LAMP-TR-156 May 2011
CS-TR-4981
UMIACS-TR-2011-08

ART IMAGES AND MULTILINGUAL SOCIAL TAGGING: A MUSEUM WITHOUT BORDERS

Irene Eleta

Computational Linguistics and Information Processing
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
ieleta@umd.edu

Abstract

The use of social tagging for crowd-sourcing the annotation of images in online collections of art is in a nascent stage, but it has the potential to bridge language borders and reach wider audiences. How much do different language communities agree with each other when tagging images of art? This exploratory quantitative study is based on a collection of 24 digital images for which tags in Spanish and English were collected. The results show that when adding a second language for tagging images of art, the proportion of agreement among taggers does not seem to change significantly with respect to only one language.

Keywords: social tag, multilingual, image description

The support of this research by the Institute of Museum and Library Services under grant LG-30-08-0117-08 is gratefully acknowledged.

Art Images and Multilingual Social Tagging: a Museum without Borders

Irene Eleta
University of Maryland
ieleta@umd.edu

Abstract

The use of social tagging for crowd-sourcing the annotation of images in online collections of art is in a nascent stage, but it has the potential to bridge language borders and reach wider audiences. How much do different language communities agree with each other when tagging images of art? This exploratory quantitative study is based on a collection of 24 digital images for which tags in Spanish and English were collected. The results show that when adding a second language for tagging images of art, the proportion of agreement among taggers does not seem to change significantly with respect to only one language.

1 Introduction

The digitization of cultural heritage works has enabled digital libraries and museums to overcome the barrier of physical location, reaching wider -even global- audiences. Digital image collections are becoming available for the general public in growing numbers thanks to the digitization and dissemination efforts of many cultural heritage institutions around the world, such as The Library of Congress, The British Museum, El Prado Museum, and The Louvre, to cite a few well-known examples. The volume of digital image collections poses many challenges for indexing, access and use.

These collections are accessed by the public through textual queries, which requires images to be annotated (metadata). Cataloguers enter the metadata of images, which typically includes the creator name, the title, institution, etc. This approach presents two problems: the large volume of digital collections makes annotation by professionals unfeasible, and only experts that know the creator or title are likely to find the images. New approaches have appeared as thriving areas of research to overcome this challenge, such as the study of user-centered indexing, automatic annotation and social tagging.

There are different levels of knowledge and interpretation that come into play when describing images of art with words. Trant (2009) suggests that social tags could be useful for bridging the "semantic gap" between curatorial language and lay terms. The social tags the community provides have the advantage of reflecting to some extent the vocabulary they would use when typing text queries and, if well harnessed, they could improve the accessibility of the images.

Another advantage of social tagging is that it does not require training to participate in the system; also, it facilitates browsing, information discovery and unanticipated uses (Mathes, 2004). There are drawbacks to this approach. For example, the social tagging environment has to stimulate participation to diminish the problem of sparse social tags in part of the collection (Sigurbjörnsson and van Zwol, 2008). The absence of standards leads to noise, such as spelling variants, misspellings, and languages not available in the system (Guy and Tokin, 2006).

This observation points to yet another challenge, which arises from the multilingual aspects of the Internet; the language barrier is more present than ever due to the language diversity of Internet users. Also, the collections are composed of art images originating from different cultures whose peoples might want to access using their own language, but often those images are annotated only in the languages of the institutions that display them abroad.

Multilingual social tagging has been studied in Flickr (Gonzalo et al., 2009), and in the context of

educational resources online in Europe (Vourikari et al., 2007); also, the PanImages system uses both automatic and crowd-sourced methods for improving multilingual search of images (Colowick, 2008). Unfortunately, multilingual social tagging has not been applied yet in image collections of museums. The potential and benefits of social tagging for access and use (or interpretation) of art images in multilingual and multicultural contexts remains unexploited.

The enrichment of the image annotations could vary from simply diminishing the problem of sparse tags by including more languages and populating the collection using machine translation, to multilingual crowd-sourced indexing. In addition to that, if the semantic diversity increases with language diversity, multilingual social tagging could add the value of different interpretations, and provide new access paths in cross-language search.

