SRC TR 87-215

Extending the Flajolet-Regnier-Sotteau Method of Trie Analysis to the Prefix Model

by

Pilar de la Torre

Extending the Flajolet–Regnier–Sotteau Method of Trie Analysis to the Prefix Model

Pilar de la Torre

Systems Research Center
The University of Maryland
College Park, MD 20742

ABSTRACT

Recent work of Flajolet, Regnier and Sotteau has presented a systematic approach to the analysis of tries through generating functions. We extend their approach to a probabilistic model that, unlike other models used for the analysis of tries, takes into account sample sets containing keys that are prefixes of other keys in the set. As the main application of this extension, we find the exact average running time of two algorithms for computing set intersections. These algorithms are based on the representation of a set of binary string keys as a *full endmarker trie* and also as a *compact endmarker trie*.

1. Introduction

Since Knuth's original analysis [Knu73], the average case performance of trie data structures has received a good deal of attention (see, for example, [Knu73, Fra77, Tra78, Reg81, Dev82, Fla83, Pla83, Gon84, Fla84, FRS85, FS86, delaT87a,b]). Recently, Flajolet, Regnier, and Sotteau have presented a systematic approach to the analysis of tries through generating functions [FRS85]. The approach consists of setting up a translation mechanism based on rules that translate trie cost functions into their generating functions of normalized expectations. The translation rules, which depend on the particular underlying probabilistic model for sets of string keys, have been obtained in [FRS85] for the *Bernoulli model* of infinitely long keys, a *biased bit model* of infinitely long keys, and for two other models also studied by Trabb Pardo in [Tra78]. These are the uniform *finite identical length keys model* $\mathcal{F}$ and its corresponding binary set-intersection model. The probability space of the model $\mathcal{F}$ consists of the n element sets of keys of length h over a fixed finite alphabet, wherein all sets are assumed to be equally probable. The probability space of the corresponding binary set-intersection model is

$$\{\,(\xi,\eta)\mid \xi,\eta\subseteq\{0,1\}^h,\ |\xi|=m,\ |\eta|=n,\ |\xi\cap\eta|=k\,\},$$

where all pairs (ξ,η) are assumed equally to be equally probable. Among other applications, [FRS85] contains the computation of the exact average time needed to intersect pairs of sets $\xi,\eta\subset\{0,1\}^h$, when ξ and η are represented by compact binary tries. The sets of keys within all the above mentioned models satisfy the *no-prefixing key restriction.* That is, no key in a sample set is a prefix of another.

The design of tries for storing sets of keys that may contain *prefixing keys* (that is, keys that are prefixes of other keys in the set), was taken up in recent work of Knott [Kno86]. The first analysis of tries that store prefixing keys was done in [delaT87a], where we computed the exact average space and time complexities of the retrieval algorithms of several trie varieties by two approaches: recurrences and direct counting. The underlying probabilistic model used was the *prefix model*, whose sample space consists of all n element sets of keys of length not greater than h composed from a fixed finite alphabet; all the sets are assumed to be equally probable.

The work in [delaT87a] raises the question of the applicability of the generating function approach of [FRS85] to the prefix model, and also to its corresponding binary *set-intersection prefix model*, whose sample space is

$$\{\,(\xi,\eta)\mid \xi,\eta\subseteq\{0,1\}^0\cup\{0,1\}^1\cup\ldots\cup\{0,1\}^h\,,\ |\xi|=m,\ |\eta|=n,\ |\xi\cap\eta|=k\,\};$$

all set pairs (ξ,η) are assumed to be equally probable. This paper provides affirmative answers to both questions by developing the appropriate generating function tools along the lines of the work presented by Flajolet, Regnier, and Sotteau in [FRS85]. As the main application, we find the exact average running time of two algorithms that compute set intersections. These algorithms are based on the representation of a set of binary string keys as a full *endmarker trie* and also as a *compact endmarker trie*.

