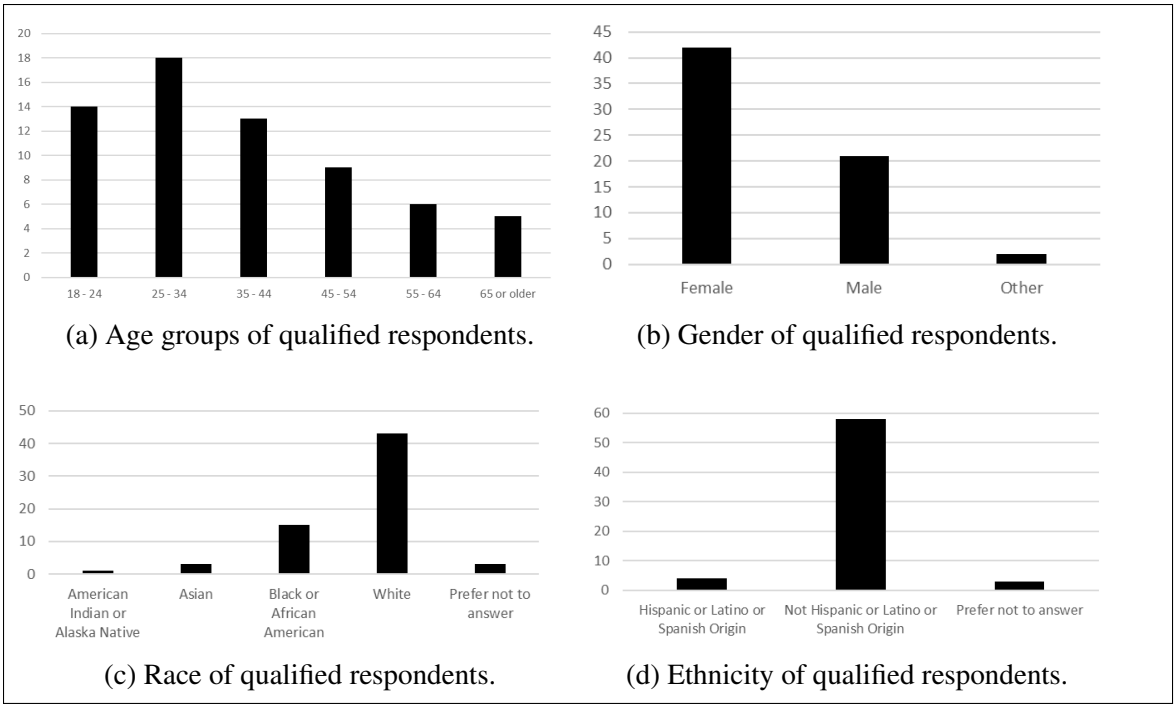
On subsequent examinations, some of what we had initially believed to be valid responses turned out to be fraudulent responses. After several rounds of filtering (e.g., removing respondents from the same IP, non-qualifying locations, etc.), we ended up with 65 responses that we determined were likely valid; some possibility remains that there were undiscovered cases of fraudulent responses that were more sophisticated in their use of fabricated emails, Virtual Private Network (VPN) technology and other approaches to hide their origins. We discuss these challenges in more detail in Section 5.6.2.

In this section, we present information about the respondents to the qualifying survey. While the primary purpose of the survey was to facilitate selection of Daybreak user study participants, the demographics and other information presented here provides additional insights into the characteristics and interests of individuals who perform change detection tasks. We were looking to select a participant set that had overlapping topical interests, so we could assign multiple participants to the same topic for comparison purposes. At the same time, we were looking to maximize demographic variety across the participant set. The factors we attempted to diversify included location, age, gender, as well as race and ethnicity.

5.5.1 Demographics of Qualified Respondents

We reviewed the demographic information provided by the 65 respondents determined to be qualified for the survey. Figure 5.4 contains subfigures representing the age, gender, race, and ethnicity of the qualified respondents.

Respondents represented a wide range of age groups, with representation in all age ranges, starting from 18-24 and through to 65 and older. We received multiple responses



(a) Age groups of qualified respondents.

(b) Gender of qualified respondents.

(c) Race of qualified respondents.

(d) Ethnicity of qualified respondents.

Figure 5.4: Demographics related to age, gender, race, and ethnicity of the qualified respondents to the Daybreak user study qualifying survey.

for each age range, with the most responses (28%) from the 25-34 age group. We received the fewest from the group of 65 or older respondents (8%).

As for gender diversity, 65% of the qualified respondents identified as female. 32% of respondents identified as male, and the remaining 3% selected "other."

The qualified respondents covered a variety of racial and ethnic backgrounds. The largest subset, 66% of the qualified respondents were white. 23% of respondents were Black or African American. 5% were of Asian background, and the remaining 2% were from an American Indian or Alaska Native background. The remainder did not provide information about their racial background. The majority of respondents (89%) were not of Hispanic, Latinx, or Spanish origin; 6% of qualified respondents were of Hispanic, Latinx, or Spanish origin; and the other 5% chose not to respond to this question.

The qualified respondents represented a variety of locations across the United States.

There were a few clusters in specific cities, to include near San Francisco, Dallas, Detroit, Boston, and Washington, DC. These clusters of locations are likely related to the advertising approach; these respondent locations align with locations of universities and organizations that received our targeted email messages.

5.5.2 Change Detection Preferences of Qualified Respondents

Through the selection survey, we were able to learn about respondents' interests in and performance of change detection tasks. Given that it was one of the qualifying questions, all qualified respondents indicated that they perform change detection tasks. Only one respondent to the qualifying survey had indicated that they do not perform change detection tasks.

The survey included a number of questions about respondents' change detection task performance, ranging from their personal topics of interest (used for selecting Daybreak study topics) to their reasons for performing change detection tasks. Respondents were allowed to select multiple responses to this questions. As indicated in Figure 5.5, 43% of responses were from respondents who perform change detection only for personal reasons. Another 9% perform change detection tasks only for professional reasons. A further 40% perform change detection tasks both for personal and professional reasons. The remaining responses (7%) indicated usage for other reasons; the explanations mentioned by respondents included keeping up with current events, career development, lifelong interest, and health. Two of the respondents indicating "other" reasons also performed change detection tasks for personal or professional reasons.

As shown in Figure 5.6, the vast majority of qualified respondents indicated that they perform change detection tasks on at least a daily basis, with 42% seeking information multiple times a day and 37% looking for updates once a day. An additional 20% sought

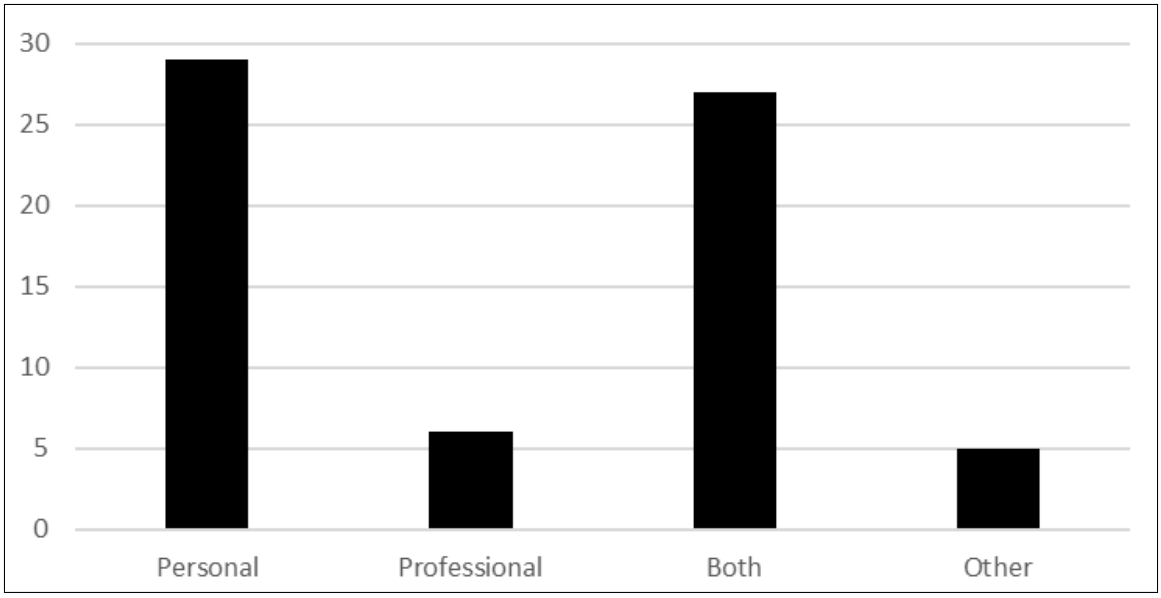


Figure 5.5: Why the respondent performs change detection tasks. Note that respondents were able to select multiple answers to this question. Responses in which a respondent selected both "personal" and "professional" are reflected in "both." Some respondents selected the "other" option in addition to one of the other responses.

information on a weekly basis. Only one respondent said they seek updates on a monthly basis, and no respondents seek updates on their topic(s) of interest less frequently than that.

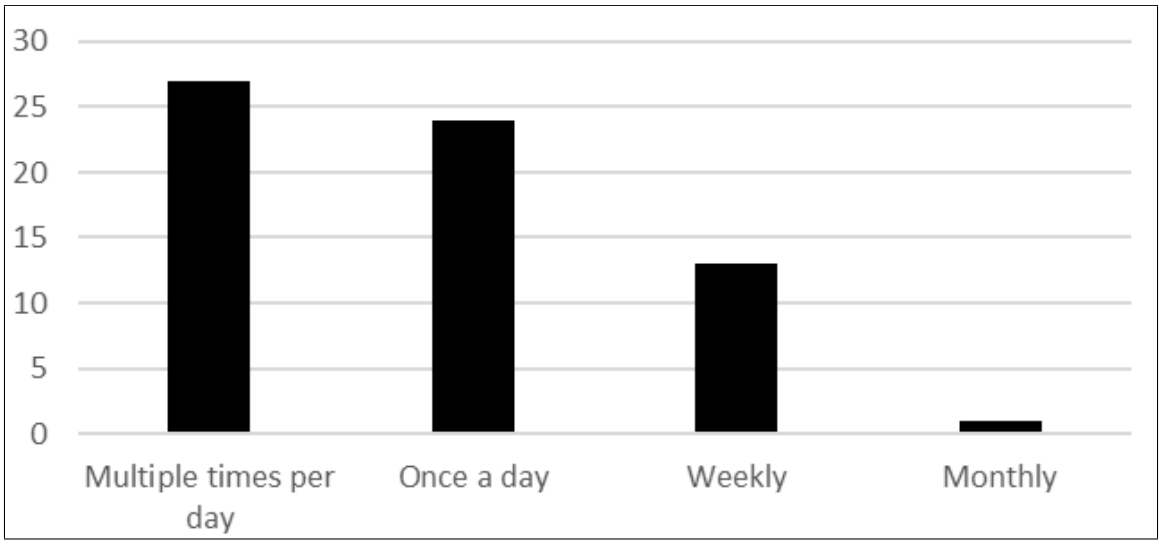


Figure 5.6: How often respondents perform change detection tasks.

The qualified respondents follow a wide range of topics of interest. In Figure 5.7 we have included the specific topics mentioned by the respondents. Some of the more frequently mentioned topics included news updates, technology, politics, business, and sports.

Figure 5.7: Overview of topics that respondents to the qualifying survey indicated that they follow.

5.6 Daybreak Study Participant Selection

To meet our goal of recruiting at least 12 participants, but not more than 30, we reviewed the qualifying survey responses to find suitable candidates. In particular, we focused on the topics of interest listed by qualifying survey respondents. We wanted to choose people who had an interest in topics that were also followed by other respondents; this would aid us in selecting topics that could be used during multiple sessions. Based on the responses to the qualifying survey, we chose five topics that appeared relevant to multiple respondents, and had sufficiently large result sets (enough new documents within a 24-hour period that the participant would not be able to review all of the documents presented on a given day). This was the point where we finalized the five topics for the study: Boston Red Sox, cryptocurrency, global health trends, satellites and space debris, and extreme weather.

We then selected participants to assign to each topic based on our assessment of similarity between a study topic and one of the respondents' self-reported topics of interest. Prospective participants were assigned to the topics that we believed were most closely aligned with their actual interests, as conveyed in the qualification survey. For instance, a participant who was interested in financial trends might be assigned to the cryptocurrency topic. Prospective participants were each assigned a participant ID to anonymize their activity. The participant ID was a combination of the 3-digit topic code (described in Section 5.3.2), plus a random 3-digit number (e.g., BBL-100 for a participant assigned to the Red Sox baseball topic).

5.6.1 User Study Sessions: From Scheduling to Completion

From the pool of 65 qualified respondents, we selected a balanced set of potential participants who indicated that they perform change detection tasks on a diverse range of topics. In our initial selection, we attempted to maximize the diversity of the participants

across genders, race and ethnicity, age, and location. In practice, not all of the respondents we had selected ended up participating; we scheduled 27 sessions, but not all resulted in a valid session. For five of the 27 scheduled sessions, the participant did not show up at the prearranged time; three did not respond to emails requesting to reschedule. Two were rescheduled, but ended up as some of the seven remaining sessions that were incomplete and could not be evaluated—including three that were believed to be fraudulent participation. When one was disqualified (fraudulent participation, no-show, etc.), we swapped in another qualified participant. The remaining 15 sessions were complete, recorded sessions in which the participant finished all Daybreak user study tasks.

The questionnaire asked for examples of specific topics followed over time, to include current events such as entertainment, product brand management, or sports. We aligned participants with topical interests, in particular identifying topics for which multiple qualified respondents expressed an interest, to minimize the burden of organizing study documents.

5.6.2 Challenges with Participant Recruitment and Selection

We encountered a number of issues when recruiting participants for the Daybreak study, in particular due to the pivot from an in-person study to a virtual study.[1] Two themes emerged: first, trust mechanisms were weaker than would be expected for an in-person study, resulting in greater coordination difficulties; and second, what seemed to be a fair reimbursement rate appears to have provided an outsized incentive for fraud. Our study revealed a number of challenges, in particular related to recruiting enough responses to maximize the diversity of our study participants. Participants were lost to both expected factors (e.g., nonresponsive people or incomplete sessions) and from a surprising number of

[1] Challenges faced during the Daybreak study were published as a poster paper in the Proceedings of the ASIS&T 2022 Conference in Pittsburgh, PA [218].

participation attempts from people outside the Institutional Review Board (IRB)-approved demographic (adults in the U.S.).

We posted about the study to research-related groups and hashtags on Facebook and Twitter. Although the posts went to a large number of people, they likely were viewed by many outside the targeted study population; this turned out to be the first link in a chain that led to fraudulent participation. We had chosen to reward participants (adults in the U.S.) at what we thought to be a fair rate ($25 for approximately 90 minutes); that rate seems to have been high enough to provide an incentive for fraud, at least in some regions. Survey fraud has been noted by others as well, particularly during the COVID-19 pandemic [36, 245, 255, 306]. Qualtrics metadata helped us identify some problematic responses, but much of our review for fraud was manual. For future studies, we recommend using different survey links for each advertising approach so that if fraud is detected in one response, others advertised in the same way can be more closely examined.

Additionally, we learned later that our invitation email messages sometimes went to a spam folder—possibly because we sent standardized, formally-worded messages to multiple recipients. In future studies, we might consider a more focused recruiting effort, relying more heavily, for example, on snowball sampling.

We were ultimately able to complete sessions with 15 participants, though we had to schedule nearly twice as many sessions (27 in all) before meeting this mark.

Due to our user study recruitment challenges, our final participant set was less evenly distributed than we had anticipated. In some cases, this resulted in participants who were less effectively matched to topics. for instance, a participant who expressed an interest in science may have been assigned to the space topic, even if they did not specifically indicate interest in space.

Some fraudulent participation was detected only after a session began. One participant connected using a mobile phone with a U.S. IP address, but when we asked them to move

to a laptop (which the qualification questionnaire had stated was needed for the study), their laptop IP was geolocated to Africa. In another case, a participant insisted that they were connected from a computer, but according to metadata, they were using a mobile phone. This precluded their ability to control the mouse, and thus prevented their participation in the study. We also saw some implausible inconsistencies between the demographic data collected with the qualification questionnaire and from responses during the actual session. We interpreted this as indicating that the participant may have provided false data on the qualification questionnaire, and then a few weeks later did not remember what they had originally submitted. IRB guidance specifically discouraged us from repeating questions during the study that we had asked during qualification. While minimizing repeating questions is excellent advice in a face-to-face setting (to minimize burden on participants), in an online study we recommend the practice as a fraud detection measure.

In one case, the same individual attempted to participate three times, having signed up under multiple names and email addresses, having provided different enough data on interests and demographic details that we selected them three times! One way of dissuading or detecting such cases would be to require that participants share video using their camera. We did not require this because we did not want to unnecessarily exclude some participants for technical reasons—such as bandwidth or lack of a camera. Our pattern-based fraud detection had failed in this case (and in the two cases described above), so we looked more closely to see if there were other cues that we had missed during qualification. We noted no instances of fraud for people using organizational email addresses (e.g., university email accounts), although those accounted only a minority of the respondents that we selected. Some email accounts using services such as Gmail or Yahoo Mail displayed additional information that suggested actual personal use (e.g., a profile photo or a chat link). Limiting participants to those whose email accounts have such markers of veracity would likely have

excluded many qualified respondents as well; however, the absence of such markers might suggest the need for additional vetting.

We also encountered the typical problems of scheduling difficulties and no-shows, though in somewhat greater numbers than would have been expected for an in-person study. After selection, we scheduled sessions using YouCanBook.Me; some selected participants did not respond to our scheduling email. Some invited participants may have been concerned that the unfamiliar domain was a phishing attempt. We might avoid some of these problems by using personal messages to schedule sessions, which is practical for a study of this size. Despite the scheduling site sending out an invitation plus an email reminder before the session, and our personal email reminder 12-24 hours in advance, some people who had signed up did not show up for their session. Our study was conducted in December 2021 and January 2022, a period when holiday schedules may have resulted in changing availability.

5.7 Details about Selected Participants

We ended up with our final set of 15 Daybreak user study participants after 27 attempts at selecting participants. This set of participants—while somewhat diverse—was less demographically diverse than originally planned. In this section we discuss the final set of participants for the Daybreak study, to include their demographics as well as their views about change detection.

5.7.1 Participant Demographics

To paint a picture of the demographics of the Daybreak user study participants, we draw from the qualifying survey as well as the pre-study questionnaire. In order to minimize the burden placed on participants, we attempted to ask demographics and personal preference

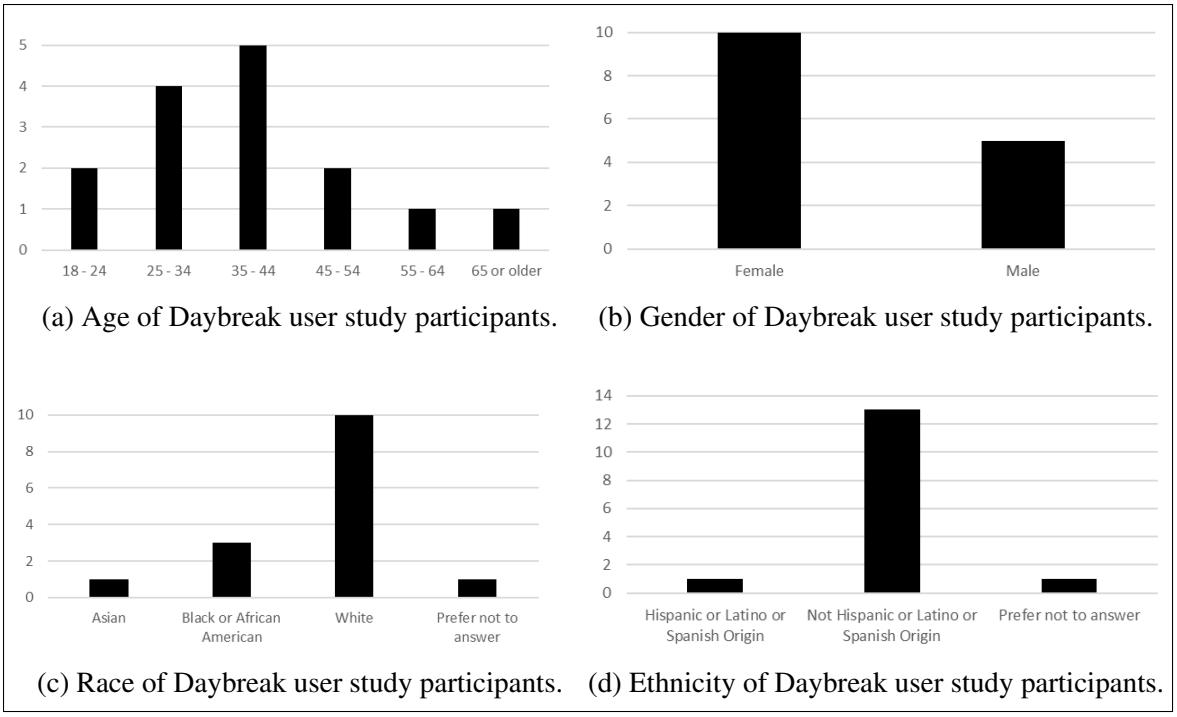
questions in only one place—either the qualifying survey or the pre-study questionnaire. We supplemented the information from the qualifying survey with a pre-study questionnaire that included additional details about the participants' background, to include education, employment, and other information. The participants' demographics for age, gender, race, and ethnicity from the qualifying survey are displayed in Figure 5.8.

Age of participants is displayed in Figure 5.8a. The largest number of user study participants were in the 35-44 (five participants) and 25-34 (four participants) age ranges, respectively. We also had participants from each of the other age ranges represented in the qualifying survey, though we had only one participant apiece for the 55-64 and 65 and older age ranges.

While we hoped for more gender diversity across our participants, our actual numbers were consistent with the response rates of qualified respondents from the selection survey; these responses are included in Figure 5.8b. Ten of our participants identified as female, and five identified as male. Future research could look at whether women are more likely to participate in change detection tasks than men, or if the imbalance was an outcome of our selection approach—or perhaps this is consistent with other research showing that women are more likely to participate in surveys and other studies [247].

The composition of race across the participant set represented less diversity than there was across the qualified respondents; these results are shown in Figure 5.8c. Within our the user study participants, ten were white, three were Black or African American, and one was Asian. One respondent chose not to disclose their race. As for ethnicity (Figure 5.8d, one study participant identified as being of Hispanic, Latinx, or Spanish origin. One did not provide information about their ethnicity. The remaining thirteen participants indicated that they are not of Hispanic, Latinx, or Spanish ethnicity.

As shown in 5.9, the US-based geographic locations of the Daybreak user study partic-



(a) Age of Daybreak user study participants.

(b) Gender of Daybreak user study participants.

(c) Race of Daybreak user study participants.

(d) Ethnicity of Daybreak user study participants.

Figure 5.8: Demographics related to age, gender, race, and ethnicity of the Daybreak user study participants.

ipants clustered around locations on the West Coast and East Coast. There were also a few participants located in the Michigan area, one in North Carolina, and one in Colorado.

The above demographic information was supplemented with a few additional questions about the participants from the pre-study questionnaire. This includes questions about the participants' education and employment.

The participants in the Daybreak study tended to have advanced degrees. Three of the participants have up to a Bachelor's degree, without graduate degrees. 11 participants have a Master's degree, professional degree, or a Doctorate. One participant has a high school diploma.

Most (11) of the Daybreak user study participants were employed on a full-time basis. Two of the participants indicated that they are students. Two others are employed on a part-time basis. The remaining participant is self-employed. The respondents represented

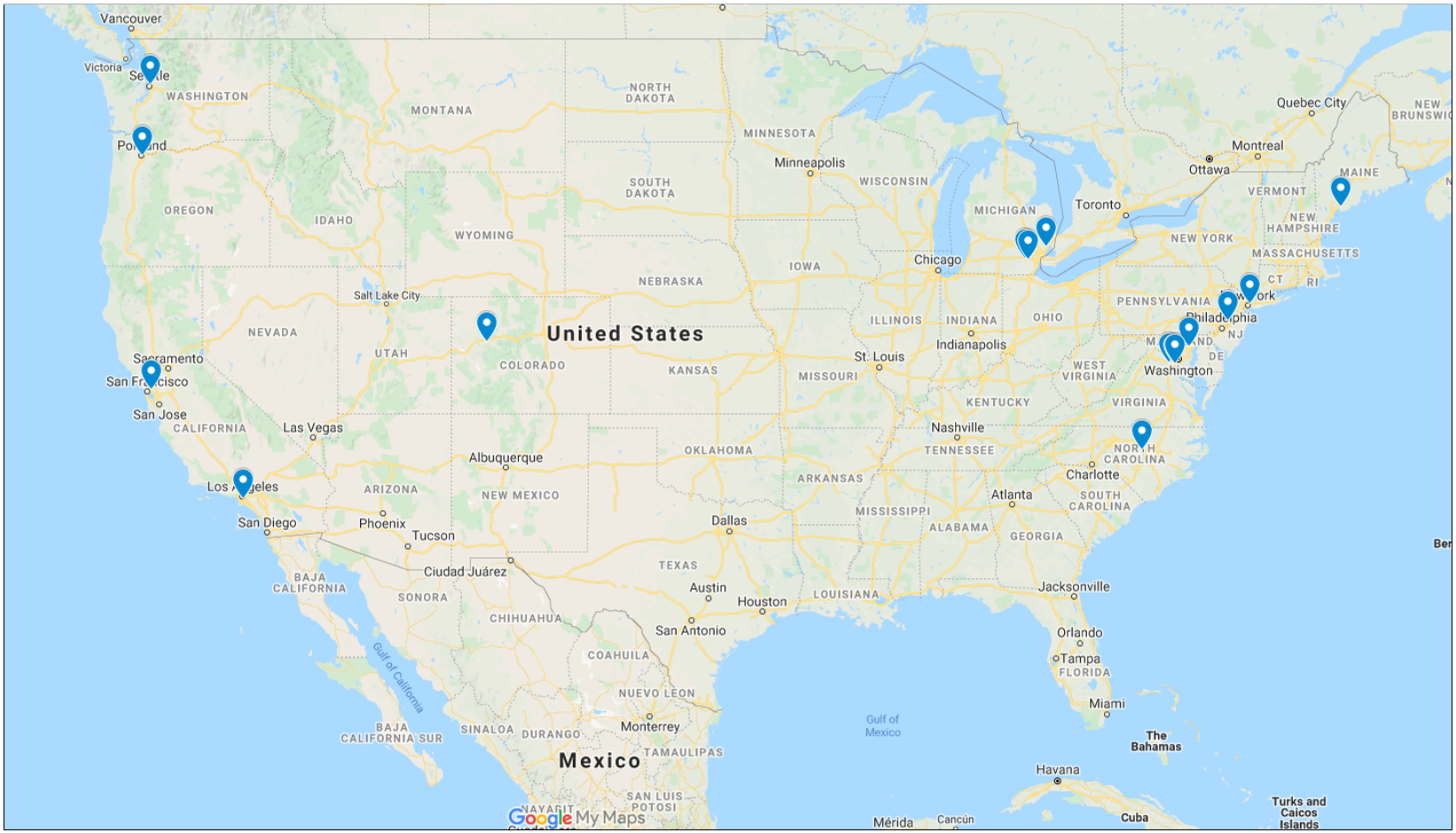


Figure 5.9: Locations of Daybreak user study participants.

a range of industries, including technology, education, government, publishing, and non-profits.

5.7.2 Insights about Participants' Performance of Change Detection Tasks

To expand upon the information provided in the qualifying survey, at the beginning of each session we asked the participant to complete a pre-study questionnaire. This allowed us to gain a better sense for the participant's background and interests, to include getting more detailed demographic information that was not necessary for the selection process.

Most Daybreak study participants had prior experience with systems that provide tagging capabilities. Three participants had no prior experience with tagging, and two participants were not certain whether they had used tagging systems previously. Examples of systems that participants had previously used included Atlas.ti, NVivo, Getty, bookmarking systems, tagging capabilities in email systems, Confluence, and social media sites (hashtags).

Based on the results of the qualifying survey, the Daybreak study participants perform change detection tasks for personal or professional reasons (Figure 5.10). Nine of the participants perform change detection tasks for both personal and professional reasons. We delved further into their reasons in the pre-study questionnaire; we asked participants to select from a variety of explanations for their interest in keeping up-to-date on a topic of interest. Curiosity was universally identified as one of the reasons for completing change detection tasks. Other common responses included entertainment, desire not to miss something interesting, and various social pressures—to include work expectations, or because friends and family are talking about the topic.

We also asked the participants in the pre-study questionnaire to note specific topics that they follow over time. While some of these aligned with the ones provided in the

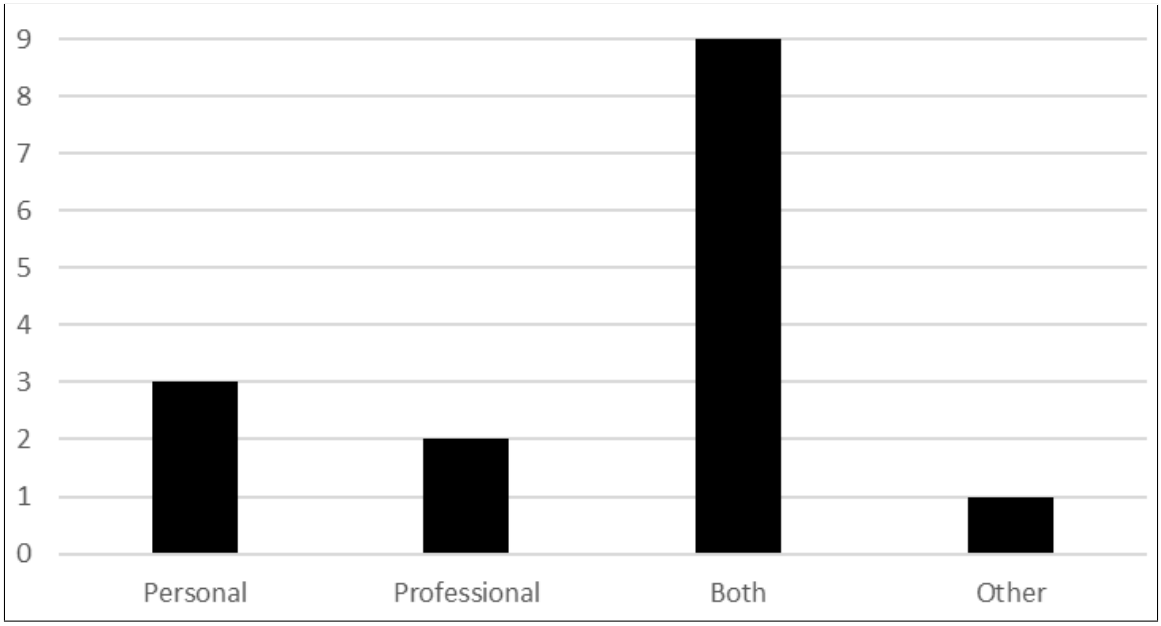


Figure 5.10: Why the participants perform change detection tasks. Note that respondents to the qualifying survey were able to select multiple answers to this question. Responses in which both "personal" and "professional" were selected are reflected in "both."

selection questionnaire, several participants provided additional topics of interest as well. This includes news about music, television shows, video games, ethics in AI, and more. When asked to estimate their level of expertise on their topic of interest, 14 of the 15 participants said they have an intermediate or better knowledge of their topic of interest. Only one participant indicated that they have only a basic level of knowledge of the topic that they follow.

This strong level of expertise was further reflected in the participants' responses to a question about the length of time that they have been following their topic of interest. Nine of the 15 participants have been following their topic for more than 5 years. Four participants have been following their topic of interest for 2-5 years. Two participants have focused on their topic for 1-2 years. No participants indicated that they have followed their topic of interest for less than a year.

These participants also look up information about their topic on a regular basis. Ac-

cording to the responses, seven of the Daybreak participants spend 10 minutes to 1 hour a day looking up information about their topic of interest. Four participants spend 1-2 hours on their topic per day. The remaining four participants spend less than 10 minutes per day catching up on their topic of interest.

When asked how recently they had searched for information about their specific topic of interest, nine of the 15 participants had searched for their topic of interest within the past day—and three of those participants had looked for updates within the hour before their session. Five participants had searched their topic within the past week, and one had looked it up within the past month. These responses were consistent with their responses to a question in the qualifying survey about how frequently they look for updates on a topic of interest, indicated in Figure 5.11.

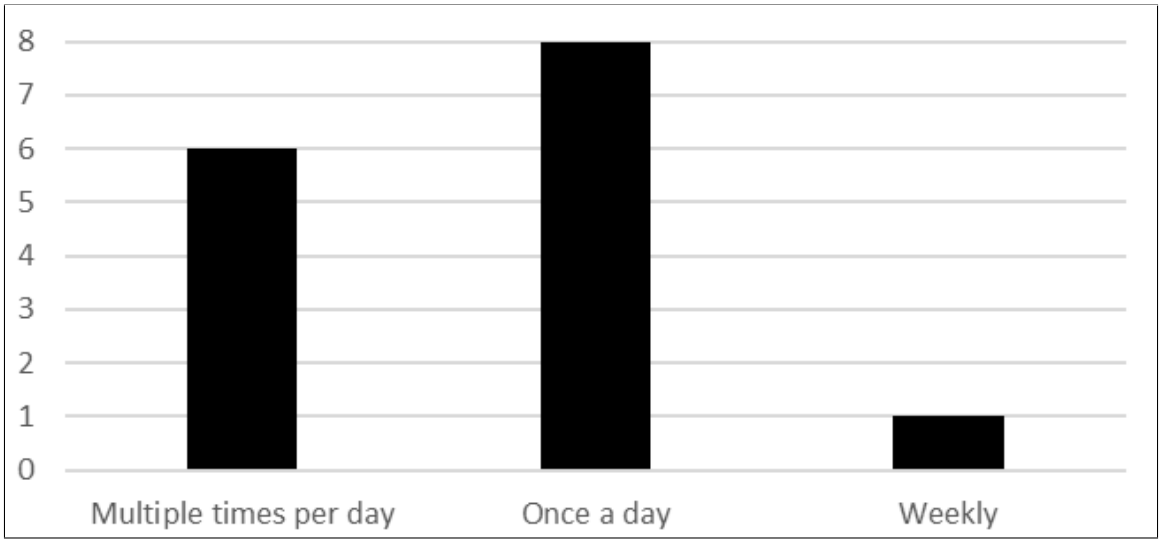


Figure 5.11: How often the participant performs change detection tasks. Note that all participants indicated that they perform change detection tasks weekly, or more frequently.

We asked the participants in a few ways to describe how they currently seek information about their topics of interest. Their responses indicated heavily search-driven approaches for seeking new information. In an open-ended question about their change detection workflow, most participants said that their process starts by searching for their topic, either in

a Web search engine (Google, Google News, or Bing) or in social media (especially Twitter, Reddit, SnapChat, or TikTok. A handful of participants use other resources, such as podcasts, browsing directly to specific curated sites of interest, as well as RSS feeds and Google News alerts. Their responses to questions emphasize this reliance on online news sites, search engines, and social media sites for gathering updates on the topic.

The participants were asked a few questions to gather additional details about their information seeking practices related to change detection tasks. These participants use a wide range of devices that the respondents typically use to perform change detection tasks. All except one of the 15 Daybreak study participants leverage their phone to get updates on their topic of interest. Nearly all (13 participants) also leverage laptop computers as part of their update process. A smaller number make use of either tablets (6 participants) or desktop computers (5 participants) to support their updates.

Consistent with results from the sort order survey in Chapter 4, the participants indicated that they perform change detection tasks at various points during the day, with the most common being updates in the evening—many respondents to that survey saw this as an end-of-day task. Similarly, as show in 5.12, the largest number (9 participants) performs change detection tasks in the evening. Note that participants were able to select multiple times of the day in their responses to this question, and may seek updates at various points throughout the day. After participants who perform change detection tasks in the evening, the next highest numbers of participants (7 participants) sought updates in the mid-to-late morning, afternoon, and night. From these responses, we did not discern an identifiable pattern or consistency in when these participants perform change detection tasks.

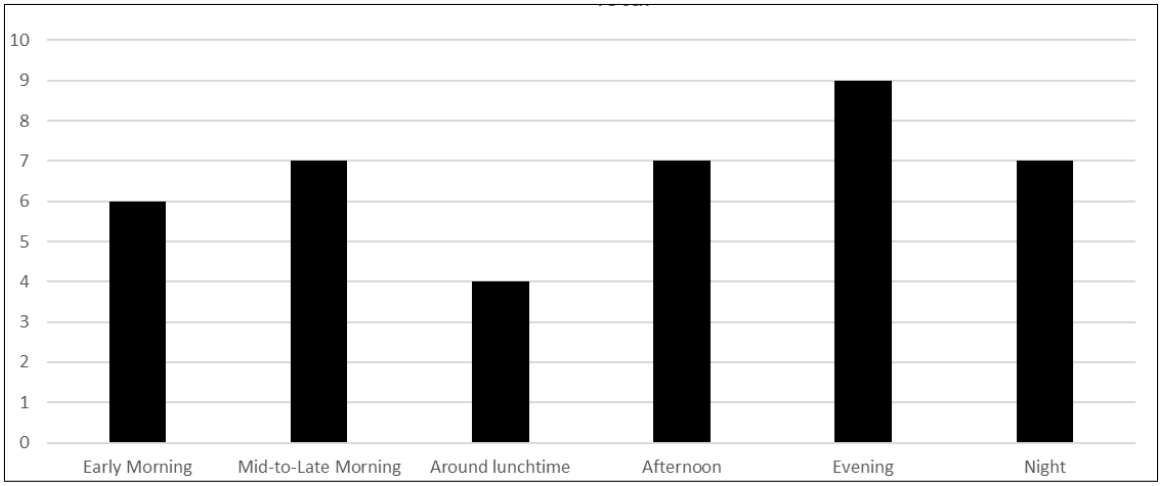


Figure 5.12: Time of the day when the respondent performs change detection tasks. Note that respondents were able to select multiple answers to this question.

5.7.3 Participants' Use of Technology

The Daybreak user study participants tend to be heavy users of computer systems. Of the 15 participants, six indicated that they spend an estimated 8-12 hours using their computer outside of work or school, which is a very large amount of computer use. Five of the participants use their computer 2-4 hours per day. An additional three participants use their computers 4-8 hours per day. Only one participant indicated that they spend 1-2 hours per day on technology use. Overall, the participants indicated large amounts of personal time spent using a computer. While this may be an accurate representation of their computer use, it is also possible that they may have misinterpreted the question as being inclusive of work or school-related computer use.

In our recruitment survey, we requested that participants have either a laptop or a desktop machine available for the study to ensure that the full system would be visible. 12 of the 15 Daybreak user study participants connected into the study from a laptop. Three participants indicated that they were using a desktop computer. The one participant who said they were leveraging a phone noted that this was as a supplement to their laptop. The majority of

Daybreak participants were on the Windows operating system. The other four participants were using Apple products. The participant who indicated use of a supplemental phone was on an Android device.

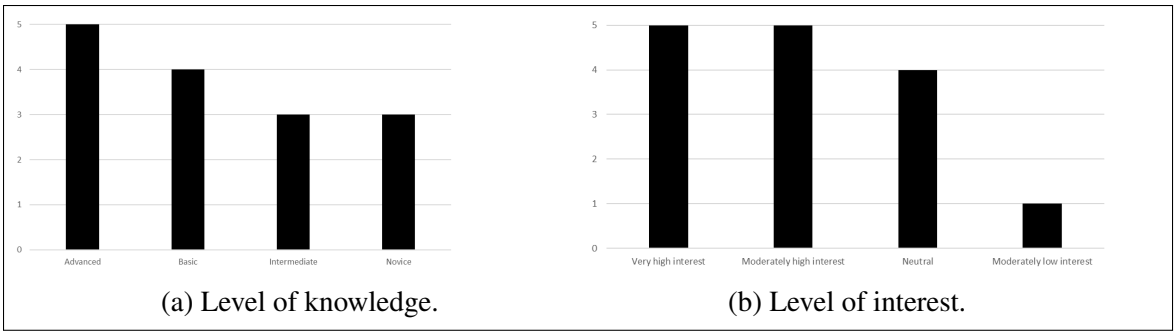
5.7.4 Assigning Participants to Daybreak Topics

We made an effort to assign people to topics that we believed would be similar to their interests. However, we were not successful in all cases. For example, a participant who said they were interested in news related to COVID would be assigned to the global health topic. Although all participants were able to complete the storytelling task effectively, the sessions that were misaligned with both knowledge and interests may differ from one in which a participant already has familiarity with the topic. Figure 5.13 covers both the knowledge and interest levels of the participants, as an indicator of the extent to which we successfully assigned participants to topics for which they have at least some knowledge and interest.

Figure 5.13a shows the participants' self-assessed familiarity with their assigned topic. As for their levels of interest in their assigned topic (Figure 5.13b) ten of the participants expressed a moderately or very high interest in the topic assigned to them. Four participants were neutral about the topic. One participant indicated that they had a low interest in their assigned topic.

5.8 Summary

Based on the GPA Change Detection Theory (Chapter 3 and the results of the sort order survey (Chapter 4), we designed a system and a user study to test whether a system can meet the change detection needs of users. To support the study, we designed and implemented the Daybreak system. Daybreak enables users to follow a topic of interest over



(a) Level of knowledge. (b) Level of interest.

Figure 5.13: Figures showing the participants' self-assessed level of familiarity and level of interest in their assigned topics for the Daybreak user study.

time, by reading documents, then adding tags and tag labels to documents. The system then would take new results and apply functionality from the GPA Change Detection Theory: group (subtopic clusters), pile (document sorting), and arrange (arranging subtopic clusters). The user study focused on testing the prototype Daybreak system, to determine the extent to which participants could leverage it to complete a change detection task. The study simulated a five-day time span, in which the user would review documents from a topic related to the participants' interests. Participants would use the system to follow a topic across five simulated days, each one representing a day's worth of new documents on their assigned topic. At the end of the task, the participant would be asked to generate an outline summarizing key developments on the assigned topic.

We used a selection survey to recruit participants for the Daybreak user study. From a set of 65 qualified respondents, we recruited 15 participants. Based on the interests of the participants, we selected five topics for the study: Red Sox baseball, cryptocurrency, global health, space, and extreme weather. We assigned three participants to each topic in order to compare different participants' behaviors within the same dataset. The results of the Daybreak user study are presented in Chapter 6.