In this exploratory quantitative study, I seek to answer a first question: How much do different language communities agree with each other when tagging images of art? And, secondly, whether this agreement (or lack thereof) depends on the type of painting. I conducted a pilot experiment using a collection of 24 images of paintings for which tags in Spanish and English were collected. I used these social tags to study the agreement between representative sets of users to draw conclusions about cross-cultural tagging behaviors.

This paper is organized as follows: I will discuss the related literature from diverse fields in section 2, including some considerations on similarity measures that are relevant to section 3, experiment design; the results will be presented in section 4, followed by the analysis in section 5. Finally, in section 6, I will discuss the next steps of this work before concluding in section 7.

2 Background and Literature Review

Pictorial semiotics explains the different ways in which humans construct meaning from images as opposed to text. In text, meaning is constructed linearly, while in images meaning appears in a holistic form, in layers (Sonesson, 1994). This key difference is critical when we rely on text queries for searching images, and on image annotations. Dillon (1999) suggested that the text describing an image helps to settle one meaning and solves the polysemy of images. This is happening when the professional cataloguer annotates the art images: other possible meanings or cultural interpretations are not present in the description, neither "findable".

Sonesson (1994) explained that this construction of meaning from images has both an individual component -studied by perceptual psychology-, and also a social and cultural component. In the context of crowd-sourcing and folksonomies, the collective and decentralized construction of knowledge is known as "Collective Intelligence" (Lévy, 1997), or "The Wisdom of the Crowds" (Surowiecki, 2005), or "distributed cognition" (Hutchins, 2000).

The study of patterns in social tagging has an important role in promoting useful applications in information access and discovery, by uncovering the "collective intelligence". Golder and Huberman (2006), among many others, determined that, rather than fostering chaotic growth, the social tags stabilize in fix proportions in the long run. Often, the vocabulary for a resource resembles a power law distribution, where there is a small group of social tags that are very frequent, and a Long Tail of diverse tags (Sigurbjörnsson and van Zwol, 2008). The tagging behavior is also related to the functionalities of the system (Lee at al., 2009), therefore the study of tagging patterns should take into account the design of social tagging systems.

One way of harnessing the "collective intelligence" in such systems is identifying the statistical consensus of the community on a particular set of tags for a resource, but without forgetting the variety of opinions in the Long Tail (Peters, 2009). To achieve this, some degree of natural language processing is needed: Klavans et al. (2011) show how computational linguistic processing, morphological and semantic analysis can help reduce the 'noise" in tag sets and impact tag clustering.

This study involves measuring the similarities of sets of tags in two different languages and in the same language, so as to determine the statistical consensus. Catutto et al. (2008) describe different measures of similarity related to semantics in social tagging research, but their survey does not include cross-language similarity of tag sets.

The Jaccard Index, as used by Olson and Wolfram (2007) for studying non-expert indexers agreement, has the advantage of being a

proportion, which facilitates interpretation and comparison. However, if the number of tags for a resource is clearly unbalanced between two languages, the index is going to be negatively biased.

Alternatively, similarity can be measured by the cosine distance between the vector representing the set of terms in English and the vector representing the set of terms in Spanish for a particular painting. This technique is used for matching short queries to documents (Kolda and O'Leary, 1998), and therefore is better suited for a variety of term set sizes. The disadvantage of the cosine distance is that comparisons might be obscured because it is not a proportion.

The asymmetric Jaccard index is used by Santos-Neto et al. (2009) for measuring the similarity of tag sets -and resources- to indentify the shared interests, the commonalities, from a particular user's perspective. This measure has the advantage of being a proportion and, when the number of tags between two sets is unbalanced, giving more weight to the shared terms. Unlike this exploratory study, if the intention is to measure the proportion of tags that are in the Long Tail of the power law distribution of tags for an image, the asymmetric Jaccard index is biased because it does not include all the vocabulary used by the most prolific taggers.