Section §2 derives the generating function translation rules corresponding to the prefix model. Section §3 illustrates the use of these rules by applying them to compute the average space and time requirements of the retrieval algorithms of two trie varieties analyzed in [delaT87b]: full endmarker tries and compact endmarker tries. Section §4 derives the translation rules for the set–intersection prefix model. Applying these rules, section §5 calculates the exact average running time of the algorithms for computing set intersections.

2. The prefix model

Let $\mathcal{A}$ be a totally ordered alphabet of m (≥ 2) symbols that we will identify with $\mathcal{A} = \{1, \ldots, m\}$, where $1 < 2 < \ldots < m$. Let $\mathcal{A}^{[h]} := \mathcal{A}^0 \cup \mathcal{A}^1 \cup \ldots \mathcal{A}^h$ be the set of all strings of length $\leq h$ composed from $\mathcal{A}$. The set of finite length strings composed from $\mathcal{A}$ will be denoted by $\mathcal{A}^*$, the set of infinitely long strings by $\mathcal{A}^\infty$, and $\mathcal{A}^\circledast := \mathcal{A}^* \cup \mathcal{A}^\infty$. For a finite set B, $\mathcal{R}_n(B)$ will denote the set of n element subsets of B, and $\mathcal{R}(B) := \bigcup_{n\geq 0} \mathcal{R}_n(B)$.

For the integer–valued parameters h, n, m, with $h, n \geq 0$ and $m \geq 2$, the probability space for the prefix model consists of the n element subsets of $\mathcal{A}^{[h]}$, which are assumed to be equally probable. We have

$$m^{[h]} := |\mathcal{A}^{[h]}| = \frac{m^{h+1} - 1}{m - 1}, \qquad \text{and} \qquad |\mathcal{R}_n(\mathcal{A}^{[h]})| = \binom{m^{[h]}}{n}.$$

Throughout this section X will denote a real–valued function of finite subsets $\xi \subseteq \mathcal{A}^{[h]}$. The expected value of $X(\xi)$ over the n element subsets $\xi \subseteq \mathcal{A}^{[h]}$ will be denoted by $E[X]$, and also by $E_{hn}[X]$ when we wish to emphasize its dependence on h and n. The sum

$$N_{hn}[X] := \sum_{\xi \in \mathcal{R}_n(\mathcal{A}^{[h]})} X(\xi)$$

is related to the expectation of X by $N_{hn}[X] = \binom{m^{[h]}}{n} E_{hn}[X]$, and will be called the *normalized expectation* of X.

2.1 *Translation rules*

To each real–valued function X of subsets $\mathcal{A}^{[h]}$ we associate its *generating function of the normalized expectations* $X^{(h)}(x)$,

$$X^{(h)}(x) : = \sum_{0 \leq n \leq m^{[h]}} N_{hn}[X]\, x^n = \sum_{\xi \in \mathcal{R}(\mathcal{A}_m^{[h]})} X(\xi)\, x^{|\xi|}.$$

Our intention is to establish rules that often help in translating a function X into its generating function $X^{(h)}(x)$. These translation rules will be formulated as properties of the operator $\mathbf{F}_h[X] := X^{(h)}(x)$, which maps real–valued functions of subsets $\mathcal{A}_m^{[h]}$ to polynomials in x.

We introduce the family of auxiliary functions $\mathbf{P}_x$, with $x \in \mathcal{A}^*$. The value of $\mathbf{P}_x$ on a subset $\xi \subset \mathcal{A}^{\circledast}$ is $\mathbf{P}_x(\xi) := \xi_x$, where $\xi_x := \{y \mid xy \in \xi\}$ (i.e., ξ_x is the set of tails of the strings of ξ that begin with x). For each $c \in \mathcal{A}$, $\mathbf{P}_c$ maps $\mathcal{R}(\mathcal{A}^{[h]})$ onto $\mathcal{R}(\mathcal{A}^{[h-1]})$. We also define the function $\mathbf{P}_\perp(\xi) := \xi \cap \{\varepsilon\}$, which maps $\mathcal{R}(\mathcal{A}^{[h]})$ onto $\mathcal{R}(\{\varepsilon\})$.