Chapter 6: Testing a System in a Real-World Change Detection Scenario

In this chapter we present results and analysis from the Daybreak study, which was described in Chapter 5, a mixed-methods study of user preferences for change detection. The goal of this study was to understand whether the prototype Daybreak system could meet the change detection needs of users, based on the components of the GPA Change Detection Theory. We ran 15 sessions, which focused on participants' experiences with the working prototype of Daybreak. We assessed whether participants were able to complete a change detection task, and looked at the extent to which the Daybreak system met their change detection needs. We designed our study to aid in understanding whether the preferred organization indicated in the sort order survey in Chapter 4 is consistent with the preferences expressed during the Daybreak user study. This research bridges the gap between our prior sort order survey—which revealed participants' general preferences for organizing data for update tasks—and the evaluation design, which describes an approach for comparing aspects of change detection-focused systems.

6.1 Data Collection and Analysis

For this study, we devised a scenario where a participant looks for daily updates on a topic that they follow over time to support generation of a notional blog post outline. All 15 of the user study sessions took place via Zoom. As described in Chapter 5, each session consisted of the following components:

- Startup and Task Familiarization
- Day 1
- Post-Day Activities
- Practice Storytelling Task
- Days 2-5
- Final Storytelling Task
- Semi-Structured Interview

For each day, the participants were able to view up to 100 historical news articles on an assigned topic to simulate a process they might use to detect changes or developments related to their topic of interest.

At the beginning of each session, the researcher (the author) provided a brief overview of the study, then asked the participant to complete a pre-study questionnaire. The participant then viewed a brief (six-minute) training video that included a description of the task and a walk-through of a Daybreak session. After the training video, the researcher then displayed the live Daybreak system, started the timer for Day 1, and granted keyboard control to the remote participant via the Zoom application. The participant was given up to eight minutes to complete document review for each of the five days in their session. Sessions were recorded for transcription and coding [94, 234]. After completing the study, participants received an Amazon gift card. Copies of the questionnaires and semi-structured interview are included in Appendix D.

6.1.1 Application of Research Methods

As noted in Section 5.1.2, we applied the framework method for the Daybreak user study. The framework method is a qualitative research method under the group of analysis methods known as thematic analysis or qualitative content analysis [90]. Its structured

approach makes it accessible and easy-to-use, and is applicable to researchers from a wide range of backgrounds, including novices. The framework method consists of two broad components: creating and then applying an analytic framework. The five main steps defined under the framework method include data familiarization, identification of the framework (in alignment with the research questions), indexing (also known as coding) data in alignment with the framework, charting and summarizing the data, and finally mapping and interpreting patterns within the data [96]. To support our use of the framework method, we leveraged Atlas.ti, a qualitative analysis tool that contains functionality for coding, searching, and analyzing study artifacts.

We implemented the steps of the framework method in the following ways, with additional details about our analysis noted in Section 6.3:

- **Familiarization:** We reviewed each session recording, and the session transcript. We compiled log files from each session, and standardized time notations to start at the beginning of sessions; this enabled us to compare the amount of time taken for different participants' sessions. We also gathered other artifacts related to the sessions, to include the questionnaires and story outlines, as well as the researcher's and notetaker's notes.

- **Framework Identification:** We developed a codebook based on our framework, which was the GPA theory. This included codes that would enable us to answer the Daybreak user study research questions (Section 5.1.4), which related to the group, pile, and arrange components of the theory, as well as a question about the combination of these components. While the main focus of the study was to gauge the utility of the components of the GPA Change Detection Theory, we also ensured that our codes also included topics beyond the research questions—reflecting the deductive and inductive processes of the framework method. These included codes related to

the overall user experience, codes related to participants' change detection processes that were not anticipated in the research questions, as well as noting important items such as noteworthy quotes from the participants. The codebook was peer reviewed by the researcher's advisor to enhance the repeatability of the coding process. This involved walking through the proposed codes (based on the change detection theory), collaboratively applying codes to one specific session, then having the researcher and advisor independently code a separate session and compare results. This resulted in a codebook (included in Appendix E) that connected the research to the GPA Change Detection Theory.

- **Indexing:** After the codebook was validated, we used Atlas.ti to review the artifacts from each participant session and apply codes. While the majority of the coding was applied to the participant transcript, we also applied codes to comments in the questionnaires and to the system logs (including tags and tag labels). We applied the process that was prototyped during the codebook development and validation process, using Atlas.ti to mark the codes for later analysis. For example, if a user commented during the semi-structured interview that they did not want to use clusters while reviewing documents, we would code that statement with something like "preferred clusters off." While the initial codebook addressed the concepts that most directly tied to the GPA Change Detection Theory, we adapted it over time to include markers for system issues, functionality requests, useful quotes, and other concepts.

- **Charting and Summarizing:** This step involves taking the coded data and arranging it into a discernible order (for instance, sorting results within a matrix). For our analysis, the main focus was on organizing the codes based on their relationship to the research questions. We also compiled codes related to system limitations and requested functionality; while these codes were not specifically focused on the the-

ory, they helped us to understand a broader range of functionality that users associate with change detection tasks. In practice, the coding and charting processes turned out to be somewhat iterative in nature. The process of consolidating and compiling information related to our initial codebook revealed gaps, which led us to make additional refinements to our codes. For example, we ended up generating a separate set of codes specifically to address RQ5.5 (Section 5.1.4, p. 126), which looked at the overall effectiveness of the Daybreak system. This involved coding as well as assessing the final story outlines generated by the participants.

- **Mapping and Interpretation:** In this phase, we looked across the study results to see how the data provided answers to the research questions. We reviewed the applied codes to identify trends; some of this included general counting of tags to assess prevalence of use (for instance, the number of participants who had applied clustering, chronological sort, etc.). Other analysis include triangulation of ideas across artifacts. Triangulation aided in identifying connections between different artifacts that related to the same observation. For instance, we looked for areas where a participant's tagging approach and semi-structured interview comments both described their behaviors while using the system.

6.1.2 Analysis of User Study Artifacts

To answer the research questions for the Daybreak user study results, we analyzed the collected artifacts. At a high level, this included the following:

- **Zoom recording:** The video contains everything following the training video, through the end of the semi-structured interview. We created session transcripts through a combination of computer-generated speech-to-text transcriptions and manual cleanup

of the transcriptions. Sessions were recorded for transcription and coding in Atlas.ti, to identify themes and triangulate concepts with other artifacts. [94, 234]

- **Surveys and questionnaires:** Participant responses were exported into spreadsheets; these were analyzed in two ways. First, where appropriate, we aggregated responses from multiple participants to understand behavior and preferences over the course of the session. For example, we looked at responses to questions such as the one about whether the participant believed they had seen "enough" data to identify trends by participant over the course of the five days, across all participants, and by individual topic. Second, narrative responses to some questions were imported into Atlas.ti for coding. This includes responses to questions about the participant's background, the most important tag label for the day, and their views about the Daybreak system.

- **Storytelling task:** We imported the final story outlines into Atlas.ti for analysis, where we analysed their summaries about what they had learned about their topic across all five days of the session. These stories were assessed for adequacy as well as for connections between the story and the participant's tag labels.

- **System logs:** These artifacts were put into spreadsheets, and analyzed in two ways. First, we performed a quantitative analysis by participant, topic, and overall for documents viewed, tags and tag labels applied (by document), subtopic clusters viewed, and other related information.

- **Researcher's and Notetaker's notes:** We used these notes as a bookmark of sorts, to identify and triangulate noteworthy parts of transcripts and log files. The second type of analysis was via coding, through Atlas.ti. We coded tag labels based on themes and interpreted participant intent—for instance, identifying cases where the participant indicated that a tag label was a procedural tag (e.g., tag labels that serve

a function other than describing the content or the participant's mental model—for example, a tag label intended to represent a lack of interest, such as "yuck").

We further address the analysis performed on various artifacts and study findings in our discussion of the research question in Section 6.3.

6.2 Overview of Daybreak User Study Results

Before addressing our research questions, we provide some general information about the participants' sessions and what we learned about the practice of change detection. This section includes a summary of the overall sessions—to include amount of time spent on each section of the user study. We also provide summaries and comparisons by topic. The goal of this section is to orient the reader to the overall results prior to diving into the answers to the research questions.

6.2.1 Description of Participant Sessions

Prior to the study, our estimated length for the Daybreak user study sessions was 90 minutes. In actuality, many sessions lasted longer than what we had anticipated, with participants' session lengths ranging from 90 (the estimate) to 140 minutes. Upon review of the data collected for each session, much of the variance for the longer sessions can be explained by the participants' more detailed descriptions of their personal use cases and change detection experiences in different parts of the session, to include the semi-structured interviews. There was also variance in the amount of time taken to complete the post-day questionnaires and the storytelling tasks.

Table 6.1 shows the amount of time each participant spent in various activities within their session. We note that there is a great deal of variation across some of these activities; in a future study, we recommend introducing a common baseline task that all participants

can perform, to help with interpretation of results. Most participants took the full eight minutes allotted for each day with the Daybreak system; one participant [HLT-409] ended a day early; they completed their Day 5 session prior to the eight-minute mark.

The participant's first view of the Daybreak system came into the video. Even though the video addressed what to do at the start of a day, the participants experienced some confusion when first granted access via Zoom to the keyboard and mouse to the Daybreak system on Day 1. As described in Chapter 5, before seeing any document results in the main workspace, the participant first had to choose their clustering option (clusters on or off) and sort order (reverse chronological or relevance ranked). In some cases, we had to prompt the participant to select a clustering and sort order to get into the system. Looking at the amount of time between the system start each day and the participant's first actions, we found that the longest amount of time between the countdown start and the participant's selection of the clustering option (the first possible action within the system) decreased over each day.

Figure 6.1 shows an example of the range of actions performed by the participant during a day in Daybreak. This data, a side-by-side comparison of two participant sessions, shows the complete set of actions performed during the eight-minute Day 2 for their respective Cryptocurrency-focused sessions. The visualization depicts both participants' complete set of actions during the eight-minute day. The indentation within the workflow is intended to show the actions taken on specific documents, and which clustering and sort mode the participant was in at the time. For instance, at the beginning of the workflow on the left, the participant was using clusters, with reverse chronological sort order. They opened the PARTNERSHIP cluster, viewed document CRYPTO-142, then tagged that document with the tag label PARTNERSHIP.

We were also interested in metrics about the numbers of documents that the participants viewed each day. Table 6.2 shows the total number of documents viewed by participant.

User Study	**Red Sox**			**Crypto**			**Health**			**Space**			**Weather**			—
Segment	BBL-201	BBL-887	BBL-927	FIN-326	FIN-455	FIN-499	HLT-409	HLT-555	HLT-913	SPC-259	SPC-471	SPC-688	WEA-093	WEA-367	WEA-842	Average
Start	0:02:55	0:02:37	0:02:12	*0:01:22*	0:01:49	**0:04:12**	0:01:58	0:02:31	0:03:26	0:01:49	0:01:47	0:02:49	0:02:46	0:02:23	0:02:37	0:02:29
Pre-study Questionnaire	0:05:05	0:05:34	0:04:24	0:04:38	0:03:28	0:04:21	**0:09:22**	0:06:04	*0:01:53*	0:06:20	0:04:02	0:05:47	0:05:55	0:08:29	0:04:24	0:05:19
Training Video	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01
Discussion	0:02:08	0:01:54	0:02:03	0:03:06	0:04:20	**0:05:25**	0:02:11	0:02:03	0:02:45	0:03:00	*0:01:29*	0:02:49	0:03:04	0:02:57	0:01:53	0:02:44
Daybreak Day 1	0:08:32	0:08:26	0:08:27	0:08:24	0:08:34	*0:08:14*	0:08:27	0:08:33	0:08:29	0:08:30	0:08:28	0:08:34	0:08:36	**0:09:00**	0:08:30	0:08:31
Questionnaire 1	0:03:02	0:03:13	0:03:11	0:03:58	0:02:48	0:02:56	**0:07:09**	0:03:21	0:02:15	0:01:44	0:01:57	0:02:20	0:05:27	0:04:13	*0:01:03*	0:03:14
Storytelling Task	0:06:09	0:03:59	0:05:52	0:05:09	*0:03:11*	0:07:23	0:05:34	0:03:44	**0:17:59**	0:03:52	0:04:21	0:12:13	0:04:40	0:08:19	0:03:25	0:06:23
Daybreak Day 2	0:08:27	0:08:36	0:08:22	0:08:46	**0:09:39**	*0:08:09*	0:08:48	0:08:34	0:08:24	0:08:43	0:08:44	0:08:22	0:08:32	0:08:42	0:08:22	0:08:37
Questionnaire 2	0:02:46	0:02:26	0:01:18	0:02:26	0:01:21	0:01:39	**0:06:33**	0:01:52	0:01:40	0:02:16	0:01:28	0:02:08	0:03:16	0:03:26	*0:01:17*	0:02:23
Daybreak Day 3	0:08:15	0:08:19	0:08:34	0:08:28	0:08:25	*0:08:04*	0:08:27	**0:10:54**	0:08:23	0:08:32	0:08:23	0:08:22	0:08:24	0:08:48	0:08:24	0:08:35
Questionnaire 3	0:03:09	0:02:36	0:01:57	0:02:04	0:02:43	0:02:12	0:03:24	0:02:06	0:01:24	0:01:36	0:01:35	0:02:07	**0:04:32**	0:03:25	*0:01:18*	0:02:25
Daybreak Day 4	0:08:14	0:08:17	0:08:22	*0:08:13*	0:08:19	0:08:18	0:08:16	0:08:17	0:08:19	0:08:21	0:08:15	0:08:14	0:08:30	0:08:39	**0:09:02**	0:08:22
Questionnaire 4	0:02:42	0:02:12	0:01:43	0:02:19	0:03:17	0:01:48	0:03:49	0:02:16	*0:01:23*	0:02:20	0:02:05	0:01:54	**0:06:29**	0:03:15	0:01:58	0:02:38
Daybreak Day 5	0:08:14	0:08:14	0:08:25	0:08:11	0:08:23	0:08:14	*0:07:52*	0:08:14	0:08:15	**0:10:34**	0:08:22	0:08:25	0:08:48	0:08:34	0:08:23	0:08:29
Questionnaire 5	0:01:38	0:01:35	0:00:58	0:01:35	0:01:55	0:01:50	**0:07:42**	0:01:25	0:00:39	0:00:59	0:01:19	*0:00:53*	0:05:05	0:02:42	0:01:14	0:02:06
Final Story	0:09:08	0:02:56	0:03:07	0:06:36	0:03:37	0:11:01	0:12:43	*0:01:19*	0:09:23	0:11:43	0:04:53	0:09:20	**0:15:25**	0:07:24	0:04:16	0:07:31
Interview	*0:08:35*	0:23:05	**0:35:04**	0:11:44	0:21:10	0:20:13	0:31:44	0:12:46	0:23:22	0:23:40	0:16:51	0:11:42	0:34:30	0:09:43	0:18:53	0:20:12
Total	1:35:00	1:40:00	1:50:00	1:33:00	1:39:00	1:50:00	**2:20:00**	*1:30:00*	1:54:00	1:50:00	*1:30:00*	1:42:00	**2:20:00**	1:46:00	1:31:00	1:46:00

Table 6.1: Lengths of time spent by each participant on Daybreak user study activities within the session. Times indicated in hours, minutes, and seconds. The times in bold represent the longest time taken by any participant to complete that user study segment. Italicized times represent the shortest time taken by any participant for that activity.

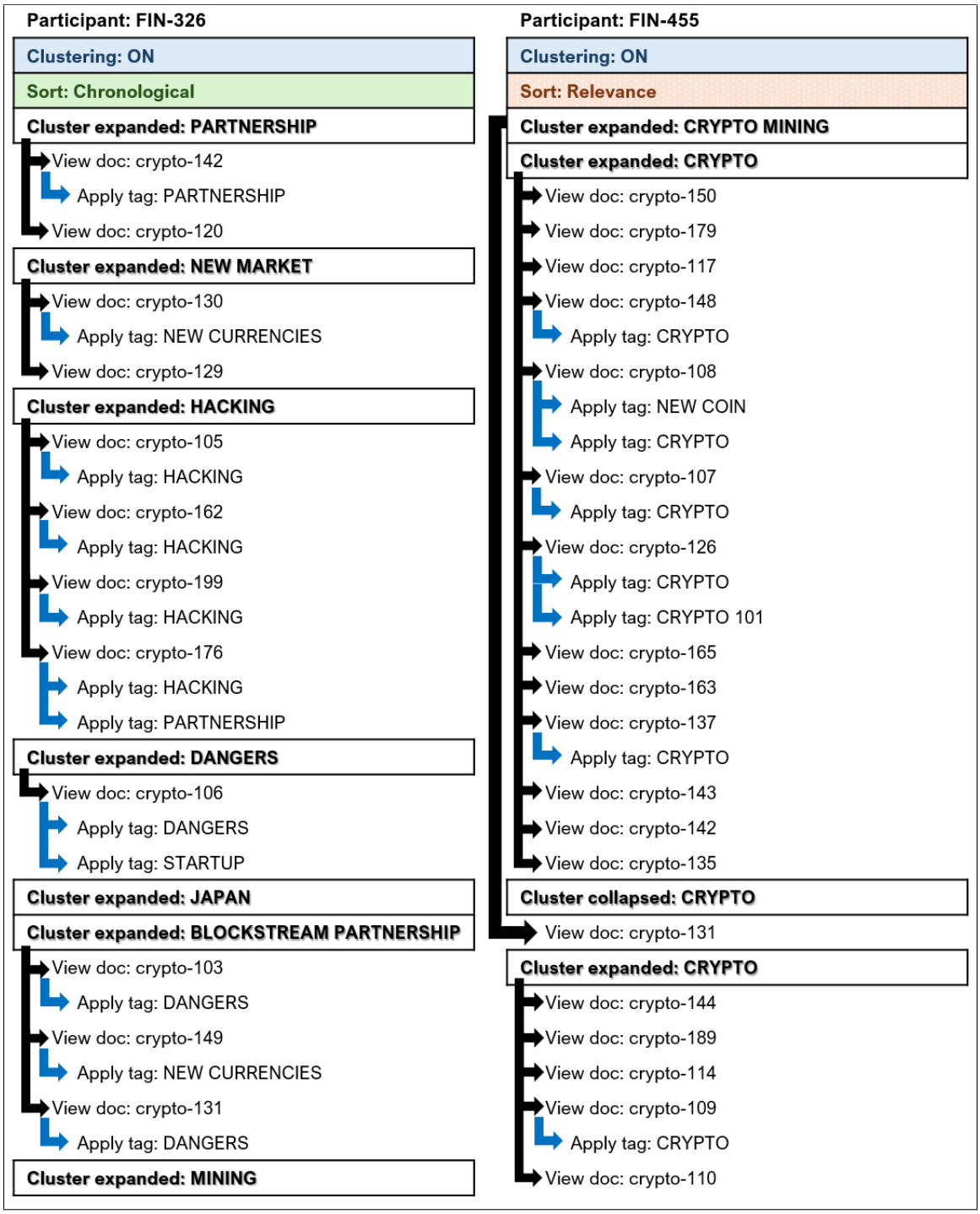


Figure 6.1: Sample of a participant workflow for the Cryptocurrency topic (Day 2). The workflow shows points where the participants took various actions within the Daybreak system. Indentation indicates the clustering mode, sort order, subtopic cluster (if applicable), documents viewed, and actions taken on the documents.

This includes cases where the participant viewed the same document multiple times—for instance, if it was relevant to multiple clusters, if they intentionally returned to the same document during that day to review the content, or if they accidentally clicked on the document link multiple times due to slow system response times or other issues. There was a wide range in numbers of documents viewed during the sessions, with an average of 15.5 documents viewed per day, close to two documents per minute. Numbers of documents viewed on a single day ranged from a lower bound of five documents (0.6 documents viewed per minute) to an upper bound of 38 documents (4.8 documents viewed per minute). For each day, the participant had up to 100 documents available to view.

Participant	Day 1	Day 2	Day 3	Day 4	Day 5	Avg
BBL-201	18	23	27	23	21	22.4
BBL-887	22	21	21	22	20	21.2
BBL-927	20	**26**	12	25	15	19.6
FIN-326	11	14	13	11	16	13.0
FIN-455	15	23	20	20	30	21.6
FIN-499	11	8	12	9	10	10.0
HLT-409	6	*5*	10	12	9	8.4
HLT-555	19	22	28	**38**	**38**	**29.0**
HLT-913	18	17	17	14	17	16.6
SPC-259	16	13	15	14	11	13.8
SPC-471	**24**	23	**31**	32	33	28.6
SPC-688	9	9	*7*	*7*	12	8.8
WEA-093	*5*	7	10	8	8	*7.6*
WEA-367	8	8	8	8	*7*	7.8
WEA-842	15	17	13	14	12	14.2

Table 6.2: Total number of documents viewed per day by each Daybreak participant. Bold values indicate the highest number of documents viewed on a day. Italicized values show the smallest number of documents viewed that day.

In Table 6.3 we show the number of tagging events for each day, by participant. This includes cases where a single document may have been tagged multiple times. In some cases this happened because the participant applied multiple tags during a single view of the document. In other cases, the participant may have viewed the document multiple times

when they encountered it in different subtopics clusters. The lowest number of tagging events per day was by WEA-367 (one tag applied on Day 3) and SPC-688 (one tag applied on Day 5). The highest number of tags applied per day was 32; this number was reached by HLT-913 on Day 1 and Day 5, and by HLT-555 on Day 4.

Participant	Day 1	Day 2	Day 3	Day 4	Day 5	Avg
BBL-201	10	13	10	11	10	10.8
BBL-887	7	12	8	9	3	7.8
BBL-927	31	**26**	3	15	9	16.8
FIN-326	12	15	11	14	15	13.4
FIN-455	13	8	10	9	9	9.8
FIN-499	17	8	9	*3*	8	9.0
HLT-409	14	25	**31**	30	24	24.8
HLT-555	16	20	24	**32**	31	24.6
HLT-913	**32**	24	23	23	**32**	**26.8**
SPC-259	11	12	7	21	14	13.0
SPC-471	15	10	14	18	17	14.8
SPC-688	9	3	2	4	*1*	3.8
WEA-093	6	22	20	14	20	16.4
WEA-367	*4*	*2*	*1*	*3*	2	*2.4*
WEA-842	13	3	15	7	9	9.4

Table 6.3: Overall number of tagging events per day by each Daybreak participant. Note that this includes all cases where labels were applied to documents; multiple tag labels may have been applied to the same document. The highest number of tag labels applied per day is indicated in bold, and the lowest number is italicized.

To understand the variations in usage of tag labels by participants, we have provided Table 6.4. This table shows the number of unique tag labels applied per day by each of the Daybreak participants. Some overlap exists across tag labels, due to the reuse of tags across multiple days. The lowest number of tag labels applied was 0; across all five days, BBL-201 tagged documents, but did not add any tag labels to the tag. The highest number of tag labels used on a single day was by HLT-409 (Day 3), when they applied 26 unique tag labels to documents. On average, participants applied 8.4 unique tag labels per day.

Looking at tag labels in a different way, Table 6.5 displays the total number of unique

Participant	Day 1	Day 2	Day 3	Day 4	Day 5	Avg
BBL-201	*0*	*0*	*0*	*0*	*0*	*0.0*
BBL-887	7	10	7	8	3	7.0
BBL-927	14	8	3	5	4	6.8
FIN-326	8	5	7	8	5	6.6
FIN-455	2	3	5	7	7	4.8
FIN-499	2	4	6	1	4	3.4
HLT-409	12	**22**	**26**	**23**	**20**	**20.6**
HLT-555	7	11	14	16	10	11.6
HLT-913	**20**	20	12	9	16	15.4
SPC-259	8	11	7	17	14	11.4
SPC-471	15	9	13	17	13	13.4
SPC-688	9	3	2	4	1	3.8
WEA-093	6	16	16	10	18	13.2
WEA-367	4	2	1	3	2	2.4
WEA-842	6	2	8	6	4	5.2

Table 6.4: Unique tag labels applied by participant per day within the Daybreak system. Note that the highest number of unique tag labels per day are indicated in bold, and the lowest number is in italics.

tag labels created by participants across all five days of their individual sessions. The average number of unique tag labels created by the Daybreak participants was 29 tag labels. The number of unique tag labels ranged from zero (BBL-201, who did not apply any tag labels) to 71 (HLT-409), more than twice the average number of tag labels.

At the end of the session, we asked each of the participants to generate an outline representing what they would include in a blog post covering what they learned over the five days within Daybreak. We provide an analysis of these stories in Section 6.3.7. Table 6.6 provides an example of a final story outline (storytelling task) generated by one of the participants assigned to the Cryptocurrency topic.

No single metric represents the full range of the participant's experience in the Daybreak system. Many documents viewed does not necessarily mean that the participant did not apply many tags. Interestingly, the two participants with the longest sessions [HLT-409 & WEA-093] (tied at 2 hours and 20 minutes apiece) viewed among the fewest documents

Participant	Unique Tag Labels
BBL-201	*0*
BBL-887	28
BBL-927	29
FIN-326	11
FIN-455	10
FIN-499	12
HLT-409	**71**
HLT-555	36
HLT-913	51
SPC-259	37
SPC-471	59
SPC-688	19
WEA-093	48
WEA-367	12
WEA-842	12
Average	29

Table 6.5: Total number of unique tag labels generated by each Daybreak participant over the course of the Daybreak session. The highest number of tag labels is indicated in bold, and the lowest number is indicated in italics.

```
Blog focus: The rise and growth of crypto with new currencies and
partnerships, but also the dangers of it.

Rise of crypto:
- More investments in the industry: new startups, partnerships with
  top companies, orgnaizations [sic] and people
- More interest in crypto from the general public
- more people are wondering how the blockchain can help them
- New assets being made: Ethereum, BTC Cash, Litecoin. Rise of
  popularity of cryptotokens
- Ethereum is especially popular

However, there are still many dangers
- Mining is not foulproof [sic]
- Recent issues with CPU hacking
- A lot of hacking historically, safety issues with storing crypto
```

Table 6.6: Example of a Daybreak final story outline created by a participant assigned to the cryptocurrency topic.

over the course of the session, with an average of 8.4 and 7.6 documents viewed per day, respectively. However, both of these participants generated large numbers of tag labels and applied many tags to documents. HLT-409 created the largest number of tag labels out of all participants, with 71 unique tag labels over the course of the five days, applying tags an average of 24.8 times per day. WEA-093 created 48 unique tags across the entire session, and applied an average of 16.4 tags per day. In a few examples of contrasting system usage, participant BBL-201 did not generate any tag labels, but viewed an average of 22.4 documents per day and applied on average 10.8 tags per day. Additionally, participant WEA-367 created 12 unique tag labels across all five days, and had on average 2.4 tagging events per day.

6.2.2 Overview of Results by Topic

In addition to differences in the ways that participants used the Daybreak system, we also saw variance from topic to topic. In this section we capture some of the themes and trends that we saw by topic. In this section, we will look at some of the ways that activities such as viewing and tagging documents differed across topics.

First, we roll up the amount of time spent per topic on the activities within a Daybreak session by topic. There are similarities in amount of time taken by participants in each topic; however, we note a few areas where there is variance. Participants assigned to the extreme weather topic spent more time completing the pre-study questionnaire. The cryptocurrency-focused participants spent more time in discussions with the researcher prior to starting Day 1. The Global Health-focused participants spent much more time on the practice storytelling task than participants assigned to other topics, whereas the Extreme Weather-focused participants spent the most time on the final storytelling task. Participants assigned to the Global Health topic spent the most time in the semi-structured interview

(22:37), with the Red Sox (22:15) and Extreme Weather (21:02) participants taking similar amounts of time. The Cryptocurrency and Space-focused participants spent an average of 17:42 and 17:24 (respectively) in the interview.

Activity	Red Sox	Crypto	Health	Space	Weather
Start	0:02:35	0:02:28	**0:02:38**	0:02:08	0:02:35
Pre-study Questionnaire	0:05:01	0:04:09	0:05:46	0:05:23	**0:06:16**
Training Video	0:06:01	0:06:01	0:06:01	0:06:01	0:06:01
Discussion	0:02:02	**0:04:17**	0:02:20	0:02:26	0:02:38
Daybreak Day 1	0:08:28	0:08:24	0:08:30	0:08:31	**0:08:42**
Questionnaire 1	0:03:09	0:03:14	**0:04:15**	0:02:00	0:03:34
Storytelling Task	0:05:20	0:05:14	**0:09:06**	0:06:49	0:05:28
Daybreak Day 2	0:08:28	**0:08:51**	0:08:35	0:08:36	0:08:32
Questionnaire 2	0:02:10	0:01:49	**0:03:22**	0:01:57	0:02:40
Daybreak Day 3	0:08:23	0:08:19	**0:09:15**	0:08:26	0:08:32
Questionnaire 3	0:02:34	0:02:20	0:02:18	0:01:46	**0:03:05**
Daybreak Day 4	0:08:18	0:08:17	0:08:17	0:08:17	**0:08:44**
Questionnaire 4	0:02:12	0:02:28	0:02:29	0:02:06	**0:03:54**
Daybreak Day 5	0:08:18	0:08:16	0:08:07	**0:09:07**	0:08:35
Questionnaire 5	0:01:24	0:01:47	**0:03:15**	0:01:04	0:03:00
Final story	0:05:04	0:07:05	0:07:48	0:08:39	**0:09:02**
Interview	0:22:15	0:17:42	**0:22:37**	0:17:24	0:21:02

Table 6.7: Average lengths of activities within the session by topic.

In Table 6.8 we list the averages by topic for some of the document and tag-related actions taken by the participants. First, we show the average number of documents viewed by topic. The Red Sox participant viewed the most documents, with an average of more than 21 documents viewed. The Extreme Weather-focused participants viewed the fewest documents, with an average of 9.9 documents viewed. Looking at average numbers of documents tagged by topic, the participants focused on Global Health tagged on average 25.4 documents, which was more than twice as many documents as participants assigned to other topics. Extreme Weather-focused participants tagged the fewest documents overall, with 9.4 documents tagged on average. The participants assigned to the Global Health topic created a significantly higher number of tag labels than participants assigned to the other

topics (52.7 tag labels on average), and applied more unique tag labels per day (15.9 tag labels) than participants assigned to other topics. In contrast, participants assigned to the Cryptocurrency topic created an average of 11 tag labels, and and applied 4.9 unique tag labels per day.

Topic	Red Sox	Crypto	Health	Space	Weather
Docs Viewed	**21.1**	14.9	18.0	17.1	*9.9*
Docs Tagged	11.8	10.7	**25.4**	10.5	*9.4*
Unique Tag Labels Applied	*4.9*	*4.9*	**15.9**	9.5	6.9
Tag Labels Created	19.3	*11.0*	**52.7**	38.3	24.0

Table 6.8: The average by topic for documents viewed, documents tagged, unique tags applied to documents per day, and number of tags created for each topic. Bold values indicate the highest for that category, and italicized values indicate the lowest.

6.3 Daybreak User Study Findings

Returning to our original goal of using the Daybreak system to test the components of the GPA theory, in this section we address the research questions that were laid out in Chapter 5. Here we tie the results of the user study to each research question. For each of the research questions, we discuss the methodologies applied, data analysis, findings, and recommendations. The recommendations section for each research question introduces potential areas for future work.

After addressing the research questions, we provide an overview of some noteworthy observations related to the practice of change detection. While not tied to the topics specifically studied, these were behaviors observed across multiple participants' sessions; they may represent areas for possible future study. Some of these behaviors include participants' actions related to browsing results as well as tagging of documents.

6.3.1 RQ5.1: Tags as Mental Model Representations

We expected that participants would leverage tagging to externalize their mental models, by creating new tag labels as they encountered documents that described new, previously unseen themes. We set up tagging and tag labels to play two roles within the Daybreak system: first, tagging was a method through which the participant could link specific documents to a tag label in support of the retrieval and reranking process for generating future days' result set. The second role was a knowledge management function, to allow the participant to save a document for the purpose of writing the outline for the final storytelling task.

RQ5.1 (Does tagging and tag label generation aid users in representing their mental model of a topic?) addresses these uses for tags. To answer this question, we looked at system logs to understand trends and patterns in participants' tagging. We also asked the participants about their use of tagging during the semi-structured interview. These results were further correlated with comments made at other times during the session, and in responses to the post-day questionnaire.

Across all sources of evidence, we found that participants applied tags to make note of and track themes that arose within documents on their topics. The use of tag labels revealed aspects of the participants' thinking about their topic and what themes mattered to them. All participants applied tags to documents, though they used different approaches for their tag labels. One participant clicked the button to tag documents without applying any specific tag labels during the study; many documents were tagged, but the system could not distinguish the participant's mental model components. There was a wide range in the number of tags applied by the participants—and some variance across topics, with the global health and space topics resulting in the most tags. As for individual tagging patterns, we saw different participant behaviors ranging from heavy readers (people who

applied few tags, but read a lot of a document before applying a tag) to heavy document taggers (people who applied multiple tags per document opened) to heavy topic taggers (people who applied a lot of tags across a variety of documents, often making a judgment about whether to tag simply by looking at document titles).

In discussions as well as questionnaires, the participants described their approaches and outcomes from applying tags in their learning process. Even the participant who did not display clustered result sets still heavily made use of tags; this participant [HLT-409] wanted to be able to find documents again when completing the storytelling task. They noted the goal of "trying to anticipate what sort of things [they] might want to pull out in the blog post."

One participant [FIN-499] noted in a post-day questionnaire, "Previously I had a taxation category but realized that there were articles that spoke of the more broad 'regulation' category. If I were truly blogging about this, I would need articles about regulations that are being created around the world in the area of cryptocurrency." Another participant [HLT-555] reinforced this idea after a few days, indicating their evolving understanding of the topic. "I started grouping the articles before I thought about how they related to the larger topic."

A participant assigned to the Red Sox topic [BBL-887] described tagging as "a competition." They went on to explain, "It felt like I needed to tag things in the beginning... So, I was looking for things to tag. ...After the first couple of documents, I realized, 'Oh, I can skip things if they're not interesting, or I opened this one by mistake." They adjusted their approach from attempting to tag all documents, to focusing only on ones that were of specific interest.

One of the participants [BBL-927] tied their evolving use of tags during the study to their everyday approaches for organizing information. "I feel vindicated that in my profes-

sional life I have encouraged the offices I work with not to make decisions about information organization quickly."

Several participants referred to ways they were adapting their strategy in their approach for reviewing and tagging documents. A participant assigned to the Space topic [SPC-259] commented, "I think my strategy is slowly improving. I should have thought ahead of time what terms are most likely to yield interesting things (supernova, duh!) and searched those out rather than simply reacting to what I saw." Another participant [HLT-093] noted about their tagging, "This is very much influenced by the number of documents that I could get through: in light of time constraints, I was tagging documents solely to populate the next day's clusters."

A common theme across the participant feedback was a desire for more flexibility in managing their set of tag labels. The participants expressed concerns about the tags they initially used, and downstream impact on the results they were shown later. They wanted to be able to evolve their tag set, adding refinements based on their understanding of the topic over time. participants expressed a desire for more flexibility to modify, delete, and update tags—including to fix misspellings.

We received particular feedback on tags from participants assigned to the global health topic, who used tags most prolifically across all the participant groups. Specific participant comments about simplifying tagging included areas such as [HLT-409] "wishing it was possible to add tags from the list. There are enough tags now that it's getting hard to remember exactly what tags I chose. That slows me down." Another participant [HLT-555] wanted to be able to tag documents more quickly, stating that it "...would be easier to tag articles if I could drag and drop existing tags, pull from collected tags, or if the system autofilled." Another participant [HLT-913] commented about "...the difficulty tagging (needing to backspace or highlight and delete the tag I just applied when I want to use multiple tags)."

There was also some confusion about the alignment between the cluster labels and tagging. During an early day, one participant assigned to the global health topic [HLT-913] said, "...One thing that's not clear to me... because it's clustered under 'global health,' it's not clear that... if it's already tagged [sic]."

Recommendation: Improvements to Tagging and Tag Label Organization

The tagging functionality in Daybreak did appear to support participants' mental model updates. However, some aspects of the interface were not straightforward to the participant. In an updated version of the Daybreak capability, we would maintain a tagging and tag labeling capability, and incorporate enhanced features to support participants' ability to apply tags and tag labels. First, to reduce confusion and further enable future users to update their mental models, the subtopic clustering and tagging portions of the interface could be updated to make it more clear that subtopic clusters are untagged until the participant explicitly tags them—and perhaps provide a faster way to add that as the tag label for the document (e.g., drag-and-drop functionality).

Building on the participants' desire to modify and adjust their tags over time as their mental model changes, we would focus on additional functionality that could aid a participant with the tag management. This could include functionality to update and correct tags (a common participant request during the interview), as well as more complex capabilities related to ontology management. We also see an opportunity for future research on aiding the participant in organizing their mental model. This could include providing the ability to merge, split, and organize their tag labels, including into hierarchical relationships. The Daybreak prototype enabled use of one level of tags. A future version could include more levels of hierarchically-linked tags. For instance, the system could support a future user focused on baseball who wants to reorganize all of the tag labels related to a team's pitchers under a "pitchers" top-level tag label.

6.3.2 RQ5.2: Participants Leveraged Grouping by Personalized Clusters

Through the participant study, we wanted to determine whether clustering the document result set would aid the participant's ability to perform change detection. We addressed this in RQ5.2 (Does organizing search results by subtopic clusters aid users in performing change detection tasks?). We wanted to use clusters to operationalize subtopics beyond a knowledge management role; instead of solely enabling tagging as a bookmarking function, to provide the participant with the ability to return to documents later, the Daybreak system enabled tagging to drive organization of future result sets. Our expectation was that participants would prefer a view that includes subtopic clusters, to allow them to see how new documents align with their personal tag labels, supporting their mental model.

In the study, all except one of the Daybreak participants displayed search results organized by clusters that were customized based on their tag labels. Through system logs we observed that most participants stayed in the clustered view through the entire study. In the Daybreak system design, the default clustering option had clustering enabled, as described in Section 5.3.1. Some participants may have simply continued using the default option rather than exploring further; this may have introduced complications into in the interpretation of the results. However, the logs revealed that participants would experimentally switch back and forth between clustered and unclustered results to view the alternatives, before ultimately settling on the clustered view. This quick return and remaining with the personalized subtopic clusters may indicate that the clustered view did have utility to the Daybreak participants, and was not simply a participant choice to stay with the default option.

We reviewed system-generated logs to determine how often the participants applied clustering. Almost universally, participants applied clustering as they reviewed their documents. For Day 1, the use of clustering was inconsequential, since it was the cold start

day and no documents could yet be organized by the participant's subtopics. For Days 2-5, across all sessions, participants were in clustered mode for a total of 94.4% of the time. Figure 6.2 shows the percent of the participant's time across Days 2-5 when the participants applied clustered vs. unclustered views. One participant actively moved away from clustering, a change that was consciously made near the start of Day 3. This participant [HLT-409] commented when they made the change, "OK, I'm going to turn the clusters off, because I didn't like what they were doing." Later they explained that they wanted the list of documents all at once, and clicking on cluster labels to view documents added too much separation.

To gather further insights into participants' interest in having results clustered based on their tag labels, we included several opportunities for the participants to comment about the clustering options in the Daybreak system. This included the post-day questionnaires and the semi-structured interview at the end of the session. We also reviewed the recording for instances where a participant commented about the Daybreak clustering feature. Several participants described the utility of clustering; it appeared that many participants accepted this personalized clustering approach as a core Daybreak feature, and talked more about how they would arrange clusters than about turning them on or off.

In one of the questionnaires, a participant [FIN-455] commented, "I loved how new articles were added to the clusters of tags that I made! It gave it a 'tailor-made' feel." During the semi-structured interview, another participant [FIN-326] noted that the customized clusters were "very important. That kind of guided the way I organized the [story outline]."

As noted earlier, only one participant stated that they intentionally did not use clusters to review search results for part of the session. When asked about this choice during the semi-structured interview, they [HLT-409] explained, "I found that the clustering made it really, really slow... really agonizingly slow. It added an extra step before you could get to the content and made it much harder to skim the content because you had at least one or

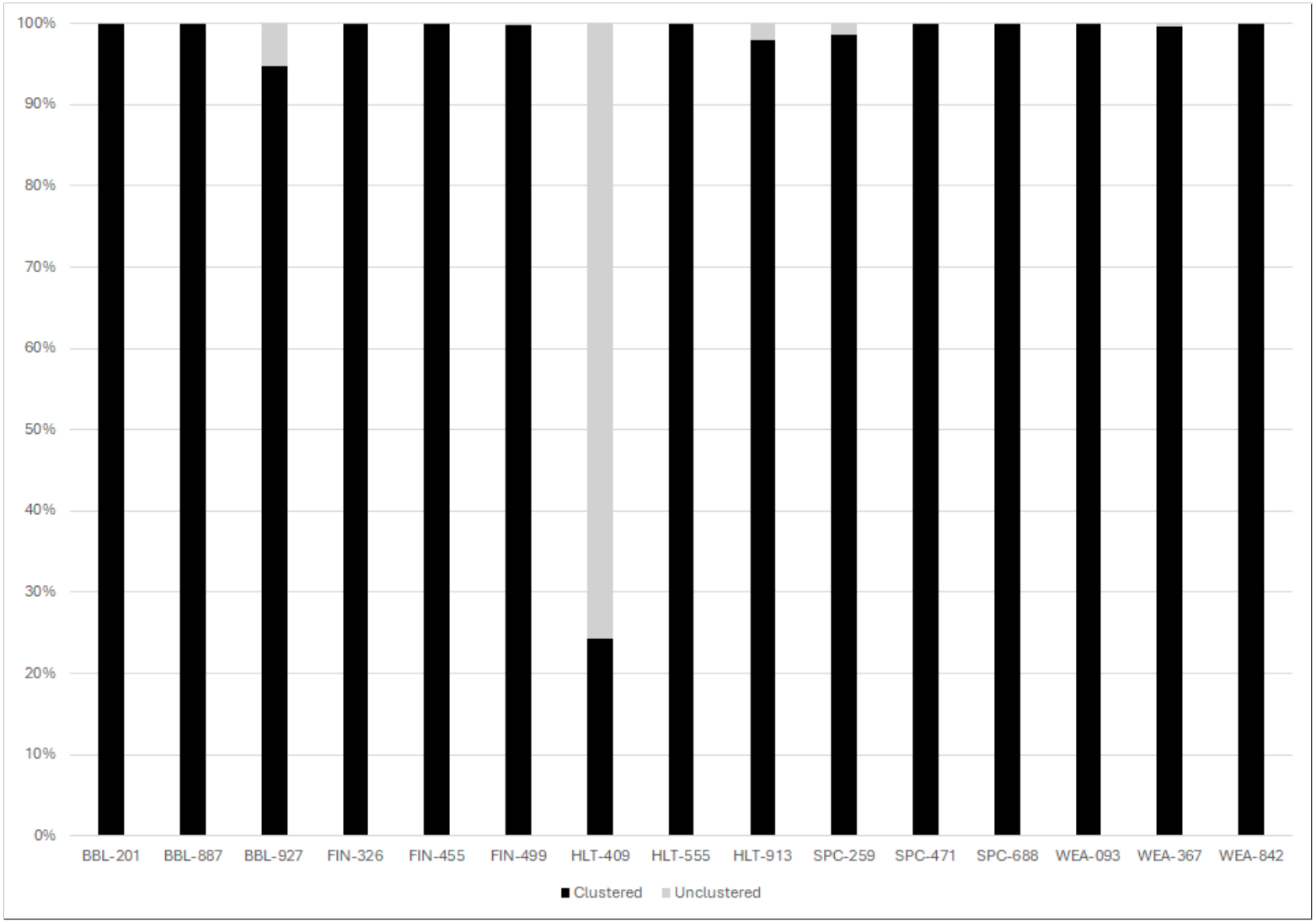


Figure 6.2: Based on system logs, percent of Days 2-5 for which participants applied clustering for viewing the day's results. Note: Day 1 has been omitted because (due to the cold start) the results for clustered and unclustered options were the same that day.

two more steps before you could see things. ...I do this type of work a lot and skim a lot of stuff; it's what I do. It's my job. And I would not be able to use a system that put those additional barriers between me and the content." Interestingly, even though this participant did not display results in clusters after Day 2, they still continued to tag documents heavily on Days 3-5. They indicated that this was to enable them to return to documents during the final storytelling task.