This work aims to be a first step, to answer a first question, upon which to build a research path that merges both multilingual social tagging and its application in museums. There are promising prospects: A study on the use of multilingual social tagging for the discovery of online learning resources in an educational repository in Europe (Vourikari and Koper, 2009) provided evidence that 23% of cross-boundary discoveries happened through the multilingual tag cloud and 5% from the lists of "travel-well" tags (with spelling similarities across languages).

In the art context, museum professionals concluded that 77% of the social tags could be potentially useful in describing the art images (Trant, 2009). Chun et al. (2006) wonder "how this new sort of engagement with museum objects might help […] draw contributors who bring a multi-cultural perspective to looking at our works of art".

3 Experiment Design

This work was designed as an exploratory quantitative study. First, I created a website to collect the social tags for art images, and designed an interface that directed the participants in a sequence: from a brief questionnaire of 5 profiling questions to successive images with a text box for tagging. The participants could not see other people's tags, neither the title nor other metadata about the painting, and the text box allowed them to enter multi-words tags. The intention was to encourage their creativity and reduce collective influence. There was a version in English for participants in the United States and a version in Spanish for Spain.

Second, I selected 24 images of paintings from the experimental *steve.museum* collection. Each image belongs to one of the following three categories: non-western paintings, western paintings with no abstraction, and western paintings with abstraction. The images appeared in different order for each participant to compensate for any cognitive associations. There was a maximum of 10 images shown in each session, to keep the participation time to a reasonable limit, and they could skip images without tagging them. Once the social tags were gathered in both languages, the following step was to analyze them separately, by language. The objective is to quantify the proportion in which taggers agreed in the vocabulary for describing an image.

The first problem was to process the multi-word tags. In those types of tags either the semantic meaning is qualified (i.e. specification) or they refer to more than one concept. For simplification of analysis, social tags were divided in their lexical components, which in this work will be called *units*. I started by applying tokenization, using spaces in between words as dividers, with the exception of compound words like "still life". After that, I eliminated articles, prepositions, and conjunctions (stop words). Then, I applied spelling correction, lower-case normalization, and manual lemmatization (i.e. plural forms into singular, verbs in infinitive).

Once I obtained the list of units per participant for each image, the next step was to quantify the vocabulary similarity between participants. I used the asymmetric Jaccard index for the reasons explained in section 2.

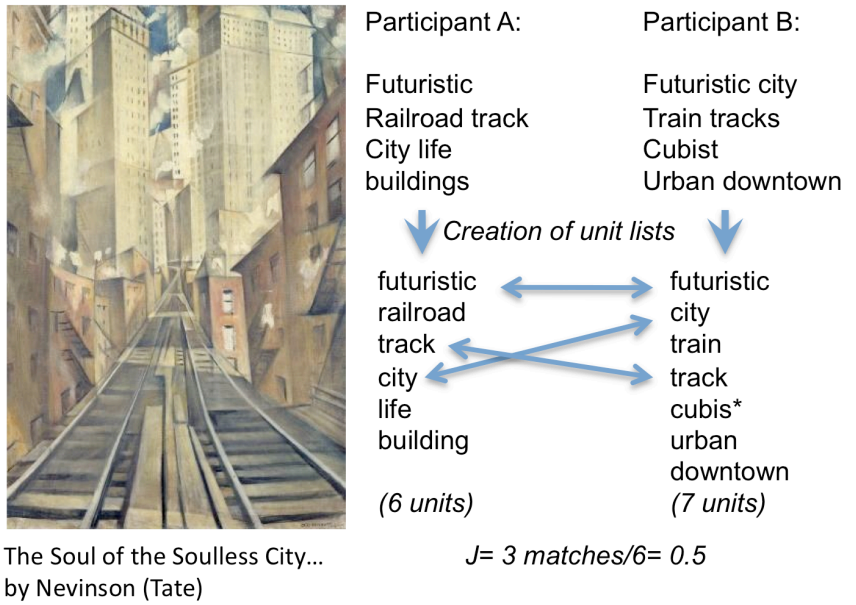

Figure 1. Quantifying tag similarity between participants using asymmetric Jaccard index