LEMMA 1. [*Additive–multiplicative rule*] *Let X, Y, Y_0, $Y_1,\ldots,Y_m$ be real-valued functions of subsets of $\mathcal{A}^{[h]}$. Then,*

(*i*) $\mathbf{F}_h[\lambda.X] = \lambda \mathbf{F}_h[X]$;

(*ii*) $\mathbf{F}_h[X+Y] = \mathbf{F}_h[X] + \mathbf{F}_h[Y]$;

(*iii*) *For* $h \geq 1$,

$$\mathbf{F}_h[(Y_1 \circ \mathbf{P}_1)\ldots(Y_m \circ \mathbf{P}_m)] = (1+x)\,\mathbf{F}_{h-1}[Y_1]\ldots\mathbf{F}_{h-1}[Y_m];$$

(*iv*) *For* $h \geq 1$,

$$\mathbf{F}_h[(Y_0 \circ \mathbf{P}_\perp)(Y_1 \circ \mathbf{P}_1)\ldots(Y_m \circ \mathbf{P}_m)] = \mathbf{F}_0[Y_0]\,\mathbf{F}_{h-1}[Y_1]\ldots\mathbf{F}_{h-1}[Y_m].$$

Proof: Properties (*i*) and (*ii*) follow directly from the definition of the operator $\mathbf{F}_h$. Since $\mathbf{F}_0[I] = 1 + x$, relation (*iii*) can be obtained from (*iv*) by taking $Y_0 = I$, with $I(\xi) := 1$. To verify (*iv*), we first observe that the partition $\mathcal{A}^{[h]} = \mathcal{A}^{[0]} \cup 1\mathcal{A}^{[h-1]} \cup \ldots \cup m\mathcal{A}^{[h-1]}$, $h \geq 1$, implies

$$\mathcal{R}(\mathcal{A}^{[h]}) = \bigcup_{\substack{\sigma^{(0)} \in \mathcal{R}(\{\varepsilon\}) \\ \sigma^{(j)} \in \mathcal{R}(\mathcal{A}^{[h-1]}) \\ j=1,\ldots,m}} \{\sigma^{(0)} \cup 1\sigma^{(1)} \cup \ldots \cup m\sigma^{(m)}\}$$

Hence, the mapping $\mathbf{P}(\xi) := \big(\mathbf{P}_\perp(\xi), \mathbf{P}_1(\xi), \ldots, \mathbf{P}_m(\xi)\big)$ defines a bijection $\mathbf{P} : \mathcal{R}(\mathcal{A}^{[h]}) \to \mathcal{R}(\{\varepsilon\}) \times \mathcal{R}(\mathcal{A}^{[h-1]}) \times \cdots \times \mathcal{R}(\mathcal{A}^{[h-1]})$. Using this bijection, the value of $\mathbf{F}_h$ at $X := (Y_0 \circ \mathbf{P}_\perp)(Y_1 \circ \mathbf{P}_1) \ldots (Y \circ \mathbf{P}_m)$ can be written as follows,

$$\mathbf{F}_h[X] = \sum_{\xi \in \mathcal{R}(\mathcal{A}^{[h]})} X(\xi)\, x^{|\xi|}$$

$$= \sum_{\substack{\sigma^{(0)} \in \mathcal{R}(\{\varepsilon\}) \\ \sigma^{(j)} \in \mathcal{R}(\mathcal{A}^{[h-1]}) \\ j=1,\ldots,m}} X\Big(\sigma^{(0)} \cup \bigcup_{1 \le j \le m} j\sigma^{(j)}\Big)\, x^{|\sigma^{(0)}|+|\sigma^{(1)}|+\cdots+|\sigma^{(m)}|}$$

$$= \Big[\sum_{\sigma^{(0)} \in \mathcal{R}(\{\varepsilon\})} Y_0(\sigma^{(0)})\, x^{|\sigma^{(0)}|}\Big] \sum_{\substack{\sigma^{(j)} \in \mathcal{R}(\mathcal{A}^{[h-1]}) \\ j=1,\ldots,m}} Y_1(\sigma^{(1)}) \ldots Y_m(\sigma^{(m)})\, x^{|\sigma^{(1)}|+\cdots+|\sigma^{(m)}|}.$$