One other participant—who had used the clustered view across all days—mentioned similar concerns about inefficiencies around getting to document contents. They suggested an alternative interface design to make this easier. They [HLT-913] said, "So you want to see the long list, with the tags available, you didn't want to have to click on the tag to see what was under it." They used an example that they had seen within Microsoft Excel, explaining that they would prefer to have an option to expand or hide all sections at once, rather than clicking on cluster labels individually to view documents.

A commonality between these two participants is the fact that they both applied a large number of tags to documents. It is possible that there is a threshold for the number of tags that are reasonable to open and close, before it becomes unwieldy to the participant. For the Daybreak interface, we made a design choice to implement subtopic clusters as an accordion, requiring the participant to click on an individual subtopic cluster heading to expand it. In part, this was intended to enable us to register clicks in the log, so we would know when a participant accessed a specific subtopic during the study. This design choice unintentionally added more of a burden than anticipated for the participant.

Recommendation: Introduce Efficiency-related Improvements into the Subtopic Clustering Design

The results to this research question appear to support the idea that participants are interested in seeing the relationship between new documents and the participant's mental

model, as externalized through their tag labels. That said, two participants reasonably noted that the current interface—which requires that the participant click on and open each cluster independently to view documents—adds inefficiencies that compound when the participant has a large number of tags. We want to maintain the subtopic clustering component that indicates alignment between the participant's tag label and new documents; that said, an updated Daybreak design could reduce the number of steps needed to transition between the subtopic cluster labels and the document results listed within the clusters. We could also explore other interface mechanisms to show alignment between documents and subtopic clusters, such as integrating the subtopic label into the overall results list.

6.3.3 RQ5.2a: Approaches for Populating Subtopic Clusters

Related to the concept of displaying personalized subtopic clusters, we included RQ5.2a (What information retrieval approaches would be effective for transforming a participant's tags into clusters of relevant documents?) to examine our baseline retrieval approach for aligning documents with subtopics. The first step in addressing the question of how to cluster documents in support of this use case was to apply a subtopic clustering approach and get participant feedback. We wanted Daybreak to customize a participant's search results by creating subtopic clusters that were based on the participant's tag labels and documents tagged. As described in Section 5.3.3, we implemented a bag-of-words retrieval approach for dividing new topic results into subtopic clusters based on the terms in the participant's tag labels as well as prominent terms from the set of documents to which the tag had been applied. The same document could be retrieved in multiple clusters if it was retrieved for both subtopics. The Daybreak prototype relied on the browser changing the color of the link to indicate whether a document had been read previously.

The basic approach that the Daybreak prototype applied for generation of subtopic

clusters held up better than expected, and produced documents that the participants considered relevant to the subtopics. As noted in RQ5.2, according to system logs, over time nearly all participants continued to leverage the result clustering capability in the system. We consider this not only a vote for the concept of subtopic clustering, but an indication that the documents in the subtopics were sufficiently relevant to keep the participant continue using them. Other indicators of the utility of this approach included the fact that in multiple instances, the participant labeled documents within a cluster with the same tag label—reinforcing the idea that the aligned documents were indeed relevant to the subtopic.

Although the approach generally worked, there were some issues with irrelevant documents in clusters. One participant [WEA-842] indicated in a questionnaire that "...it would be helpful to be able to move articles between and out of tags. It is slightly bothersome when articles do not match their label."

We also noted other feedback about the documents that were showing up within subtopic clusters. While they did come up sporadically, we heard relatively few comments about misaligned articles that a participant didn't think were relevant to the subtopic cluster. We received far more comments about near-duplicate documents. Even though we had filtered out exact duplicate documents, the collection contained a number of articles that appeared to be syndicated; while there were differences between the documents (e.g., adding localized color commentary about a baseball game as a chapeau to the syndicated article), they were similar enough to cause the participant to be concerned. One of the Red Sox participants [BBL-927] said, "it felt like I saw a lot of duplicate stuff, particularly about that '14 to 1' game. ...Choosing date [sorting] sort of corrected that because it tended to have duplicates of stuff that was a wrap up of yesterday fall further down, because it had probably been filed early in the morning." One of the participants on the Space topic [SPC-259] commented in a questionnaire, "...I think this search was less effective. I think it's getting stuck on a narrower range of things, with lots of repeat articles."

In general, participants wanted better indicators that they had already seen certain documents. Three of the participants stated that they disliked the fact that a document could be in multiple clusters. They wanted it to be clearer that they had previously read the document while reviewing a different cluster.

Recommendation: Subtopic Alignment and Improved Near-Duplicate Handling

The Daybreak system could apply information retrieval approaches that would produce better document alignment with subtopics. We see a number of opportunities to enhance the document-to-subtopic cluster alignment within the Daybreak system. We would experiment with other information retrieval or natural language processing approaches for better clustering of subtopics, including methods based on the Scatter-Gather approach [201], which was designed to aid with populating clusters in a meaningful way. In the Scatter-Gather approach, a set of documents is clustered into semantically similar groups ("scatter" phase). Summaries of the groups (e.g., representative keywords) are presented to the user, who decides which ones they want to see; these groups are then combined ("gather" phase) into a single subcollection [110]. Since our approach involves maintaining the connection between the documents and the subtopic clusters, we might stop at the "gather" phase, and leverage keywords or summaries produced by a Large Language Model (LLM) to inform the user about the contents. Improvements to the population of clusters with documents might reduce cases where participants believe documents assigned to the cluster were not relevant to that cluster. This could include running a future study comparing clustering approaches to determine which approach is preferred by participants.

Another approach for improving the set of documents in the subtopic clusters is to further address the issue of duplicate and near-duplicate documents. Although we had removed exact duplicates in the initial Nexis Uni search, we did not address the issue of near-duplicates, to include syndicated news with modifications, or where an updated

version of an article has been published. This is problematic for users who want to see new content, especially for topics with a storyline that heavily relies on syndication (e.g., sports articles). An updated version of the Daybreak system could display very similar documents together as a set (for instance, if 80% or more of the documents are identical). The user would then be able to decide which of the documents to view—for example, they might want to view the most recent version of the article, which might present a more complete view of the situation.

Also, the Daybreak system could highlight or summarize the parts of the document that most directly relate to the subtopic cluster. This approach could aid with the near-duplicate issue as well as situations where the same document appears in multiple clusters. For near-duplicates, if they are displayed as a set, the highlighted relevant portions of the document could be displayed as a preview, to allow the user to gauge which documents might provide a new or unique angle to the story. In the case of documents that appear in multiple subtopic clusters, within each cluster, the system could display a preview highlighting what in the document relates to that specific subtopic—enabling the user to see why the document is relevant to multiple subtopics.

6.3.4 RQ5.3: Cluster Sort Ordering Preferences May Tie to Specific Goals

For the Daybreak study, we wanted to understand participants' sort order preferences when reviewing documents, to determine whether there are any discernible patterns in sort order preference related to change detection tasks. Based on the GPA Change Detection Theory, we expect that the sort order should support writing a story outline describing the time period studied. According to the theory, documents have to be presented in some order to the participant. In this study we focused on two options: relevance ranking (without explaining the specific approach to the participant) and reverse chronological order.

Coming into the study, we expected to see that change detection participants have a strong preference for reverse chronological sort over relevance sort, because reverse chronological order allows them to see the story as it evolves over time, as respondents had indicated they preferred in the sort order survey (Chapter 4). We believed that reverse chronological sorting of documents would be more effective for showing the participant how the topic was changing over time. We incorporated this idea into RQ5.3 (Does organizing search results within subtopic clusters in some sort order aid users in performing change detection tasks?).

To address this research question, we built into the Daybreak system an ability to select from one of two sort orders, presented as radio buttons in the following arrangement: reverse chronological (date) or relevance sort, presented in that order to the participant; participants appeared to be comfortable switching between the two options, and we did not see any indications that participants were biased toward clicking on the first option they encountered.

After selecting one of the two options, the results would be presented in the order selected. If the participant had previously selected clustering, the results in each cluster would be sorted using the selected sort order. In cases where clustering was turned off, the full result set would be sorted according to the participant's selection. The reverse chronological sort was based on the publication date and timestamp of the article, whether clusters were on or off. Relevance was based on the Indri default relevance ranking algorithm, which is a combination of language modeling and inference-based approaches [257]. Relevance scores across different searches are not comparable; as a result, we could not interleave subtopic cluster rankings to generate an unclustered sort order. Instead, we used single ranking as a simplifying function. When subtopic clusters were in use, the relevance score for a document within the cluster was based on the document's relevance to the subtopic

query. In cases where subtopic clusters were turned off, the relevance was calculated based on the topic terms.

We expected that the participant would want to have the documents arranged in a meaningful way. Based on theories of reading comprehension we expected that a participant would want to organize text in a story-like way (e.g., organizing around a timeline), emphasizing the difference between recency-weighted relevance ranking and unaltered reverse chronological sort.

To understand the participants' sort order preferences, we first looked at the system-generated logs to determine how often the participants selected reverse chronological option or relevance ranking, and participant statements to determine effectiveness. In Figure 6.3 we show the percent of time that each participant spent in each of the two sort order options. Based on the results of the social media survey detailed in Chapter 4, we expected a stronger signal in favor of reverse chronological sort. However, analysis of the system logs shows that the participants were somewhat flexible in their sort preferences, with a weak preference for chronological sort order in support of the change detection task. Participants applied reverse chronological sort approximately 56% of the time across all five days. Seven of the 15 participants spent at least 60% of their time in reverse chronological sort order. Two participants applied each sort option for approximately half the time. The remaining six participants chose relevance ranking for the majority of their session. There were a number of cases where participants would spend some time in one sort order (for instance, starting in reverse chronological), then switching to the other option after a few minutes of reviewing documents.

Another way we reviewed the data was by looking at the amount of time per day that users applied chronological vs. relevance sort ordering. Table 6.9 shows the percentage of time per day that each participant spent in chronological sort order. While some participants applied chronological sort consistently from day to day, we found that many participants

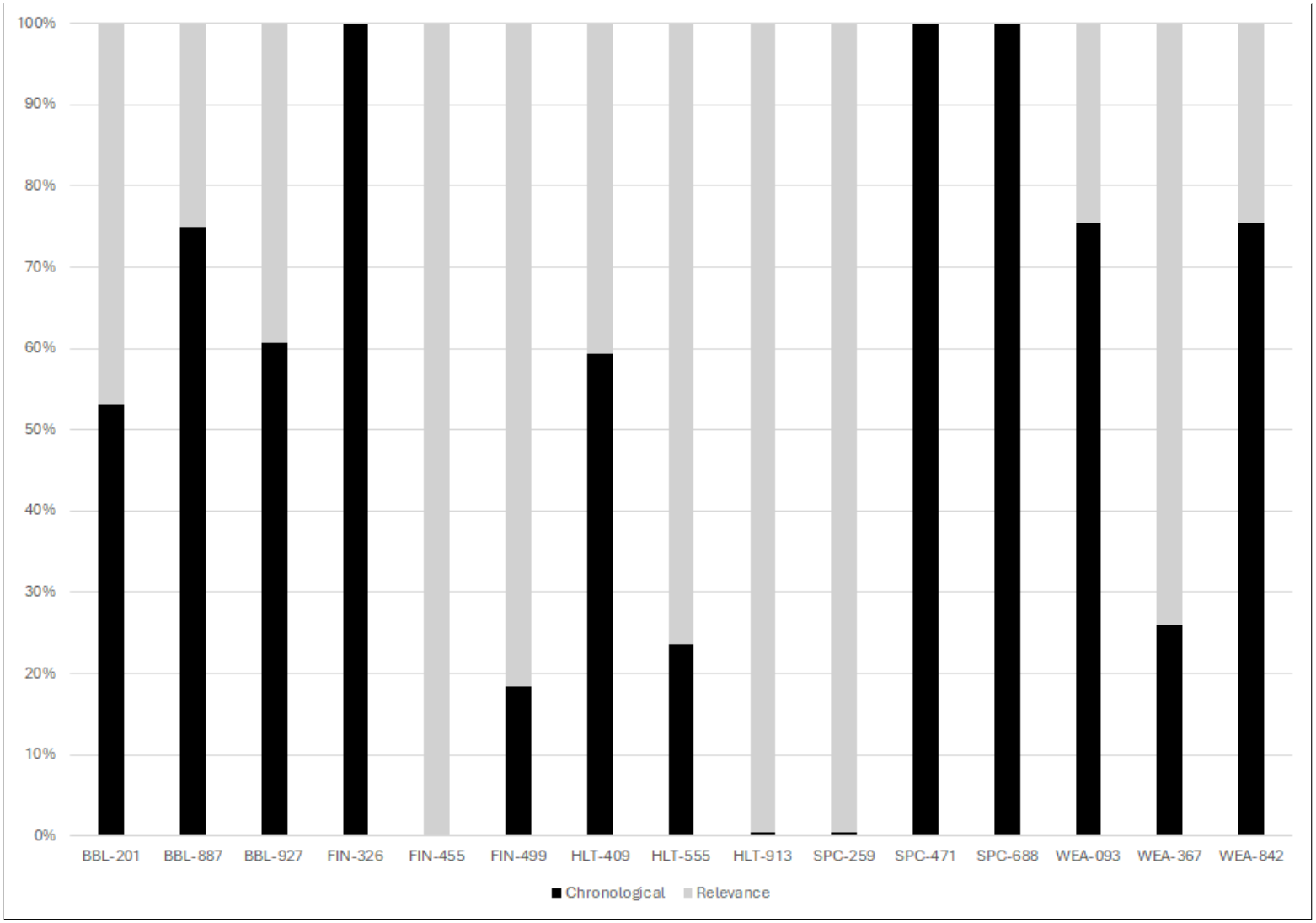


Figure 6.3: Based on system logs, percentage of Days 1-5 for which participants applied reverse chronological vs relevance sort for viewing the day's results. Unlike with clustering (where clustered and unclustered results were identical on Day 1), the different sort order options produced unique results for all five days; as a result, all days' results are shown here.

tended to leverage chronological sort in earlier days, and on later days switched to using relevance ranking instead.

Participant	Day 1	Day 2	Day 3	Day 4	Day 5	% by User
BBL-201	100%	52%	67%	52%	41%	63%
BBL-887	100%	100%	0%	100%	100%	80%
BBL-927	1%	79%	64%	99%	0%	49%
FIN-326	85%	100%	100%	100%	100%	97%
FIN-455	61%	0%	0%	0%	0%	12%
FIN-499	65%	73%	0%	0%	0%	28%
HLT-409	0%	0%	100%	100%	32%	46%
HLT-555	56%	1%	96%	0%	0%	31%
HLT-913	100%	2%	0%	0%	0%	20%
SPC-259	100%	2%	0%	0%	0%	20%
SPC-471	100%	100%	100%	100%	100%	100%
SPC-688	100%	100%	100%	100%	100%	100%
WEA-093	100%	100%	100%	100%	0%	80%
WEA-367	2%	0%	3%	0%	100%	21%
WEA-842	100%	100%	100%	100%	2%	80%
% by Day	71%	54%	55%	57%	38%	55%

Table 6.9: Percent of time that the user spent in chronological sort order, by day.

Additionally, we looked at the correlation between the participant's familiarity with or interest in their assigned topic and their use of chronological sort ordering. The familiarity and interest levels were obtained in the Post-Day 1 questionnaire; the participants self-assessed both their familiarity level and their interest in the topic. The correlation was made with the amount of time spent in chronological sort order. As shown in Figure 6.4, we noted a weak negative correlation between topic familiarity and use of chronological sort, and a very weak positive correlation between topic interest and use of chronological sort, neither of which was statistically significant, based on linear regression.

We believe there are three potential factors leading to the participants' flexibility in their sort order use. First, sort order choice might be dependent on what the participant is trying to accomplish in the moment rather than favoring one specific sort option. Second,

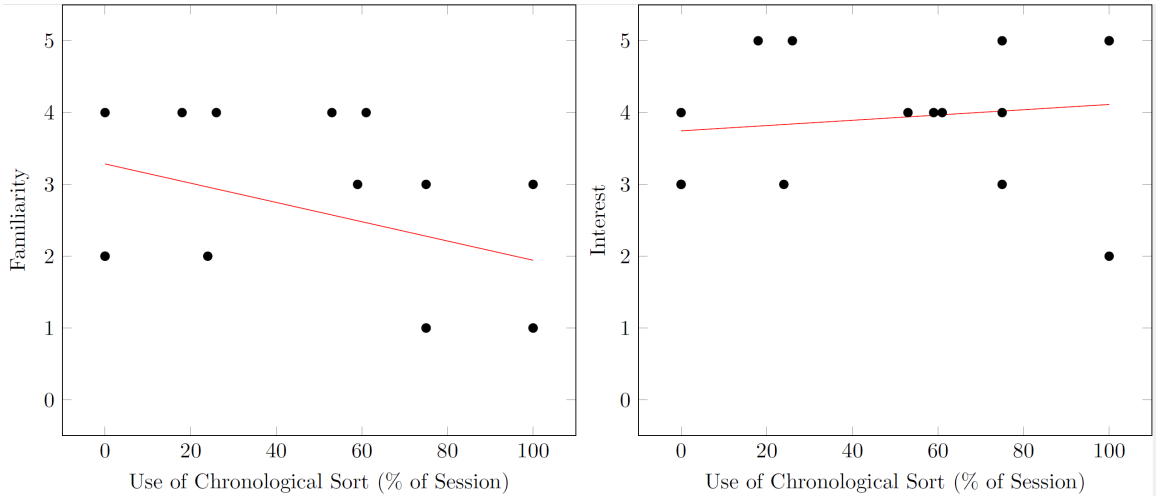


Figure 6.4: Correlations between the participant's self-assessed familiarity or interest in their assigned topic and their use of chronological sort orders. There was a weak negative correlation between user familiarity in the topic and use of chronological ordering, and a very weak positive correlation between user interest and chronological ordering. Neither result was statistically significant.

flaws in the topic assignment process meant that participants may not have been familiar enough with their topic; change detection is about meeting the need of a participant who is familiar with the topic, and less focused on getting a novice user up to speed. It is possible that because some of the participants were unfamiliar with the topic, relevance ranking may have been more useful, and the task requires higher baseline of familiarity than was present. Additionally, some participants may have been experimenting with the system throughout the five days, and did not get into a consistent pattern that might have been applied in a system they actually use from day to day. Finally, there were not clear enough story lines in some of the topics we selected. The Red Sox topic featured different pre-game, during-game, and post-game reporting, so chronological ordering showed meaningful changes to the plot. However, the other topics contained a broader range of themes, many of which did not have as clear of a flow or a connection to one another. For example, the cryptocurrency topic contained a mix of information about new coins, security issues, and policy decision that had little to no connection to one another. As a result, from the perspective of the

participant there might not have been a noticeable flow for documents in other topics when sorted chronologically.

To further understand the Daybreak participants' sort order preferences, we coded information that they provided in the post-day questionnaire, their comments about sort orders made while using the Daybreak system, and items from the semi-structured interview—where we specifically asked their sort order preference. While some participants expressed a strong preference for reverse chronological sort, many did not. Some explanation emerged in questionnaire responses and during the semi-structured interviews—a number of participants indicated that if they were following a topic of personal interest from day to day, and were relatively up-to-date on that topic, they would likely prefer reverse chronological sort. However, since they were learning a topic adjacent to their topic of interest, they wanted to use relevance sort to get a general sense of the topic first. When asked about their sorting choices in the interview, three of the participants indicated that if they knew the topic better, they would have preferred a chronological sort. They saw relevance as a better approach for getting a baseline understanding of the topic that they were still learning.

Several responses supported the idea that the participants were assigned to topics that they were not familiar with, which may have impacted their sort preferences. One participant [HLT-409] said, "In my own searching, I almost always do date first. I'm not doing this for [this study] because I use relevance when the topic [is one] that I'm unfamiliar with." Later, this participant also mentioned that they use date sort so they can identify information that they had already seen—for instance, that sort order made near-duplicate documents stand out. In an apparent reference to the older date range of the historical documents used for the study, another participant [HLT-555] mentioned, "If I'm not looking for super current information, then I want the relevance ranking."

A number of participants indicated that they prefer reverse chronological order for change detection tasks. One such participant [SPC-471] brought up their current work-

flow: “Even when I use Google News and whatnot I always sort by time. I never sort by relevance... I like to see what’s the most current pieces of information that are discussed and then work backwards from there.”

Personal preference came up in a number of responses. As one participant [FIN-326] said, “I think it’s just personal preference; I like to look at the most recent stuff first, especially since this was a topic I was following for five days straight. If I was coming to this topic after a month or a week later, I would have sorted by relevance to see... the main things.”

Overall, four participants stated that they had a preference for relevance ranking during the user study. For example, one of the Red Sox participants [BBL-201] noted in a questionnaire, “Relevance seemed more important than the date here. When sorting by date it seemed like I had a lot of the same responses. I was getting multiple versions of the same story,” referring to near-duplicates and other similar documents. During the interview, another participant [FIN-455] talked about their preference for socially curated information when they are looking for new information in Reddit. “...When you sort by new, you just get a hodgepodge of things that have nothing to do with what you’re looking at... just low effort posts. When you go to relevance, it’s already kind of vetted in a way, either by other people or by the algorithm. So by the time it gets to you, it’s vetted. When you saw it [sorted] by new it is kind of like, ‘OK, someone just posted this, the mods haven’t gotten to it... Downvotes haven’t gotten to it... I would like it to be vetted by people who have the same interest as me. So that’s why I instantly went to relevant... I like that the curation component of it.” Such a curation aspect could also aid a user in dealing with duplicate or near-duplicate content, which could be filtered out by the community.

Recommendation: Flexibility in Sort Options

While participants’ preferences for chronological ordering was a weak one, the findings

related to chronological sort are not incompatible with the GPA Change Detection Theory; however, more research is needed to test out chronological ordering alongside other sort options to determine what additional sort orders are useful for change detection, and whether there is a need to assign a default sort order for change detection tasks. It would be useful to build in flexibility for the user to select from a range of sort orders, and set their individual defaults based on their sort preferences. For updated versions of the system, it makes sense to continue to include a variety of sort order options, including both chronological and relevance sort (making it clear how relevance is defined). These preferences could be tested in future studies.

Some adjustment is needed to address the issue of assigning participants to topics; in particular, future research would be needed to determine how to assign participants to their assigned topic—preserving the learning component, without forcing them to learn a completely new topic. Based on the change detection theory, we believe there does need to be some understandable sort order for the documents; we initially expected a strong preference for chronological sort orders due to the opinions expressed in the sort order survey from Chapter 4. However, this may be operationalized in different ways depending on the participant's prior knowledge of the topic and overall goals.

6.3.5 RQ5.4: Rarity Not Effective Proxy for Subtopic Importance

Participants identify certain subtopics as being more important than other subtopics within a time window; subtopic importance evolves over time. We looked at this issue with RQ5.4 (Does arranging subtopic clusters in some order aid users in performing change detection tasks?). With this question, we looked at both subtopic importance (things that the participant wants to see) as well as unimportance (subtopics that the participant may have a low interest in, or that are not relevant). We assumed that subtopics, like documents,

must be presented in some order, and that participants want to have important subtopics identified first.

To test this research question, we built into the Daybreak system a single approach for organizing subtopics: as a simplifying approach, we sorted the subtopic clusters by rarity—organizing a cluster so that the ones populated with with the fewest documents appear first, and the most populated clusters are last—as a proxy for subtopic importance. In cases where multiple clusters had the same number of documents, we sorted those clusters alphabetically. Applying the assumption that less common subtopics provide some advantage, we expected that the participant would find some value in less common clusters. Can document rarity serve as a proxy for subtopic importance? We anticipated that participants would aid in addressing this question, and have additional feedback about other characteristics that can further inform a definition of subtopic importance and how it evolves over time.

In the Daybreak prototype, we did not provide the participant with an ability to change the subtopic ordering. We only tested in this study whether the participant leveraged clustering; while not using clustering could be a potential indicator that they do not like the subtopic ordering, we did not have participant behavior data from system logs that would clearly address this question. In order to compare potential subtopic ordering approaches more effectively, we also needed a formal evaluation approach. We further address this in Chapter 7, where we design and test such an approach for comparing subtopic orderings.

Our main source of information for this research question came from the semi-structured interviews with the participants. The participants sent a strong signal that rarity was not an adequate proxy for subtopic importance. However, this aspect of the study revealed factors that would be useful in future studies of subtopic importance. This includes the utility of having exposure to the full range of personalized subtopics with relevant documents for that day. Participants also provided input about potential manual and automated approaches for

organizing clusters, as well as how to treat subtopics that the participant identifies as being unimportant to them, and want to have them hidden or removed.

The post-day questionnaires also contained some information relevant to this research question. Each day, we asked the participant to indicate which of the tags were most important that day. The participants' responses for each of the five days are included in Table 6.10. The participants' responses support the idea that subtopic importance varies from day to day. Participants treat certain subtopics as being more important than other subtopics within a time window, and this importance level changes along with changes to the topic.

Two of the 15 participants correctly identified without prompting that the subtopic clusters were sorted by rarity, with the least populated cluster first. We explained the subtopic clustering ordering choice to all of the participants, then requested their opinion on the approach used in the Daybreak prototype. When asked whether ordering clusters by rarity was useful, the Daybreak participants nearly universally said it was not. Many participants preferred the opposite—showing the most populated clusters at the top. Other suggestions included allowing participants to pin certain subtopics, having the system figure out importance automatically using machine learning, or simply alphabetizing the clusters. In general, ideas varied on how to define subtopic importance. For instance, anytime that specific topic appears, users might want it to be placed in a prominent position. For instance, a baseball fan may always want to see a subtopic cluster about pitcher Max Scherzer at the top.

A number of participants described options for gauging subtopic importance that were based on counts of various sorts. Most participants indicated that prevalence—the opposite of our definition of rarity—would be useful; participants wanted to make sure they did not miss subtopics that many people were discussing. For example, one Red Sox-focused participant [BBL-887] noted that "pitching" was the most important topic one day, because "it

Participant	Day 1	Day 2	Day 3	Day 4	Day 5
BBL-201	Red Sox	Red Sox	Red Sox	n/a	n/a
BBL-887	collapse	pitching	pitching	pitching	pitching
BBL-927	results	hutchison	not sure	analysis	standings
FIN-326	hacking	new currencies	New currencies, partnerships, and hacking	rise of crypto	Rise of crypto and hacking/dangers
FIN-455	crypto, crypto mining	Crypto	crypto activism	crypto future	crypto future
FIN-499	cryptocurrency, crypto hog	layman	Investment strategy	access	regulation
HLT-409	community health	wellness	poverty	outbreaks	emtech
HLT-555	health care	health care	health care, prescription drugs	flu	flu, tuberculosis
HLT-913	global health	infectious disease	infectious disease	infectious disease	infectious disease
SPC-259	n/a	space photography	space photography	simulation	not sure
SPC-471	USPTO	Remote sensing	Asteroid	SpaceX	China
SPC-688	Mars	Space exploration	SpaceX	Space Innovation	China in Space
WEA-093	climate change	climate change	climate change	emissions	climate change
WEA-367	drought	Storms	Climate Change	Humans	Food shortage
WEA-842	global	other	methodology	Resilience	Food/ag

Table 6.10: Tag labels identified by the participant as being the most important each day. These responses are verbatim from the participant responses, including participants' indications when they did not wish to identify a specific subtopic as most important.

was dominant in the stories." Another participant [SPC-688] mentioned "China in Space" as the most important topic one day. Their explanation was that "There were a lot of articles of China doing something in space." Alternatively, other kinds of counts might be valuable. A global health participant [HLT-555] identified "health care, prescription drugs" as the top tags on a specific day. Their explanation was that "I used these tags the most during this session." Whether or not there were many documents available on the subtopic, the fact that the participant repeatedly tagged with that label may be an indicator of interest.

Other participants provided more subjective descriptions of subtopic importance, highlighting the complexity of automatic approaches. For some, the information did not need to be important—descriptors such as "interesting" came into the discussion. A participant who covered the extreme weather topic [WEA-093] said, "...My bias is towards interesting, more than important. I'm sure progress in Paris would have mattered to climate change, right? Extinction of butterflies or caves, graffiti. It's like, 'Oh, that's interesting.' Even if it's not terribly important. ...I tend to gravitate towards interesting, and if it's really interesting, then I'll try to figure out if or why it's important." Other topics can catch people's attention serendipitously. For example, a participant focused on cryptocurrency [FIN-455] found "crypto activism" to be important. They explained, "I saw a lot of articles, but the one that stood out to me was the one about how crypto could be used for good instead of just profit." An additional participant used potential emerging impact as a reason for identifying a subtopic as important. They [FIN-455] explained that "crypto future" was important because, "Looking at the articles, I saw that predictions about what would happen in the future were incredibly valuable and would change the way that I, and others, thought about cryptocurrency." Giving the system knowledge of both the participant's mental model and interests could help to match the subtopic ordering to their preferences.

Finally, some participants were task-focused. For example, one of the global health-focused participants [HLT-913] listed as a day's most important topic "infectious disease"

because it "seems most applicable to my 'global health' assignment." If a participant does have a specific research assignment—for instance, an article—it makes sense that information related to that topic would be deemed more important.

Regardless of the order, several participants noted that they liked having the ability to look across the range of new subtopics for the day, as it helped them understand what to expect in the documents. For instance, a participant [HLT-913] said, "The clustering primed me before I reviewed each article, which I think cognitively helped me a little bit."

A few participants specifically addressed the idea of subtopics that were unimportant to them. Four participants specifically applied tags that were intended to reduce or hide uninteresting results. They used tag labels such as "yuck" or "irrelevant." One of these four participants applied tag labels that they later indicated were meant to denote something uninteresting—for instance, tagging documents with the label "Regina" in the Red Sox topic, since it corralled document results on the Regina Red Sox (a team unrelated to the Boston Red Sox), to allow the participant to ignore that cluster. Another participant [HLT-555] chose to do nothing with irrelevant documents. "I don't even want to categorize it because it's not relevant at all. That's not something I would use. So then I just didn't tag it."

Recommendation: Apply customizable subtopic importance approaches.

The way we defined rarity for the Daybreak study was very specific (e.g., a cluster containing a small number of documents), and may not have adequately represented the idea of a rare subtopic. Issues such as an overly specific set of search terms could unintentionally limit the number of documents matched to a subtopic cluster; the actual concept represented might not be rare, there may simply be only few documents about that topic at that time. While the participants strongly rejected the application of rarity as defined in this study, further research could test other interpretations of rarity in cluster arrangement

options. For example, the user could manually identify certain subtopics as ones that they would expect to see infrequently—in particular, indicating subtopics that they would want to know about. For instance, in the context of baseball, the user might identify "triple play" (a very rare occurrence) as a subtopic that does not come up often. Whether there is one document about the triple play or hundreds, the user would want to know about it.

Further research is needed into defining subtopic importance, whether through automated or manual approaches. As an initial test case, the system could provide an ability to pin certain subtopics at the top, to ensure that the participant comes across those first. This might prove useful in particular for some rare themes that come up only infrequently. The participant-provided suggestions (prevalence, tag frequency, etc.) could serve as starting points for a future test of subtopic importance ordering approaches. It seems that flexibility is warranted; the answer may not a single subtopic, but buckets of important subtopics. We provide further treatment of this research question in Chapter 7, comparing subtopic ordering options using a repeatable, automated evaluation.

Additional future research could be done on the utility of being able to see the range of topics available on a given day, to see how that insight alone aids from the perspective of their mental model. Additionally, because the list of subtopics could get long over time, there may be a need for further research on arranging or clustering the subtopics themselves to make it easier for the participant to understand the range of what has changed on their topic.

An updated version of the system should also have some way to indicate unimportant subtopics, along with the appropriate actions to take—for instance, whether they should be hidden or removed.

6.3.6 RQ5.4a: Placement of Uncategorized Subtopic Cluster

One of the assumptions behind the change detection concept is that the participant is looking for new subtopics or themes. It is unlikely that the participant has a comprehensive set of tag labels for everything that could be relevant to the topic. As a result, there is some ongoing need for a category that captures things outside of the participant's existing externalized mental model. We address this issue with RQ5.4a (How should the system handle documents that do not fit in any existing subtopic cluster?).

We operationalized this concept within the Daybreak system by creating a catch-all category that we labeled "Other." This cluster contained all documents relevant to the topic that were not assigned to the participant's personalized subtopic clusters. We did not apply the rarity sorting to this cluster; instead, we placed the "Other" cluster at the bottom of the result list. To understand how participants discovered new subtopics via the "Other" cluster, we looked at system logs showing when participants opened the "Other" cluster, and when they tagged documents that they had seen in the "Other" cluster—in particular, cases when they added a previously unidentified tag. We were particularly interested in understanding how looking at documents in the "Other" cluster led to identification of new subtopics, such as new themes that had not arisen on prior days.

We note based on the logs that many participants did not open the "Other" cluster, and some who did open it used it only sporadically from day to day. Most participants did make at least some use of the "Other" cluster. On one day, a participant indicated that they believed "Other" was the most important cluster [WEA-842, Day 2]. One participant did not notice it, and six participants stated that they had intended to use it, but were unable to find time to do so. This mixed usage may have been due to the fact that it was always placed last in the subtopic list. Participants might benefit from having potential new themes displayed more prominently in the participant interface.

We reviewed the post-day questionnaire and interviews, to determine the extent to which participants expressed an interest in or made comments about the "Other" cluster. Participants generally liked the idea of having a cluster that captured new developments or things that they had not thought to tag previously. Five participants mentioned that they saw the "Other" cluster as a good place to start their day's review. One participant [BBL-927] said "I feel like 'Other' is ... exciting in its way, like when you get a grab bag and you're like, 'I don't know what's in this, but I'm going to buy it... because, maybe the most exciting thing is in 'Other' today. Not everyone is motivated by surprise, so that may not be as big a selling point to other people. But I did find 'Other' was sort of a helpful tool for me to figure out where I had left stuff out of tags. But also, I went there just to make sure that I wasn't missing anything in the other groups."

One of the Red Sox participants [BBL-201] found that they used the "Other" cluster more than expected. "I was surprised that I ended up using it. ...You're just finding those kind of random articles that were in there, like the one with the people they arrested for selling tickets at too-high prices... You're writing for people who are following all the main stories, so you don't want to just keep repeating what's on ESPN or MLB."

Another participant [SPC-259] described a potential role for the "Other" category in breaking out of a filter bubble. They said, "I think it's always interesting to have a little bit of untamed variety to browse just because you don't want to get too much of an overly focused lens. Need a little bit of randomness in that to keep you to keep your your view broad, right?"

Three of the participants noted that, over time, they saw fewer articles being assigned to the "Other" category. Two observed that after a few days, the content they cared about more was being pulled into labeled clusters, and the information in the "Other" category was not as directly useful to them. As one participant [SPC-471] pointed out, "[I] still want to make sure I manually tag additional articles to feel comfortable. However, it was

reassuring to see that articles tagged as 'Other' were less relevant to my topic than those automatically tagged into an existing cluster." Over time, clusters were more tuned to the participant's interests, leaving less interesting items in the "Other" category.

While the participants responded positively to the concept, the question of placement of the "Other" category did not receive a clear answer. One participant [FIN-499] started with the "Other" category on at least a few days, potentially indicating that it could have higher placement in the list; three other participants expressed an interest in starting with this cluster as well. "It was interesting which articles appeared in which categories. I looked in the 'Other' section first because I thought many uncategorized articles would be grouped there." However, another participant [FIN-455] preferred its current location. "Having 'Other' at the bottom was very helpful because it gave me kind of the digital space to to handle the tags that I had created first. I think that was the best setup... a lot of people's attention spans are just at an all time low, so if you put 'Other' first people going to be like, 'oh, too much labor, I'm not dealing with that.'"

Recommendation: Prominent placement of documents unaligned with available subtopic clusters

Given participants' interest in seeing new developments and not getting locked into updates based solely on prior tags, documents that are relevant to the topic but unaligned with current subtopic clusters should get a prominent placement in the UI. The Daybreak system put it into a single "Other" cluster at the bottom of the results. There could be alternative options beyond adding it to either the top or the bottom of the list; it could be placed to the side, or in another location to distinguish it from the personalized subtopic clusters. The key is to make it easier to find within the interface, to ensure that new, breaking themes are not missed.

The study also revealed that it might be helpful to apply an organizing approach within

the "other" category. Even though the documents in this cluster may not have been directly related to any tag labels that the participant had previously used, they may be interested in documents within that cluster; participants who reviewed that cluster sometimes found documents that led them to add new tag labels or think about other potential topics for their audience. For instance, one participant [BBL-201] took interest articles about third-party ticket resale fraud in the "Other" cluster.

It may be helpful to the participant to organize the documents within the "Other" category so that similar themes are together. This question of how to organize the documents in the "Other" cluster is noted as an area for future research. If automated approaches are applied, it could be useful to study whether users are influenced by labels suggested automatically.

6.3.7 RQ5.5: Daybreak Supports Detection of Topic-relevant Changes

After reviewing what the Daybreak participant study revealed individually about the group, pile, and arrange components of the theory, we wanted to see the extent to which the combination of these features aided in completion of a change detection task. To explore this idea, we included RQ5.5 (Does the system help users develop and externalize mental models?). We addressed this question in two ways. First, we assessed the extent to which the participants (enabled by the system) successfully expressed what they learned about their topic over the five days. Second, we looked at how comfortable the participants were with the system, and how their comfort level changed over the course of the five days.

We created the storytelling task as the primary method for understanding what the participant learned during their document review process over the five-day session. This was an opportunity for the participant to externalize the components of their mental model that were developed across the session. To understand what the participant learned during the

session, we reviewed the tags, the participants' most important tags, the final story outline, and the interview. This aided us in understanding how the participants' tag labels evolved from day to day, and the role they played in creation of the story. To accomplish this assessment, we used a multi-pass approach where we looked at different data elements separately: first, we coded the tags; next, we looked at the alignment between the story and tags; then we coded the interview; finally, we performed a review of the story outline. For each pass, we used the same order, traversing the list of participant identifiers alphabetically. This made it easier to spot trends by assigned topic.

To determine the extent to which the Daybreak system helped participants learn about a topic, we started by looking at how the participants' tags evolve in relation to the themes from the document set over the five days. Table 6.11 shows the participants' creation of new tag labels, as well as tag label reuse. With the exception of the participant who did not apply any tag labels, all other participants created new tag labels on at least three days—revealing a changing understanding of the topic over the course of the session. All except two participants reused at least some of their tag labels, including usage that carried across multiple days. The introduction of implied sub-subtopics (tag labels that contain an implied hierarchy) also reflect the participant's learning. For instance, a participant may have started with tag "crypto," and later added "crypto hacking" to represent a more specific subset. This is another potential area for future research, where we determine whether a multi-word phrase is intended as a contiguous phrase, or if a user is attempting to distinguish between levels within a hierarchical relationship.

After this assessment, we identified alignment between the content of the story outline and the tag labels. This included seeking out matches (including plurals, acronyms, etc.) as well as conceptually similar examples (e.g., "developed countries" in the outline mapped to "US" or "Canada" in the tag labels). While we did not ask participants specific questions in the interview about why they chose to reference specific tag labels in the story, we observed

Participant	Tag Creation	Tag Reuse	Hierarchical
BBL-201	Day 1	n/a	no
BBL-887	All 5 days	yes	yes
BBL-927	All 5 days	yes	no
FIN-326	Days 1-3	yes	yes
FIN-455	All 5 days	yes	yes
FIN-499	All 5 days	yes	no
HLT-409	All 5 days	yes	no
HLT-555	All 5 days	yes	no
HLT-913	Days 1-4	yes	yes
SPC-259	All 5 days	yes	no
SPC-471	All 5 days	yes	yes
SPC-688	All 5 days	no	yes
WEA-093	All 5 days	yes	yes
WEA-367	All 5 days	no	no
WEA-842	Days 1-4	yes	no

Table 6.11: Tag label generation and reuse by Daybreak participants, as well as use of potentially hierarchical tag labels (multi-word tag labels that may indicate conceptually linked subtopics at different levels of specificity).

that some of the never-mentioned tag labels include procedural tags (tags intended to serve a separate function, such as noting that a document is irrelevant or "blog fodder"). Table 6.12 shows the percentage of a participant's tags that were either mentioned directly or indirectly in the the story outline. A direct mention is one where the literal tag label was included in the outline. For example, the tag label might be (Red Sox baseball player) "Ortiz," who is mentioned by name in the story outline. An indirect mention is one in which the exact terms from the tag label are not used, but a related concept is used in its place. For example, for the Ortiz example above, the tag label could be "Big Papi" (a nickname for Ortiz), but the reference in the outline mentioned "Ortiz." An example of an indirect mention for the cryptocurrency task might be a reference in the article to "crypto security," when the tag label used was "crypto hacking." The final story outlines were coded to include both direct and indirect references.

We also reviewed how the tag labels that the participants identified as important were

Participant	Tags Used in Story	Total Tags	Percent
BBL-201	0	0	n/a
BBL-887	6	28	21%
BBL-927	8	29	28%
FIN-326	9	11	82%
FIN-455	6	10	60%
FIN-499	6	12	50%
HLT-409	37	71	52%
HLT-555	2	36	6%
HLT-913	21	51	41%
SPC-259	22	37	60%
SPC-471	15	59	25%
SPC-688	6	19	32%
WEA-093	15	48	3%
WEA-367	7	12	58%
WEA-842	8	12	67%

Table 6.12: Fraction of each participant's tag labels that were referenced in the story outline, either directly or indirectly. A direct reference is when a participant specifically mentions the term(s) from the tag label in the final story outline. An indirect reference would be one where the participant mentions a term that was not the literal tag label, but that would map to it—for example, if the outline mentioned "developed countries" and the tag label was "UK" or "Canada."

referenced in the story outlines, shown in 6.5. All participants referenced important tag labels from at least two of the five days in their final stories. Note that this includes direct (the literal term) and indirect (synonyms or related terms) references to the important tag labels, as well as cases where the participant listed the same tag label as being important on multiple days. The one exception is the participant who did not apply tag labels [BBL-201]. That said, BBL-201 did list "Red Sox" (the topic label) in response to the question about the most important subtopic in multiple post-day questionnaires; he did mention this phrase in the final outline, which is counted here in the tally.