The asymmetric Jaccard index in this analysis is the number of equal units that two taggers used for describing an image divided by the total number of units of the tagger that used fewer words. The equation is: J (A, B) = (A ∩ B)/ A, where A represents the units of the participant that used fewer words to describe an image, and B represents the units used by the other participant. This index gives the proportion of agreement from 0 to 1 between the two taggers. If the image had more than two taggers, I calculated the asymmetric Jaccard index pair wise, and used the mean. Figure 1 illustrates the extraction of units per participant and the computation of the asymmetric Jaccard index.

The last step was to compare the similarity values across languages to see if the results were different than within the same language group. For each image, I compared the units between all the pairs of participants where one was from Spain and one from the US. When the units were the translation of one another, in the context of the painting, they counted as a match. This process gave 4 or 6 indexes per image, depending on the number of persons that tagged the image, and I used the mean for the analysis.

One consideration in this quantification process was that when matching units in the same language I did not count synonyms to be the same unit (i.e. *cloudy* was considered different than *overcast*). However, the cross-language matching is similar to synonym matching within a language. I compared the results of keeping synonyms separate within a language (Method A) and merging them (Method B). Table 1 shows that there is little difference in the asymmetric Jaccard indexes between the two methods. This supports the comparability between groups of the same languages and across-language.

	Method A	**Method B**
Spanish (Spain)	0.24	0.25
English (US)	0.32	0.33

Table 1. Mean asymmetric Jaccard index for methods A and B.

4 Results

There were 7 participants from Spain (Spanish speakers) and 7 from the US (English speakers). All participants use Internet several times a day and have a university degree. The engagement with art ranges in the Spanish sample from no particular interest in art (1), interested in art (5), to expert (1); in the group from the US, it ranges from no particular interest in art (3) to interested in art (4).

Table 2 shows a summary of results of the asymmetric Jaccard index per sample (Spanish, English, and cross-language).

	Spanish	English	Cross-lang.
Mean	0.242	0.316	0.250
St.Dev.	0.279	0.244	0.141

Table 2. Mean and Std. deviation of asymmetric Jaccard index per sample.

The Wilcoxon signed ranks test for paired samples did not provide support for the rejection of the hypothesis that the sample means are equal: p value (Spanish vs. English) = 0.376; p value (Spanish vs. cross-language) = 0.323; p value (English vs. cross-language) = 0.315.

Finally, table 3 shows a tendency of lower agreement between taggers of paintings with some degree of abstraction compared to the non-abstract paintings:

Art type	Sp	En	Crosslang
Non-western	0.155	0.445	0.300
Non-abstract	0.384	0.322	0.297
Abstract	0.107	0.252	0.172

Table 3. Mean asymmetric Jaccard index per image category and per sample.

5 Analysis

The results imply that, when adding a second language for tagging images of Art, the proportion of agreement among taggers does not seem to change significantly with respect to only one language. However, the results cannot show whether the consensus words in English are translations of the consensus words in Spanish, and whether they are the same as the cross-language consensus words. Only a qualitative analysis could shed light into this question.

The relatively small proportion of agreement between taggers across all samples, ranging from 24% of consensus words by Spanish participants to 31% of consensus words by US participants, is consistent with the observations by Golder and Huberman (2006), Sigurbjörnsson and van Zwol, (2008), which were discussed in section 2. Their small number and popularity make these consensus words potential candidates for indexing terms. But as Peters (2009) warns: they might be too general and other specific and relevant terms might be hidden in the Long Tail.

In the remaining 69%-76% of words in this experiment, there is space for diverse opinions and perceptions. A cursory qualitative analysis provides two examples where the art images inspire a different interpretation between Spanish and American taggers. A painting of Polynesian women by Gaugain (Ia Orana Maria) and a painting of the Blue Period of Picasso (La Vie) received no tags related to religion in English, while in Spanish there were tags related to Christianity, such as *Virgin Mary*, *Adam and Eve*.