Noting that the second factor can be expressed as a product, and also

$$\sum_{\sigma \in \mathcal{R}(\{\varepsilon\})} Y_0(\sigma) x^\sigma = Y_0(\emptyset) + Y_0(\{\varepsilon\})\, x = \mathbf{F}_0[Y_0],$$

we deduce

$$\mathbf{F}_h[X] = \mathbf{F}_h[Y_0] \prod_{1 \le j \le m} \sum_{\sigma^{(j)} \in \mathcal{R}(\mathcal{A}^{[h-1]})} Y_j(\sigma^{(j)})\, x^{|\sigma^{(j)}|}$$

$$= Y_0^{(0)}(x)\, Y_1^{(h-1)}(x) \ldots Y_m^{(h-1)}(x).$$

The verification of the lemma is now complete. □

LEMMA 2. [*Initialization rule*] *Let* $I(\xi) := 1$, *and* $C(\xi) := |\xi|$. *Then*

(*i*) $\mathbf{F}_h[I] = (1 + x)^{m^{[h]}}$;

(*ii*) $\mathbf{F}_h[C] = m^{[h]}\, x\, (1 + x)^{m^{[h]}-1}$;

(*iii*) *If* $X(\xi) = \delta_{|\xi|,p}$ *then* $\mathbf{F}_h[X] = \binom{m^{[h]}}{p} x^p$.

Proof: These relations can be established by direct counting. Relation (*ii*), for instance, results by noticing that the number of n element subsets of $\mathcal{A}^{[h]}$ is equal to $\binom{m^{[h]}}{n}$, whereby

$$\mathbf{F}_h[C] = \sum_{\xi \in \mathcal{R}(\mathcal{A}^{[h]})} |\xi|\, x^{|\xi|} = \sum_{n \geq 0} n \binom{m^{[h]}}{n} x^n = m^{[h]}\, x\, (1+x)^{m^{[h]}-1}.$$

□

THEOREM 3. *Let X and Y be real–valued functions of subsets of $\mathcal{A}^{[h]}$.*

(*i*) *If $X(\xi) = Y(\xi \cap \{\varepsilon\})$ then $X^{(h)}(x) = (1+x)^{m^{[h]}-1}\, Y^{(0)}(x)$;*

(*ii*) *If $X = Y \circ \mathbf{P}_c$, with $c \in \mathcal{A}$, then $X^{(h)}(x) = (1+x)^{m^h} Y^{(h-1)}(x)$;*

(*iii*) *Let $r_X(\xi) := X(\xi) - X(\xi_1) + \cdots + X(\xi_m)$. Then,*

$$X^{(h)}(x) = r_X^{(h)}(x) + m\,(1+x)^{m^h}\, X^{(h-1)}(x). \qquad (1)$$

Proof: Property (*i*) can be verified by applying (*iv*) of Lemma 1 to

$$X = (Y \circ \mathbf{P}_\perp) \prod_{j \in \mathcal{A}} I \circ \mathbf{P}_j,$$

and then property (*i*) of Lemma 2. In order to prove (*ii*), we write

$$X \circ \mathbf{P}_c = (I \circ \mathbf{P}_\perp)(Y \circ \mathbf{P}_c) \prod_{c \neq j \in \mathcal{A}} I \circ \mathbf{P}_j,$$

and from Lemma 1, Lemma 2, and property (i) we deduce

$$\begin{aligned}
\mathbf{F}_h[X \circ \mathbf{P}_c] &= \mathbf{F}_h[(I \circ \mathbf{P}_\perp)\,(Y \circ \mathbf{P}_c) \prod_{c \neq j \in \mathcal{A}} I \circ \mathbf{P}_j],\\
&= \mathbf{F}_0[I]\,\mathbf{F}_{h-1}[Y] \prod_{c \neq j \in \mathcal{A}} \mathbf{F}_{h-1}[I]\\
&= (1+x)^{(m-1)m^{[h-1]}+1}\, Y^{(h-1)}(x)\\
&= (1+x)^{m^h}\, Y^{(h-1)}(x).
\end{aligned}$$