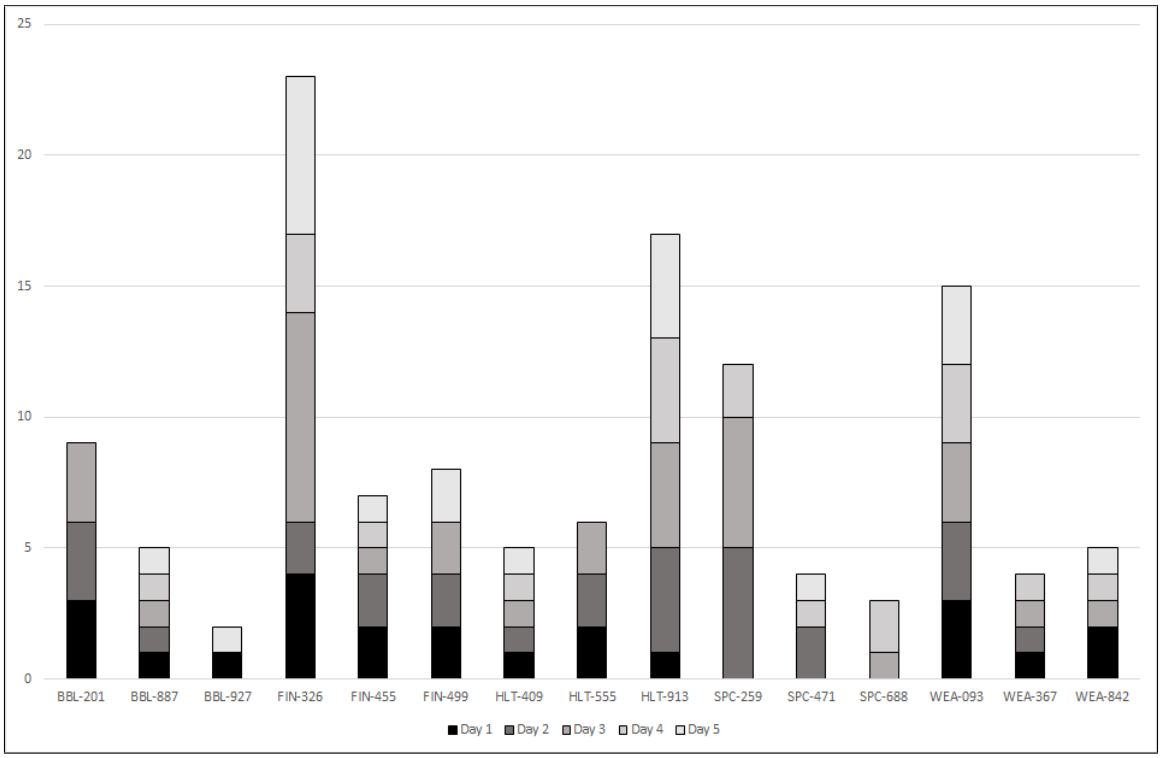


Figure 6.5: Usage of tag labels identified as important by participants in the final story outline. This includes both direct and indirect mentions of an important tag label.

We assessed the participants' final story outlines based on two factors: length and specificity. Length was judged based on the relative number of lines included in the story; options were short, medium, and long. Specificity refered to the amount of detail avail-

able within the outline, whether specific or general. Table 6.13 displays the assessment of each of the story outlines. While all 15 participants wrote final story outlines that related to their topic, two of the outlines were short and general in nature, and likely could have been written by someone with a basic awareness of the topic who did not view the documents within the session. The remaining 13 participants drafted specific outlines that were assessed as being either medium or long. We note that the participant [BBL-201] who tagged documents but did not apply tag labels successfully wrote a detailed outline for the storytelling task. These outlines reflected the information from the documents that the participant viewed within the Daybreak system for the session. For the participants who produced general outlines, we reviewed the participants' tag labels from the session. These participants' tag sets were specific to the themes of the document collection, and indicated likely learning. The brevity of the outline could be explained by study fatigue. Story outline preparation was the last task in a long study; some participants may have hit a mental saturation point before that point. For example, one of the participants who produced a general outline [HLT-555] spent 3:45 minutes drafting the unevaluated practice storytelling task after Day 1. However, for the final storytelling task this participant spent only 1:19 minutes on their final outline; this was the shortest amount of time spent by any of the participants.

Table 6.14 provides two final story examples: one is a short, general story, and the other is a long, specific story. The second example is noteworthy because the participant [BBL-201] was the sole participant who tagged documents, but did not apply tag labels for his tags. Additional research would be needed to determine the extent to which the learning process is supported or enhanced through a participant's action of externalizing the mental model with tag labels.

The study also included some moments of serendipitous discovery for the participants. Even though the documents were at least a few years old, participants were introduced

Participant	Length	Specificity
BBL-201	Long	Specific
BBL-887	Short	Specific
BBL-927	Medium	Specific
FIN-326	Long	Specific
FIN-455	Medium	Specific
FIN-499	Long	Specific
HLT-409	Medium	Specific
HLT-555	Short	General
HLT-913	Long	Specific
SPC-259	Long	Specific
SPC-471	Long	Specific
SPC-688	Medium	Specific
WEA-093	Long	Specific
WEA-367	Short	General
WEA-842	Medium	Specific

Table 6.13: Results from the assessment of each participant's final story outlines.

to ideas that they had not previously encountered. For example, one of the participants assigned to the cryptocurrency topic indicated an interest in the concept of crypto activism (application of cryptocurrency to issues such as human rights or environmental activism). They had not previously been aware of this aspect of cryptocurrency use, and indicated that it was an area that they wanted to explore further.

We also wanted to determine how the participant's comfort level with the system—especially related to completing the task—had changed over time, as another approach for determining the effectiveness of the system. Most participants indicated that the system did help them in understanding what was happening on a day to day basis. All participants were at least partially successful in completing the change detection task, and 13 of the 15 participants drafted an outline of the events that happened over the course of the five days, as reflected in the documents. Participants quickly caught on to the task of tagging documents, and all except one added tag labels to their document tags. Their tagging approaches evolved over the five-day sessions, applying many tags on the first day, but

Short, general outline: HLT-555

```
Health Care
- Need
- Health care reform efforts
- Affordable Care Act
```

Long, specific outline: BBL-201

```
Start with a review of the week for the Red Sox.
    - The 14-1 victory over the Blue Jays
    - The 6-4 loss to the Blue Jays
    - The 8-0 Loss to the Blue Jays
    - The 6-4 loss to the Rays

A review of the standings and where the Red Sox are in the
standings.

A review of the milestones this week---David Ortiz tying and
passing Yaz on the homerun list.

A review of the David Price vs David Ortiz matchup from previous
years. A review of Ortiz and his health.

A look to the future with players likely to be traded this year
or signed in the off season. A mention of the possible opening
day for the following season

Mention the Red Sox Ring Raffle for the University of Maine.
```

Table 6.14: Examples of final story outlines produced by Daybreak user study participants at different assessment levels. One was assessed as short and general, and the other was assessed as long and specific.

over time reducing their new tag generation volume as their familiarity with the topic and task increased. Seven participants showed a drop of at least five tagging events per day between Day 1 and Day 2.

Some participants indicated that the Daybreak task resembles things they do on a regular basis. For instance, one participant [HLT-409] stated that "This is really quite a lot like the work I do daily. Environmental scanning is a big part of the work I do." Environmental scanning is a use case similar to change detection in which an individual seeks information on behalf of an organization, looking for issues that could affect the organization's decision-making process [7].

Based on input from the participants in the post-day questionnaires, we found that the participants became more familiar and comfortable with system and task over time. Figure 6.6 shows participants' self-assessments of the extent to which they had seen enough documents from day to day. After Day 1, the majority of participants indicated that they had not seen enough documents on the topic. However, over the course of the session this changed; by Day 5, 60% of the Daybreak participants said they had seen enough documents on their topic.

The combination of the GPA components led to participants being able to complete the change detection task successfully, by externalizing their learning and updated mental model into a story. All but two participants created stories that represented their learning. These participants' tagging approaches indicate that they successfully followed changes in the topic over time; however, the story they produced was more of an explanation of prior knowledge, less tuned to the specific documents reviewed in the task.

The participants also described how the system helped to customize the results to their specific interests. One of the Red Sox participants [BBL-927] said, "I like the idea of being able to differentiate between the particular components of things I like." This participant went on to compare their interests with that of their spouse, who are both are baseball fans.

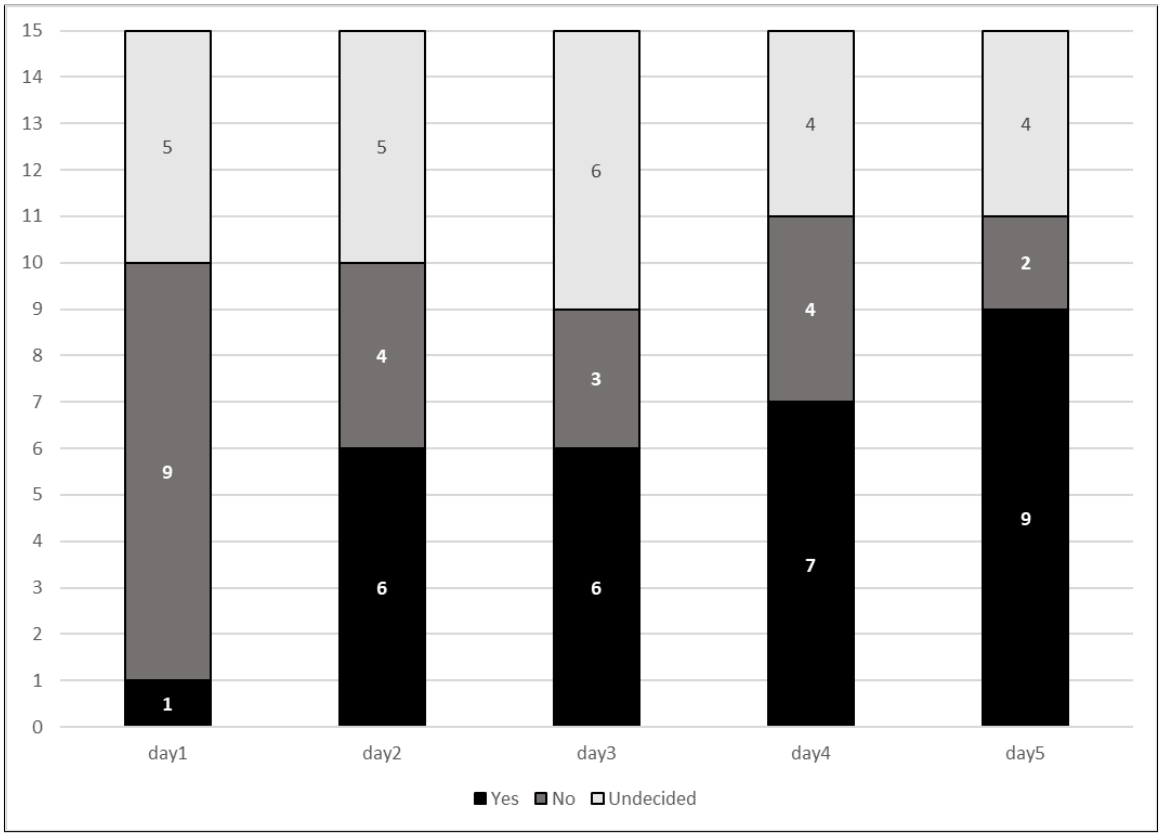


Figure 6.6: Daybreak study participants' views on whether they had seen enough documents, by day.

The participant tends to be more interested in the rules and the economics of baseball. Their spouse, however, is more interested in the "hot stove"—a reference to the baseball offseason, when the free agent market starts to heat up. This participant liked the fact that the Daybreak system allowed them to focus on the aspects of the baseball that they find more interesting—which was one of the goals for the Daybreak system.

Recommendation: Improvements to Daybreak User Experience

While tuning and adjustments are needed in specific aspects of the system, in general, the participants were largely successful at performing a change detection task within the Daybreak system. Future research could focus on enhancing the system based on findings from this study, to enable more functionality to help with a change detection task. This could include generated summaries for subtopic clusters (for instance, generated by a Large Language Model) or options that show individual documents contributed to new information on the subtopic [305]. A future study could devise additional methods (beyond our assessment of a story outline) to determine whether a participant learned, and whether their usage of the improved system result in improvements in participants' comfort level and understanding of the topic they follow.

6.3.8 Observations about the Practice of Change Detection

During our research, we observed a number of participant behaviors that were noteworthy, but that do not directly tie to the research questions. In particular, this includes participants' behaviors and reactions related to the eight-minute time limit for each day, and the (potentially related) amount of time participants spent on different actions each day that they used the Daybreak system. We note the recall-oriented nature of the change detection task; the survey results from Chapter 4 revealed that respondents generally preferred

to see "all" data. To what extent did the participants strive to see every document, driven by a quest for digital omniscience?

Even though participants were largely successful in completing their change detection task, we found that multiple participants revealed concerns about missing important information—a sort of digital "fear of missing out" (FOMO). The time limit was intended to keep the overall session within a reasonable range; while we expected it to add some pressure to the task, some of the comments indicated that it may have exceeded our intended level of pressure, possibly driving participant behavior in directions that they might not take in real life. One participant [SPC-259] noted, "With no time limitation, I definitely would have thought a little bit more, but I think I was just like, 'I have to hit upon something I can do quickly and just go with it.'"

This time pressure also tied in with the concept of "enough"—whether the participants felt they had seen sufficient documents to understand what was happening related to their assigned topic. One participant discussed the ongoing nature of the change detection task, tying together the concepts of seeing "enough" and FOMO (e.g., "satisficing" behavior") [FIN-455]. "It never ends. None of these things ever end. So I think that's the answer about whether or not I looked at enough. Once I hit like 15 articles, I was like, 'Alright.' But being an internet addict like I am, I knew I would just keep going regardless because I didn't want to be left out of whatever was happening online."

We observed a variety of participant behaviors during the session that may be worth further study. Based on the setup of this study, it would be difficult to determine whether these behaviors represent the participants' normal behavior when performing change detection tasks, or whether some of these actions represent adjustments and shortcuts taken due to the time constraints. We separated these behaviors into four categories: 1. attention and assumptions, 2. tagging urgency, 3. tag label volume, and 4. tagged document volume. While these areas were noted as interesting observations, given the potential conflating factor of

the time constraint and the fact that they were not directly tied to the research questions, we did not formally code for these activities; instead, we note them here as potential areas for future research.

1. **Attention and Assumptions:** This category covers where the participant spent the majority of time and attention before taking another action. There were three areas where participants tended to apply their attention: viewing the list of subtopic clusters, viewing document metadata such as document titles, and viewing documents. Participant indicators of these actions include scrolling up and down through the respective section, and pausing to read a specific section before clicking on something else. The particular area of focus of the participant is potentially significant because their attention can indicate assumptions and other issues. For instance, a participant may make an assumption based on a subtopic cluster label that no documents in the cluster are relevant; alternatively, a participant might decide whether or not to open a document based on a document title or publisher. Such assumptions could have an impact on whether or not the participant identifies information that they might need. The differing approaches for applying could also serve as a topic for future research on behaviors related to learning and assumptions. Such a study could look more deeply into different discovery and learning approaches, perhaps under different time constraints—for instance, understanding why some participants spent time on reviewing titles whereas others focused on spending their time reading documents.

2. **Tagging Urgency:** We noted two general behaviors relating to the urgency with which participants applied tags. Some participants opened a document and immediately began adding tags—as well as multiple tag labels—for each document they opened. It was not clear that these participants found the document useful; rather, they appeared to be documenting many of the concepts they saw in the document.

In contrast, other participants waited and read through a document (albeit sometimes quickly) prior to deciding whether to tag the document, and at the end of their reading tagged the document with one or more tag labels. This behavior may have shifted over time, with participants tagging more frantically on Days 1-2, and relaxing their tagging as they became more comfortable with the task.

3. **Tag Label Volume:** This concept relates to the number of tags that a participant applied per document. Some participants added a large number of tag labels per document, whereas others added only a single tag label to represent a document. While the most prolific users of tag labels may have also applied tagging urgency, there may be differences between these two sets of participants. Further research could look at what drives someone to apply tag labels in a variety of circumstances. As an example, as discussed in Section 6.2.1, the two participants who applied the most tag labels also viewed among the fewest documents. In general, the global health participants tended more than other topics to apply multiple tags to a single document, to represent various themes they saw in the text.

4. **Tagged Document Volume:** Finally, we observed that some participants were more judicious than others about the number of documents that they tagged. Some participants appeared to tag all or most of the documents that they viewed. However, others only tagged a few documents. A future study could look at these behaviors, to better understand the differences in these behaviors.

As one participant [WEA-093] indicated in a post-day questionnaire, "Even with clusters, I find that I want [to] read titles, skim documents, read and tag them and don't have the time to do so, with the lurking fear that my inability to do so is going to hurt me—in terms of situational awareness—over the coming days." Even taking into consideration the time-related pressure, over the course of the five days, we expected that the participants

would become increasingly comfortable with their information seeking approach, and optimize their use of the time to identify sufficient documents representing changes on their topic. These observed behaviors and tradeoff decisions made during these sessions revealed complexity in the participants' learning approaches, and potential knowledge management preferences.

6.4 Challenges and Limitations

The Daybreak user study had a complicated setup consisting of many moving parts—a manually-driven prototype interface, Zoom calls requiring sharing keyboard access with the participant, and more. This manual approach felt at times like a Rube Goldberg machine while sessions were being run; it was a complicated task to line everything up correctly and execute the pieces for each session. While the overall study was completed and produced useful results, we identified a number of areas for improvements. We have organized these issues into categories: challenges while running the study, biases and related issues, and usability issues.

6.4.1 Challenges while Running the Study

We encounted a number of issues and challenges during the study. In this section we address issues with system setup, Zoom issues, and system malfunctions.

The Daybreak system contained multiple pieces behind the scenes due to the way that we had chosen to implement it as a bare-bones prototype. The subtopic customization by participant per day proved to be especially challenging. The pilot sessions aided in hardening that component sufficiently that there were no crashes or errors within the main sessions. Even so, each session contained a real risk of something going wrong; a future

study of this sort could focus on implementing the system in a more robust way that requires less interaction from the researcher.

The Daybreak user study included a large number of activities for the participant to complete, with five separate simulated days and other activities surrounding them. This complicated series of tasks, plus the imposed time crunch within each day, were noted by the participants as being stressful. In fact, one participant repeatedly asked whether the study was actually a psychological study where we were studying the participant's reaction to pressure or stress, or something else that was not our intended and stated intent.

Fraudulent participation caused a good deal of wasted effort, but the majority of our participants were indeed real people who had seriously intended to participate in the study. Some of these authentic participants had issues connecting to the Zoom call or lacked experience using Zoom features. One factor that we had not anticipated was that the small Zoom video window obscured portions of our interface as seen on the participant's screen. We adapted after the first few sessions and started offering brief instruction on minimizing or moving that small window out of the way. In the future, we would design our screen layout to include an unused place where that window can be parked.

Our choice to run our instrumented prototype locally on our laptop, sharing keyboard and mouse control with participants so that they could manipulate the system themselves, generally worked well. Latency was noticeable at times, but tolerable. Some participants did comment on this in the semi-structured interview at the end of their session. If we were to do this again we would want to test the system under a broad range of latency conditions (e.g., by using distant VPN servers) prior to the main study.

We had only one computer malfunction, when the researcher's computer froze near the end of a session and had to be rebooted. This happened after Day 5 with the Daybreak system, so it did not interrupt any of the hands-on elements. We had planned for such an eventuality, and had designated the observer as a Zoom co-host. The observer kept the par-

ticipant informed as we worked through this brief delay. No data was lost due to the system crash. We originally included the observer simply as a note-taker, but there were a number of ways in which the assistant was helpful, including dealing with the unexpected crash. The observer also helped to identify general themes related to the participant's actions and identified points of note to review during the analysis. We could have added additional responsibilities to the observer. In one participant session that was not included in the final results, we neglected to start the recording; this resulted in the inability to use the results from that session, due to a major missing artifact. Had we thought of it in advance, the observer could have verified that recordings were started for all sessions.

Even though we started with a demographically balanced set of selected participants, the challenges described earlier in this chapter resulted in a set of participants that was less demographically diverse than previously envisioned. This could result in results that are biased toward a certain subset of participants, which may not be as generalizable as hoped to a broader audience. This is a tradeoff of qualitative research; however, the results of the study did provide helpful initial view into the participants' change detection practices.

After Day 1 was complete, the post-day questionnaire included a number of extra questions—beyond what was in later questionnaires—to understand the alignment between the assigned topic and their knowledge and interests. Figure 5.13a showed the participants' knowledge levels about their topics of interest. Nine of the 15 participants had an intermediate or advanced knowledge of the topic. The remaining six have only a novice or basic understanding of the topic assigned to them. As noted in Section 6.3.4, a misalignment between participants and assigned topics means that the participant's observed behavior might differ from the way they would complete the task if it was a topic for which they have more familiarity, or that they follow regularly. This came out in some of the interview responses—for instance, one participant commented that they preferred relevance sort for this study because they were still learning the topic, but in a real situation they would likely

prefer reverse chronological ordering. In a future study, more dimensions of information could be used to assign participants to topics—for instance, there could be a further validation step in advance of the main study through which the participant could indicate their interest in the topic, as an opportunity to identify cases where the participant may have less interest in the topic than anticipated.

One of the most common concerns raised by participants related to the perceived large amount of data to review within only eight minutes. While participants generally appeared to adapt to time constraints over time several participants repeatedly mentioned their stress levels in the post-day questionnaires and in the final interview. While we intended to cause the participant to feel a time crunch, there is a concern that the high stress levels may have led some participants to change their behavior, such as taking shortcuts that they would not have done in an ordinary situation. For instance, they might have made assumptions about a document's relevance based on titles, and tagged documents without reading the text. This is an issue that could be further studied as future research.

Distracted by Storytelling Task: In at least one case, a participant was distracted by the fact that the storytelling task was about writing an outline for a blog post. They mentioned at a number of points that the Daybreak use case was not something they could relate to. They said, "It was hard for me to put myself in that situation and that sort of scenario, so I got really focused on tagging in the articles and less focused on, like, 'is this information useful for me or is it not useful for me?'" They noted that a use case focused on their personal interests, not the scenario of finding information for the blog, "seemed like a completely different task."

6.4.2 Biases and Other Impacts

The Daybreak user study was intended to be a starting point to enable us to understand how a participant might use Daybreak to perform a change detection task. However, we note that there were some specific issues and limitations that may have impacted our results.

As mentioned during the analysis of the Daybreak user study results, there are possible issues related to the topics selected, and the extent to which there are clear stories available in the document collection. The approach of choosing an arbitrary date range for a sufficiently large document set (100 documents per day) may not have been successful in capturing themes that a participant could follow within the topic. Of the five topics, only the Boston Red Sox topic had a clear, coherent story line over the five days (a baseball team playing a variety of other baseball teams). There were other stories mixed in (e.g., a soccer game at Fenway Park, charity activities, etc.), but the consistent theme was the Boston Red Sox and the games they played during that period of time. For the other topics, there may have been hints of themes (UK extreme weather, healthcare discussions, etc.), but there were fewer consistent threads for those topics. As a result, the startup learning burden for those four topics may have been higher, and the participant likely had less-developed mental models compared to the ones that already baseball-interested participants possessed relating to a baseball team playing baseball games. For future studies we can correct this issue by ensuring that all topics have some clear storyline running through them across the dates studied. Additionally, all participants viewed fewer than 40 documents on each simulated eight-minute day; a future version of a change detection study could likely get by with fewer documents in the collection.

Some participants expressed concerns about the system producing a myopic view based on their prior tags. One participant [WEA-093] noted, “I felt like I narrowed my aperture or reinforced a narrow aperture when that’s not necessarily what I was going for. And in

some ways, it could be a function that actually what I would need are multiple Daybreak agents." Future research could investigate approaches to ensure that the Daybreak system was not unintentionally introducing filter bubbles and other issues.

6.4.3 Usability Issues

We categorized user interface issues, and have divided this section into two categories: system issues and system requirements. The following items were bugs or system issues encountered by the participants:

- **System responsiveness:** One drawback of using Zoom for the user study was that it introduced some latency issues into the user study. In some cases, a participant would click on something, but the system wouldn't respond in the time they expected; they would click again, resulting in multiple selections of the same document and other issues.

- **Adobe Acrobat Reader caching issues:** We encountered two issues with Acrobat Reader caching information, resulting in unexpected participant interactions. Our document viewer frame used Acrobat Reader to display documents. While the first document would open at the top of the document, subsequent documents would open at the same point in the document; this often meant that the participant would have to manually scroll back to the top of the document to read it. A second caching-related issue only happened on one topic. One participant attempted to highlight a passage of interest with Acrobat Reader. We were unable to identify a way to remove that highlight from the document set without overwriting the file; as a result, subsequent participants for that subtopic also saw that highlight. We also note that the participant's interest in using the built-in highlighting indicates interest in a highlighting function for a future version of the system.

We tracked participants' requests for additional functionality that was requested throughout the sessions. For a future study, participant behavior might vary in a more fully built-out version of a Daybreak interface with robust features to support tagging, tag labeling, and other capabilities. Additionally, we found it important to look at user interface issues because there could be aspects of the system that unintentionally distracted the participant from the overall task or caused confusion. The most common interface and functionality feedback included the following:

- **Dated look and feel:** While the Daybreak interface was intentionally built to be a simplified version of an interface, multiple participants commented on the dated look and feel and lack of functionality in the system. A future version of the system could be implemented in a more robust, modern framework to remove the distraction of the dated look and feel of the software.

- **Tag label reuse:** The Daybreak system prototype used in the study required that participants retype an entire tag label in order to apply the same tag to multiple documents. Participants expressed an interest in shortcuts and simplified approaches for associating more documents to the same tag label, to include drag-and-drop.

- **Ontology management:** Participants expressed an interest in having additional functionality available for tag label modifications and ontology generation. For instance, they wanted to be able to modify their tag labels. This included correcting misspellings or adjusting labels as they learned more about the topic. They also were interested in splitting or merging their tags, which represent more sophisticated ontology management than was possible within the Daybreak prototype.

- Knowledge management across documents: The Daybreak user study included the ability to return to previously tagged documents only for the final storytelling task.

A number of participants expressed an interest in returning to documents at other times. Additionally, they were interested in returning to documents from prior days at any point in their system usage. While Daybreak was developed with information retrieval in mind (finding new documents), the knowledge management aspect of the system of being able to return to a document was of interest to participants.

- **Knowledge management within documents:** The Daybreak prototype enabled participants to indicate that a document was of interest, but did not provide functionality for indicating what specific areas of the document were of interest. Participants expressed an interest in features such as highlighting and commenting, which would allow them to indicate what was of interest or what to return to in the document later. In addition to the requested knowledge management role, functions such as highlighting could also aid in the information retrieval process—specifically, for generating queries or machine learning models that prioritize or focus on those terms, rather than all terms across the entire document.

6.5 Summary

In this user study, the participants performed a series of tasks to get updates on a topic and to identify important developments. The user study focused on how well various aspects of the Daybreak system supported the underlying change detection theory—factors needed to effectively learn what has changed on the topic of interest. This study was designed based on prior work—in particular, the sort order survey of social media users to understand how they would like to see data organized in support of change detection and other tasks. These results inform the design of the change detection evaluation approach presented in Chapter 7. There we discuss an evaluation design for comparing subtopic clustering orders, building upon what we learned from the Daybreak user study.

Chapter 7: Evaluation for Change Detection

How can you tell which version of a system provides the best result set ordering for users? Hosting user studies—such as our our virtual Daybreak system user study—can be beneficial for understanding user behavior in the system. However, the field of information retrieval often uses automated evaluation approaches to compare system performance against a task. These types of approaches determine how well each system meets core user needs for a task, expanding beyond the user's statements and actions in the user study. To address this, we have developed evaluation approaches that could enable a comparison of the effectiveness of subtopic ordering for change detection. In this chapter we design an evaluation approach for change detection systems, focused primarily on subtopic importance. We apply Spearman's rank correlation coefficient for comparing the relative system orders for subtopic clusters, to determine which methods for organizing the subtopic clusters might best align with user interests. The discussion of the evaluation approach is followed up with two example implementations, one applied to a manual session by the author, and the other based on the annotated participant data from the Daybreak study.

7.1 Evaluation Approach

Automated information retrieval system evaluations enable comparisons between systems that use different approaches, and make it possible to understand the state of the art. Unlike in-person user studies, these types of evaluations are abstractions of the user need,

translated into a mathematical formula. The evaluation approach should be general enough to enable a comparison between human-driven, computer-driven, and hybrid systems. This chapter includes two example evaluations—a sample session completed by the author, and the Daybreak user study results from Chapter 6. These examples are not intended to present any results as definitive findings; rather, they show how an evaluation can be used to gauge the extent to which the Daybreak system's method of arranging subtopic clusters meets the change detection needs of users. For this evaluation approach we evaluate a range of potential subtopic clustering orderings for the Daybreak system, including document rarity, the approach applied within Daybreak for the user study.

For this use case, we apply a listwise comparison. While a listwise approach applied to a complete document set could be cumbersome, we focus here on the ordering of subtopics only; the order of the documents within the subtopic cluster would be organized separately from the subtopic arrangement. To perform this comparison we apply an evaluation measure to determine subtopic orderings that align with user subtopic rankings. This chapter describes our evaluation approach for subtopic importance, with the goal of enabling evaluation of the effectiveness of a system designed to meet the change detection needs of users and compare systems' approaches for organizing subtopic clusters.

7.1.1 Assumptions

Given that the change detection task focuses on the information needs of an individual user, the set of ideal rankings is tied to the preferences of one user. We assume that the set of relevance judgments related to subtopic importance is representative of the user's interests, and that they have identified at least the most important subtopics, from their individual perspective.

Additionally, we assume that the document ordering within a subtopic cluster is not rel-

evant to the subtopic cluster order evaluation. Document retrieval and sorting are separate, and the document sort approach (e.g., chronological vs. relevance ranked) is not tied to the subtopic cluster ordering.

The evaluation approach should not penalize the system for not including subtopics that were not yet discovered by the user. We expect the user to generate new tag labels over time; the Daybreak system is not set up to predict future or emerging tag labels. As a result, these subtopics are scored on or before the user creation date, but are part of the evaluation on future days.

7.1.2 Characteristics of a Subtopic Importance Measure

To be able to compare subtopic cluster orderings, we need an annotated subtopic list for which the user has stated their interests (rank 1 is most important, rank n is least important), which represents the user's ideal subtopic ordering (the ground truth data for this evaluation). We can compare that ideal list with a range of system orderings to determine which approach yields results most closely aligned with the user's ideal arrangement. To compare the two lists, we need a measure with the following characteristics:

- The evaluation measure should enable a comparison between the user's ideal subtopic rankings and results generated by a change detection system.
- The resulting score from the evaluation measure value goes up when the subtopic importance order provided by the system better aligns with the ranked subtopic importance list from the user, and down if it does not align as well with the ideal results.
- The evaluation measure provides an ability to show the direction of correlation between the ideal ranking and system ranking.

- There is an identifiable change in the score if one of the ranked lists change (i.e., either the system or user modifies their ranking of subtopics)
- The evaluation measure penalizes the system ranked list if it omits some subtopics that were known to be important to the user on a prior day.
- The evaluation measure should be capable of handling ties if multiple items have the same rank.
- The result from the measure is summarized in a single value.

7.1.3 Evaluation Measure: Spearman's Rank Correlation Coefficient

To meet the goals of the evaluation approach, we have selected Spearman's rank correlation coefficient. This measure enables a comparison between two ranked lists. Rather than comparing the values of the list, as is done by measures such as Pearson's correlation coefficient, Spearman's rank correlation coefficient focuses on comparing the relative rank ordering of the values in the two lists. The equation for Spearman's rank correlation coefficient is shown as equation 7.1. We follow our description of the calculation process with an example.

$$\rho = 1 - \left(\frac{6}{n(n^2-1)}\right) * \sum d_i^2 \tag{7.1}$$

To calculate Spearman's rank correlation coefficient, we start with two ranked lists: the ideal ranked subtopic list, and the system-generated ranked subtopic list. For this correlation, we look at the relative rankings between the two lists. Rather than referring to the subtopic, we focus on the rank of the subtopic. For instance, suppose the ideal rank order includes the subtopic labels [loss, security, injuries]. These would be ranked [1, 2, 3] for

the calculation. If the system produces a ranking of [security, injuries, loss], the rank for these subtopics would be [2, 3, 1].

After we have the numbered ranks for each list, we compute $\sum d_i^2$, which is the sum of the squares of the differences between the ranks of the two items in each paired set. This is multiplied by the constant $\frac{6}{n(n^2-1)}$, where n is the total number of items in the ranked list. The value of 6 is a normalization factor to ensure that the resulting score is between -1 and 1.

As an example, we look at the ideal and system subtopic orderings from one of the Daybreak sample cases. For the sample calculation we take the first 4 ranks ($n = 4$) from BBL-927, Day 5; we compare the ideal ranks [1, 2, 3, 4] with the system's rarity rankings [1, 4, 2, 3] for their rank order within the system result set.

We calculate the the square differences between the ranks as follows:

ideal rank	system rank	difference	square
1	1	1 - 1 = 0	$0^2 = 0$
2	4	2 - 4 = -2	$-2^2 = 4$
3	2	3 - 2 = 1	$1^2 = 1$
4	3	4 - 3 = 1	$1^2 = 1$

For this example, the sum of the squares is 6. For this use case, we have 4 pairs; with n = 4, we compute a Spearman's rank correlation coefficient for the example.

$$\rho = 1 - \left(\frac{6}{60}\right) * 6 = 0.40 \tag{7.2}$$

Spearman's rank correlation coefficient is able to handle ties between two items of the same rank. In these cases, the items receive the rank that is the average of the ranks that they would have been in. For instance, if the values in ranks 1, 2, and 3 are the same (tied),

the ranks used to compute the differences are (1 + 2 + 3)/3. The result, a rank of 2, would be used in the rank difference calculation for all three ranks.

One limitation of this measure is that it is designed to handle variability in the sets. If one of the sets contains ties at all values—that is, all of the ideal ranks or all of the system values are the same—it is unable to calculate a Spearman's rank score.

7.2 Example: Manually Generated Change Detection Dataset

Now that we have selected an evaluation measure for the task, we need a set of human-generated ground truth data and system results. To test the evaluation approach on sample data, we manually annotated documents from the Daybreak user study to enable further testing of potential systems to rank subtopic importance. Specifically, we reused the documents from the Red Sox topic. These results are to further demonstrate the utility of the evaluation process, and provide insights about potential sort orders suitable for the change detection task.

For subtopic label generation, the author ran the Red Sox task in the Daybreak system for five days, to be consistent with the approach used with the Daybreak user study in Chapter 6. For these sessions, the task was performed without a time limit. This resulted in significantly more tags and tag labels per day than the participants in the Daybreak study generated. Total numbers of tags generated in this session are included in Table 7.1. While new tag labels were created on each day of the session, the number of new tag labels created per day decreased over the course of the session in favor of reusing existing tag labels.

Immediately after tagging documents each day, ground truth (the ideal rank) was generated. The process for creating the ideal rank was to take the tag labels for that day and ranked each one as being of high, medium, or low importance for the day. Some subtopics were also identified as being "irrelevant" (e.g., documents about a different Red Sox team

Action	day1	day2	day3	day4	day5	Average	Total
Tags Created	**93**	63	30	27	18	46.2	231
Tags Applied	235	**712**	539	625	528	527.8	2,639

Table 7.1: Total number of tags created and applied per days in the author's untimed session. "Tags created" represents new tag labels generated that day. "Tags applied" includes all tag labels added to a document's tag, including new as well as reused tags. Note that in many cases, multiple tags were applied to a single document.

Tag Label	day1	day2	day3	day4	day5	Total
RED SOX VS BLUE JAYS	5	48	50	**56**	41	200
RED SOX LOSS	0	2	42	**50**	40	134
ORTIZ	3	**43**	27	37	10	120
BIG PAPI	3	**43**	25	34	10	115
HOME RUNS	1	**46**	31	22	1	101
RECAP	11	11	6	**16**	14	58
BOX SCORES	9	11	9	12	**15**	56
BIG PAPI RECORD	1	**35**	11	8	0	55
RED SOX WIN	17	**37**	1	0	0	55
FENWAY	3	8	**19**	10	9	49

Table 7.2: Top ten most frequently applied tag labels across the five days in the author's Daybreak session. The most common tag label for each day is indicated in bolded terms.

than the Boston Red Sox, references to red socks, etc.). After generating the subtopic labels through the Daybreak task, we then ranked the entire list. This process consisted of first marking each subtopic label (starting with the group of subtopics identified as having high relevance) from 1 to n. We did not allow for any ties in the high, medium, or low categories. However, we applied the same bottom rank both to subtopics marked irrelevant as well as to subtopics not present on a given day. The resulting list from this process is what we used as the author's ideal ranking set.

As an example of the tag labels applied and their corresponding rankings, Table 7.2 lists the top 10 most prevalent tag labels applied by the author during the Daybreak session. Table 7.3 includes the ideal ranking corresponding with these tag labels for each day.

Tag Label	Day 2	Day 3	Day 4	Day 5
RED SOX VS BLUE JAYS	8	**1**	3	7
RED SOX LOSS	18	**2**	2	3
ORTIZ	**3**	7	4	8
BIG PAPI	**2**	6	5	10
HOME RUNS	**6**	9	14	19
RECAP	**42**	61	75	64
BOX SCORES	**43**	62	77	69
BIG PAPI RECORD	**1**	5	6	71
RED SOX WIN	7	**4**	79	71
FENWAY	**10**	30	39	32

Table 7.3: Ideal rank by day for each of the Top 10 most frequently used tag labels from the author's Daybreak session, where 1 represents the highest ranked document. For each tag label, we have indicated that label's best rank across the five days in bold text. Note that Day 1 ranks have been omitted because the first day we scored subtopic cluster orderings was on Day 2.

7.2.1 Evaluation Approach for Author's Session

The version of the Daybreak system used in Chapter 6 only enabled subtopic ordering by rarity (least populated subtopic cluster presented first). To compare the rarity-based ordering with other options, we reranked the subtopics in a variety of ways to see what other sort options might align with the ground truth rankings. The system-generated subtopic orders that we evaluated included the following:

- **Rarity (default Daybreak sort order):** Subtopic cluster containing the fewest documents is listed first.
- **Prevalence:** The opposite of rarity, in which the most populated subtopic is displayed first.
- **Document recency:** The subtopic containing the most recent document is first.
- **Document maturity:** The subtopic containing the oldest document is listed first.

Note that this ordering is not the opposite of document recency; a single subtopic cluster could contain both the oldest and newest document.

- **Subtopic recency:** The most recently created subtopics are at the top, sorted chronologically by the time they were created.
- **Subtopic maturity:** The oldest subtopics are listed first, with the remainder sorted by reverse chronological order.
- **Tag frequency:** Organized based on the number of times a tag label was applied across sessions, with most prevalent tag label listed first.
- **Tag infrequency:** Tag labels that were applied with the lowest frequency are listed first.
- **Term frequency:** Sorted based on the number of times the topic term is present within the document, with documents containing more instances of the topic term listed first. More specifically, the topic terms that we used to calculate term frequency were Red Sox, cryptocurrency, health, space, and weather for the respective topics' document sets.
- **Alphabetical:** Subtopic clusters are arranged alphabetically by tag label. This is included as an example of an arbitrary ordering.

We ran a separate comparison between the ideal ordering and the system ordering for each day. Because paired sets of values are needed to calculate Spearman's rank correlation coefficient, for each day's comparison we limited our analysis only to tag labels that were created prior to the current day. By using the tag labels available to the system that day, we did not penalize or reward the system for finding or not finding subtopics that a user had not tagged before that day.

System Ordering	Author
tag frequency	**+0.48**
prevalence	+0.34
document recency	+0.26
document maturity	+0.25
topic tf	+0.24
subtopic recency	0.00
subtopic maturity	0.00
alphabetical	-0.04
rarity	-0.34
tag infrequency	-0.48

Table 7.4: The Spearman's rank correlation coefficients for the author's session across Days 3-5.

7.2.2 Results from Author's Daybreak Session

For the author's session, the most effective system orderings were based on tag frequency and prevalence. Rarity was the least effective ordering for this session. For Spearman's rank correlation coefficient, a system ordering was considered effective if it had a high score (closer to 1). The top ordering have a moderate positive correlation with the ideal rank. The full set of results are presented in Table 7.4.

Use of a head-weighted version of the Spearman's rank correlation coefficient, such as the one described by Bailey, et al. may aid in reducing the impact of low-ranked tag labels on the score [21]. This could be a topic for future research, to include identifying the threshold for which the number of tags for which a head weighted measure would be advisable.

7.3 Annotated Data: Daybreak User Study Results

To further demonstrate the proposed evaluation approach, we applied the methodology to the results from the Daybreak user study (Chapter 6). This serves as a second example of how a change detection-focused evaluation approach could work, and provides further

insights into how subtopic importance might be defined. We did not specifically ask the participants to rank more than their top subtopics for a given day. The example discussed in this section makes substantial assumptions about the participants' interests. Given that participants did not specifically identify more than three top subtopics per day, this section is based on assumptions about their interests, and is not intended to present our results as definitive findings. That said, the results may provide some broad insights into RQ5.4 (Section 5.1.4), in particular related to the effectiveness of rarity (the default ordering in the Daybreak system) compared to other possible subtopic cluster ordering options.

In order to perform an evaluation, we need ground truth data—which we refer to as the "ideal" ordering of subtopics. However, we had only an incomplete set of information about what subtopics were most important to the users on each day. When completing the user study, the between-days questionnaire included a question about which tag label was most important for that day. Participants noted up to three subtopics for the day; some participants did not identify important tag labels on some specific days. We applied recency as a secondary heuristic to interpret participant behavior for the ranking process. This enabled us to produce an ordered list of which subtopics they might have considered more important on a given day. We leveraged this information in the process of turning the participants' tag labels into a ranked list. Given that the underlying task is change detection (identifying what is new or different on the topic of interest), we set up the rankings with the concept of recency in mind. That is, we assumed that the participant has a higher interest in something that is new or different. This interpretation of the participants' rankings added potential bias toward recently created or recently applied tags, which comes out in the example results. In contrast to the author's ideal ranking, which was a list of 1-*n*, the assumptions used to generate the ground truth data (ideal ranking) for the Daybreak user study data resulted in ties at certain ranks. For example, using our recency heuristic, multiple tag labels had most recently been applied on the same day.