Another finding of this study is that paintings with some degree of abstraction tend to have lower indexes of agreement than the other types, both in the same language samples and in the cross-language sample. This finding suggests that this type of paintings inspires more diverse interpretations than non-abstract paintings.

Regarding limitations of this study, the decision to apply tokenization to the multi-word tags relies on the assumption that the social tags have compositional meaning; the implications of not meeting this assumption in all cases should be studied.

If the intention is to account for tags that are in the Long Tail of the power law distribution of tags for an image, the asymmetric Jaccard index is biased because it does not include all the vocabulary used by the most prolific taggers. In that case, it would be preferable to use the symmetric Jaccard index or the cosine distance value, as discussed in section 2.

6 Next Steps

The next step after this exploratory study is to increase the sample of participants for both languages. As a consequence of the higher volume of tags, more processing tasks will need to be automated. Also, I will include more images of paintings in each category, with a wider variety of themes, which could provide richer results about the differences across art categories and cultural perceptions.

In addition to the quantitative analysis, a qualitative study will seek to clarify whether consensus words are different or not across all

three samples (Spanish, English, and cross-language).

It would be interesting to compare the results using other similarity measures, and study the social tags with a focus on diversity, in the Long Tail, with the purpose of complementing the consensus words in a rich and meaningful way.

Finally, as a separate project, it would be helpful to investigate the implications if the assumption of compositional meaning for multi-word tags is not met in all cases.

7 Conclusion

The large volume of museums' online image collections poses many challenges for indexing and searching. These collections are accessed by the public through textual queries, which requires images to be annotated. Moreover, the language barrier is more present than ever due to the language diversity of Internet users.

Social tagging is a promising approach because it has the advantages of crowd-sourced indexing, user-oriented vocabulary -which might improve access and discovery-, and it provides the possibility of adding multiple languages.

In this exploratory quantitative study, I used a collection of 24 images of paintings for which tags in Spanish and English were collected. I analyzed these social tags to study the agreement between Spanish taggers, US English speaking taggers, and across both languages. Also, I studied the agreement index across categories of paintings.

The goal of this study is to quantify how much different language communities agree with each other when tagging images of art, and whether this agreement depends on the type of painting.

The results show that, when adding a second language for tagging images of art, the proportion of agreement among taggers does not seem to change significantly with respect to only one language. The relatively small proportion of agreement between taggers in all samples is consistent with the findings of previous studies in folksonomies. However, it becomes clear that a qualitative study is needed to determine whether consensus words are different or not across all three samples.

Another finding of this study is that paintings with some degree of abstraction tend to have lower indexes of agreement than the other types, which suggests that abstract paintings inspire more diverse interpretations.

Consensus words are potential candidates for indexing terms in multiple languages; on the other hand, the Long Tail of tags might hide more specific terms and cultural interpretations that could add access paths to the images. In the case of art collections, whether the social tags that promote cross-language discovery are hidden down the tail of a power law distribution is a question that a qualitative study could answer.

Despite its limitations, this ongoing work constitutes a first step for understanding the patterns of bilingual social tagging in collections of art images, and seeks to uncover exploitable strengths for its application in museums.

Acknowledgments

This study owes much to Dr. Jennifer Golbeck, my adviser, for her continuous guidance, to Dr. Douglas W. Oard, for his insightful comments on the earlier stages of this study, to Dr. Judith Klavans, for sharing her knowledge and inspire me within the context of the T3 research project (Text, Tags and Trust), and to my sponsor, Fulbright, for the economic support. I also want to thank the enthusiastic participants of the experiment.

References

Catutto C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. *The Semantic Web - ISWC 2008. Lecture Notes in Computer Science*, 5318, 615-631.

Colowick, S.M. (2008, March). Multilingual Search with PanImages. *Multilingual*, 19(2), 1-3.