To prove (iii), we first apply Lemma 1 to $X = r_X + \sum_{c\in\mathcal{A}}(X\circ\mathbf{P}_c)$, and then with the help of (ii) we deduce

$$\begin{aligned}\mathbf{F}_h[X] &= \mathbf{F}_h[r_X] + + \sum_{c\in\mathcal{A}} \mathbf{F}_h[X\circ\mathbf{P}_c] \\ &= \mathbf{F}_h[r_X] + m\,(1+x)^{m^h}\,\mathbf{F}_{h-1}[X],\end{aligned}$$

which completes the proof of the theorem. □

The following lemma, the *iteration lemma for the model $\mathcal{F}$*, was given in [FRS85] for solving the recurrences satisfied by the generating functions of normalized expectations with respect to the model $\mathcal{F}$. This lemma also provides the general solution to the recurrences (compare recurrence (1)) emerging in connection with the prefix model.

LEMMA 4. (*Flajolet-Regnier-Sotteau*) [*Iteration rule*] *Let $A_1, \ldots, A_h$ and $B_0, \ldots, B_h$ be polynomials. The solution to the recurrence $z_0 = B_0$,*

$$z_h = A_h z_{h-1} + B_h \qquad (h > 0), \tag{2}$$

is $z_h = \sum_{0\le j\le h} \left[B_j \prod_{j+1\le k\le h} A_k\right]$.

Proof: The standard procedure for solving linear recurrences of this type yields the claimed solution. □

THEOREM 5. *Let X be a real-valued function of subsets of $\mathcal{A}^{[h]}$ and let $r_X(\xi) := X(\xi) - X(\xi_1) - \ldots - X(\xi_m)$. Then,*

$$X^{(h)}(x) = \sum_{0\le j\le h} m^{h-j}\,(1+x)^{m^{[h]}-m^{[j]}}\, r_X^{(j)}(x). \tag{3}$$

Proof: By (*iii*) of Theorem 3, $X^{(h)}(x)$ satisfies recurrence (1). We solve (1) with the help of Lemma 4, wherein we take $B_h = r_X^{(h)}(x)$ and $A_h = m(1+x)^{m^h}$. Since $\prod_{j+1 \le k \le h} m(1+x)^{m^h} = m^{h-j}(1+x)^{m^{[h]}-m^{[j]}}$, the theorem follows. □

Remark. An 'average values' version of identity (3) was proved in [delaT87a,b] by an approach based on recurrence equations.

3. Analysis of endmarker tries

This section presents the data structures used by the set intersection algorithms that will be presented later in §5. *Endmarker tries*, which are natural adaptations of the original tries of de la Brandais [delaB59] and Fredkin [Fred60] for the purpose of storing sets of keys that may contain prefixing keys, have been analyzed in [delaT87a,b]. Applying the generating function tools of §2, we shall now rederive the exact average space and time requirements of the retrieval algorithms of *full endmarker tries* and *compact endmarker tries.*

Tries are implementations of the *prefix tree.* Let ξ be a finite set of strings composed from the totally ordered alphabet $\mathcal{A} = \{a_1, \ldots, a_m\}$, whose characters are ranked by $rank(a_i) = i$, $1 \le i \le m$. The set $pref(\xi) := \{x \mid xz \in \xi\}$ of prefixes of the elements of ξ has a natural tree structure. The set of nodes of this tree is $pref(\xi)$, wherein the length zero string (denoted by ε) is the root node; the i–th subtree a node $x \in pref(\xi)$, $1 \le i \le m$, consists of those strings of $pref(\xi)$ that begin with xa_i, i.e. $\{xa_iw \mid xa_iw \in \xi\}$. This