We applied the following approach to generate the ideal ranked list:

- The tag label(s) that the participant indicated in the post-day questionnaire as the most important for that day received the highest rank. Exception: if all of the most important subtopics were first created that day, the most important tag label(s) from the prior day were used.
- The next highest rank(s) went to the subtopics identified as important in prior days' questionnaires, from most recent to oldest day (e.g., for Day 5, after adding the Day 4 top ranked item the next would be the most important subtopic(s) from Day 3, then Day 2, etc.).
- After that, the next highest [tied] rankings would go to subtopics created most recently. For instance, if it is Day 5, we would include subtopics that were new on Day 4 next, followed by ones created on Day 3, and so forth. Rather than using specific creation times to assign rankings, we considered all subtopics created on the same day to be tied at the same rank.

Table 7.5 contains an example of the ideal rankings for Red Sox user BBL-927 on Day 3. In this example, the tag label HUTCHISON (ranked first) was identified by the user as the most important tag label on the Day 2, and RESULTS (ranked second) was most important on Day 1. The tag labels that are tied for the third ranking were created on the same day. The lowest-ranked tag labels were either identified as irrelevant or not present within the result set that day.

7.3.1 Comparing Rankings

The next step in the evaluation process is to compare the ideal ranking order with the subtopic orders produced by a system. We apply Spearman's rank correlation coefficient

Ideal Rank	Tag Label
1	HUTCHISON
2	RESULTS
3	14 1
3	7 22
3	LOSS
3	RECORD
3	WIN
8	BOSTON
8	BOXSCORE
8	DREW
8	FARRELL
8	IRRELEVENT
8	LACKEY
8	LESTER
8	NAVA
8	NON GAME
8	ORTIZ
8	REGINA
8	ROSS
8	SCHEDULE ONLY

Table 7.5: Sample ideal ranking for Daybreak participant BBL-927 for Day 3, with tag labels presented as worded by the participant. This ideal rank starts with the tag label(s) that the participant indicated were most important, followed by the most recently created tag labels. The tag labels ranked last (in this case, tied at position 8) were either identified as irrelevant or not present that day.

to ascertain which system performs better at ordering subtopic clusters in a way that aligns with needs of change detection users. We generated the same set of system ordering types as applied for the author's session to produce system rankings for subtopics, to test whether another order might better suit the needs of the Daybreak study participants. The Daybreak system generated subtopic clusters as a pre-processing step for all days, even if a participant chose not to leverage clustering for certain days; we used those generated orderings here, even if the participant did not choose to display them during the session.

Given our underlying assumptions that led us to organize the ground truth data based on the day when the subtopic was created, we would expect the subtopic recency ordering to result in a high correlation with the ideal ranking order. In contrast, we expect the subtopic maturity ordering to perform poorly, as more recently created subtopics are ranked at the bottom.

As with the author's session, for each session here, we compared the ideal ordering with the system ordering—this time, for each participant, on each day. As mentioned previously, we did not include subtopics that the participant had not yet discovered or labeled that day in our orderings. Subtopics that the system did not find on that day were ranked at the bottom of the list.

Similar to the approach used for the author's session, we did not include a comparison for Day 1. That was the cold start day; participants came into the first day with no subtopics, and began creating them that day. The first day when subtopics were available for a comparison was Day 2, the first day when the Daybreak system generated subtopic clusters based on users' tags and tag labels. We ran our Spearman's rank correlation coefficients on Days 3-5, because the Day 2 results only looked at the system results from tags created on Day 1, when the participant was still gaining familiarity with the system. By limiting our comparison to Days 3-5 we had a larger number of tags for the comparison.

Of the 15 total sessions from the Daybreak study, we used 13 for the comparison. One

participant did not add labels to their tags; they assigned tags, but all tag labels were left blank. As a result, no comparison was possible for that participant. The other participant not included in the analysis had only applied one tag on each day, which meant no score was possible for the Spearman's rank correlation coefficient for the tag frequency and tag infrequency orderings across the session. The remaining 13 participants' data covers Days 3-5 for all system orderings. If we had included Day 2 results, an additional three participants' results would have been eliminated. The first would be eliminated because the two system subtopics on Day 2 had the same term frequency, so the Spearman's rank correlation coefficient was not calculable. The second participant only had results from one subtopic returned on Day 2, so no Spearman's rank correlation coefficients could be calculated for any of the system orderings. For the third participant, the ideal ranking approach resulted in a tie for a single value in the tag frequency and infrequency orderings, preventing calculation of the Spearman's rank correlation coefficient.

7.3.2 Correlation Results for Daybreak User Study

Table 7.6 includes the results of the Spearman's rank correlation coefficients for the evaluated twelve sessions for each of the ten system orderings. We also performed paired two-tail t-tests to determine the statistical significance of these results. We first compared rarity results to every other ranking, since that is the current Daybreak system approach. After applying Bonferroni's correction, only two orderings (subtopic recency and tag frequency) performed statistically significantly better than rarity at $p \leq 0.05$. We also compared the subtopic recency results to every other system ordering, and after applying Bonferroni's correction to those results, rarity, tag infrequency, and subtopic maturity were statistically significantly worse than subtopic recency.

While subtopic recency and tag frequency were the top two system orderings for the sort

order study results, subtopic maturity and tag infrequency scored the lowest. Additional research would be needed to determine whether the ranks of system orderings are a result of the assumptions applied in the generation of ground truth data for the Daybreak study, or if there are other factors—such as personal preferences or differences in applications of tags in timed vs. untimed tasks—leading to such a difference.

The system ordering that best aligned with the Daybreak participants' ideal system ordering was based on the recency of the creation of the tag label. The system that aligned least well with the ideal order was subtopic maturity (tags that were created on an earlier day). Rarity (ordered from the cluster containing the fewest documents to the most documents) turned out to be the third least effective approach for organizing documents for the study participants. In a full evaluation, we could compare a broader range of system orderings than the ten selected for this study.

Table 7.7 shows a comparison between the author's session and the Daybreak user study results from Chapter 6. The rarity-based approach for subtopic organization ranked low for both the author's session and the Daybreak user study results. This appears to be consistent with the results for RQ5.4 (Section 5.1.4), as discussed in Chapter 6, where participants expressed negative views regarding leveraging rarity as a proxy for subtopic importance.

As Table 7.7 also shows, the best system ordering for the average Spearman's rank correlation coefficent differs between the two: for the author's session, tag frequency outperformed the alternatives. However, the highest average Spearman's rank correlation coefficient for the user study results was the subtopic recency system ordering. This difference can likely be explained by the way we crafted the ideal ranks based on the user study results; our approach for interpreting rankings gave preferential treatment to recently used tag labels, and the correlation results reflect that bias. As previously noted, this evaluation is intended as an example of a comparison, and further research and input from participants would be needed to characterize the ideal rank.

Topic	**Red Sox**		**Crypto**			**Health**			**Space**			**Weather**			**p**	
system ordering	BBL-887	BBL-927	FIN-326	FIN-455	FIN-499	HLT-409	HLT-555	HLT-913	SPC-259	SPC-471	SPC-688	WEA-093	WEA-842	Average	Rarity	Best Ordering
subtopic recency	**+0.62**	**+0.79**	+0.45	-0.70	-0.23	**+0.69**	**+0.46**	**+0.69**	**+0.72**	**+0.34**	**+0.44**	**+0.68**	**+0.42**	**+0.41**	***0.00***	—
tag frequency	-0.11	+0.37	**+0.78**	+0.61	+0.42	-0.10	+0.33	+0.18	-0.03	+0.29	+0.43	+0.23	+0.07	+0.27	***0.00***	0.30
document recency	+0.26	+0.16	+0.26	+0.60	+0.05	+0.30	+0.24	-0.16	+0.19	+0.10	-0.17	+0.28	+0.17	+0.18	*0.01*	0.08
document maturity	+0.15	+0.27	+0.21	**+0.73**	+0.15	+0.03	+0.22	-0.27	-0.26	+0.31	-0.32	+0.11	-0.11	+0.09	0.11	*0.03*
topictf	+0.32	-0.02	+0.23	+0.61	**+0.66**	+0.09	+0.24	-0.42	-0.02	-0.02	-0.13	-0.18	-0.23	+0.09	0.15	*0.04*
prevalence	+0.18	+0.11	+0.02	+0.62	-0.02	+0.17	+0.02	-0.34	-0.02	+0.32	-0.19	+0.25	-0.08	+0.08	0.11	*0.02*
alphabetical	+0.25	+0.01	-0.15	+0.57	+0.51	-0.06	-0.12	+0.04	-0.15	-0.19	-0.70	+0.39	+0.39	+0.06	0.25	*0.03*
rarity	-0.17	-0.11	-0.02	-0.62	+0.02	-0.17	-0.02	+0.34	+0.02	-0.32	+0.19	-0.25	+0.08	-0.08	—	***0.00***
tag infrequency	+0.11	-0.37	-0.78	-0.61	-0.42	+0.10	-0.33	-0.18	+0.03	-0.29	-0.43	-0.23	-0.07	-0.27	0.07	***0.00***
subtopic maturity	-0.62	-0.79	-0.45	+0.70	+0.23	-0.69	-0.46	-0.69	-0.72	-0.34	-0.44	-0.68	-0.42	-0.41	*0.02*	***0.00***

Table 7.6: Spearman's rank correlation coefficients for each of the Daybreak sessions containing complete results across Days 3-5, sorted with the highest average score at the top. Statistical significance was assessed by comparing each system ordering to rarity as well as to subtopic recency. Tests resulting in statistical significance at the 0.05 level are indicated in italics. Tests that remained statistically significant after applying a Bonferroni correction are in bold and italics.

System Ordering	Author	User Study	User Study Rank
tag frequency	**+0.48**	+0.27	2
prevalence	+0.34	+0.08	6
document recency	+0.26	+0.18	3
document maturity	+0.25	+0.09	4
topic tf	+0.24	+0.09	5
subtopic recency	0.00	**+0.41**	1
subtopic maturity	0.00	-0.41	10
alphabetical	-0.04	+0.06	7
rarity	-0.34	-0.08	8
tag infrequency	-0.48	-0.27	9

Table 7.7: Comparison of the Spearman's rank correlation coefficients for the author's session with the average scores for the system orderings from the Daybreak study participants, from Days 3-5. The system orderings are sorted from highest to lowest based on the scores from the author's sessions. The "User Study Rank" column indicates the order this system ordering received for the user study results, with 1 representing the one with the highest score. The highest scoring system orderings for the author's session and user study results are also represented in bold.

Another difference between the author's session and the Daybreak user study sessions were in the time limits. The author's session was not time limited, whereas there was an eight-minute time limit for each day in the Daybreak user study sessions. Some of the difference in the system ordering preferences may be attributed to the fact that the lack of a time limit allowed for many more tags and tag labels being applied across the author's five-day session, and greater reuse of tag labels across the session than were observed in the user study sessions.

7.4 Recommended Approach for Future Study

For implementation of an evaluation, we would apply a similar approach for data collection to the one used in the Daybreak study (Chapters 5 and 6). We would include more focus on getting participant rankings of subtopics by importance for each day, with the goal of producing a reusable dataset for evaluation. Here we detail the steps that we would apply to generate more robust system evaluation information.

To annotate the documents with subtopic labels, we would have the participants perform change detection sessions with an updated version of the Daybreak system over a series of days, as done previously, to generate a list of tag labels represented in the collection for that day—both through creation of new tag labels and reuse of tag labels created on prior days. The Daybreak study included five days, similar to a work week; we would continue to use five days as the lower bound for the evaluation to ensure that the participant is comfortable with the system and the tagging process, without creating an overwhelming experience for the participant. The maximum time for each day was set at eight minutes for the study in Chapter 6; we would increase this to 10 minutes to provide participants additional flexibility to review and tag documents. At the end of each day, the participant will be asked to indicate which subtopic(s) were most important that day.

For the Daybreak user study described in Chapters 5 and 6, after Day 1 (the cold start day) we arranged the subtopic clusters by rarity, with the least prevalent cluster shown first. Even though we made some broad assumptions in our construction of ground truth, after combining the low scores of rarity ranking with the comments provided by participants during the interview, we would not leverage rarity-based sort as the sole option for subtopic arrangement. Instead, we would include a variety of subtopic cluster arrangement options, to include the higher-scoring orderings from the sample evaluation (e.g., tag frequency and subtopic recency, the highest-scoring system orderings from the author's session and user study results, respectively).

Once we have collected the participants' data, we will need to ascertain their ideal ranking for the subtopics. To do so, we will need to refresh their memory of the individual days. After all five days, we will replay sessions to the participant at a faster speed, calibrated based on the preference of the individual participant; the goal of this step is to remind them of what documents they saw and what actions they took, albeit at a compressed rate. After each day is replayed, we will ask them to review the complete list of all subtopics applied to

documents that day, to bin the documents into categories based on high, medium, and low importance. They will also have the opportunity to note any subtopics that are irrelevant (e.g., tag label applied only to bin and hide uninteresting documents). Once these subtopics are binned, we will have the participant sort at least the high importance subtopics for the day, to get rankings for those items. The participant can have the option to rank medium and low importance subtopics; any items not sorted will be treated as having a tied rank.

As an alternative to ranking subtopics from 1 to n, the participant could perform pairwise comparison between the values of the limited list of subtopics within the bin. The idea of pairwise comparison was first introduced experimentally by Thurstone in 1927, which he called the "law of comparative judgment" [270]. Pairwise comparison could enable the participant to get to an ordering for the set of labels using an approach that can be less of a cognitive burden on the participant, by only requiring a comparison of two items at a time rather than trying to sort an entire list. Similar approaches have been applied for generating relevance judgments for information retrieval evaluations [48]. That said, it is important to keep the number of items in a pairwise comparison fairly low, to prevent the participant from having to make a large number of comparisons, though some researchers have shown the possibility of making judgments based on smaller numbers of comparisons [147]. For tag labels identified as being in uninteresting, rather than requiring the participant to rank irrelevance in some way, we would automatically tie all of these subtopic labels for the lowest rank.

Another improvement would be to include the use of a head-weighted variant of Spearman's rank correlation coefficient, or another correlation measure. This would reduce the impact to the score of lower-ranked items on the score. Similar approaches have been applied to Kendall's tau, an alternative correlation measure. This includes research by Yilmaz, et al. and Gao and Oard [91, 302].

Through whichever ranking method is applied, we end up with a ranked list of the

participant's ideal subtopic importance order for each day. Once the results are generated, they should be compared by using an error rate adjustment. While we selected Bonferroni's correction for our example analysis, this represents a fairly conservative approach. It is designed to reduce false positives, but may result in false negatives. Other alternatives could be applied in a future evaluation to avoid some of the limitations of Bonferroni's corrections, such as the Holm or Hochberg correction [180], or the False Discovery Rate correction [286].

7.5 Summary

This chapter described an evaluation approach for comparing a system subtopic ranking to a user's ideal importance ranking. To calculate a subtopic importance score, we applied Spearman's rank correlation coefficient to determine the relative effectiveness of various approaches for sorting subtopics against a user's preferred ordering. We applied this approach to a session by the author as well as to the Daybreak user study results, to show the comparison process and some sample results. Future work in this space would include hosting a new set of Daybreak sessions to gather data and complete subtopic rankings that could be used in a formal evaluation based on the approaches identified in this study. In Chapter 8 we conclude the dissertation by revisiting the set of studies performed, addressing contributions, and discussing limitations and future work that would build on the change detection studies.

Chapter 8: Conclusion and Future Directions

Throughout this dissertation we have explored the concept of change detection, and how a system might better enable a user to more quickly update their mental model related to the topic they follow over time. Our findings can aid in improving systems for users who follow a specific topic over time. The intended audience for this individual information seeking task includes users who have sufficient interest (whether personally- or professionally-motivated) that they typically review large quantities of information to understand developments related to their topic of interest. The goal is to improve systems that are used for change detection so that they more effectively support a user's day-to-day activities for expertise development on a topic, through focused and organized display of information.

We addressed the change detection problem from a number of angles. First, we developed the GPA Change Detection Theory, which posits that a human-driven, machine-driven, or hybrid system could better address users' change detection needs by grouping documents by theme, organizing the groups into piles, then arranging the piles in some order based on the preferences of the user. Next, we conducted a survey of social media users to start to understand the prevalence of change detection relative to other tasks performed in social media sites. We found that change detection is a recognized and common task on social media sites. From there, we built upon the findings from the survey by developing a prototype change detection system—the Daybreak system—which focused on reviewing news articles about a topic of interest. We ran a user study with the Daybreak system to

determine whether the system could aid participants in completing a change detection task. Finally, we devised an evaluation approach to enable comparison between system orderings for change detection—in particular, focusing on arranging subtopic clusters.

8.1 Review of Research

In this dissertation, we have addressed the topic of change detection as an individual information seeking task—for a user who has followed a topic over time, how can we provide information in a way that enables them to update to their mental model quickly? To address this research area, we applied mixed methods approaches to perform research into the following themes:

- **Theory Development:** We created a theoretical framework for change detection, which we call the Group-Pile-Arrange (GPA) Change Detection Theory.
- **Survey:** We conducted a survey to study respondents' sort order preferences within social media, to understand the prevalence of change detection use cases and compare what sort orders respondents preferred for each use case.
- **Daybreak System Development:** We designed and built Daybreak, a prototype system that addresses the change detection needs of users.
- **Daybreak User Study:** We planned and conducted a user study to test the Daybreak system on real-world change detection scenarios; applied the Framework Method to analyze and interpret results.
- **Evaluation Approach:** We devised an automated method for comparing subtopic arrangements produced by change detection systems.

8.2 Findings for Research Questions

To frame our research, we generated a series of research questions related to change detection tasks. In this section we review the questions and provide our findings related to each of these research questions.

Chapter 4: Sort Order Survey

RQ4.1: How prevalent is change detection? (introduced in Section 4.3.2, addressed in Section 4.4.1) Based on the sort order study, we found that all of the studied use cases (change detection, experiential, search, and browsing) were prevalent use cases within social media among the respondents. Change detection was the second most prominent use case for these respondents, with 66% indicating that they perform change detection tasks in social media.

RQ4.2: Would users accept clustering as an approach for organizing posts? (introduced in Section 4.3.2, addressed in Section 4.5.1) Our survey results showed that respondents are interested in having results clustered for change detection use cases; we did not specifically define what was meant by clustering in the survey. We leveraged respondents' positive responses toward clustering for change detection when we designed the Daybreak user interface; the interpretation we applied for clustering was based on users' tag labels.

RQ4.3: How do users prefer to have results sorted for a change detection? (introduced in Section 4.3.2, addressed in Section 4.5.2) For change detection in particular, we found that respondents preferred to see social media organized in chronological order. This is in contrast with other use cases, such as search, where respondents prefer to have results organized by relevance.

RQ4.4: How many posts do respondents feel they need to see when performing a change detection task? (introduced in Section 4.3.2, addressed in Section 4.5.4) Respon-

dents described a general fear of missing out on useful information. For change detection tasks, respondents were particularly interested in seeing all posts related to certain accounts that they follow on social media. After that, they were interested in the "best" posts (which was not specifically defined) related to the topic they follow. The top responses were split between wanting to see all posts related to the topic, to 5-10 documents per theme (subtopic) related to the topic.

Chapters 5 and 6: Daybreak User Study

RQ5.1: Does tagging and tag label generation aid users in representing their mental model of a topic? (introduced in Section 5.1.4, addressed in Section 6.3.1) We found that participants heavily made use of tagging for tracking themes and changes related to the topics in the study. Even the sole participant who intentionally turned off clustering for three of the five days continued tagging, and made use of the tags for knowledge management purposes when drafting the final story outline. We also note that the sole user who tagged documents without applying tag labels was successful in completing the storytelling task.

Questions related to the "Group" concept:

RQ5.2: Does organizing search results by subtopic clusters aid users in performing change detection tasks? (introduced in Section 5.1.4, addressed in Section 6.3.2) Participants overwhelmingly leveraged subtopic clusters when reviewing results, according to system logs. A few users indicated that the current Daybreak clustering interface (which required the participant to click on a subtopic heading to view documents within the cluster) was too burdensome; alternative implementations could be explored to maintain the cluster labels but make it easier for the participant to view the documents within the cluster.

RQ5.2a: What information retrieval approaches would be effective for transforming a user's tags into clusters of relevant documents? (introduced in Section 5.1.4, addressed in Section 6.3.3) The Daybreak system leveraged a bag-of-words technique for assigning new documents into subtopic clusters. The technique was based on the subtopic label as well as top terms from the documents that had been tagged with that label. While the bag-of-words approach provided reasonable results, other techniques—to include machine learning approaches—might provide better matches between documents and subtopics.

Question related to the "Pile" concept:

RQ5.3: Does organizing search results within subtopic clusters in some sort order aid users in performing change detection tasks? (introduced in Section 5.1.4, addressed in Section 6.3.4) Based on the sort order study, we expected that participants would primarily leverage reverse chronological sort ordering. However, we found the results to be much more mixed. This may be due in part to imperfect matching between participants and topics; some participants indicated that they prefer relevance sorting when they are learning a topic (which they said was the case for their assigned topic), but in general would prefer chronological ordering if it was a topic they knew well and were following over time. It appears that participants prefer a mix of sort options, so they can adjust their sort based on the specific task at the moment. Additional research could dive deeper into the potential link between specific sort orders (to include chronological sort) for change detection tasks, when focused on a topic for which the participant has expertise and interest.

Questions related to the "Arrange" concept:

RQ5.4: Does arranging subtopic clusters in some order aid users in performing change detection tasks? (introduced in Section 5.1.4, addressed in Section 6.3.5) Based on participant feedback, rarity was not viewed as an appropriate proxy for subtopic im-

portance; several participants noted that prevalence (the reverse of rarity) was more likely to be useful. That said, the participants did support the idea of arranging subtopic clusters in order of importance. The approaches they described for defining importance ranged from manually pinning certain topics to having a machine learning approach that predicted subtopic alignment with participants' interests.

RQ5.4a: How should the system handle documents that do not fit in any existing subtopic cluster? (introduced in Section 5.1.4, addressed in Section 6.3.6) For the prototype, the "Other" category contained documents that were not aligned with other clusters. While most participants agreed that this cluster was important for discovering new themes, in our time-constrained user study, many participants did not have sufficient time to use the cluster; in a few cases, the participants did not even see the "other" cluster. We recommend placing this cluster in a prominent location outside of the personalized list, to ensure that participants see it and understand that it may contain new developments that do not relate to items they have previously tagged.

Question regarding the GPA Change Detection Theory:

RQ5.5: Does the system help users develop and externalize mental models? (introduced in Section 5.1.4, addressed in Section 6.3.7) We found that the Daybreak system—which contained capabilities to group (cluster), pile (sort order), and arrange (subtopic importance, by rarity) document results—did enable participants to complete change detection tasks, as evidenced by participants' ability to generate a story outline detailing what they learned in their session. Over the course of the five-day session, participants became more effective at identifying new developments, and became more confident that they were seeing enough documents to understand what was happening on their topic.

8.3 Contributions

This research into the topic of change detection sought to build a deeper understanding of user needs. We were able to build that understanding of change detection, to include users' interests and ways that a system could assist the user in accomplishing their goals. Some of the specific contributions provided by this dissertation research include:

- **Connections:** We performed cross-disciplinary research that combined a number of research methods for our exploration of change detection. By looking at change detection from multiple angles, we were able to see a rich view of this use case. We gained insights into users' needs, organization preferences while performing change detection tasks, as well as an understanding of system components that aid users for change detection.

  - By studying change detection as an arc of a single research program, we were able to start with a foundational theory that we devised, the GPA Change Detection Theory, then perform two studies that tested the theory and explored users' interests and preferences. The studies revealed that the GPA Change Detection Theory served as a reasonable starting point for organizing information in support of change detection tasks; study participants were successful in completing a change detection task using the Daybreak system, which was organized based on the theory. The results from these studies also informed the initial evaluation design.

  - This research included studies about change detection practices in social media (sort order survey) and news articles (Daybreak user study). Based on our findings, the GPA Change Detection Theory appears to apply to both content types, though the social media survey respondents expressed a stronger sort or-

der preference for chronological ordering than the Daybreak study participants. This provides an initial understanding of change detection for both sources. It can serve as a foundation on which further research can be performed to identify similarities and differences in change detection needs across these and other sources of information.

  - During the process of generating the theory and performing our studies, we drew connections between concepts from many fields that have relevance to change detection. This included information retrieval concepts such as reranking and evaluation approaches, theory development from information studies, system and user study design from human-computer interaction, coherence relations from reading comprehension research, and a range of other concepts drawn from cognitive psychology, behavioral economics, and more. This is a benefit of information studies research, which is a field that encourages cross-disciplinary research, as well as integration of ideas and concepts from multiple fields. As a result, we developed a theory that is at the intersection of multiple disciplines.

- **Theoretical:** We developed the GPA Change Detection Theory primarily as a foundation for our research. This served as an organizing approach that addressed a gap in research. It also served as the organizing principle for the Daybreak system and user study. This new method for thinking about change detection could also be used by other researchers, who can test and build upon these ideas to identify additional approaches to enable users to get updates on topics of interest.

  - The GPA Change Detection Theory describes an approach for organizing results on a topic of interest in a way that enables a user to update their mental model more quickly: grouping the documents by theme, piling them in some

order, and arranging the piles into a meaningful order for the user. We designed this approach to be sufficiently general that it could be applied to scenarios supported by people, computers, or some combination of the two.

  - We drew upon real-world use cases as examples of possible applications of the GPA change detection theory.

- **Survey Administration:** Our survey on users' sort order preferences in social media aided in filling gaps related to change detection and how it relates to usage of social media. The survey served as a method for comparing change detection with other social media use cases: experiential (following a live event as it happens), browsing, and search. This enabled us to understand how respondents' preferences for sort orders differ depending upon their task.

  - The survey provided some initial insights into the prevalence of change detection; though we do not generalize beyond the survey respondents, their responses provide a foundation for further understanding of how broad the change detection use case might be.

  - The sort order survey revealed insights about clustering (group) and sorting (pile) preferences of users that could be operationalized by a system focused on the change detection use case.

  - Our research contrasted change detection with other use cases in social media, which could provide social media system developers with insights into identifying distinct user behaviors and organizing posts to enable users to accomplish their intended specific task.

- **User Study with a Prototype System:** To delve deeper into users' change detection activities, we designed and built the Daybreak system, which was designed to test

the components of the GPA Change Detection Theory in practice. The system was leveraged in a user study for simulated change detection task. Participants were assigned to one of five topics (Red Sox, cryptocurrency, global health, space, or extreme weather), and reviewed news articles during five simulated days.

  - The Daybreak prototype system enables users to review documents and add tags and tag labels for a document. This system also includes change detection functionality for displaying search results: grouping (subtopic clusters), piling (sorted documents, in either reverse chronological or relevance ranked order), and arranging the groups (organizing subtopic clusters by rarity—least populated cluser first).
  - We designed and conducted a user study through which participants performed a simulated five-day session in the Daybreak system. After completing their review of relevant documents, they drafted an outline for a blog article summarizing what they learned. While completing their change detection tasks, the participants heavily made use of clustering. Their sort order use was more varied, with a weak preference for chronological ordering. Users did not view rarity as defined in the study to be a helpful proxy for subtopic importance.

- **Evaluation Design:** In order to be able to compare system results in the future, we designed an automated evaluation approach focused on comparing subtopic cluster orders. For this evaluation, we applied a correlation measure. We ran a sample evaluation for the change detection use case based on a manual run as well as the Daybreak user study results.

  - Our automated evaluation process leverages Spearman's correlation coefficient to compare system arrangements for subtopic clusters. We demonstrated how

to compare an ideal list of subtopic clusters (based on user input) with a range of system orderings.

  - In the sample evaluation based on the author's session and the results from the Daybreak user study, we found that a head-weighted version of this measure (e.g., one that provides higher weight to results near the top of the list) might be preferable, to better handle cases of irrelevant results.

- **Code:** To support this research, we created a number of artifacts that can serve as exemplars for other researchers' research related to change detection. The code is available on GitHub.[1] It is intended as research code, not production-ready code; some of the components required manual setup and tuning during execution.

  - **Daybreak:** The main code artifact was the prototype Daybreak system.[2] This includes the HTML and Javascript-based user interface, which includes the tagging, clustering, and sorting components. The Daybreak back-end is Python-based, which includes code to retrieve results for clusters with the Indri search engine, and generate the custom subtopic clusters based on users' tag labels.

  - **Evaluation:** Another code artifact is the Python code used for the evaluation process.[3] This code leverages Spearman's rank correlation coefficient to compare the subtopic cluster orders produced by various systems.

8.4 Limitations

As with any research initiative, choices were made during the process that resulted in limitations in the conclusions that can be drawn. These range from intentional decisions

[1] The code artifacts are available at the following site: https://github.com/change-detection/
[2] The Daybreak system code is available at https://github.com/change-detection/daybreak/
[3] The evaluation code is available at https://github.com/change-detection/cd_eval/

based on time and resource constraints to unanticipated issues that surfaced during the study. We note these limitations here, along with their implications.

A limitation related to the GPA Change Detection Theory presented in Chapter 3 is the focus on lists as an organization approach for change detection systems. While this data structure is useful for the constructs in this dissertation, it does not take into account the broader range of data structures that could enable a user to understand what has changed on their topic of interest.

In the sort order study described in Chapter 4, we received fewer than 200 responses to the survey. This is likely due to the complexity of the survey, which contained a combination of demographic questions, general questions about all use cases studied, as well as detailed questions about specific use cases. While we did receive valuable information that shaped our understanding of change detection, the limited number of respondents may mean that we received feedback principally from people with preexisting strong feelings about sort orders in social media; additional research may be needed to understand the prevalence of change detection more broadly.

Participants for both the sort order study from Chapter 4 and the Daybreak user study described in Chapters 5 and 6 were limited to US-based adults aged 18 and older. For both studies, this means we are unable to generalize the findings to other user demographics without conducting additional research. Additionally, the advertising strategy for both studies included targeted messages to potential respondents in academic fields such as information studies, library science, and political science. There may be selection bias related to the advertising approach that remained uncharacterized.

A significant limitation was in the pivot in the type of data studied between the sort order study from Chapter 4 and the Daybreak user study described in Chapters 5 and 6. The survey on sort order preferences focused on social media posts, whereas the Daybreak user study looked at news articles. This change was in part to provide more signal

in the search functions; a longer document such as a news article tended to contain more signal indicating that the document was relevant to the participant's assigned topic. Selecting relevant social media would have been more difficult both in determining topicality as well as assembling sufficient posts to represent complex changes to the topic. While this decision was intentional, it raises the possibility that social media users have different expectations in change detection scenarios than users reviewing news articles. Additional research would be needed to determine the extent of these differences. We do note that many respondents to the sort order survey indicated that they often use social media to find news, which led to our expectation that the results would at least be consistent.

The Daybreak user study described in Chapters 5 and 6 had two important limitations. First, because we narrowed the topic space from the interests of all participants to five topics that were somewhat related to participants' interests, we ended up with an imperfect assignment of participants to topics. This may have resulted in observed behaviors that would have been different with a better topic assignment. For instance, several participants indicated during the unstructured interview that they would have preferred reverse chronological sorting over relevance ranking if it was a topic that they were more familiar with, or that they had been following regularly. Since our intent was to create those conditions within the study, in some cases we may have ended up with results more representative of novice users who are still learning a topic.

The second limitation related to the Daybreak study was in the document selection for the five topics. The Red Sox topic was the only one with a clear story: one specific team, the Boston Red Sox, was traveling to play games against specific teams. This included game recaps and analysis related to those specific activities. The other selected topics lacked such clear stories. There were some common threads that played out across the five days of stories, such as the release of new cryptocurrencies. However, the Red Sox document set was the only one confirmed to have a consistent theme across the entire five days. A

future study could focus less on numbers of documents and more on identifying documents around some tangible event or activity; this would enable the participant to focus on a smaller number of themes for the study, rather than trying to understand a broad range of disconnected activity all loosely connected to the topic.

The main limitation of the evaluation design described in Chapter 7 was a lack of data collected as ground truth information. In the Daybreak user study, we only gathered the top subtopic per day from the participants. We would need additional information—more complete ideal subtopic orderings—to run a complete evaluation using this approach. While we were able to apply some assumptions and demonstrate an evaluation approach as an example, we were not able to use our results to draw conclusions about system orderings.

8.5 Future Work

Through this research we identified a variety of opportunities to build on the research with additional studies, to further understand users' needs and preferences when performing change detection tasks. The results presented in this dissertation, as well as future research, could aid in providing users with systems that are more optimized for understanding changes that are related to a topic of interest. In this section we propose a series of future work ranging from improvements to the Daybreak prototype to exploration of additional relevant concepts, to include understanding users' thought processes as they are reviewing documents when getting updates. In this section, we detail some of the potential future studies that we view as logical next steps for furthering understanding of change detection and how to meet user needs for this use case.

8.5.1 Change Detection Research Areas

In addition to system design improvements, there is an opportunity to build on a number of aspects of change detection that are not specifically tied to a user interface. One such study could focus on implementation of the evaluation design described in Chapter 7. Future research could focus on building this out to a formal study that would result in a reusable evaluation dataset and a formal comparison of system design options. For instance, such an evaluation could be proposed for a shared-task evaluation venue such as the annual Text REtrieval Conference (TREC).

Future research could expand the GPA Change Detection Theory to represent a broad range of data structures. This could include supporting data structures such as trees, results visualized as clusters, maps, and other approaches for arranging information in a meaningful way to the user.

Subtopics play a critical role within the GPA Change Detection Theory. Through this research, we have observed that a user's interest in a specific subtopic is not static; users identified different themes as being more important on different days. A follow-on study could look at stability of subtopic importance: are there certain subtopics that are stickier than others—ones that the user would always want to see displayed first? How dynamic are users' interests in specific subtopics? How does this affect approaches for arranging subtopic clusters?

Given our identified limitation in our pivot from social media to news articles, a future study could formally test several questions: Are there differences between the change detection needs of social media users vs. readers of news articles? To what extent might a specialized system be needed to address each one? Future research could also evaluate the applicability of change detection concepts to other areas where people are looking for updates, such as in their email inboxes. For each of these, more detailed studies could be

performed to look at specific disciplines, to determine whether change detection activities are different depending on the topic the user follows.

During the Daybreak user study, we identified a number of user behaviors that would benefit from further study. For instance, when reviewing clustered results, participants spent their time in one of three ways: browsing subtopic clusters, browsing document titles and other metadata, and reading documents. Future research could build on this, to gain a deeper understanding about these behaviors—for instance, why some participants spend more time browsing titles than viewing documents. How do they decide whether to view a certain cluster, or open a certain document? What assumptions affect these decisions? How do judgments based on a document's title impact their ability to discover relevant changes related to their topic?

Another approach for introducing external information would be to study the extent to which it is useful to the user to bring in historical data as background information. Such research could focus on defining points within the user interface where such supplemental information would be beneficial. Using the analogy of a sports broadcast booth, how might a change detection user benefit from having information at their fingertips that would provide historical context about what is happening on their topic. The goal would be to aid the user in addressing additional questions, such as why certain changes have taken place.

One idea that may merit future study relates to the temporal nature of change detection, and the extent to which chronological ordering (as opposed to other orderings) is tied into the change detection process: for instance, if a user starts with the most recent information about the topic, then traces a path back to where the story ended the last time they checked, forming a sort of temporal continuum. Future research could look at people's desire for completeness, and the extent to which there is a temporal concept related to documents being "in order" vs. "out of order." This could also address the potential recall-oriented aspect of change detection; in both the social media sort order survey and the Daybreak

user study, users expressed a desire not to miss something (a sort of fear of missing out). Is there a notable preference for chronological ordering, compared to ordering options that are time-weighted but not directly chronological? How strong is this preference? What other approaches are beneficial (e.g., would users accept relevance with recency weighting)? How do users identify and characterize gaps, such as missing information related to their topic?

We may wish to avoid cases where a system like Daybreak fits information too closely to the interests of the user, which could unintentionally exclude information and introduce filter bubbles. Future research could look at ways to introduce serendipity into the process, to include information peripheral to the user's interests that would not be directly retrieved based on the query that approximates the user's interest. Adjacent information could be introduced in a similar way to the "other" category; while the "other" category is about capturing on-topic documents that don't align with subtopics, another category could display documents that are conceptually similar but that do not meet the query of the user. A change detection-focused filter bubble study could focus on providing information adjacent to the topic, to see whether adding such information contributes to understanding, adds to serendipitous discovery, or serves as a distraction.

Another assumption made in the Daybreak study was that change detection is an individual process, and that users would benefit from externalizing their personal mental model for use in the change detection system. A follow-on study could compare individual and collaborative approaches for getting updates on a topic (for instance, if a study or work team were looking at similar angles to a topic). Such a study could also draw comparisons between individual tag labeling compared to team tag labels, building on research into folksonomies and related topics.

8.5.2 Recommended Daybreak System Improvements

Another area for future work relates to improvements to the Daybreak system. Based on our findings, we would update the Daybreak to make improvements and address lessons learned. Proposed updates to the Daybreak system include:

1. **Search customization and refinement:** Rather than being given a preassigned topic, we could allow the user run a search and produce a result set customized to their specific interests; while this might add complexity in comparing users' actions, it might better mirror users' real-world change detection behaviors. Additionally, we could enable the user to modify or adjust their query over time, to be able to study the extent to which a change detection user maintains vs. adjusts their query over time, as their understanding of the topic evolves.

2. **Knowledge management:** Based on participants' feedback, we could add more knowledge management features. These could include capabilities that also aid in our understanding of a user's specific interests. For example, a highlighting capability could provide insights into what specific portions of the document are most salient to the user.

3. **Subtopic organization:** We would provide the user with simpler, more straightforward functionality for tagging documents, including for adding hierarchical subtopic labels. Additionally, we would include functionality to allow the user to refine their tags—to include modifying tags, splitting or combining tags—and to make other changes to their tag ontology. Not only were these commonly requested by users of the Daybreak study, but they will provide additional opportunities to capture aspects of the user's evolving mental model; this would enable further understanding of how a user's knowledge changes over time, and how it is reflected back to the system.

Future research could also explore implied hierarchies in tags; for instance, could the system infer an ontology based upon the user's set of tag labels?

4. **Subtopic summarization:** While summaries may play a role in an interface, the information need of the user we have focused on is not likely to be satisfied with a brief summary or overview on its own. Large Language Models (LLMs) such as ChatGPT have grown in popularity during the time this dissertation was being written. These capabilities enable interactive chats with users, and enable a range of natural language activities such as summarization of content [300]. We would look at future opportunities to apply LLMs as aids to the change detection process, without taking away from the user's ability to view a variety of documents. We see two plausible uses for LLMs in the context of change detection:

    - **Update Summary:** The first option would be to leverage LLMs or other technologies to summarize the new documents on each subtopic, to give the user a sense for what has changed on that subtopic [92]. We have made an assumption related to change detection users that they are interested in their topic to such an extent that we believe they would not stop at the summary, but would read individual documents as well to bolster their understanding of the changes to their topic. The ability to add LLM-generated summaries customized to their mental models would allow us to test this assumption, by seeing if they read the summaries and move on or read summaries as well as documents.
    - **Attributed Summary:** We could also leverage LLMs to aid in identifying specific documents or parts of documents that are relevant to a subtopic. This could tie in with current research into explainable AI, which includes concepts such as providing sourcing for statements [144].

5. **Cold start:** The first day in the current version of the system is the first time when the system learns about the user's specific interests. Future system functionality could focus on getting specific information about users' interests earlier—for instance, identifying prior to Day 1 about what subtopics they care about. Two options include the following:

    - **Stereotypes:** Not all stereotypes are negative. The concept of stereotypes is used in recommender systems to aid in cold start situations. This approach involves using a representation of typical interests for a topic, which later will adjust based on the user's actions [5, 6]. For instance, people who focus on baseball might break down their topic in a similar way, and might typically be interested in—for example—recent game news over special interest stories about a specific player.
    - **User input:** The user could provide a small number of terms that come to mind related to the topic, which could be included within the clustering approach starting on Day 1. Pairwise comparisons could help to understand which aspects users prefer within a topic [220]. Also, if the user has difficulty in specifying aspects of topic that are of interest, this could help to indicate in advance that the user may not be a good fit for that topic.

6. **Subtopic organization:** System updates could focus on adding more approaches for sorting the clusters in alignment with the user's preferences. One of the strong signals from the Daybreak study was that rarity was not a suitable proxy for subtopic importance. An updated system could include a combination of manual—e.g., pinning specific subtopics—and system-driven approaches—e.g., subtopic recency, subtopic frequency, or Machine Learning (ML) options for predicting which subtopics would be of greatest interest to the user. For example, an ML-driven approach could include

supervised topic modeling approaches to map documents to relevant topics [178]. Word or vector embeddings could also be an approach for matching documents to existing subtopics [4, 15].

7. **Improved treatment of documents on new themes:** In the version of the Daybreak system used in the study, no attempt was made to organize the documents that did not align with the user's tags; these were simply put into separate cluster labeled "other." A future version of the system could apply ML such as topic modeling to organize by theme the documents that aren't related to the subtopics. Additionally, a number of users did not notice the "other" category, because it was at the end of the subtopic cluster list, and they never scrolled all the way to the end of the list during their session. Future research could explore a variety of approaches for displaying the "other" cluster, to make it more prominent for the user; the interface approach could also add indicators for the user to show the difference between the "other" cluster and the ones generated based on their tag labels.

8.6 Closing Thoughts

This dissertation research focused on understanding and meeting the needs of users engaged in change detection tasks—trying to find updates on topics of interest. We performed mixed methods research to understand users' needs and how to improve the processes through which they find updates. Our research included creating the GPA Change Detection Theory, conducting a study on users' sort order preferences in social media, designing and building the Daybreak change detection system, hosting a user study with the Daybreak system, and devising an evaluation approach for comparing system orderings. Based upon our findings, GPA theory is potentially useful for looking at this problem; additional research into change detection could aid in devising systems that reduce the fric-

tion in finding updates on topics of interest, in order to better meet users' change detection needs.

Circling back to the initial problem that change detection attempts to address, we have a user who wants to know what is happening on a specific topic of interest. This user has a deep enough interest that they are willing to invest time on a regular basis to deepen their understanding and catch up on the topic. This dissertation research was intended both to understand the user's interest and needs, as well as to identify opportunities to create and enhance systems to help the user. Through this research, we have found that there are opportunities to provide specific systems or system improvements that would help them get updates on topics that they follow. The group-pile-arrange approach described in the GPA Change Detection Theory appears to provide a beneficial framework for organizing information to make it easier for people to identify and track what information has changed in relation to their topic.