Chun, S., Cherry, R., Hiwiller, D., Trant, J., & Wayman, B. (2006). Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums. *Museums and the Web 2006*. http://www.archimuse.com/mw2006/papers/wyman/wyman.html

Dillon, G.L. (1999). *Art and the Semiotics of Images: Three Questions About Visual Meaning*. Retrieved, March 21, 2011, from http://faculty.washington.edu/dillon/rhethtml/signifiers/sigsave.html

Golder, S.A. & Huberman, B.A. (2006). The Structure of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 198-208.

Gonzalo, J. Peinado, V, Clough, P., & Karlgren, J. (2009). Overview of iCLEF 2009: Exploring Search Behaviour in a Multilingual Folksonomy environment. *Working Notes for the CLEF 2009 Workshop.* http://www.clef-campaign.org/2009/working_notes/CLEF2009WN-Contents.html

Guy, M. & Tokin, E. (2006, January) Tidying up tags. *D-Lib Magazine*, vol. 12(1). http://www.dlib.org/dlib/january06/guy/01guy.html

Hutchins, E. (2000). *Distributed Cognition*. Retrieved on March 21, 2011, from http://www.telelearning-pds.org/coa/distributed_cognition.pdf

Klavans, J., Guerra, R., LaPlante, R., Stein, R., & Bachta, E. (2011). Taming Social Tags: Computational Linguistic Analysis of Tags for Images in Museums. Computer Science Report No. CS-TR-4980, University of Maryland Institute for Advanced Computer Studies Report No. UMIACS-TR-2011-07, Language and Media Processing Laboratory Report No. LAMP-TR-155.

Computer Science Report No. CS-TR-

Kolda, T.G. & O'Leary, D.P. (1998). A Semidiscrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval. *ACM Transactions on Information Systems*, 16(4), 322-346.

Lee, C.S., Goh, D.H., Razikin, K., & Chua A.Y.K. (2009). Tagging, Sharing and the Influence of Personal Experience. *Journal of Digital Information*, 10(1). http://journals.tdl.org/jodi/article/view/275/275

Lévy, P. (1997). *Collective Intelligence: Mankind's Emerging World in Cyberspace*. (R. Bononno, Trans.). Cambridge, MA: Perseus Books.

Mathes, A. (2004) *Folksonomies - Cooperative Classification and Communication through Shared Metadata*. Retrieved on March 21, 2011, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Olson, H.A. & Wolfram, D. (2007). Syntagmatic Relationships and Indexing Consistency on a Larger Scale. *Journal of Documentation*, 64(4), 602-615.

Peters, I. (2009). *Folksonomies. Indexing and Retrieval in Web 2.0*. (P. Becker, Trans.).Berlin: De Gruyter Saur.

Santos-Neto, E., Cordon, D., Andrade, N., Iamnitchi, A., & Ripeanu, M. (2009). Individual and Social Behavior in Tagging Systems. *Proceedings of the 20th ACM conference* HT '09, 183-192.

Sigurbjörnsson, B. & van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. *Proceedings of WWW '08*, 327–336

Sonesson, G. (1994). Pictorial semiotics, Gestalt Theory, and the ecology of perception. *Semiotica*, 99 (3-4), 319–440.

Surowiecki, J. (2005). *The Wisdom of Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. New York: Anchor Books.

Trant, J. (2009). *Tagging, Folksonomy and Art Museums: Results of steve.museum's research* (Technical report). Retrieved on March 21, 2011, from http://conference.archimuse.com/jtrants/stevemuseum_research_report_available

Vourikari, R. & Koper, R. (2009). Evidence of Cross-boundary Use and Reuse of Digital Educational Resources. *International Journal of Emerging Technologies in Learning*, 4(4). http://dspace.ou.nl/handle/1820/1709

Vourikari, R., Ochoa, X., & Duval, E. (2007). Analysis of User Behavior on Multilingual Tagging of Learning Resources. *Proceedings of the 1st Workshop on Social Information Retrieval for Technology Enhanced Learning*, SIRTEL'07. CEUR, 307, 6-17.