In the end, we return to the foundational concept driving change detection: we have an individual user with an interest in a topic, and some understanding of that topic as represented by their mental model. As the wizard Dallben said to the young main character Taran in Lloyd Alexander's novel *The Book of Three*, "We learn more by looking for the answer to a question and not finding it than we do from learning the answer itself." Reflecting a personal learning journey that is similarly experienced by a user engaged in change detection, Dallben continues, "If you grow up with any kind of sense... you will very likely reach your own conclusions. They will probably be wrong," he quipped, referring to Taran's youthful interpretations of the world. "However, since they will be yours, you will feel a little more satisfied with them" [8]. The goal of this dissertation has been to provide users with a customized view of their topic of interest based on their own mental model—a view that is truly theirs—to enable them to detect changes and follow the ebbs and flows of their topic over time.

Appendix A: IRB Approval for Survey on Users' Sort Order Preferences

In this appendix we include the University of Maryland, College Park IRB approval documentation for the sort order survey discussed in Chapter 4.

UNIVERSITY OF MARYLAND
INSTITUTIONAL REVIEW BOARD

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	December 18, 2018
TO:	Douglas Oard
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[1350198-1] Sort Orders in Social Media
REFERENCE #:	
SUBMISSION TYPE:	New Project
ACTION:	DETERMINATION OF EXEMPT STATUS
DECISION DATE:	December 18, 2018
REVIEW CATEGORY:	Exemption category *# 2, Waiver of consent, 45CFR46.117(c)(1).*

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has determined this project is EXEMPT FROM IRB REVIEW according to federal regulations.

We will retain a copy of this correspondence within our records.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Maryland College Park (UMCP) IRB's records.

Appendix B: Survey about Users' Sort Order Preferences in Social Media

This appendix includes the complete question set from the social media sort order survey that is covered in Chapter 4.

B.1 Section 1: Consent Form

Q1: Project Title: Sort Orders in Social Media

Purpose of the Study: This research is being conducted by Kristine Rogers at the University of Maryland, College Park under the supervision of Dr. Douglas W. Oard. We are inviting you to participate in this research project because you have indicated that you use social media apps and websites. The purpose of this research project is to understand your preference for how posts should be ordered depending on what you are doing on social media: following live events, following a topic over time, browsing, or searching. You will be asked to complete your responses for at least one of the four options.

Procedures: You will participate in an online survey on sort orders in social media usage. The survey will take 7-10 minutes for the demographic section and first set of sort order questions; after completing one section, you may be given the option to answer questions for additional scenarios. Questions for each additional scenario takes approximately 5 minutes to complete. If you complete the main survey plus all optional sections, we anticipate that the survey will take 25-30 minutes. All materials will be analyzed at a later point.

After completing the survey, you will be given the option to provide your email address to be entered into a raffle for one of five $25 Amazon gift cards. Your email address will not be associated with your survey results.

Potential Risks and Discomforts: There are no known direct risks inherent in participating in this survey.

Potential Benefits: There are no direct benefits from participating in this research. However, in the future, other people might benefit from technologies that are developed based on the findings.

Confidentiality: We will not collect your name or other identifying information as part of this survey. Pseudonyms will be used for any quotes used in the research.

If we write a report or article about this research project for publication in academic journals, your identity will be protected to the maximum extent possible. General information about the findings will be published. Information may be shared with governmental authorities if you or someone else is in danger or if we are required to do so by law.

Those participants who choose to be entered into the contest to win one of the Amazon gift cards will be asked to provide their email address separately from the survey; these email addresses will not be associated with the survey results.

Right to Withdraw and Questions: Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you stop taking the survey before completing at least one of the sections on sort orders you may not be eligible for the gift card raffle.

If you are an employee or student at the University of Maryland, your employment status or academic standing will not be positively or negatively affected by your decision to participate in the study.

If you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the student:

Kristine Rogers
University of Maryland
College of Information Studies
Computational Linguistics and Information Processing Lab
3126 AV Williams Building
8223 Paint Branch Dr
College Park, MD 20740
E-mail: sort.order.study@gmail.com
Telephone: 301-405-7590

Participant Rights: If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

University of Maryland College Park
Institutional Review Board Office
1204 Marie Mount Hall
College Park, Maryland 20742
E-mail: irb@umd.edu
Telephone: 301-405-0678

This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.

Q2: Statement of Consent:

Selecting "I agree" indicates that you are at least 18 years of age, located in the US, and use social media; you have read this consent form or have had it read to you; your

questions have been answered to your satisfaction and you voluntarily agree to participate in this research study.

◯ I agree

◯ I disagree

B.2 Section 2: Screening Questions

Q3: What is your age?

◯ Under 18

◯ 18 - 24

◯ 25 - 34

◯ 35 - 44

◯ 45 - 54

◯ 55 - 64

◯ 65 - 74

◯ 75 - 84

◯ 85 or older

Q4: Are you located in the US?

◯ Yes

◯ No

Q5: Are you an active social media user?

◯ Yes

◯ No

B.3 Section 3: Demographic Questions

Display This Question If Q4 = Yes
Q6: Please enter your US zip code.
[Text Response]

Q7: What is your gender?

◯ Female

◯ Male

◯ Other

◯ Prefer not to answer

Q8: I identify my ethnicity as: (select all that apply)

☐ Asian

☐ Black or African American

☐ Hispanic/Latinx

☐ Native American

☐ Pacific Islander

☐ White/Caucasian

☐ Prefer not to answer

☐ Other (please specify)

Q9: What is the highest degree or level of school you have completed? (If you're currently enrolled in school, please indicate the highest degree you have received.)

◯ Less than a high school diploma

◯ High school degree or equivalent (e.g. GED)

◯ Some college, no degree

◯ Associate degree (e.g. AA, AS)

◯ Trade/technical/vocational training (e.g. culinary school)

◯ Bachelor's degree (e.g. BA, BS)

◯ Master's degree (e.g. MA, MS, MEd)

◯ Professional degree (e.g. MD, DDS, DVM)

◯ Doctorate (e.g. PhD, EdD)

Q10: What is your current employment status?

◯ Full-time employment

◯ Part-time employment

◯ Unemployed

◯ Self-employed

◯ Homemaker

◯ Student

◯ Retired

◯ Other

Q11: What is your industry?

[Text Response]

Q12: Outside of work or school, how much time do you spend online (phone, tablet, computer) on an average day?

◯ Less than 10 minutes per day

◯ 10 minutes to 1 hour per day

◯ 1-2 hours per day

◯ 2-4 hours per day

◯ 4-8 hours per day

◯ 8-12 hours per day

◯ More than 12 hours per day

B.4 Section 4: Background

Q13: In this survey, we are focused on the following social media sites:

- Facebook
- Instagram
- LinkedIn
- Pinterest
- Reddit
- Slack
- Snapchat
- Twitter

Q14: Do you use any social media sites other than the ones listed above?

◯ Yes

◯ No

Display This Question If Q14 = Yes
Q15: What other social media site(s) do you use?

[Text Response]

Q16: How often do you use the following social media sites?

Frequency of Use

	Hourly	Daily	Weekly	Monthly	Yearly	Rarely	Never
Facebook	◯	◯	◯	◯	◯	◯	◯
Instagram	◯	◯	◯	◯	◯	◯	◯
LinkedIn	◯	◯	◯	◯	◯	◯	◯
Pinterest	◯	◯	◯	◯	◯	◯	◯
Reddit	◯	◯	◯	◯	◯	◯	◯
Slack	◯	◯	◯	◯	◯	◯	◯
Snapchat	◯	◯	◯	◯	◯	◯	◯
Twitter	◯	◯	◯	◯	◯	◯	◯
Response to Q15	◯	◯	◯	◯	◯	◯	◯

Display This Question If Q5 = Yes
Q17: How do you typically access social media sites? (select all that apply)

☐ Phone

☐ Tablet

☐ Laptop computer

☐ Desktop computer

☐ Other (please specify)

Q18: Do you use social media to follow a specific topic or theme over a long period of time (example: following a television show over the course of a season, finding news about a celebrity, or following year-round developments relating to a sports team)?

◯ Yes

◯ No

Display This Question If Q18 = Yes
Q19: Please provide an example of a specific topic that you regularly look up on social media:

[Text Response]

Q20: Do you ever use social media to follow or interact users and posts during an event (for instance, "live tweeting"), such as a television show, sporting event, or political event?

◯ Yes

◯ No

Display This Question If Q20 = Yes
Q21: Please provide an example of a specific event that you followed or interacted with on social media:

[Text Response]

Q22: Do you ever use social media to browse content, without a specific reason for browsing (example: browsing current events, looking for interesting content, following a news organization, seeing what my friends and family are up to)?

◯ Yes

◯ No

Display This Question If Q22 = Yes
Q23: Please provide an example of the type of information you expect to see when you browse:

[Text Response]

Q24: Do you ever use social media to run searches for specific information?

◯ Yes

◯ No

Display This Question If Q24 = Yes
Q25: Please provide an example of a specific search you have run on social media:

[Text Response]

Display This Question If Q18 = No And Q20 = No And Q22 = No And Q24 = No
Q26: What do you do on social media?

[Text Response]

B.5 Section 5: Change Detection

Q27: Category: Following a topic over time

In this section, you will answer questions about your use of social media to follow a topic over a long period of time.

Q28: You indicated that you follow specific topics over long periods of time. What categories of topics do you follow over time? (select all that apply)

☐ Art

☐ Brands/Fashion

☐ Celebrity information

☐ Financial information

☐ Health

☐ Movies

☐ Music/Musicians

☐ News and politics

☐ Outdoors

☐ Sports

☐ Television shows

☐ Video games

☐ Other (please specify)

Q29: Earlier you said that you follow this topic on social media: [Response to Q19]

Please answer the following questions based on this topic.

Q30: When was the last time you looked up this topic?

◯ Within the past hour

◯ Within the past day

◯ Within the past week

◯ Within the past month

◯ Within the past six months

◯ Within the past year

◯ More than a year ago

[Carry Forward Unselected Choices from Q16]
Q31: Which social media sites have you used to follow your specific topic over time? (select all that apply)

☐ Facebook

☐ Instagram

☐ Pinterest

☐ Snapchat

☐ LinkedIn

☐ Twitter

☐ Slack

☐ Reddit

☐ [Response to Q15] *(Display This Choice If "What other social media site(s) do you use?" Text Response Is Not Empty)*

Q32: Is there a particular time of day when you tend to look for information on this topic on social media? (select all that apply)

☐ Early Morning

☐ Mid-to-Late Morning

☐ Around lunchtime

☐ Afternoon

☐ Evening

☐ Night

Q33: How would you categorize your level of knowledge on the topic that you follow over time?

◯ I'm still learning about it (Basic)

◯ I know a lot of the terminology (Novice)

◯ I could explain to a friend what was going on (Intermediate)

◯ I could teach a 1-hour class on it (Advanced)

◯ I could teach a college level-course on this topic (Expert)

Q34: How long have you been following this topic?

◯ Less than a week

◯ Between a week and a month

◯ 1-6 months

◯ 6-12 months

◯ 1-2 years

◯ 2-5 years

◯ More than 5 years

Q35: How much time per day do you spend getting updates on this topic?

◯ Less than 10 minutes per day

◯ 10 minutes to 1 hour per day

◯ 1-2 hours per day

◯ 2-4 hours per day

◯ More than 4 hours per day

Q36: Why do you follow specific topics over time? (select all that apply)

☐ Curiosity

☐ Entertainment

☐ Don't want to miss out on something interesting

☐ Friends are talking about it

☐ Family is talking about it

☐ Required by school

☐ Required by work

☐ Other (please specify)

Q37: How satisfied are you with the way social media systems organize posts when you are trying to get updates on a topic you follow regularly?

◯ Very Unsatisfied

◯ Unsatisfied

◯ Neutral

◯ Satisfied

◯ Very Satisfied

Q38: How would you like to have results sorted when you follow a topic over a long period of time on social media?

[Text Response]

Q39: We are considering new ways to organize social media data. Please answer the following questions about how you'd like posts to be organized when you follow a topic over a long period of time.

Which posts do you prefer to see at the top of your feed when you look for updates on the topic?

◯ Newest

◯ Oldest

◯ Closest to my topic of interest

◯ Most popular

◯ Most unique or unusual

◯ Other (please specify)

Q40: If you could have your posts grouped together while getting updates on the topic you follow regularly, which option would you prefer:

◯ Grouped by accounts I follow

◯ Grouped by themes or developments relating to the topic

◯ Grouped by verified or credible accounts

◯ No grouping

◯ Other (please specify)

Q41: When getting updates on the topic you follow, would you like to see:

◯ All posts by people I follow

◯ Only the best posts by people I follow

◯ All posts related to the topic I follow

◯ Only the best posts related to the topic I follow

◯ All posts about themes or developments that happen on the topic

◯ Only the best posts related to each theme or development

Q42: How many posts should you see on the same theme or development while looking for updates on your topic?

◯ 1

◯ 2-4

◯ 5-10

◯ 11-15

◯ More than 15, but not every post

◯ All posts on the new theme

Q43: How do you decide when you've seen enough posts on the topic? In other words, at what point are you "updated"?

◯ When I run out of time

◯ When I have read all of the relevant posts

◯ When I have found all of the relevant themes or developments

◯ When I have answered my questions

◯ Other (please specify)

Display This Question If Q20 = Yes Or Q22 = Yes Or Q24 = Yes
Q44: Would you like to answer questions about sort orders for another category of social media use?

◯ Yes

◯ No

B.6 End Condition: Raffle Signup

Q45: Would you like to enter a raffle for one of five $25 Amazon gift cards?

◯ Yes

◯ No

B.7 Section 6: Live Events

Q46: Category: Live events

In this section, you will answer questions about your use of social media to follow or interact with a live event.

Q47: What type(s) of events have you interacted with on social media while they were happening (for example, "live tweeting")? (choose all that apply)

☐ Award ceremony

☐ Charity event

☐ Conference or talk

☐ Movie

☐ Music event

☐ Live streaming video

☐ Political event

☐ Sporting event (in person)

☐ Sporting event (broadcast)

☐ Sports championship game

☐ Television show

☐ Theater event

☐ Video game

☐ Other (please specify)

Q48: You previously indicated that you have interacted with the following event through social media: [Response to Q21]

Please answer the following questions based on this topic.

[Carry Forward Unselected Choices from Q16]
Q49: Which of the following social media sites have you used to follow or interact with an event? (select all that apply)

☐ Facebook

☐ Instagram

☐ Pinterest

☐ Snapchat

☐ LinkedIn

☐ Twitter

☐ Slack

☐ Reddit

☐ [Response to Q15] *(Display This Choice If "What other social media site(s) do you use?" Text Response Is Not Empty)*

Q50: Is there a particular time of day when you tend to follow a specific event on social media? (select all that apply)

☐ Early Morning

☐ Mid- to Late Morning

☐ Around Lunchtime

☐ Afternoon

☐ Evening

☐ Night

Q51: When you use social media for these types of events, do you: (select all that apply)

☐ Browse posts of my friends and people I follow

☐ Browse other users' posts

☐ Interact with (reply, retweet/repost) posts of my friend and people I follow

☐ Interact with (reply, retweet/repost) other users' posts

☐ None of the above

☐ Other (please specify)

Q52: How much time per day do you spend following this event?

◯ Less than 10 minutes per day

◯ 10 minutes to 1 hour per day

◯ 1-2 hours per day

◯ 2-4 hours per day

◯ More than 4 hours per day

Q53: How satisfied are you with the way social media systems organize posts when you are following a live event?

◯ Very Unsatisfied

◯ Unsatisfied

◯ Neutral

◯ Satisfied

◯ Very Satisfied

Q54: How would you like to have results sorted when you follow a live event on social media?

[Text Response]

Q55: We are considering new ways to organize social media data. Please answer the following questions about how you'd like posts to be organized when you follow a live event on social media.

Which posts do you prefer to see at the top of your feed when you are following a live event?

◯ Newest

◯ Oldest

◯ Closest to my topic of interest

◯ Most popular

◯ Most unique or unusual

◯ Other (please specify)

Q56: If you could have your posts grouped together while following an event, which option would you prefer:

◯ Grouped by accounts I follow

◯ Grouped by themes or developments relating to the topic

◯ Grouped by verified or credible accounts

◯ No grouping

◯ Other (please specify)

Q57: When getting updates on the event, would you like to see:

- ◯ All posts by people I follow
- ◯ Only the best posts by people I follow
- ◯ All posts related to the topic I follow
- ◯ Only the best posts related to the topic I follow
- ◯ All posts about themes or developments that happen on the topic
- ◯ Only the best posts related to each theme or development

Q58: How many posts should you see on the same theme or development while following an event?

- ◯ 1
- ◯ 2-4
- ◯ 5-10
- ◯ 11-15
- ◯ More than 15, but not every post
- ◯ All posts on the new theme

Q59: How do you decide when you've seen enough posts on the topic? In other words, at what point are you "updated"?

- ◯ When I run out of time
- ◯ When I have read all of the relevant posts
- ◯ When I have found all of the relevant themes or developments
- ◯ When I have answered my questions
- ◯ When the event is over
- ◯ Other (please specify)

Display This Question If Q24 = Yes Or Q22 = Yes
Q60: Would you like to answer questions about sort orders for another category of social media use?

◯ Yes

◯ No

B.8 Section 7: General Browsing

Q61: Category: Browsing

In this section, you will answer questions about your use of social media to browse content when you don't have a specific goal or objective.

Q62: What categories of topics do you prefer to see when you browse social media without a specific goal? (select all that apply)

☐ Art

☐ Brands/Fashion

☐ Celebrity information

☐ Financial information

☐ Health

☐ Movies

☐ Music/Musicians

☐ News and politics

☐ Outdoors

☐ Sports

☐ Television shows

☐ Video games

☐ Weather

☐ Other (please specify)

Q63: You previously used the following example of a topic you might expect to see when browsing social media: [Response to Q23]

Please answer the following questions based on this topic.

[Carry Forward Unselected Choices from Q16]
Q64: Which of the following social media sites have you used to browse without a specific question or goal? (select all that apply)

- ☐ Facebook
- ☐ Instagram
- ☐ Pinterest
- ☐ Snapchat
- ☐ LinkedIn
- ☐ Twitter
- ☐ Slack
- ☐ Reddit
- ☐ [Response to Q15] *(Display This Choice If "What other social media site(s) do you use?" Text Response Is Not Empty)*

Q65: Is there a particular time of day when you tend to browse social media without a specific goal? (select all that apply)

- ☐ Early Morning
- ☐ Mid- to Later Morning
- ☐ Around lunchtime
- ☐ Afternoon
- ☐ Evening
- ☐ Night

Q66: Why do you browse social media? (select all that apply)

☐ Curiosity

☐ Entertainment

☐ Don't want to miss out on something interesting

☐ Friends are talking about it

☐ Family is talking about it

☐ Required by school

☐ Required by work

☐ Other (please specify)

Q67: When browsing social media, would you like to see:

◯ All posts by people I follow

◯ Only the best posts by people I follow

◯ All posts related to the topic I follow

◯ Only the best posts related to the topic I follow

◯ All posts about themes or developments that happen on the topic

◯ Only the best posts related to each theme or development

Q68: How satisfied are you with the way social media systems organize posts when you are browsing without a specific goal?

◯ Very Unsatisfied

◯ Unsatisfied

◯ Neutral

◯ Satisfied

◯ Very Satisfied

Q69: How would you like to have results sorted when you browse social media?

[Text Response]

Q70: We are considering new ways to organize social media data. Please answer the following questions about how you'd like posts to be organized when you browse social media.

Which posts do you prefer to see at the top of your feed when you browse?

◯ Newest

◯ Oldest

◯ Closest to my topic of interest

◯ Most popular

◯ Most unique or unusual

◯ Other (please specify)

Q71: If you could have your posts grouped together while browsing, which option would you prefer:

◯ Grouped by accounts I follow

◯ Grouped by themes or developments relating to the topic

◯ Grouped by verified or credible accounts

◯ No grouping

◯ Other (please specify)

Q72: When browsing, would you like to see:

◯ All posts by people I follow

◯ Only the best posts by people I follow

◯ All posts related to the topic I follow

◯ Only the best posts related to the topic I follow

◯ All posts about themes or developments that happen on the topic

◯ Only the best posts related to each theme or development

Q73: How many posts should you see on the same theme or development while browsing?

◯ 1

◯ 2-4

◯ 5-10

◯ 11-15

◯ More than 15, but not every post

◯ All posts on the new theme

Q74: How do you decide when you've seen enough posts? In other words, at what point are you done browsing?

◯ When I run out of time

◯ When I have read all of the relevant posts

◯ When I have found all of the relevant themes or developments

◯ When I have answered my questions

◯ Other (please specify)

Display This Question If Q24 = Yes
Q75: Would you like to answer questions about sort orders for another category of social media use?

◯ Yes

◯ No

B.9 Section 8: Ad Hoc Search

Q76: Category: Searching

In this section, you will answer questions about running searches in social media sites.

Q77: What types of searches do you typically run on social media? (select all that apply)

- ☐ Look for a specific user's profile
- ☐ Information about or pages for organizations or places
- ☐ Look up news articles
- ☐ Find events
- ☐ Research current events
- ☐ Find more information about a recent event
- ☐ Find a trending topic
- ☐ Other (please specify)

Q78: You previously provided the following example of a search you ran on social media: [Response to Q25]

Please answer the following questions based on this topic.

[Carry Forward Unselected Choices from Q16]
Q79: Which social media sites have you used to run searches for specific data? (select all that apply)

- ☐ Facebook
- ☐ Instagram
- ☐ Pinterest
- ☐ Snapchat
- ☐ LinkedIn
- ☐ Twitter

☐ Slack

☐ Reddit

☐ [Response to Q15] *(Display This Choice If "What other social media site(s) do you use?" Text Response Is Not Empty)*

Q80: Is there a particular time of day when you tend to run searches like this on social media? (select all that apply)

☐ Early Morning

☐ Mid- to Late Morning

☐ Around lunchtime

☐ Afternoon

☐ Evening

☐ Night

Q81: How much time per day do you spend searching social media?

◯ Less than 10 minutes per day

◯ 10 minutes to 1 hour per day

◯ 1-2 hours per day

◯ 2-4 hours per day

◯ More than 4 hours per day

Q82: Why do you search social media? (select all that apply)

☐ Curiosity

☐ Entertainment

☐ Don't want to miss out on something interesting

☐ Friends are talking about it

☐ Family is talking about it

☐ Required by school

☐ Required by work

☐ Other (please specify)

Q83: How satisfied are you with the way social media systems organize posts in response to your search?

◯ Very Unsatisfied

◯ Unsatisfied

◯ Neutral

◯ Satisfied

◯ Very Satisfied

Q84: How would you like to have results sorted when you run a search on social media?

[Text Response]

Q85: We are considering new ways to organize social media data. Please answer the following questions about how you'd like posts to be organized when you run a search on social media.

Which posts do you prefer to see at the top of your feed after you run a search?

◯ Newest

◯ Oldest

◯ Closest to my topic of interest

◯ Most popular

◯ Most unique or unusual

◯ Other (please specify)

Q86: If you could have your posts grouped together after a search, which option would you prefer:

- ◯ Grouped by accounts I follow
- ◯ Grouped by themes or developments relating to the topic
- ◯ Grouped by verified or credible accounts
- ◯ No grouping
- ◯ Other (please specify)

Q87: When looking through search results on social media, would you like to see:

- ◯ All posts by people I follow
- ◯ Only the best posts by people I follow
- ◯ All posts related to the topic I follow
- ◯ Only the best posts related to the topic I follow
- ◯ All posts about themes or developments that happen on the topic
- ◯ Only the best posts related to each theme or development

Q88: How many posts should you see on the same theme or development in your search results?

- ◯ 1
- ◯ 2-4
- ◯ 5-10
- ◯ 11-15
- ◯ More than 15, but not every post
- ◯ All posts on the new theme

Q89: How do you decide when you've seen enough posts on the topic? In other words, at what point are you "updated"?

- ◯ When I run out of time
- ◯ When I have read all of the relevant posts
- ◯ When I have found all of the relevant themes or developments
- ◯ When I have answered my questions
- ◯ Other (please specify)

B.10 Section 9: Other

Q90: Category: General Social Media Use

In this section, you will answer questions about your use of social media.

Q91: What categories of topics do you prefer to see on social media? (select all that apply)

- ☐ Art
- ☐ Brands/Fashion
- ☐ Celebrity information
- ☐ Financial information
- ☐ Health
- ☐ Movies
- ☐ Music/Musicians
- ☐ News and politics
- ☐ Outdoors
- ☐ Sports
- ☐ Television shows
- ☐ Video games

☐ Weather

☐ Other (please specify)

Q92: Is there a particular time of day when you tend to use social media? (select all that apply)

☐ Early Morning

☐ Mid- to Later Morning

☐ Around lunchtime

☐ Afternoon

☐ Evening

☐ Night

Q93: Why do you use social media? (select all that apply)

☐ Curiosity

☐ Entertainment

☐ Don't want to miss out on something interesting

☐ Friends are talking about it

☐ Family is talking about it

☐ Required by school

☐ Required by work

☐ Other (please specify)

Q94: When using social media, would you like to see:

◯ All posts by people I follow

◯ Only the best posts by people I follow

◯ All posts related to the topic I follow

◯ Only the best posts related to the topic I follow

◯ All posts about themes or developments that happen on the topic

◯ Only the best posts related to each theme or development

Q95: How satisfied are you with the way social media systems organize posts?

◯ Very Unsatisfied

◯ Unsatisfied

◯ Neutral

◯ Satisfied

◯ Very Satisfied

Q96: How would you like to have results sorted on social media?

[Text Response]

Q97: We are considering new ways to organize social media data. Please answer the following questions about how you'd like posts to be organized.

Which posts do you prefer to see at the top of your feed when you browse?

◯ Newest

◯ Oldest

◯ Closest to my topic of interest

◯ Most popular

◯ Most unique or unusual

◯ Other (please specify)

Q98: If you could have your posts grouped together, which option would you prefer:

◯ Grouped by accounts I follow

◯ Grouped by themes or developments relating to a topic

◯ Grouped by verified or credible accounts

◯ No grouping

◯ Other (please specify)

Q99: Would you like to see:

◯ All posts by people I follow

◯ Only the best posts by people I follow

◯ All posts related to the topic I follow

◯ Only the best posts related to the topic I follow

◯ All posts about themes or developments that happen on the topic

◯ Only the best posts related to each theme or development

Q100: How many posts should you see on the same theme or development?

◯ 1

◯ 2-4

◯ 5-10

◯ 11-15

◯ More than 15, but not every post

◯ All posts on the new theme

Q101: How do you decide when you've seen enough posts? In other words, at what point are you done using social media?

◯ When I run out of time

◯ When I have read all of the relevant posts

◯ When I have found all of the relevant themes or developments

◯ When I have answered my questions

◯ Other (please specify)

Appendix C: IRB Approval for Daybreak User Study

In this appendix we include the University of Maryland, College Park IRB approval documentation for the Daybreak user study discussed in Chapters 5 and 6.

UNIVERSITY OF MARYLAND
INSTITUTIONAL REVIEW BOARD

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	November 10, 2021
TO:	Douglas Oard
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[1701773-1] User study: Applying Daybreak change detection system to real-world situations
SUBMISSION TYPE:	New Project
ACTION:	APPROVED
APPROVAL DATE:	November 10, 2021
REVIEW TYPE:	Expedited Review
REVIEW CATEGORY:	Expedited review category *# 7, Waiver of written consent, 45CFR46.116(f)(3).*

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a project design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

Prior to final approval of this project scientific review was completed by the IRB Member reviewer.

This submission has received Expedited Review based on the applicable federal regulations.

This project has been determined to be a MINIMAL RISK project.

Please remember that informed consent is a process beginning with a description of the project and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the project via a dialogue between the researcher and research participant. Unless a consent waiver or alteration has been approved, Federal regulations require that each participant receives a copy of the consent document.

Please note that any revision to previously approved materials must be approved by this committee prior to initiation. Please use the appropriate Amendment forms for this procedure.

All UNANTICIPATED PROBLEMS involving risks to subjects or others (UPIRSOs) and SERIOUS and UNEXPECTED adverse events must be reported promptly to this office. Please use the appropriate reporting forms for this procedure. All FDA and sponsor reporting requirements should also be followed. All NON-COMPLIANCE issues or COMPLAINTS regarding this project must be reported promptly to this office.

Please note that all research records must be retained for a minimum of seven years after the completion of the project.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

Appendix D: Daybreak User Study Artifacts

This appendix includes the complete set of artifacts from the Daybreak study described in Chapters 5 and 6. This includes the selection survey questions, pre-study questionnaire, post-day questionnaires, and semi-structured interview questions.

D.1 Daybreak Selection Survey

In this section we present the complete Daybreak selection survey.

D.1.1 Consent Statement

Q1: Consent to Participate

Project Title

Daybreak Change Detection Study

Purpose of the Study

This research is being conducted by Kristine Rogers at the University of Maryland, College Park under the supervision of Dr. Douglas W. Oard. We are inviting you to participate in this research project because you are at least 18 years of age, met all other eligibility criteria, and have indicated that you follow a specific topic over time. Many people follow

consistent topics over time, whether for professional reasons or due to personal interest. We refer to this as "change detection" – getting an update on a topic to determine what has changed since the last time they checked. We are interested in understanding how best to organize search results for individuals who regularly get updates on a topic.

This questionnaire will be used to determine whether you are eligible to participate in an evaluation of a system called Daybreak, which shows news articles related to a topic similar to one you follow. The user has the ability to categorize interesting documents using tags. In subsequent days, the system will find other documents related to the user's tags. The purpose of this research project is to understand your preference for organizing new articles, to make it easier for you to understand what has changed on your topic since the last time you looked it up. Users who qualify and are selected for the main study will receive a $25 Amazon gift card in compensation after they have tested the system.

Procedures

This is a selection questionnaire to determine whether you qualify for the main study. The selection questionnaire consists of a series of questions to determine whether you follow specific topics over time. The survey also includes basic demographic questions to be used to select a diverse range of participants for the main study. If selected for the study, we will contact you using the email address provided in the survey.

Potential Risks and Discomforts

There are no known or foreseeable risks inherent in participating in this study, with the exception of a breach of confidentiality. The breach of confidentiality risk will be mitigated by limiting access to the information created as part of the study to members of the research team. All questionnaires will be stored within the UMD Qualtrics system. If you do not

qualify or are not selected, we will delete your response information within 1 week of making the determination. Participants may choose to terminate the study at any time.

Potential Benefits

There are no direct benefits from participating in this research. However, in the future, other people might benefit from any technologies that are developed based on the findings. This study explores change detection tasks in an attempt to understand how users prefer to get updates on tasks that they follow over time. Results from this study can inform this field, and lead to development of capabilities that make it easier for users to complete these types of tasks.

Confidentiality

We will only collect your name and other identifying information to schedule the main session and provide your compensation. We will use a participant ID for all other parts of this study.

The questionnaire results will be stored in the University of Maryland's Qualtrics account, and stored on the researcher's password-protected laptop for analysis. For users selected for the main study, demographic information provided in this questionnaire will be used in analysis of study results.

If we write a report or article about this research project for publication in academic journals, your identity will be protected to the maximum extent possible. General information about the findings will be published. Information may be shared with governmental authorities if you or someone else is in danger or if we are required to do so by law. Pseudonyms will be used for any quotes used in the research.

30 days after publications are complete, all raw data collected in this study will be deleted.

Right to Withdraw and Questions

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time.

If you are an employee or student at the University of Maryland, your employment status or academic standing will not be positively or negatively affected by your decision to participate in the study.

If you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the Principal Investigator:

Dr. Douglas Oard
University of Maryland
College of Information Studies
Patuxent 1109C
4161 Fieldhouse Dr
College Park, MD 20742
E-mail: oard@umd.edu
Telephone: 301-405-7590

Participant Rights

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

University of Maryland College Park
Institutional Review Board Office
1204 Marie Mount Hall

College Park, Maryland, 20742

E-mail: irb@umd.edu

Telephone: 301-405-0678

This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.

Statement of Consent

Selecting "I agree" indicates that you are at least 18 years of age, located in the US, and perform change detection tasks; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study.

Participants can print or save the consent form using the browser's print or save functions. Additionally, the consent form for both the qualification survey and the pre-study questionnaire include the Principal Investigator's contact information for respondents who would like to request a copy of the consent statement.

◯ I agree

◯ I do not agree

Skip To End of Survey If Q16 = I do not agree

D.1.2 Screening Questions

Q2.1: What is your age?

◯ Under 18

◯ 18 - 24

◯ 25 - 34

◯ 35 - 44

◯ 45 - 54

◯ 55 - 64

◯ 65 or older

Skip To Q2.9 If Q2.1 = Under 18

Q2.2: Are you located in the US?

◯ Yes

◯ No

Skip To Q2.9 If Q2.2 = No

Display This Question If Q2.2 = Yes
Q2.3: Please enter your US zip code

[Text Response]

Q2.4: Some people have topics that they follow over a long period of time. For example, they might read articles about the topic every day, to learn the latest news. Is this something you do?

◯ Yes

◯ No

Skip To Q2.9 If Q2.4 = No

Q2.5: Please provide an example of one or more specific topics that you follow on a regular basis:

[Text Response]

Q2.6: Why do you follow this topic? (select all that apply)

☐ Personal reasons (i.e. hobbies)

☐ Professional reasons (i.e. it relates to my work or career development)

☐ Other (please specify)

Q2.7: How often do you look up information about this topic?

◯ Multiple times per day

◯ Once a day

◯ Weekly

◯ Monthly

◯ Yearly

◯ Less than yearly

Q2.8: Do you have access to a laptop or desktop computer connected to a working video camera, and an Internet connection? (required for the virtual session)

◯ Yes

◯ No

Skip To Q2.9 If Q2.8 = No

Display This Question If Q2.1 = Under 18 Or Q2.2 = No Or Q2.4 = No Or Q2.8 = No
Q2.9: Thank you for your interest in this study! Unfortunately, you do not qualify for participation at this time. Please click on the [>] button to exit the survey.

Skip To End of Survey If Q2.9 Displayed

D.1.3 Demographic Questions

Q3.1: What is your gender?

◯ Female

◯ Male

◯ Other

◯ Prefer not to answer

Q3.2: I identify my race as: (select all that apply)

☐ American Indian or Alaska Native

☐ Asian

☐ Black or African American

☐ Native Hawaiian or Other Pacific Islander

☐ White

☐ Prefer not to answer

Q3.3: Please specify your ethnicity:

◯ Hispanic or Latino or Spanish Origin

◯ Not Hispanic or Latino or Spanish Origin

◯ Prefer not to answer

D.1.4 Respondent Contact Information

Q4.1: To allow us to contact you regarding the study, please provide your name:

[Text Response]

Q4.2: Please provide your email address:

[Text Response]

D.1.5 Confirmation

Q5.1: Thank you for your response! If you are selected for this study, we will contact you at the email address you provided with additional details. Please click on the [>] button to complete the survey.

Skip To End of Survey If Q5.1 Displayed

D.2 Daybreak Pre-Study Questionnaire

In this section we present the complete Daybreak pre-study questionnaire, completed by participants prior to starting the session.

D.2.1 Consent Statement

Q1.1: Consent to Participate

Project Title

Daybreak Change Detection Study

Purpose of the Study

This research is being conducted by Kristine Rogers at the University of Maryland, College Park under the supervision of Dr. Douglas W. Oard. We are inviting you to participate in this research project because you are at least 18 years of age, met all other eligibility criteria, and have indicated that you follow a specific topic over time. Many people follow consistent topics over time, whether for professional reasons or due to personal interest. We refer to this as "change detection" – getting an update on a topic to determine what has changed since the last time they checked. We are interested in understanding how best to organize search results for individuals who regularly get updates on a topic.

This study is an evaluation of system called Daybreak, which shows news articles related to a topic similar to one you follow. The user has the ability to categorize interesting

documents using tags. In subsequent days, the system will find other documents related to the user's tags. The purpose of this research project is to understand your preference for organizing new articles, to make it easier for you to understand what has changed on your topic since the last time you looked it up.

Procedures

You will participate in a virtual study, hosted via the University of Maryland's Zoom system, to test a system called Daybreak. The entire study may take up to 120 minutes (2 hours) to complete. Two researchers will be present during the study; one will interact with you, and the other will be there to take notes. Throughout the study, you can skip any question you do not wish to answer.

The study includes the following segments:

- First, you will fill out an initial questionnaire. After the initial questionnaire, recording of the session will begin. Audio and video for the session will be recorded and saved securely using the University of Maryland Panopto system.
- You will complete five 8-minute sessions using the Daybreak system. The Daybreak system will show you documents related to a specific topic. You can use the system to read and tag documents of interest. Between each session is a brief questionnaire.
- At the end of the survey you will complete a storytelling task in which you will write an outline that lists key events or themes across the five sessions. You will be able to view previously tagged documents when you draft the outline.
- The study will end with an interview where the researchers will ask about your experience performing the tasks with the Daybreak system.

Compensation

In compensation for your time, you will receive a $25 gift card if you complete the full study. If you withdraw early, you may be eligible for a $10 gift card in competition if you complete at least one "day" of the exercise. Users who withdraw prior to completing one "day" will not be eligible for compensation.

Potential Risks and Discomforts

There are no known or foreseeable risks inherent in participating in this study, with the exception of a breach of confidentiality. The breach of confidentiality risk will be mitigated by limiting access to the information created as part of the study to members of the research team. Zoom recordings will be stored in the UMD Panopto system. All questionnaires will be stored within the UMD Qualtrics system. The user-generated tags and logs will be stored on the student's password-protected laptop and backed up in a UMD UMIACS account that only the student and PI can access.

Another potential risk is of boredom or frustration on the part of the user, due to the length of the study (up to 120 minutes), which may be considered too long by some users. To mitigate this risk, a break will be offered approximately halfway through the study. Participants may choose to terminate the study at any time. Potential Benefits There are no direct benefits from participating in this research. However, in the future, other people might benefit from any technologies that are developed based on the findings. This study explores change detection tasks in an attempt to understand how users prefer to get updates on tasks that they follow over time. Results from this study can inform this field, and lead to development of capabilities that make it easier for users to complete these types of tasks.

Confidentiality

We are not collecting any new personally identifying information in this section of the study. The information you provided in the qualifying study (your name and email address)

will only be used to provide your compensation after completion of the study. We will use a participant ID for all other parts of this study. The key mapping identifying information to participant IDs will be stored on a thumb drive and only accessed by the researcher. This data will be deleted 30 days after the compensation has been provided.

The questionnaire results will be stored in the University of Maryland's Qualtrics account, and stored on the researcher's password-protected laptop for analysis. The recording of the Zoom session will be recorded in the University of Maryland's Panopto system. All other notes will be stored on password-protected computers of the research team members.

If we write a report or article about this research project for publication in academic journals, your identity will be protected to the maximum extent possible. General information about the findings will be published. Information may be shared with governmental authorities if you or someone else is in danger or if we are required to do so by law. Pseudonyms will be used for any quotes used in the research.

30 days after publications are complete, all raw data collected in this study (to include questionnaire results, Zoom recordings, and notes about sessions) will be deleted.

Right to Withdraw and Questions

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you leave the session early (prior to completion of the interview) you may not be eligible for the $25 Amazon gift compensation.

If you are an employee or student at the University of Maryland, your employment status or academic standing will not be positively or negatively affected by your decision to participate in the study.

If you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the Principal Investigator:

Dr. Douglas Oard

University of Maryland

College of Information Studies

Patuxent 1109C

4161 Fieldhouse Dr

College Park, MD 20742

E-mail: oard@umd.edu

Telephone: 301-405-7590

Participant Rights

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

University of Maryland College Park

Institutional Review Board Office

1204 Marie Mount Hall

College Park, Maryland, 20742

E-mail: irb@umd.edu

Telephone: 301-405-0678

This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.

Statement of Consent

Entering your participant ID indicates that you are at least 18 years of age, located in the US, and perform change detection tasks; you have read this consent form or have had

it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study.

Participants can print or save the consent form using the browser's print or save functions. Additionally, the consent form for both the qualification survey and the pre-study questionnaire include the Principal Investigator's contact information for respondents who would like to request a copy of the consent statement.

Q1.2: Please enter your Participant ID to indicate your consent to participate in this study:

[Text Response]

D.2.2 Demographic Questions

Q2.1: What is the highest degree or level of school you have completed? (If you're currently enrolled in school, please indicate the highest degree you have received.)

- ◯ Less than a high school diploma
- ◯ High school degree or equivalent (e.g. GED)
- ◯ Some college, no degree
- ◯ Associate degree (e.g. AA, AS)
- ◯ Trade/technical/vocational training (e.g. culinary school)
- ◯ Bachelor's degree (e.g. BA, BS)
- ◯ Master's degree (e.g. MA, MS, MEd)
- ◯ Professional degree (e.g. MD, DDS, DVM)
- ◯ Doctorate (e.g. PhD, EdD)

Q2.2: What is your current employment status?

◯ Full-time employment

◯ Part-time employment

◯ Unemployed

◯ Self-employed

◯ Homemaker

◯ Student

◯ Retired

◯ Other

Q2.3: What is your industry?

[Text Response]

Q2.4: Outside of work or school, how much time do you spend online (phone, tablet, computer) on an average day?

◯ Less than 10 minutes per day

◯ 10 minutes to 1 hour per day

◯ 1-2 hours per day

◯ 2-4 hours per day

◯ 4-8 hours per day

◯ 8-12 hours per day

◯ More than 12 hours per day

D.2.3 Topics of Interest

Q3.1: What categories of topics do you follow over time? (select all that apply)

☐ Art

☐ Brands/Fashion

☐ Celebrity information

☐ Financial information

☐ Health

☐ Movies

☐ Music/Musicians

☐ News and politics

☐ Outdoors

☐ Sports

☐ Television shows

☐ Video games

☐ Other (please specify)

Q3.2: Please describe your current process for finding new information on topics you follow.

[Text Response]

Q3.3: Please provide an example of one or more specific topic(s) that you have followed for a long time:

[Text Response]

Q3.4: How long have you been following this topic?

◯ Less than a week

◯ Between a week and a month

◯ 1-6 months

◯ 6-12 months

◯ 1-2 years

◯ 2-5 years

◯ More than 5 years

Q3.5: How would you categorize your level of knowledge on the topic that you follow over time?

◯ I'm still learning about it (Basic)

◯ I know a lot of the terminology (Novice)

◯ I could explain to a friend what was going on (Intermediate)

◯ I could teach a 1-hour class on it (Advanced)

◯ I could teach a college level-course on this topic (Expert)

Q3.6: How much time per day do you spend getting updates on this topic?

◯ Less than 10 minutes per day

◯ 10 minutes to 1 hour per day

◯ 1-2 hours per day

◯ 2-4 hours per day

◯ More than 4 hours per day

Q3.7: Is there a particular time of day when you tend to look for information on this topic on social media? (select all that apply)

☐ Early Morning

☐ Mid-to-Late Morning

☐ Around lunchtime

☐ Afternoon

☐ Evening

☐ Night

Q3.8: When was the last time you looked up this topic?

◯ Within the past hour

◯ Within the past day

◯ Within the past week

◯ Within the past month

◯ Within the past six months

◯ Within the past year

◯ More than a year ago

Q3.9: What resources have you used to follow your specific topic over time? (select all that apply)

☐ Paper newspaper or magazines

☐ Online news sources

☐ Online blogs

☐ Social media sites

☐ Other (please specify)

Q3.10: What websites or systems do you use to follow a topic over time? (select all that apply)

☐ Google Search

☐ Twitter

☐ Facebook

☐ Bing

☐ Google News

☐ I do not use websites or systems to follow topics

☐ Other (please specify)

Q3.11: Have you previously used systems that allow you to tag, categorize, or label data?

◯ Yes

◯ No

◯ Not sure

Display This Question If Q3.11 = Yes
Q3.12: What system(s) have you used to tag, categorize, or label data?

[Text Response]

Q3.13: What device(s) do you typically use to follow your topic(s) of interest? (select all that apply)

☐ Phone

☐ Tablet

☐ Laptop computer

☐ Desktop computer

☐ Other (please specify)

Q3.14: Why do you follow specific topics over time? (select all that apply)

☐ Curiosity

☐ Entertainment

☐ Don't want to miss out on something interesting

☐ Friends are talking about it

☐ Family is talking about it

☐ Required by school

☐ Required by work

☐ Other (please specify)

D.2.4 Technology Use

Q4.1: What kind of device are you using to connect into this study?

- ◯ Desktop computer
- ◯ Laptop computer
- ◯ Tablet
- ◯ Cell phone
- ◯ Other (please specify)

Q4.2: What operating system are you using?

- ◯ Windows
- ◯ Apple/Mac
- ◯ Android
- ◯ Linux
- ◯ Other (please specify)

Q4.3: What is your monitor size?

- ◯ Small
- ◯ Medium
- ◯ Large
- ◯ Or, list monitor dimensions (if known)

D.3 Daybreak Post-Day 1 Questionnaire

Q1.1: What is your participant ID?

[Text Response]

Q1.2: What topic are you covering for this study?

[Text Response]

Q1.3: How would you categorize your level of knowledge on the topic that you're following for this study?

◯ I'm still learning about it (Basic)

◯ I know a lot of the terminology (Novice)

◯ I could explain to a friend what was going on (Intermediate)

◯ I could teach a 1-hour class on it (Advanced)

◯ I could teach a college level-course on this topic (Expert)

Q1.4: What is your level of interest in the topic you're following for this study?

◯ Very high interest

◯ Moderately high interest

◯ Neutral

◯ Moderately low interest

◯ Very low interest

◯ Dislike this topic

◯ Other (please specify)

Q1.5: Think about the list of tags/categories you generated for the most recent session. Which category was the most important for this session?

[Text Response]

Q1.6: Briefly explain why the topic you mentioned was the most important for this session.

[Text Response]

Q1.7: How many documents did view in this session?

- ◯ 0-5
- ◯ 6-10
- ◯ 11-15
- ◯ 16-20
- ◯ 21-25
- ◯ 26+

Q1.8: Were you able to review enough documents during the session?

- ◯ Yes
- ◯ No
- ◯ Undecided

Q1.9: How did you define ”enough” in the prior question?

[Text Response]

Q1.10: Please share any brief thoughts you have on the most recent session.

[Text Response]

D.4 Daybreak Post-Days 2-5 Questionnaire

Q1.1: What is your participant ID?

[Text Response]

Q1.2: Think about the list of tags/categories you generated for the most recent session. Which category was the most important for this session?

[Text Response]

Q1.3: Briefly explain why the topic you mentioned was the most important for this session.

[Text Response]

Q1.4: How many documents did view in this session?

- ◯ 0-5
- ◯ 6-10
- ◯ 11-15
- ◯ 16-20
- ◯ 21-25
- ◯ 26+

Q1.5: Were you able to review enough documents during the session?

- ◯ Yes
- ◯ No
- ◯ Undecided

Q1.6: Please share any brief thoughts you have on the most recent session.

[Text Response]

D.5 Semi-Structured Interview Script

Post-Study Semi-Structured Interview Questions

- Tell us about any past experience you have had with tagging or labeling documents. Was the ability to label documents useful in completing the task?
- Look at the list of categories you created over the five sessions. Which one(s) were most important to your overall story?
- Knowing what you know now, what (if anything) would you do differently in the way you organized/tagged your data?
- Did the system assist you in getting updates on your topic? How/how not?
- Please compare the Daybreak system with your current approach for finding updates on the topic you follow. Did this system help you more or less than your current approach?
- During each day, were you able to review enough documents to get an understanding of what was happening on the assigned topic?
- Was it useful to organize new documents by category?
- Please describe your process for reviewing documents. Did your categorization approach change or evolve over the course of the sessions? If so, how?
- How did you decide that a document was important enough to categorize?
- Within the clusters, what document sort order did you prefer for this task? Why?
- Did your definition of "enough" documents viewed change over the course of the study?
- How do you think the clusters were sorted/organized?
- After they answer the prior question, explain that the subtopics were sorted by rarity (least common subtopic first). Was this approach useful? How would you improve the organization of the clusters?
- What makes one cluster more important than another?
- Did you review uncategorized documents? If so, how important was that set of documents?
- Did you review the list of categories when writing your story? Why/why not? If so, was it useful to have that information available?

- Did you return to and review individual documents when writing your story? Why/ why not? If so, was it useful to have that information available?
- How would you enhance the Daybreak system to help you get updates more effectively?
- Do you have any additional comments that you would like to make about the Daybreak system or other aspects of the study?
- Before we wrap up, do you have any questions about the study?

Appendix E: Artifacts from Daybreak Coding and Analysis

This appendix contains information that was used or generated during the Daybreak coding and analysis process. This includes the Daybreak codebook, a sample of a coded session, a sample log file, as well as a number of tables based on logs and other data collected during Daybreak sessions.

E.1 Daybreak Codebook

In this section, It includes the Daybreak codebook used for annotating unstructured data in the session—in particular, during the semi-structured interview. These codes are organized by theme, and were agreed upon during the code peer review process. The codes that map to the GPA Change Detection Theory directly are "Cluster" (represents "Group"), "Doc Sort" (represents "Pile"), and "Tag Sort" (represents "Arrange").

As far as application of the codebook, the semi-structured interview transcripts were the heaviest coded documents. That said, other artifacts of the study also received codes—this includes comments made by the participant during the course of the study, the outlines drafted during the final storytelling task, and the post-day questionnaire responses.

On the following pages, we have included the complete Daybreak codebook.

Section	Code	Additional	User Action	Example	Comment
CLUSTER	no clustering preference		user mentions that they did not have a preference over whether clusters were on or off	during the interview, the user mentions that they did not have a preference over whether clusters were turned on or off	
CLUSTER	preferred clusters off		user mentions that they preferred to have clusters turned off	during the interview, the user mentions that they preferred to have clusters turned off	
CLUSTER	preferred clusters on		user mentions that they preferred to have clusters turned on	during the interview, the user mentions that they preferred to have clusters turned on	
CLUSTER	user clustering strategy		user discusses their approach for clustering documents	during the interview, the user describes their use of the different clustering options and what drove changes	
DOC SORT	no sort preference		user mentions that they had no preference about document sort orders	during the interview, the user says that they did not have a preference between chronological and relevance-based sorting, or they switched between the two options	
DOC SORT	preferred chronological sort		user mentions that they preferred chronological sorting	during the interview, the user says that they preferred chronological sorting of documents	
DOC SORT	preferred relevance sort		user mentions that they preferred relevance-based sorting	during the interview, the user says that they preferred relevance-based sorting of documents	
DOC SORT	user sort strategy		user discusses their approach for sorting documents	during the interview, the user describes their use of the different sort orders and what drove changes	
FILTER	document filtering approach		user articulates the approach that they used for filtering out documents that they did not find interesting or useful	during the interview, the user describes their approach for using tags on topics that they want to filter out, for instance with labels such as “uninteresting” or “irrelevant”	
FILTER	document filtering preference		user describes how they would like to be able to filter through documents	during the interview, the user describes how they would like to be able to filter documents	Ideas people have about how the system could be adjusted, or how they would like it to behave
FILTER	duplicate document issues		user discusses issues with duplicate documents	the user mentions something about issues relating to having duplicate documents in their results, such as “I changed my sort order because there were too many duplicates”	

Section	Code	Additional	User Action	Example	Comment
FILTER	repetition in document themes	tag name	user mentions repetitive themes	the user mentions something about issues relating to seeing themes repeat throughout the documents. For instance, they might mention that a lot of game recaps were showing up in their document set.	Tag name only if mentioned
MISC	anomaly detection		user expresses interest in identifying anomalies or unusual information	during the interview, the user makes a comment about wanting to be able to find edge cases or anomalies	
MISC	cold start		user mentions the cold start problem	User makes a comment such as "I wish the system already knew what my interests were."	
MISC	exemplar system	name of system(s)	user provides an example of another system that they use that allows them to tag or organize documents	User lists one or more systems that they leverage to tag or organize documents (e.g. Pinterest, Confluence, JIRA, and others)	
MISC	good quote		user makes an interesting point	during the session, the user makes a comment that provides insights, or might be usable in the analysis or writeup	
MISC	system behavior issue		user makes a comment about functionality that did not work as expected	While using the system or during the interview, the user mentions something that didn't work correctly, such as a new document loading in the middle, not at the top	
MISC	system functionality request		user makes a comment about something they wished the system would do	While using the system or during the interview, the user mentions additional functionality that they wanted the system to perform, such as being able to modify a prior tag label	
MISC	system performance issue		user makes a comment about the speed or slowness of the system	While using the system or during the interview, the user comments about the performance of the system (response time, delays, things that took longer than they expected)	
MISC	wants query refinement		user mentions that they want to refine the query	Comments about wishing they could refine the query. Might ask whether it is possible to change the main query term or refine the query.	
OTHER	neutral about other category		user was neutral about whether it was important to review the "Other" category	during the interview, the user did not express a preference about whether it was important to review the "Other" category	

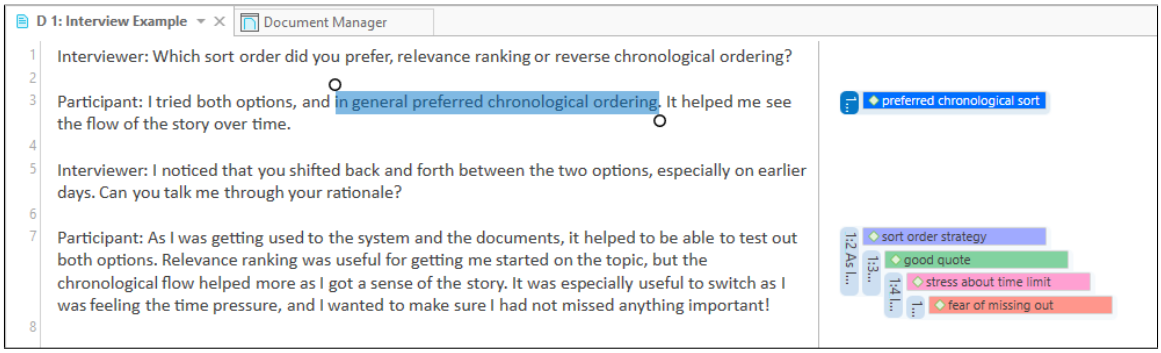
Section	Code	Additional	User Action	Example	Comment
OTHER	other category was important		user thought it was important to review the “Other” category	during the interview, the user mentions that it was important to review the “Other” category	
OTHER	other category was not important		user thought it was not important to review the “Other” category	during the interview, the user mentions that they did not think it was important to review the “Other” category	
OTHER	reviewing other category		user discusses their approach for leveraging the “Other” category	during the interview, the user describes their use of the “Other” category for discovering new subtopics	
OVERALL	does not meet change detection needs		user says system did not meet change detection needs	during the interview, the user specifically says that this system would not help them with change detection tasks	
OVERALL	meets change detection needs		user says system met change detection needs	during the interview, the user specifically says that this system would help them with change detection tasks	
OVERALL	unclear on meeting change detection needs		user says system might meet change detection needs	during the interview, the user seems ambivalent about whether this system would help them with change detection tasks	
PROCESS	fear of missing out		user mentions being concerned about missing something relevant to them	user voices their concerns about missing information, such as “I was afraid I’d miss an important document”	this code can be used during a day, or during the interview section
PROCESS	mentions time limit		user mentions the time constraint	User says something like “the time is passing quickly” or “only 2 minutes left?!”	this code can be used during a day, or during the interview section
PROCESS	stress about time limit		user expresses stress over the time constraint	The user makes a comment like “I don’t feel like I have enough time to finish this!”	this code can be used during a day, or during the interview section
PROCESS	use of document titles		user discusses their approach for reviewing document titles	during the interview, the user discusses their approach on how they reviewed titles	
PROCESS	viewing documents		user discusses their approach for deciding which documents to view	user discussed how they decided which document(s) to open	
STORY	flow of overall story		user mentions the flow of stories in the articles	User mentions the flow of stories or themes across the articles they reviewed. For instance, they might mention that a specific player kept coming up through the course of the week.	
STORY	prep for storytelling task		user discusses their approach for preparing to write the outline for the storytelling task	User discusses what they were doing to come up with items to include in the storytelling task. For instance, the user might mention that they used a specific tag to indicate that they wanted to mention that item in the story.	

Section	Code	Additional	User Action	Example	Comment
STORY	reviewing cluster labels during storytelling task		user discusses their use (or lack of use) of the clusters during the storytelling task	User discusses their choice to refer to or not refer to cluster labels (tags) while they completed the storytelling task	
STORY	writing story from memory		user does not refer to past documents when writing the story	User writes the story without clicking on any of the cluster labels, and does not view either document titles or contents	
STORY	reviewing documents during storytelling task		user discusses their use (or lack of use) of the documents during the storytelling task	User discusses their choice to refer to or not refer to documents while they completed the storytelling task	
TAG	document tagging approach		user articulates their approach for deciding when to tag a document	during the interview, the user describes their threshold for deciding that a document was worth tagging	
TAG	document tagging preference		user describes how they would like to be able to tag documents	during the interview, the user describes how they would like to be able to tag documents, or adjust their tags	Ideas people have about how the system could be adjusted, or how they would like it to behave
TAG	evolution of tagging		user discusses how their use of tagging evolved during the session	during the interview, the user describes how they adjusted or adapted their application of tags to documents over the course of the study	Can include how they wished they could have done something different in tagging approach
TAG	philosophy of tagging		user discusses their views on tagging	during the interview, the user discusses their views of tagging and how it helps them organize information	
TAG	subtopic is important	tag name	user makes a comment about the importance of a specific subtopic	While reading through and tagging documents, user makes a comment out loud about a tag being important	
TAG	subtopic saturation	tag name	Tag is still useful, but reached saturation on subtopic	user decides not to tag one that is too similar to prior document. Mmay comment about already tagging similar documents.	
TAG	tag indicates lack of interest	tag name	User applies a tag indicating lack of interest in document or subtopic	Document label like “junk” or “uninteresting”	only mark the first occurrence
TAG SORT	automated subtopic importance suggestion		user suggests algorithmic approach for sorting by subtopic importance	during the interview, the user describes an automated method for sorting subtopics by importance	

Section	Code	Additional	User Action	Example	Comment
TAG SORT	did not identify subtopic sort order		user did not figure out what the cluster sort order was	When asked if they noticed how the clusters were sorted, did not correctly answer by rarity (least populated cluster first)	
TAG SORT	disliked subtopics sorted by rarity		user indicated that they did not like having subtopics sorted by rarity	during the interview, the user says that they did not like having subtopics sorted by rarity	
TAG SORT	identified subtopic sort order		user figured out what the cluster sort order was	When asked if they noticed how the clusters were sorted, correctly answered that they were sorted by rarity (least populated cluster first)	
TAG SORT	liked subtopics sorted by rarity		user indicated that they liked having subtopics sorted by rarity	during the interview, the user says that they liked having subtopics sorted by rarity	
TAG SORT	manual subtopic importance suggestion		user suggests manual approach for sorting by subtopic importance	during the interview, the user describes a manual method for sorting subtopics by importance, such as pinning certain subtopics	
TAG SORT	skeptical about automating subtopic importance		user does not think it is possible to automate subtopic importance	specifically says it is not possible to have the system determine subtopic importance. Explanations could include that the tag label was not stable from day to day	

E.2 Sample of a Coded Session

The following example shows how the Daybreak data was coded. In this case, the coded data is from a fictitious semi-structured interview based on the types of content discussed in actual sessions.



E.3 Sample Daybreak Log File

This is an example of the log file generated by the Daybreak system from a session by the author on the Boston Red Sox topic. It includes the date and timestamps for participant actions, the day (1-5), the specific action taken by the participant, which document was impacted (if applicable), and which tag was used (if applicable).

	A	B	C	D	E	F
1	date	time	action	docno	tag	day
2	10/7/2023	12:40:28 PM	Day-2-start			day2
3	10/7/2023	12:40:32 PM	clustered			day2
4	10/7/2023	12:40:33 PM	chronological			day2
5	10/7/2023	12:40:37 PM	click-subtopic-header		ROCKY MOUNTAIN RED SOX	day2
6	10/7/2023	12:40:39 PM	viewed	sox-205		day2
7	10/7/2023	12:40:47 PM	tagged	sox-205	ROCKY MOUNTAIN RED SOX	day2
8	10/7/2023	12:40:49 PM	click-subtopic-header		MILB	day2
9	10/7/2023	12:40:50 PM	viewed	sox-206		day2
10	10/7/2023	12:40:54 PM	tagged	sox-206	MILB	day2
11	10/7/2023	12:41:08 PM	tagged	sox-206	MINOR LEAGUE	day2
12	10/7/2023	12:41:10 PM	click-subtopic-header		RED SOCKS	day2
13	10/7/2023	12:41:11 PM	viewed	sox-232		day2
14	10/7/2023	12:41:19 PM	tagged	sox-232	RED SOCKS	day2
15	10/7/2023	12:41:25 PM	click-subtopic-header		UNWRITTEN RULES	day2
16	10/7/2023	12:41:26 PM	viewed	sox-228		day2
17	10/7/2023	12:41:32 PM	tagged	sox-228	UNWRITTEN RULES	day2
18	10/7/2023	12:41:45 PM	tagged	sox-228	LAWSUIT	day2
19	10/7/2023	12:41:55 PM	click-subtopic-header		RED SOX VS RAYS	day2
20	10/7/2023	12:41:56 PM	viewed	sox-226		day2
21	10/7/2023	12:42:16 PM	tagged	sox-226	RED SOX VS RAYS	day2
22	10/7/2023	12:42:28 PM	tagged	sox-226	UPCOMING GAMES	day2

E.4 Daybreak Data Analysis Artifacts

This section includes tables describing sessions and analysis from the Daybreak user study sessions.

E.4.1 Estimates of Documents Viewed

The following table shows the participants' estimated numbers of documents viewed per day as indicated in the post-day questionnaires, compared with the actual numbers of documents that they viewed each day, and whether the participant's estimate was accurate, low (below the number they actually viewed) or high (above the number of documents viewed).

Participant	Day	Perceived Viewed	Actual Viewed	Estimate
BBL-201	day1	11 to 15	18	Low
BBL-201	day2	11 to 15	23	Very Low
BBL-201	day3	16 to 20	27	Very Low
BBL-201	day4	16 to 20	23	Low
BBL-201	day5	21 to 25	21	Accurate
BBL-887	day1	6 to 10	22	Very Low
BBL-887	day2	11 to 15	21	Very Low
BBL-887	day3	11 to 15	21	Very Low
BBL-887	day4	16 to 20	22	Low
BBL-887	day5	11 to 15	20	Very Low
BBL-927	day1	11 to 15	20	Low
BBL-927	day2	16 to 20	26	Low
BBL-927	day3	6 to 10	12	Low
BBL-927	day4	11 to 15	25	Very Low
BBL-927	day5	16 to 20	15	High
FIN-326	day1	11 to 15	11	Accurate
FIN-326	day2	11 to 15	14	Accurate
FIN-326	day3	16 to 20	13	High
FIN-326	day4	11 to 15	11	Accurate
FIN-326	day5	16 to 20	16	Accurate

Participant	Day	Perceived Viewed	Actual Viewed	Estimate
FIN-455	day1	11 to 15	15	Accurate
FIN-455	day2	11 to 15	23	Very Low
FIN-455	day3	11 to 15	20	Very Low
FIN-455	day4	16 to 20	20	Accurate
FIN-455	day5	11 to 15	30	Very Low
FIN-499	day1	16 to 20	11	Very High
FIN-499	day2	11 to 15	8	High
FIN-499	day3	11 to 15	12	Accurate
FIN-499	day4	6 to 10	9	Accurate
FIN-499	day5	11 to 15	10	High
FIN-940	day1	11 to 15	14	Accurate
FIN-940	day2	6 to 10	15	Very Low
FIN-940	day3	6 to 10	16	Very Low
FIN-940	day4	11 to 15	18	Low
FIN-940	day5	11 to 15	15	Accurate
HLT-409	day1	6 to 10	6	Accurate
HLT-409	day2	6 to 10	5	High
HLT-409	day3	11 to 15	10	High
HLT-409	day4	11 to 15	12	Accurate
HLT-409	day5	16 to 20	9	Very High
HLT-555	day1	11 to 15	19	Low
HLT-555	day2	11 to 15	22	Very Low
HLT-555	day3	16 to 20	28	Very Low
HLT-555	day4	16 to 20	38	Very Low
HLT-555	day5	16 to 20	38	Very Low
HLT-913	day1	16 to 20	18	Accurate
HLT-913	day2	11 to 15	17	Low
HLT-913	day3	21 to 25	17	High
HLT-913	day4	26+	14	Very High
HLT-913	day5	26+	17	Very High
SPC-259	day1	11 to 15	16	Low
SPC-259	day2	11 to 15	13	Accurate
SPC-259	day3	11 to 15	15	Accurate
SPC-259	day4	16 to 20	14	High
SPC-259	day5	11 to 15	11	Accurate
SPC-471	day1	11 to 15	24	Very Low
SPC-471	day2	11 to 15	23	Very Low
SPC-471	day3	16 to 20	31	Very Low
SPC-471	day4	16 to 20	32	Very Low
SPC-471	day5	16 to 20	33	Very Low

Participant	Day	Perceived Viewed	Actual Viewed	Estimate
SPC-688	day1	6 to 10	9	Accurate
SPC-688	day2	6 to 10	9	Accurate
SPC-688	day3	0 to 5	7	Low
SPC-688	day4	6 to 10	7	Accurate
SPC-688	day5	0 to 5	12	Very Low
WEA-093	day1	0 to 5	5	Accurate
WEA-093	day2	0 to 5	7	Low
WEA-093	day3	0 to 5	10	Very Low
WEA-093	day4	6 to 10	8	Accurate
WEA-093	day5	6 to 10	8	Accurate
WEA-367	day1	0 to 5	8	Low
WEA-367	day2	0 to 5	8	Low
WEA-367	day3	0 to 5	8	Low
WEA-367	day4	6 to 10	8	Accurate
WEA-367	day5	0 to 5	7	Low
WEA-842	day1	6 to 10	15	Very Low
WEA-842	day2	16 to 20	17	Accurate
WEA-842	day3	21 to 25	13	Very High
WEA-842	day4	21 to 25	14	Very High
WEA-842	day5	16 to 20	12	High

E.4.2 Daybreak Participants' Session Lengths

Participant	Session Length
BBL-201	1:35
BBL-887	1:40
BBL-927	1:50
FIN-326	1:33
FIN-455	1:39
FIN-499	1:50
HLT-409	2:20
HLT-555	1:30
HLT-913	1:54
SPC-259	1:50
SPC-471	1:30
SPC-688	1:42
WEA-093	2:20
WEA-367	1:46
WEA-842	1:31

E.4.3 Daybreak Participants' Session Lengths, by Topic

Topic	Average Length	Max	Min
Cryptocurrency	1:40	1:50	1:33
Extreme Weather	1:52	2:20	1:31
Global Health	1:54	2:20	1:30
Red Sox Baseball	1:41	1:50	1:35
Space	1:40	1:50	1:30

E.4.4 Daybreak Documents Viewed by Each Participant Per Day

Participant	Day 1	Day 2	Day 3	Day 4	Day 5	Total
BBL-201	18	23	27	23	21	112
BBL-887	22	21	21	22	20	106
BBL-927	20	26	12	25	15	98
FIN-326	11	14	13	11	16	65
FIN-455	15	23	20	20	30	108
FIN-499	11	8	12	9	10	50
HLT-409	6	5	10	12	9	42
HLT-555	19	22	28	38	38	145
HLT-913	18	17	17	14	17	83
SPC-259	16	13	15	14	11	69
SPC-471	24	23	31	32	33	143
SPC-688	9	9	7	7	12	44
WEA-093	5	7	10	8	8	38
WEA-367	8	8	8	8	7	39
WEA-842	15	17	13	14	12	71

E.4.5 Daybreak Participants' Total Numbers of Unique Tag Labels

Participant	Tag Labels
BBL-201	0
BBL-887	28
BBL-927	29
FIN-326	11
FIN-455	10
FIN-499	12
HLT-409	71
HLT-555	36
HLT-913	51
SPC-259	37
SPC-471	59
SPC-688	19
WEA-093	48
WEA-367	12
WEA-842	12

E.4.6 Daybreak Average Numbers of Tag Labels by Topic

Topic	Average Tag Labels	Max	Min
Cryptocurrency	11.0	12.0	10.0
Extreme Weather	24.0	48.0	12.0
Global Health	52.7	71.0	36.0
Red Sox Baseball	19.0	29.0	0.0
Space	38.3	59.0	19.0

E.4.7 Daybreak Participants' Unique Tag Labels by Topic

Cryptocurrency

ACCESS * AI * ANALYSIS * BITCOIN * BITCOIN ADOPTION * BITCOIN DANGERS * BITCOIN ENVIRONMENTAL IMPACT * BITCOIN INTERNATIONAL * BITCOIN OPPORTUNITY * BITCOIN VALUE * BLOCKSTREAM PARTNERSHIP * BUSINESSES GETTING INVOLVED * CPU HOG * CRYPTO * CRYPTO 101 * CRYPTO ACTIVISM * CRYPTO FUTURE * CRYPTO MINING * CRYPTO STOCKS * DANGERS * ETHEREUM * GLINT CURRENCY APP * HACKING * HACKS * HELPFUL * HISTORY * INVESTMENT STRATEGY * JAPAN * LAYMAN * MINING * NEW COIN * NEW CURRENCIES * NEW MARKET * PARTNERSHIP * QASH TOKEN * REGULATION * RISE OF CRYPTO * RISK * RUSSIA * SECURITY THREAT * SILVERTOKEN * STARTUP * TAXATION * YUCK

Extreme Weather

AFP * AUSTRALIA * BUTTERFLIES * CALIFORNIA * CAVE WALLS * CHINA * CLEAN ENERGY * CLIMATE CHANGE * CLIMATE CHANGE DEFORESTATION * CONSUMPTION * COP21 * DISASTER * DIVESTITURE * DROUGHT * EL NINO * EMISSIONS * ENVIRONMENT * EPA * EU * EXPRESSONLINE * EXTINCTION * FLOODS * FOOD AG * FOOD PRODUCTION * FOOD SHORTAGE * FORECAST * FOREST * FORESTS * FOSSIL FUELS * GLOBAL * GOP * HEAT * HEAT RELEASE * HEATWAVE * HUMANS * INDIA * INDOOR AIR QUALITY * JET STREAM * KATIE MCGINTY * MASS EXTINCTION * METHODOLOGY * OCEAN * OCEANS * PARIS SUMMIT * PENNSYLVANIA * POLICY * POLLUTION * POPULATION * PROTESTS * RAIN * RESILIENCE * RICH COUNTRIES * RISK * SOCIAL * STOCK WATCH * STORMS * TECHNOLOGY * THE AGE * THEGUARDIAN * UK * UN * UNITED KINGDOM * UNITED STATES * UNITEDSTATES * WORLD FOOD PROGRAM * WORLDOCEANASSESSMENT * XINHUA * ZAMBIA

Global Health

AFFORDABLE CARE ACT * AFGHANISTAN * AFRICA * ANIMAL HEALTH * ASIA * ATHLETES * AUSTRALIA * BIOBANK * BIRTH RATE * BLOG FODDER * BOOK AUTHOR * CAMBODIA * CANADA * CANCER * CHAD * CHILDHOOD * CHINA * CHRONIC DISEASE * CHRONIC ILLNESS * COMMUNITY BUILDING * COMMUNITY HEALTH * DENMARK * DEPRESSION * DEVICES * DIABETES * DIET * DIET INDUSTRY * DRUG RESISTANCE * EASTERN EUROPE * EHEALTH * EHR * EMERGING DISEASES * EMERGING RESEARCH * EMTECH * EPIDEMIOLOGY * EUROPE * EVENTS * FIRST NATIONS * FITNESS * FLU * FLUFF * FOOD * FOOD

SAFETY * FOOD SUPPLY * FOODBORNE ILLNESS * FRANCE * FUNDING * GERMANY * GLOBA * GLOBAL * GLOBAL HEALTH * GMO * GOVT * GRANT FUNDING * HEALTH * HEALTH CARE * HEALTH EHEALTH * HEALTH INSURANCE * HEALTH LAW * HEALTH OUTREACH * HEALTH REFORM * HEALTHCARE * HIT * HIV * HOMICIDE * HONG KONG * HOSPITALS * HUMAN TRAFFICKING * IMMUNOTHERAPY * INDIA * INDIGENOUS * INFECTIOUS DISEASE * INSURANCE * INVESTING * ISRAEL * LEADERSHIP ORG * LIFESTYLE * LIQUOR SALES * LISTERIA * LOBBYING * MENTAL HEALTH * MENTAL ILLNESS * MEXICO * MHEALTH * MIDDLE EAST * MILITARY * NEURO * NONPROFIT * NURSING HOMES * OCCUPATIONAL HEALTH * OPINION POLLING * OSHA * OUTBREAK * PAKISTAN * PARTISAN POLITICS * PEDIATRICS * PHARMA * POPULATION * POVERTY * PR * PRESCRIPTION DRUGS * PRESS RELEASE * PREVENTION * PTSD * PUBLIC HEALTH * QATAR * REFUGEES * REGS * RESEARCH FINDINGS * RUSSIA * SAFETY * SCOTLAND * SELF TRACKING * SENIORS * SEX ABUSE * SEXUALITY * SINGAPORE * SOCIAL JUSTICE * SOCIAL MEDIA * SPECIAL EVENT * STIGMA * STUDENTS * SUPPLY CHAIN * TECH * TECHNOLOGY * TOURISM * TRAUMA * TUBERCULOSIS * UAE * UK * UNITED STATES * UNKNOWN * UNRELATED * US * VACCINE * VETERANS * VIDEO GAMES * VIOLENCE * VIRUS * WEIGHT LOSS * WELLNESS * WORKPLACE WELLNESS

Red Sox Baseball

14 1 * 7 22 * AFTER BLOWOUT LOSS * ANALYSIS * BASEBALL VS OTHER * [BLANK] * BOGAERTS * BOSTON * BOX SCORE * BOX SCORE LOSS * BOX SCORES * BOXSCORE * BROCKSTAR * COLLAPSE * DREW * FARRELL * HAPP * HOLT * HUTCHISON * IRRELEVENT * JBJ * LACKEY * LESTER * LOSS * LOSS PITCHING * LOSS PITCHING PLAYOFF HOPES * LOSS PLAYOFF HOPES * LOSS RECAP * NAVA * NEWS * NON GAME * ORTIZ * PITCHING * PITCHING AFTER BLOWOUT LOSS * PITCHING LOSS * PITCHING PLAYOFF HOPES * PITCHING PUMMEL * PITCHING SAFETY * PLAYOFF HOPES * PUMMEL * PUMMEL PITCHING * REBUILD * RECORD * RECORDS * RECORDS PUMMLE * RED SOX * REGINA * RESULTS * ROSS * RULES * SCHEDULE ONLY * SOX HISTORY * STANDINGS * TOMASE * UNWRITTEN RULES LAWSUITS * WIN

Space

ARMSTRONG * ASTEROID * ASTEROID IMPACT * ASTEROID IMPACTS * BIOSTATISTICS * BJS * BLACK HOLE * BLANKET * CALI BOI TIP * CARNIVAL * CHINA * CHINA IN SPACE * CONNECTARAGRO * CONNEX ONE * COSMIC ORIGINS * CROPS * CUBESAT * CYCLONE * DEEP SPACE * DNA * DRONES * EACHPAI * ECLIPSE * ELON MUSK * ESA ASTEROID * ETHIOPIA * FACEBOOK * FALCON 9 * FCC * GLOBALSTAR * GNSS * GRAVITATIONAL WAVE * HUBBLE * HUGHES * INDIA * INFOPRO DIGITAL * INMARSAT * INTERNET * IOT *

IRRELEVANT * ISS * JAPAN * KIRK * LAUNCH SERVICE * LONG MARCH 11 * LOW COST * MACHINE LEARNING * MARS * MAXAR * METEROID STRIKE * MICRO MOON * MICROSCOPY * MOON * MOZAMBIQUE * NANOSATELLITE * NASA * NASA ASTEROID * NASA BUDGET * NASA MARS * NASA TELESCOPE * NEUTRON STAR * NRO * OPTICAL * PRIVATE INDUSTRY * REMOTE SENSING * RIO * ROBOTICS * ROCKET * SATCOM * SATELLITE * SATELLITES * SCIENCE FESTIVAL * SIMULATION * SMC * SPACE COAST * SPACE EXPLORATION * SPACE INNOVATION * SPACE PHOTOGRAPHY * SPACE SIMULATION * SPACE STATION * SPACE VEHICLE * SPACE WARFARE * SPACEX * SPACEX DELAY * SPECTRUM AUCTION * STAR FORMATION * STEM * STRATOSPHERIC * SUPERNOVA * SURVEILLANCE * TECHNOLOGY IMPROVEMENTS * TELESCOPE * UFO * USPTO TRADEMARK * WEATHER * WIFI * WORLDVIEW * YAHSAT INTERNET BRAZIL * YOUTH

E.4.8 Percentage of Tag Labels Applied Per Document

The following table shows the tag labels applied per each document for all participants. The participants near the top of the list had a smaller, potentially more curated set of tags that were applied to tagged documents. The participants near the bottom of the list applied multiple tag labels per document.

Participant	Tag Labels	Number of Tagged Docs	Tag Labels Per Document
BBL-201	0	53	0.0%
FIN-326	11	46	23.9%
FIN-455	10	37	27.0%
FIN-499	12	35	34.3%
WEA-842	12	35	34.3%
HLT-555	36	100	36.0%
BBL-927	29	56	51.8%
BBL-887	28	38	73.7%
HLT-913	51	58	87.9%
SPC-688	19	19	100.0%
WEA-367	12	12	100.0%
SPC-259	37	36	102.8%
SPC-471	59	53	111.3%
WEA-093	48	26	184.6%
HLT-409	71	30	236.7%

E.4.9 Percentage of Tags Applied Per Document Viewed

The following table includes the number of documents tagged per document viewed. Participants near the top of the list looked at multiple documents, but were more judicious in their application of tags; they applied tags to a relatively small percentage of documents viewed. Participants near the bottom of the list applied tags to a large percentage of the documents they viewed. Note that some participants may have been selective about which documents they viewed (e.g., choosing which documents to open based on the documents' titles).

Participant	Tagged Docs	Docs Viewed	Tags Per Document
FIN-455	37	95	38.9%
SPC-471	53	130	40.8%
BBL-887	38	87	43.7%
WEA-367	12	26	46.2%
SPC-688	19	38	50.0%
BBL-201	53	99	53.5%
WEA-842	35	61	57.4%
SPC-259	36	53	67.9%
BBL-927	56	77	72.7%
WEA-093	26	35	74.3%
FIN-499	35	46	76.1%
HLT-555	100	119	84.0%
HLT-913	58	68	85.3%
HLT-409	30	34	88.2%
FIN-326	46	51	90.2%

Bibliography

[1] Russell L. Ackoff. From Data to Wisdom. *Journal of Applied Systems Analysis*, 16(1):3–9, 1989.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[3] George A. Akerlof. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

[4] Jabir Alshehabi Al-Ani and Maria Fasli. Probabilistic Relational Supervised Topic Modelling using Word Embeddings. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2035–2043, December 2018.

[5] Nourah A. Al-Rossais. Improving Cold Start Stereotype-Based Recommendation Using Deep Learning. *IEEE Access*, 11:145781–145791, 2023. Conference Name: IEEE Access.

[6] Nourah A. Al-Rossais. Warming Up from Extreme Cold Start using Stereotypes with Dynamic User and Item Features. In *ACM Conference on Recommender Systems*, pages 103–108, 2023.

[7] Kendra Albright. Environmental Scanning: Radar for Success. *The Information Management Journal*, 38, January 2004.

[8] Lloyd Alexander. *The Book of Three*. Number 1 in The Chronicles of Prydain. 1964.

[9] Hesham Allam, Michael Bliemel, Omar Al Amir, Sandra Toze, Kavita Shah, and Enas Shoib. Collaborative Ontologies in Social Tagging Tools: A Literature Review of Natural Folksonomy. In *Seventh International Conference on Information Technology Trends (ITT)*, pages 126–130, November 2020.

[10] James Allan, Jaime G. Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study Final Report. January 1998. Publisher: Carnegie Mellon University.

[11] Dorit Alt. Students' Social Media Engagement and Fear of Missing Out (FOMO) in a Diverse Classroom. *Journal of Computing in Higher Education*, 29(2):388–410, August 2017.

[12] Licia Amichi, Aline Carneiro Viana, Mark Crovella, and Antonio A.F. Loureiro. Understanding Individuals' Proclivity for Novelty Seeking. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '20, pages 314–324, New York, NY, USA, November 2020. Association for Computing Machinery.

[13] Alia Amin, Jacco van Ossenbruggen, Lynda Hardman, and Annelies van Nispen. Understanding Cultural Heritage Experts' Information Seeking Needs. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 39–47, New York, NY, USA, 2008. ACM.

[14] Robert Anderson. The Rashomon Effect and Communication. *Canadian Journal of Communication*, 41(2):249–270, May 2016. Publisher: University of Toronto Press.

[15] Dimo Angelov. Top2Vec: Distributed Representations of Topics, August 2020. arXiv:2008.09470 [cs, stat].

[16] Cristiano Antonelli. The Business Governance of Localized Knowledge: An Information Economics Approach for the Economics of Knowledge. *Industry and Innovation*, 13(3):227–261, September 2006.

[17] Jaime Arguello. Aggregated Search. *Foundations and Trends in Information Retrieval*, 10(5):365–502, March 2017.

[18] Jaime Arguello, Wan-Ching Wu, Diane Kelly, and Ashlee Edwards. Task Complexity, Vertical Display and User Interaction in Aggregated Search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '12*, page 435, Portland, Oregon, USA, 2012. ACM Press.

[19] Aristotle. *Poetics*. 335 BCE.

[20] Brooke Auxier and Monica Anderson. Social Media Use in 2021, April 2021.

[21] Paul Bailey, Ahmad Emad, Ting Zhang, Qingshu Xie, and Emmanuel Sikali. Weighted and Unweighted Correlation Methods for Large-Scale Educational Assessment: wCorr Formulas, April 2018. Publisher: American Institutes for Research.

[22] Sarah Baker and Rosalind Edwards. How Many Qualitative Interviews is Enough? Expert Voices and Early Career Reflections on Sampling and Cases in Qualitative Research. *National Centre for Research Methods Review Paper*, 2017.

[23] Frank Bannister and Dan Remenyi. Multitasking: The Uncertain Impact of Technology on Knowledge Workers and Managers. *Electronic Journal of Information Systems Evaluation*, 12(1):pp1-12, January 2009.

[24] Sarit Barzilai and Yoram Eshet-Alkalai. The Role of Epistemic Perspectives in Comprehension of Multiple Author Viewpoints. *Learning and Instruction*, 36:86–103, April 2015.

[25] Mirit Barzillai, Jenny Thomson, Sascha Schroeder, and Paul van den Broek. *Learning to Read in a Digital World*. John Benjamins Publishing Company, Amsterdam/Philadelphia, Netherlands, 2018.

[26] Enric Bas. Horizon Scanning. In Enric Bas, editor, *Sharing and Collaborative Economy: Future Scenarios, Technology, Creativity and Social Innovation*, SpringerBriefs in Economics, pages 13–39. Springer International Publishing, 2022.

[27] Marcia J. Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, January 1989. Publisher: MCB UP Ltd.

[28] Marcia J. Bates. Information and Knowledge: An Evolutionary Framework for Information Science. *Information Research: An International Electronic Journal*, 10(4), July 2005.

[29] Katinka Beker, Dietsje Jolles, Robert F. Lorch, and Paul van den Broek. Learning from Texts: Activation of Information from Previous Texts During Reading. *Reading and Writing*, 29(6):1161–1178, June 2016.

[30] N.J. Belkin, R.N. Oddy, and H.M. Brooks. ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2):61–71, January 1982. Publisher: MCB UP Ltd.

[31] David Bennet and Alex Bennet. The Depth of Knowledge: Surface, Shallow or Deep? *VINE*, 38(4):405–420, October 2008.

[32] Bettina Berendt and Christoph Hanser. Tags are not metadata, but "just more content" – to some people. *ICWSM*, 2007.

[33] Ofer Bergman, Noa Gradovitch, Judit Bar-Ilan, and Ruth Beyth-Marom. Folder Versus Tag Preference in Personal Information Management. *Journal of the American Society for Information Science and Technology*, 64(10):1995–2012, 2013.

[34] Theo JD Bothma and Henning Bergenholtz. "Information Needs Changing over Time": A Critical Discussion. *South African Journal of Libraries and Information Science*, 79(1), July 2013.

[35] Paul Bouchaud, David Chavalarias, and Maziyar Panahi. Crowdsourced Audit of Twitter's Recommender Systems. *Scientific Reports*, 13(16815), October 2023.

[36] Vitaly Brazhkin. "I Have Just Returned from the Moon": Online Survey Fraud. *Supply Chain Management: An International Journal*, 25(4):489–503, March 2020.

[37] Bruce K. Britton and Arthur C. Graesser. *Models of Understanding Text*. Psychology Press, February 2014.

[38] Andrei Broder. A Taxonomy of Web Search. *ACM SIGIR Forum*, 36(2):3, September 2002.

[39] Marc Bron, Jasmijn van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. Aggregated Search Interface Preferences in Multi-session Search Tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 123–132, New York, NY, USA, 2013. ACM.

[40] Jerome Bruner. The Narrative Construction of Reality. *Critical Inquiry*, 18(1):1–21, 1991.

[41] Michael Buckland and Fredric Gey. The Relationship between Recall and Precision. *Journal of the American Society for Information Science*, 45(1):12–19, 1994.

[42] Michael K. Buckland. Information as Thing. *Journal of the American Society for Information Science*, 42(5):351–360, 1991.

[43] Michael K. Buckland. What Is a "Document"? *Journal of the American Society for Information Science*, 48(9):804–809, 1997.

[44] Sarah Burton. Duty of Expression. *Resonance Magazine*, November 2003.

[45] Vannevar Bush. As We May Think. *The Atlantic Monthly*, July 1945.

[46] Tony Buzan and Barry Buzan. *The Mind Map Book*. Pearson Education, 2006.

[47] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136, Corvalis Oregon USA, June 2007. ACM.

[48] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or There. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 16–27, Berlin, Heidelberg, 2008. Springer.

[49] Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. Evaluating Retrieval over Sessions: The TREC Session Track 2011-2014. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 685–688, New York, NY, USA, 2016. ACM.

[50] Deb Chachra. *How Infrastructure Works: Inside The Systems That Shape Our World*. Penguin Random House, October 2023.

[51] Praveen Chandar, Brian St. Thomas, Lucas Maystre, Vijay Pappu, Roberto Sanchis-Ojeda, Tiffany Wu, Ben Carterette, Mounia Lalmas, and Tony Jebara. Using Survival Models to Estimate User Engagement in Online Experiments. In *Proceedings of the ACM Web Conference 2022*, pages 3186–3195, New York, NY, USA, April 2022.

[52] Sam Charrington. Reinforcement Learning for Personalization at Spotify with Tony Jebara. The TWIML AI Podcast with Sam Charrington.

[53] Jing Chen. Information Theory and Market Behavior. *SSRN Electronic Journal*, 2004.

[54] Michelene T. H. Chi, Robert Glaser, and Marshall J. Farr, editors. *The Nature of Expertise*. L. Erlbaum Associates, Hillsdale, N.J, 1988.

[55] Chun Wei Choo. The Art of Scanning the Environment. *Reframing Environmental Scanning*, 7, 2003.

[56] Andrea Civan, William Jones, Predrag Klasnja, and Harry Bruce. Better to organize personal information by folders or by tags?: The devil is in the details. *Proceedings of the American Society for Information Science and Technology*, 45(1):1–13, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/meet.2008.1450450214.

[57] Dan PA Clark and Davide Bruno. Time Is of the Essence: Exploring Temporal and Spatial Organisation in Episodic Memory. *Quarterly Journal of Experimental Psychology*, 74(8):1406–1417, August 2021.

[58] Andy Cockburn, Carl Gutwin, Joey Scarr, and Sylvain Malacria. Supporting Novice to Expert Transitions in User Interfaces. *ACM Computing Surveys*, 47(2):31:1–31:36, November 2014.

[59] Harry Collins. Studies of Expertise and Experience. *Topoi*, July 2016.

[60] Carmela Comito, Agostino Forestiero, and Clara Pizzuti. Word Embedding Based Clustering to Detect Topics in Social Media. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 192–199, New York, NY, USA, October 2019. Association for Computing Machinery.

[61] W. S. Cooper. A Definition of Relevance for Information Retrieval. *Information Storage and Retrieval*, 7(1):19–37, June 1971.

[62] William S. Cooper. On Selecting a Measure of Retrieval Effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, April 1973.

[63] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 4th edition, December 2014.

[64] Scott Counts and Kristie Fisher. Taking It All In? Visual Attention in Microblog Consumption. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[65] Daniel Coyle. *The Culture Code*. Penguin Random House, 2018.

[66] K. J. W. Craik. *The Nature of Explanation*. University Press, Macmillan, Oxford, England, 1943.

[67] Lorrie Faith Cranor and Rebecca N. Wright. Influencing Software Usage. In *Proceedings of the Tenth Conference on Computers, Freedom and Privacy*, pages 45–55, Toronto Ontario Canada, April 2000. ACM.

[68] John W. Creswell. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Ltd., 2nd edition, 2003.

[69] Danny Crichton. RSS is undead.

[70] Brenda Dervin. Sense-making Theory and Practice: An Overview of User Interests in Knowledge Seeking and Use. *Journal of Knowledge Management*, 2(2):36–46, 1998.

[71] Brenda Dervin. Dervin's Sense-making Theory. In *Information Seeking Behavior and Technology Adoption: Theories and Trends*, pages 59–80. IGI Global, 2015.

[72] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, and Fernando Diaz. Towards Recency Ranking in Web Search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 11–20, New York, NY, USA, February 2010. Association for Computing Machinery.

[73] Larry Dossey. FOMO, Digital Dementia, and Our Dangerous Experiment. *Explore: The Journal of Science and Healing*, 10(2):69–73, March 2014.

[74] Karen M. Drabenstott. Do Nondomain Experts Enlist the Strategies of Domain Experts? *Journal of the American Society for Information Science and Technology*, 54(9):836–854, 2003. Project.

[75] Erwan Dujeancourt and Marcel Garz. The Effects of Algorithmic Content Selection on User Engagement with News on Twitter. *The Information Society*, 39(5):263–281, October 2023.

[76] Georges Dupret and Mounia Lalmas. Absence Time and User Engagement: Evaluating Ranking Functions. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 173–182, New York, NY, USA, February 2013.

[77] Miles Efron, Peter Organisciak, and Katrina Fenlon. Improving Retrieval of Short Texts through Document Expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 911–920, New York, NY, USA, August 2012. Association for Computing Machinery.

[78] K. Anders Ericsson, Ralf T. Krampe, and Clemens Tesch-Römer. The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review*, 100(3):363–406, July 1993.

[79] Thomas Hylland Eriksen. *Tyranny of the Moment: Fast and Slow Time in the Information Age*. Pluto Press, London, 2001.

[80] Michelle Falter, Aaron A. Arenas, Gordon W. Maples, Chelsea T. Smith, Lisa J. Lamb, Michael G. Anderson, Elizabeth M. Uzzell, Laura E. Jacobs, Xavier L. Cason, Tiara A. N. Griffis, Megan Polzin, and Nada Z. Wafa. Making Room for Zoom in Focus Group Methods: Opportunities and Challenges for Novice Researchers (During and Beyond COVID-19). *Forum: Qualitative Social Research*, 23(1), January 2022. Number: 1.

[81] Weiguo Fan, Michael D. Gordon, and Praveen Pathak. An Integrated Two-Stage Model for Intelligent Information Routing. *Decision Support Systems*, 42(1):362–374, October 2006.

[82] Hui Fang and ChengXiang Zhai. An Exploration of Axiomatic Approaches to Information Retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 480–487, New York, NY, USA, 2005. ACM.

[83] Jon Fingas. Alto Mail is shutting down now that AOL is part of Oath, October 2017.

[84] Jonathan G. Fiscus and George R. Doddington. Topic Detection and Tracking Evaluation Overview. In James Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 17–31. Springer US, Boston, MA, 2002.

[85] Linda Flower, Victoria Stein, John Ackerman, Margaret J. Kantz, Kathleen McCormick, and Wayne Peck. *Reading-to-Write: Exploring a Cognitive and Social Process*. Oxford University Press, 1990.

[86] Linda S. Flower and John R. Hayes. Problem-Solving Strategies and the Writing Process. *College English*, 39(4):449–461, 1977.

[87] Allen Foster and Nigel Ford. Serendipity and Information Seeking: An Empirical Study. *Journal of Documentation*, 59(3):321–340, June 2003.

[88] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. Newsjunkie: Providing Personalized Newsfeeds Via Analysis of Information Novelty. In *Proceedings of the 13th International Conference on World Wide Web*, pages 482–490, New York, NY, USA, May 2004. Association for Computing Machinery.

[89] Vincent Galante. Design and Dialectic, April 2018. Publisher: OCAD University.

[90] Nicola K. Gale, Gemma Heath, Elaine Cameron, Sabina Rashid, and Sabi Redwood. Using the Framework Method for the Analysis of Qualitative Data in Multi-disciplinary Health Research. *BMC Medical Research Methodology*, 13(1):117, September 2013.

[91] Ning Gao and Douglas Oard. A Head-Weighted Gap-Sensitive Correlation Coefficient. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 799–802, Santiago Chile, August 2015. ACM.

[92] Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. Abstractive vs. Extractive Summarization: An Experimental Review. *Applied Sciences*, 13(13):7620, January 2023. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[93] Tarleton Gillespie. The Relevance of Algorithms. January 2014.

[94] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Observations. Aldine Publishing, Chicago, 1967.

[95] Johann Wolfgang von Goethe. *Faust*. 1831.

[96] Laurie J. Goldsmith. Using Framework Analysis in Applied Qualitative Research. *Qualitative Report*, 26(6):2061–2076, June 2021.

[97] Daniel Goleman. *Focus: The Hidden Driver of Excellence*. 2013.

[98] Gregory Goth. Turning Data Into Knowledge. *Communications of the ACM*, 53(11):13–15, November 2010.

[99] John D. Gould and Clayton Lewis. Designing for Usability: Key Principles and What Designers Think. *Communications of the ACM*, 28(3):300–311, March 1985.

[100] GPAI 2022. Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action, November 2022.

[101] Jay Graber. Algorithmic choice, March 2023.

[102] Arthur C. Graesser, Danielle S. McNamara, and Max M. Louwerse. What Do Readers Need to Learn in Order to Process Coherence Relations in Narrative and Expository Text. *Rethinking Reading Comprehension*, 82:98, 2003.

[103] Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013. Publisher: [Oxford University Press, Society for Political Methodology].

[104] David A. Grossman and Ophir Frieder. *Information Retrieval: Algorithms and Heuristics*, volume 15. Springer Science & Business Media, 2004.

[105] Michael Gruninger, Olivier Bodenreider, Frank Olken, Leo Obrst, and Peter Yim. Ontology Summit 2007-Ontology, taxonomy, folksonomy: Understanding the distinctions. *Applied Ontology*, 3(3):191, 2008.

[106] A. Gulli. The Anatomy of a News Search Engine. In *The 14th International Conference on World Wide Web*, pages 880–881, New York, NY, USA, May 2005. Association for Computing Machinery.

[107] Susan Haack. Post "Post-Truth": Are We There Yet? *Theoria*, 85(4):258–275, 2019.

[108] Christina Haas and Linda Flower. Rhetorical Reading Strategies and the Construction of Meaning. *College Composition and Communication*, 39(2):167–183, 1988.

[109] Judith M. Harackiewicz and Chris S. Hulleman. The Importance of Interest: The Role of Achievement Goals and Task Values in Promoting the Development of Interest. *Social and Personality Psychology Compass*, 4(1):42–52, February 2010.

[110] Marti A. Hearst, David R. Karger, and Jan O. Pedersen. Scatter/Gather as a Tool for the Navigation of Retrieval Results. In *Working Notes of the AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, Cambridge, MA, 1995.

[111] Karl G. Heider. The Rashomon Effect: When Ethnographers Disagree. *American Anthropologist*, 90(1):73–81, 1988.

[112] Jean Helms Mills, Amy Thurlow, and Albert J. Mills. Making Sense of Sensemaking: The Critical Sensemaking Approach. *Qualitative Research in Organizations and Management: An International Journal*, 5(2):182–195, August 2010.

[113] Jonathan Hendrickx and Heritiana Ranaivoson. Why and How Higher Media Concentration Equals Lower News Diversity – The Mediahuis Case. *Journalism*, 22(11):2800–2815, November 2021.

[114] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research*, 27(4):591–608, March 2017. Publisher: SAGE Publications Inc.

[115] C. B. Hensley. Selective Dissemination of Information (SDI): State of the Art in May, 1963. In *Proceedings of the 1963 Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 257–262, New York, NY, USA, May 1963. Association for Computing Machinery.

[116] Suzanne Hidi and William Baird. Strategies for Increasing Text-Based Interest and Students' Recall of Expository Texts. *Reading Research Quarterly*, 23(4):465–483, 1988.

[117] Suzanne Hidi and K. Ann Renninger. The Four-Phase Model of Interest Development. *Educational Psychologist*, 41(2):111–127, 2006.

[118] Nathan O. Hodas and Kristina Lerman. Attention and Visibility in an Information Rich World. *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, July 2013.

[119] Donna L. Hoffman and Thomas P. Novak. Flow Online: Lessons Learned and Future Prospects. *Journal of Interactive Marketing*, 23(1):23–34, February 2009.

[120] Kartik Hosanagar. Usercentric Operational Decision Making in Distributed Information Retrieval. *Information Systems Research*, 22(4):739–755, 2011.

[121] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc., 2021.

[122] John Hulland. Conceptual Review Papers: Revisiting Existing Research to Develop and Refine Theory. *AMS Review*, 10(1):27–35, June 2020.

[123] Ferenc Huszár, Sofia Ira Ktena, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic Amplification of Politics on Twitter, December 2021. Proceedings of the National Academy of Sciences.

[124] Sitwala Imenda. Is There a Conceptual Difference between Theoretical and Conceptual Frameworks? *Journal of Social Sciences*, 38(2), October 2017.

[125] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys*, 47(4):67:1–67:38, June 2015.

[126] Giacomo Inches, Mark James Carman, and Fabio Crestani. Investigating the Statistical Properties of User-Generated Documents. In Henning Christiansen, Guy De Tré, Adnan Yazici, Slawomir Zadrozny, Troels Andreasen, and Henrik Legind Larsen, editors, *Flexible Query Answering Systems*, Lecture Notes in Computer Science, pages 198–209, Berlin, Heidelberg, 2011. Springer.

[127] Julie A. Jacko. *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition*. CRC Press, May 2012.

[128] Ken Jennings. *Brainiac: Adventures in the Curious, Competitive, Compulsive World of Trivia Buffs*. Random House Publishing Group, October 2007.

[129] Tao Jia, Dashun Wang, and Boleslaw K. Szymanski. Quantifying Patterns of Research-Interest Evolution. *Nature Human Behaviour*, 1(4):1–7, March 2017.

[130] Clay A. Johnson. *The Information Diet: A Case for Conscious Consumption*. O'Reilly Media, Inc., January 2012.

[131] Philip Nicholas Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, 1983.

[132] Natalie A. Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society*, 16(1), 2011.

[133] Larry Kahaner. *Competitive Intelligence: How To Gather Analyze And Use Information To Move Your Business To The Top*. Simon and Schuster, 1997. Google-Books-ID: K3QfGoGSzmoC.

[134] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, October 2011.

[135] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives*, 5(1):193–206, March 1991.

[136] Sanjay Kairam, Meredith Morris, Jaime Teevan, Dan Liebling, and Susan Dumais. Towards Supporting Search over Trending Events with Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):283–292, 2013. Number: 1.

[137] Min Jeong Kang, Ming Hsu, Ian M. Krajbich, George F. Loewenstein, Samuel M. McClure, Joseph Tao-yi Wang, and Colin F. Camerer. The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory. *SSRN Electronic Journal*, 2008.

[138] Minhyung Kang and Young-Gul Kim. A Multilevel View on Interpersonal Knowledge Transfer. *Journal of the American Society for Information Science and Technology*, 61(3):483–494, March 2010.

[139] Abraham Kaplan, editor. *The Conduct of Inquiry: Methodology for Behavioural Science*. Routledge, New Brunswick, N.J. London, 1964.

[140] Jeremy Kaplan. You've got Mail! AOL launches new Alto mail program, October 2012.

[141] Stephanie Kelter, Barbara Kaup, and Berry Claus. Representing a Described Sequence of Events: A Dynamic View of Narrative Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):451–464, 2004. Place: US Publisher: American Psychological Association.

[142] Pooja Kherwa and Poonam Bansal. Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), July 2019.

[143] Søren Kierkegaard. *Philosophical Fragments, or, a Fragment of Philosophy*. 1844.

[144] Doha Kim, Yeosol Song, Songyie Kim, Sewang Lee, Yanqin Wu, Jungwoo Shin, and Daeho Lee. How Should the Results of Artificial Intelligence be Explained to Users? - Research on Consumer Preferences in User-Centered Explainable Artificial Intelligence. *Technological Forecasting and Social Change*, 188:122343, March 2023.

[145] Kyung-Sun Kim and Sei-Ching Joanna Sin. Perceived Usefulness of Social Media Features/Elements: Effects of Coping Style, Purpose and System. *Proceedings of the Association for Information Science and Technology*, 54(1):722–723, 2017.

[146] Kyung-Sun Kim, Sei-Ching Joanna Sin, and Yuqi He. Information Seeking through Social Media: Impact of User Characteristics on Social Media Use. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–4, 2013.

[147] W. W. Koczkodaj and J. Szybowski. Pairwise Comparisons Simplified. *Applied Mathematics and Computation*, 253:387–394, February 2015.

[148] Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Aggregated Search: A New Information Retrieval Paradigm. *ACM Computing Surveys*, 46(3):1–31, January 2014.

[149] Andreas Krapp. Structural and Dynamic Aspects of Interest Development: Theoretical Considerations from an Ontogenetic Perspective. *Learning and Instruction*, 12(4):383–409, August 2002.

[150] Michael Kubovy. Lives as Collections of Strands: An Essay in Descriptive Psychology. *Perspectives on Psychological Science*, 15(2):497–515, March 2020.

[151] Nikolas Kuhlen and Andrew Preston. News Concentration, June 2023.

[152] Juhi Kulshrestha, Muhammad Zafar, Lisette Noboa, Krishna Gummadi, and Saptarshi Ghosh. Characterizing Information Diets of Social Media Users. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):218–227, 2015.

[153] Mounia Lalmas. Aggregated Search. In Massimo Melucci and Ricardo Baeza-Yates, editors, *Advanced Topics in Information Retrieval*, The Information Retrieval Series, pages 109–123. Springer, Berlin, Heidelberg, 2011.

[154] Lars Bo Larsen, Tina Øvad, Lene Nielsen, and Marta Larusdottir. Remote User Testing: Experiences and Trends. *Human-Computer Interaction – INTERACT 2021*, V:579–583, August 2021.

[155] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. Chapter 1 - Introduction to HCI research. In Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser, editors, *Research Methods in Human Computer Interaction (Second Edition)*, pages 1–24. Morgan Kaufmann, Boston, January 2017.

[156] Jeffery A. LePine and Adelaide Wilcox King. Developing Novel Theoretical Insight from Reviews of Existing Theory and Research. *Academy of Management Review*, 35(4):506–509, October 2010. Publisher: Academy of Management.

[157] Lukas Lerche. Using Implicit Feedback for Recommender Systems: Characteristics, Applications, and Challenges, 2016.

[158] David M. Levy. Topics in Document Research. In *Proceedings of the ACM Conference on Document Processing Systems*, pages 187–193, Santa Fe, New Mexico, United States, 1988. ACM Press.

[159] Stephan Lewandowsky, Ronald E. Robertson, and Renee DiResta. Challenges in Understanding Human-Algorithm Entanglement During Online Information Consumption. *Perspectives on Psychological Science*, 2023.

[160] LexisNexis. Nexis Uni: Academic Research Tool for Universities & Libraries, January 2024.

[161] Kwan Hui Lim and Amitava Datta. A Topological Approach for Detecting Twitter Communities with Common Interests. In Martin Atzmueller, Alvin Chin, Denis Helic, and Andreas Hotho, editors, *Ubiquitous Social Media Analysis*, Lecture Notes in Computer Science, pages 23–43, Berlin, Heidelberg, 2013. Springer.

[162] Lyes Limam, David Coquil, Harald Kosch, and Lionel Brunie. Extracting User Interests from Search Query Logs: A Clustering Approach. In *2010 Workshops on Database and Expert Systems Applications*, pages 5–9, August 2010. ISSN: 2378-3915.

[163] Linguistic Data Consortium. TDT 2004: Annotation Manual, August 2004.

[164] Live Talks LA. V.E. Schwab in Conversation with J. Elle, September 2023.

[165] Mark Lochrie and Paul Coulton. Mobile Phones as Second Screen for TV, Enabling Inter-audience Interaction. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, page 1, Lisbon, Portugal, 2011. ACM Press.

[166] George Loewenstein. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological bulletin*, 116(1):75, 1994.

[167] Valentina Maccatrozzo. Burst the Filter Bubble: Using Semantic Web to Enable Serendipity. In *The Semantic Web – ISWC 2012*, pages 391–398. Springer, 2012.

[168] Craig Macdonald, Charles Clarke, and Jun Wang. The 1st International Workshop on Diversity in Document Retrieval. *ACM SIGIR Forum*, 45(2):87–93, January 2012.

[169] Helen Macdonald. The Things I Tell Myself When I'm Writing About Nature, August 2020.

[170] I. Scott MacKenzie. *Human-Computer Interaction: An Empirical Research Perspective*. Elsevier, January 2024. Google-Books-ID: f1vbEAAAQBAJ.

[171] Ken Manktelow and Man Cheung Chung. *Psychology of Reasoning: Theoretical and Historical Perspectives*. Psychology Press, September 2004.

[172] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008.

[173] James G. March and Herbert A. Simon. Cognitive Limits on Rationality. In *Organizations*. Wiley-Blackwell, Cambridge, Mass., USA, 2nd edition edition, May 1993.

[174] Eric Margolis and Stephen. Laurence. *Concepts: Core readings*. MIT Press, Cambridge, Mass., 1999.

[175] M. E. Maron and J. L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, 7(3):216–244, July 1960.

[176] Vivian B. Martin. The Relationship between an Emerging Grounded Theory and the Existing Literature: Four phases for consideration | Grounded Theory Review. *Grounded Theory Review*, (2/3), June 2006.

[177] Tumelo Maungwa and Ina Fourie. Competitive Intelligence Failures: An Information Behaviour Lens to Key Intelligence and Information Needs. *Aslib Journal of Information Management*, 70(4):367–389, January 2018.

[178] Jon McAuliffe and David Blei. Supervised Topic Models. *Advances in Neural Information Processing Systems*, 20, 2007.

[179] Lori McCay-Peet and Elaine G. Toms. The Serendipity Quotient. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4, 2011.

[180] Matthew J. McLaughlin and Kristin L. Sainani. Bonferroni, Holm, and Hochberg Corrections: Fun Names, Serious Changes to P Values. *PM&R*, 6(6):544–546, 2014.

[181] Massimo Melucci and Ricardo Baeza-Yates. *Advanced Topics in Information Retrieval*. Springer Science & Business Media, June 2011.

[182] Robert K. Merton. *On Theoretical Sociology: Five Essays, Old and New*. Free Press, October 1967.

[183] Marvin Minsky. A Framework for Representing Knowledge. *MIT AI Laboratory*, June 1974.

[184] Ana M. Moreno, Ahmed Seffah, Rafael Capilla, and Maria-Isabel Sánchez-Segura. HCI Practices for Building Usable Software. *Computer*, 46(4):100–102, April 2013. Conference Name: Computer.

[185] Kaweh Djafari Naini, Ricardo Kawase, Nattiya Kanhabua, Claudia Niederée, and Ismail Sengor Altingovde. Those Were the Days: Learning to Rank Social Media Posts for Reminiscence. *Information Retrieval Journal*, August 2018.

[186] Peter M. Nardi. *Doing Survey Research: A Guide to Quantitative Methods*. Routledge, January 2018.

[187] Raymond S. Nickerson. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2):175–220, June 1998.

[188] Douglas W. Oard and Philip Resnik. Support for interactive document selection in cross-language information retrieval. *Information Processing & Management*, 35(3):363–379, May 1999.

[189] Judith S. Olson and Wendy Kellogg, editors. *Ways of Knowing in HCI*. Springer, New York, 2014.

[190] Eli Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, May 2011. used.

[191] P. Davud Pearson and Gina Cervetti. Fifty Years of Reading Comprehension Theory and Practice. In P. David Pearson and Elfrieda H. Hiebert, editors, *Research-Based Practices for Teaching Common Core Literacy*, pages 1–24. January 2015.

[192] William R. Penuel, Barry J. Fishman, Britte Haugan Cheng, and Nora Sabelli. Organizing Research and Development at the Intersection of Learning, Implementation, and Design. *Educational Researcher*, 40(7):331–337, October 2011.

[193] Adam Perer and Ben Shneiderman. Systematic Yet Flexible Discovery: Guiding Domain Experts Through Exploratory Data Analysis. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, pages 109–118, New York, NY, USA, 2008. ACM.

[194] Jay Peters. Bluesky rolls out feeds with custom algorithms, May 2023.

[195] Karen E. Pettigrew and Lynne (E.F.) McKechnie. The Use of Theory in Information Science Research. *Journal of the American Society for Information Science and Technology*, 52(1):62–73, 2001.

[196] Joe Pinsker. A Behavioral Economist Tries to Fix Email, March 2017. The Atlantic.

[197] Paul R. Pintrich. The Role of Goal Orientation in Self-Regulated Learning. In Monique Boekaerts, Paul R. Pintrich, and Moshe Zeidner, editors, *Handbook of Self-Regulation*, pages 451–502. Academic Press, San Diego, January 2000.

[198] Peter Pirolli and Stuart Card. Information Foraging in Information Access Environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 51–58, Denver, Colorado, United States, 1995. ACM Press.

[199] Peter Pirolli and Stuart Card. Information Foraging. *Psychological Review*, 106(4):643–675, October 1999.

[200] Peter Pirolli and Stuart K. Card. Information Foraging Models of Browsers for Very Large Document Spaces. In *Proceedings of the working conference on Advanced visual interfaces*, pages 83–93, L'Aquila Italy, May 1998. ACM.

[201] Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 213–220, Vancouver, British Columbia, Canada, 1996. ACM Press.

[202] Matthew Pittman and Brandon Reich. Social Media and Loneliness: Why an Instagram Picture May be Worth More Than a Thousand Twitter Words. *Computers in Human Behavior*, 62:155–167, September 2016.

[203] Plato. *The Republic*. C. H. Kerr, 375 BCE.

[204] Philip M. Podsakoff, Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, 88(5):879, October 2003.

[205] Rüdiger F. Pohl. *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Psychology Press, December 2012.

[206] Michael I. Posner. Attention as a Cognitive and Neural System. *Current Directions in Psychological Science*, 1(1):11–14, 1992.

[207] Neil Postman. *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. Penguin Books, 1985.

[208] Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1169, August 2016.

[209] Chandra Prabha, Lynn Silipigni Connaway, Lawrence Olszewski, and Lillie R. Jenkins. What is Enough? Satisficing Information Needs. *Journal of Documentation*, 63(1):74–89, January 2007.

[210] Manuel A. Pérez-Quiñones. Personal Information Management. *Interactions*, 24(2):14–15, February 2017.

[211] Haoliang Qi, Mu Li, Jianfeng Gao, and Sheng Li. Information Retrieval for Short Documents. *Journal of Electronics (China)*, 23(6):933–936, November 2006.

[212] Matthew Rabin and Richard H. Thaler. Anomalies: Risk Aversion. *Journal of Economic Perspectives*, 15(1):219–232, March 2001.

[213] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. Getting to Know You: Learning New User Preferences in Recommender Systems. In *Proceedings of the 7th International Conference on Intelligent User Interfaces*, IUI '02, pages 127–134, New York, NY, USA, January 2002. Association for Computing Machinery.

[214] Berry G. Richards, Christine M. Roysdon, and Sharon M. Siegler. Manual Sdi Services in an Academic Library. *Science & Technology Libraries*, 2(1):31–42, October 1981.

[215] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 42–49, New York, NY, USA, November 2004. Association for Computing Machinery.

[216] Gary L. Rogers. *Nitrifier Populations and Kinetics in Selected Aquaculture Water Reuse Biofilters*. Ph.D., Colorado State University, United States – Colorado, 1984. ISBN: 9798205176293.

[217] Kristine M. Rogers. User Preferences for Organizing Social Media Feeds. In Gabriele Meiselwitz, editor, *Social Computing and Social Media: Design, User Experience and Impact*, Lecture Notes in Computer Science, pages 185–204, HCI International Conference, 2022. Springer International Publishing.

[218] Kristine M. Rogers, Emily Trubey, and Douglas W. Oard. A User Study in a Pandemic: Some Lessons Learned. *Proceedings of the Association for Information Science and Technology*, 59(1):785–787, 2022.

[219] Rebecca A. Rogers. *Using Rogers's Theory of Perceived Attributes to Address Barriers to Educational Technology Integration*. Ph.D., Walden University, United States – Minnesota, 2007. ISBN: 9780549200789.

[220] Lior Rokach and Slava Kisilevich. Initial Profile Generation in Recommender Systems Using Pairwise Comparison. *IEEE Transactions on Systems, Man, and Cybernetics*, 42(6):1854–1859, November 2012. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).

[221] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic Objects in Natural Categories. *Cognitive Psychology*, 8(3):382–439, July 1976.

[222] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic Keyword Extraction from Individual Documents. In *Text Mining*, pages 1–20. John Wiley & Sons, Ltd, 2010.

[223] Joshua S. Rubinstein, David E. Meyer, and Jeffrey E. Evans. Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):763–797, August 2001.

[224] Douglas. Rushkoff. *Present shock: When everything happens now*. Current, New York, New York, U.S.A., 2013.

[225] Ruth Page. Seriality and Storytelling in Social Media. *Storyworlds: A Journal of Narrative Studies*, 5:31, 2013.

[226] Ian Ruthven. Interactive Information Retrieval. *Annual Review of Information Science and Technology*, 42:43–92, 2008.

[227] Neil Sadler. *Fragmented Narrative: Telling and Interpreting Stories in the Twitter Age*. Routledge, 2021.

[228] Mukesh Kumar Saini, Fatimah Al-Zamzami, and Abdulmotaleb El Saddik. Towards Storytelling by Extracting Social Information from OSN Photo's Metadata. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, WISMM '14, pages 15–20, New York, NY, USA, 2014. ACM.

[229] Gerard Salton. The Evaluation of Automatic Retrieval Procedures—Selected Test Results Using the SMART System. *American Documentation*, 16(3):209–222, 1965.

[230] Gerard Salton and Christopher Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523, January 1988.

[231] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Computer Science Series. McGraw-Hill, New York, 1983.

[232] William Samuelson and Richard Zeckhauser. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1(1):7–59, March 1988.

[233] Elizeu Santos-Neto, David Condon, Nazareno Andrade, Adriana Iamnitchi, and Matei Ripeanu. Individual and Social Behavior in Tagging Systems. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 183–192, New York, NY, USA, June 2009. Association for Computing Machinery.

[234] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in Qualitative Research: Exploring its Conceptualization and Operationalization. *Quality & Quantity*, 52(4):1893–1907, July 2018.

[235] Reijo Savolainen. Berrypicking and Information Foraging: Comparison of Two Theoretical Frameworks for Studying Exploratory Search. *Journal of Information Science*, 44(5):580–593, October 2018.

[236] H. Schiefele, Andreas Krapp, Manfred Prenzel, A. Heiland, and H. Kasten. *Principles of an Educational Theory of Interest*. August 1983.

[237] Ulrich Schiefele. The Influence of Topic Interest, Prior Knowledge, and Cognitive Capabilities on Text Comprehension. In Jules M. Pieters, Klaus Breuer, and P. Robert-Jan Simons, editors, *Learning Environments: Contributions from Dutch and German Research*, Recent Research in Psychology, pages 323–338. Springer, Berlin, Heidelberg, 1990.

[238] Leonard Seabrooke. Epistemic Arbitrage: Transnational Professional Knowledge in Action. *Journal of Professions and Organization*, 1(1):49–64, March 2014.

[239] Paul Seedhouse and Müge Satar. The Rashomon Effect: Which Features of a Speaker's Talk do Listeners Notice? *Classroom Discourse*, 14(1):1–23, June 2021.

[240] M. Shajalal, M. Z. Ullah, A. N. Chy, and M. Aono. Query Subtopic Diversification Based on Cluster Ranking and Semantic Features. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application*, pages 1–6, August 2016.

[241] Claude E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[242] Claude Elwood Shannon. *A Symbolic Analysis of Relay and Switching Circuits*. Thesis, Massachusetts Institute of Technology, 1940.

[243] David Shenk. *Data Smog: Surviving the Information Glut*. HarperCollins Publishers, New York, NY, USA, 1997.

[244] Herbert A. Simon. Theories of Decision-Making in Economics and Behavioral Science. *The American Economic Review*, 49(3):253–283, 1959.

[245] Swarndeep Singh and Rajesh Sagar. A Critical Look at Online Survey or Questionnaire-based Research Studies During COVID-19. *Asian Journal of Psychiatry*, 65:102850, November 2021.

[246] Kirk W. Smith. KWSnet Search Tools Index, 2001.

[247] William G. Smith. Does Gender Influence Online Survey Participation? A Record-Linkage Analysis of University Faculty Online Survey Response Behavior. Technical report, June 2008. Publication Title: Online Submission ERIC Number: ED501717.

[248] Jimmy Soni and Rob Goodman. *A Mind at Play*. July 2018.

[249] Daniel Sousa, Luís Sarmento, and Eduarda Mendes Rodrigues. Characterization of the Twitter @Replies Network: Are User Ties Social or Topical? In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, pages 63–70, New York, NY, USA, October 2010. Association for Computing Machinery.

[250] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[251] Liz Spencer, Jane Ritchie, Jane Lewis, and Lucy Dillon. Quality in Qualitative Evaluation: A framework for assessing research evidence, 2003.

[252] Meir Sternberg. Telling in Time (I): Chronology and Narrative Theory. *Poetics Today*, 11(4):901–948, 1990.

[253] Meir Sternberg. Telling in Time (II): Chronology, Teleology, Narrativity. *Poetics Today*, 13(3):463–541, 1992.

[254] Robert J. Sternberg. Novelty-seeking, Novelty-finding, and the Developmental Continuity of Intelligence. *Intelligence*, 5(2):149–155, April 1981.

[255] Andie Storozuk, Marilyn Ashley, Véronic Delage, and Erin Maloney. Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks. *The Quantitative Methods for Psychology*, 16:472–481, May 2020.

[256] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, Mckane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Transactions on Recommender Systems*, November 2023.

[257] Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. Indri: A Language Model-based Search Engine for Complex Queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.

[258] Robert I. Sutton and Barry M. Staw. What Theory is Not. *Administrative Science Quarterly*, 40(3):371–384, September 1995.

[259] Henri Tajfel, John C. Turner, William G. Austin, and Stephen Worchel. An Integrative Theory of Intergroup Conflict. *Organizational Identity: A Reader*, 56(65):9780203505984–16, 1979.

[260] Toru Takaki, Atsushi Fujii, and Tetsuya Ishikawa. Associative Document Retrieval by Query Subtopic Analysis and its Application to Invalidity Patent Search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pages 399–405. ACM, 2004.

[261] Bin Tan, Yuanhua Lv, and ChengXiang Zhai. Mining Long-Lasting Exploratory User Interests from Search History. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1477–1481. ACM, 2012.

[262] Robert S. Taylor. The Process of Asking Questions. *American Documentation*, 13(4):391–396, 1962.

[263] Robert S. Taylor. Question-Negotiation and Information-Seeking in Libraries. Technical report, Defense Technical Information Center, Fort Belvoir, VA, July 1967.

[264] Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development*, pages 163–170, New York, NY, USA, July 2008. Association for Computing Machinery.

[265] Philip E. Tetlock. Expert Political Judgment: How Good Is It? How Can We Know? - New Edition. In *Expert Political Judgment*. Princeton University Press, August 2017.

[266] Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Superforecasting: The art and science of prediction. Crown Publishers/Random House, New York, NY, US, 2015. Pages: 340.

[267] Richard H. Thaler. Behavioral Economics. *Journal of Political Economy*, 125(6):1799–1805, December 2017. Publisher: The University of Chicago Press.

[268] Philip S. Thomas. Environmental Scanning - The State of the Art. *Long Range Planning*, 13(1):20–28, February 1980.

[269] Paul Thompson. Satisficing or the Right Information at the Right Time: Artificial Intelligence and Information Retrieval, a Comparative Study in Medicine and Law. In *Medical Applications of Artificial Intelligence*, page 8. 2013.

[270] L. L. Thurstone. A Law of Comparative Judgment. *Psychological Review*, 34(4):273–286, 1927.

[271] Moises C. Torrentira, Jr. Online Data Collection as Adaptation in Conducting Quantitative and Qualitative Research During the COVID-19 Pandemic. *European Journal of Education Studies*, 7(11), 2020.

[272] Jan Treur. Mental Models in the Brain: On Context-Dependent Neural Correlates of Mental Models. *Cognitive Systems Research*, 69:83–90, October 2021.

[273] Daniel Trielli and Nicholas Diakopoulos. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems*, pages 1–15, New York, NY, USA, May 2019. Association for Computing Machinery.

[274] William M. K. Trochim. An Introduction to Concept Mapping for Planning and Evaluation. *Evaluation and Program Planning*, 12(1):1–16, January 1989.

[275] Twitter. How to use the Explore tab, 2022. Twitter Help.

[276] US Census Bureau. QuickFacts: United States, 2021.

[277] Elke van der Meer, Frank Krüger, and Antje Nuthmann. The Influence of Temporal Order Information in General Event Knowledge on Language Comprehension. *Journal of Psychology*, 213(3):142–151, July 2005.

[278] Lieke LF van Lieshout, Floris P de Lange, and Roshan Cools. Why So Curious? Quantifying Mechanisms of Information Seeking. *Current Opinion in Behavioral Sciences*, 35:112–117, October 2020.

[279] Jeanne Van Oosten. A Unified View of Topic. In *Annual Meeting of the Berkeley Linguistics Society*, volume 10, pages 372–385, 1984.

[280] Jeanne Hillechiena Van Oosten. *The Nature of Subjects, Topics and Agents: A Cognitive Explanation*. University of California, Berkeley, 1984.

[281] Reinout Van Rees. Clarity in the Usage of the Terms Ontology, Taxonomy and Classification. *CIB Report*, 284(432):1–8, 2003.

[282] Kurt VanLehn. Cognitive Skill Acquisition. *Annual Review of Psychology*, 47(1):513–539, 1996.

[283] Saúl Vargas and Pablo Castells. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the fifth ACM conference on Recommender systems*, RecSys '11, pages 109–116, New York, NY, USA, October 2011. Association for Computing Machinery.

[284] Lara Varpio, Elise Paradis, Sebastian Uijtdehaage, and Meredith Young. The Distinctions Between Theory, Theoretical Framework, and Conceptual Framework. *Academic Medicine*, 95(7):989, July 2020.

[285] Ike Vayansky and Sathish A. P. Kumar. A Review of Topic Modeling Methods. *Information Systems*, 94:101582, December 2020.

[286] Koen J. F. Verhoeven, Katy L. Simonsen, and Lauren M. McIntyre. Implementing False Discovery Rate Control: Increasing Your Power. *Oikos*, 108(3):643–647, 2005.

[287] Ellen M. Voorhees. The Philosophy of Information Retrieval Evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, Lecture Notes in Computer Science, pages 355–370, Berlin, Heidelberg, 2002. Springer.

[288] Karin Wahl-Jorgensen and Thomas Hanitzsch. *The Handbook of Journalism Studies*. Routledge, January 2009.

[289] Dieter Wallach and Sebastian C. Scholz. User-Centered Design: Why and How to Put Users First in Software Development. In Alexander Maedche, Achim Botzenhardt, and Ludwig Neer, editors, *Software for People: Fundamentals, Trends and Best Practices*, Management for Professionals, pages 11–38. Springer, Berlin, Heidelberg, 2012.

[290] Fulai Wang and Jim Greer. Retrieval of Short Documents from Discussion Forums. In Robin Cohen and Bruce Spencer, editors, *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 339–343, Berlin, Heidelberg, 2002. Springer.

[291] Karl E. Weick. What Theory is Not, Theorizing Is. *Administrative Science Quarterly*, 40(3):385, September 1995.

[292] Gene Weingarten. *One Day*. Penguin Random House, 2019.

[293] David A. Whetten. What Constitutes a Theoretical Contribution? *Academy of Management Review*, 14(4):490–495, October 1989.

[294] Ryen W. White, Bill Kules, and Steven M. Drucker. Exploratory Search. *Communications of the ACM*, 49(4):37, 2006.

[295] Ryen W. White and Resa A. Roth. Defining Exploratory Search. In Ryen W. White and Resa A. Roth, editors, *Exploratory Search: Beyond the Query—Response Paradigm*, Synthesis Lectures on Information Concepts, Retrieval, and Services, pages 9–23. Springer International Publishing, Cham, 2009.

[296] Robert C. Wilson, Andra Geana, John M. White, Elliot A. Ludvig, and Jonathan D. Cohen. Humans Use Directed and Random Exploration to Solve the Explore-Exploit Dilemma. *Journal of Experimental Psychology*, 143(6):2074–2081, December 2014.

[297] T. D. Wilson. The Transfer of Theories and Models from Information Behaviour Research into Other Disciplines, June 2020. Publisher: University of Boras.

[298] Thomas D. Wilson. Human Information Behavior. *Informing science*, 3(2):49–56, 2000.

[299] WIRED Insider. How to Use Gmail Like a Pro. *Wired*, June 2019.

[300] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, May 2023.

[301] Wei-Li Wu, Bi-Fen Hsu, and Ryh-Song Yeh. Fostering the Determinants of Knowledge Transfer: A Team-level Analysis. *Journal of Information Science*, 33(3):326–339, June 2007.

[302] Emine Yilmaz, Javed Aslam, and Stephen Robertson. A New Rank Correlation Coefficient for Information Retrieval. July 2008.

[303] Anita L. Zeidler and John R. Surber. Understanding Topic, Structure, and Importance of Information in a Visual and Verbal Display. *The Journal of Experimental Education*, 67(2):114–132, 1999.

[304] Pengyi Zhang and Dagobert Soergel. Towards a Comprehensive Model of the Cognitive Process and Mechanisms of Individual Sensemaking. *Journal of the Association for Information Science and Technology*, 65(9):1733–1756, September 2014.

[305] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, January 2024.

[306] Ziyi Zhang, Shuofei Zhu, Jaron Mink, Aiping Xiong, Linhai Song, and Gang Wang. Beyond Bot Detection: Combating Fraudulent Online Survey Takers. In *Proceedings of the ACM Web Conference 2022*, pages 699–709, New York, NY, USA, April 2022. Association for Computing Machinery.

[307] Meng Zhu and Rebecca K. Ratner. Scarcity Polarizes Preferences: The Impact on Choice Among Multiple Items in a Product Class. *Journal of Marketing Research*, 52(1):13–26, February 2015.

[308] Gabriel Zoran. Towards a Theory of Space in Narrative. *Poetics Today*, 5(2):309–335, 1984. Publisher: [Duke University Press, Porter Institute for Poetics and Semiotics].

[309] Rolf A. Zwaan, Carol J. Madden, and Robert A. Stanfield. Time in Narrative Comprehension. *Psychology and Sociology of Literature*, pages 71–86, 2001.

[310] Rolf A. Zwaan, Joseph P. Magliano, and Arthur C. Graesser. Dimensions of Situation Model Construction in Narrative Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(2):386, 1995.

[311] Rolf A. Zwaan and Gabriel A. Radvansky. Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, 123(2):162, 1998.