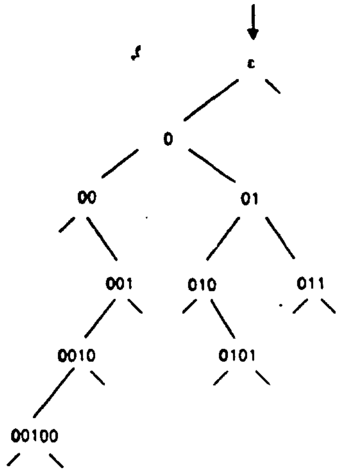


Figure 1. Prefix tree built from the set of keys $s = \{00100, 0101, 011\}$.

m–ary tree is called the *prefix tree* of ξ with respect to the alphabet $\mathcal{A}$ and will be denoted by $t(\xi)$.

The paths on the prefix tree $t(\xi)$ can be naturally mapped into the strings composed of symbols from $\mathcal{A}$. An empty path is mapped to the length zero string ε. A (possibly infinitely long) path $p = v_1, v_2, \ldots$, where v_{i+1} is the l_i–th son of v_i, is mapped to $a_{l_1} a_{l_2} \ldots \in \xi$. This correspondence defines an injective mapping between the maximal paths of $t(\xi)$ and the keys of ξ. This mapping is bijective precisely when no key in ξ is a prefix of another. The subset of all prefixing keys of ξ (that is, keys that are prefixes of other keys of ξ) will be denoted by $\mathit{prefixingkeys}(\xi) := \{\, k \in \xi \mid k \in \mathit{pref}\,(\xi - \{k\}\,\}$.

A finite set of keys $\xi \subset \mathcal{A}^{\circledast}$, which may include prefixing keys, can be easily encoded to yield a suitable representation of ξ as the set of maximal paths of an $(m+1)$–ary tree. This can be attained by attaching a symbol $\perp \notin \mathcal{A}$, the *endmarker*, to the end of the prefixing keys of ξ. In the resulting set of keys, $\xi[\perp] := (\xi - \mathit{prefixingkeys}(\xi)) \cup \{x\perp \mid x \in \mathit{prefixingkeys}(\xi)\}$, no key is a prefix of another. The *endmarker prefix tree* of ξ is the $(m+1)$–ary prefix tree, $t(\xi[\perp])$, of $\xi[\perp]$ with respect to the alphabet $\mathcal{A}^{\perp} := \{\perp\} \cup \mathcal{A}$, where the ranking in $\mathcal{A}$ has been extended by $rank(\perp) = 0$ (compare Figure 2).

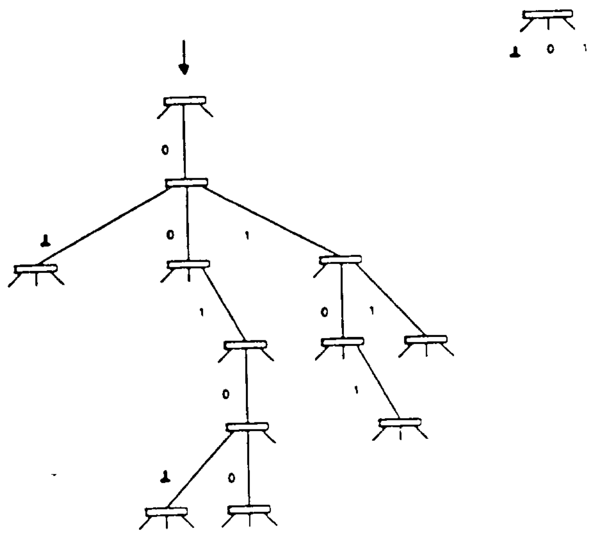
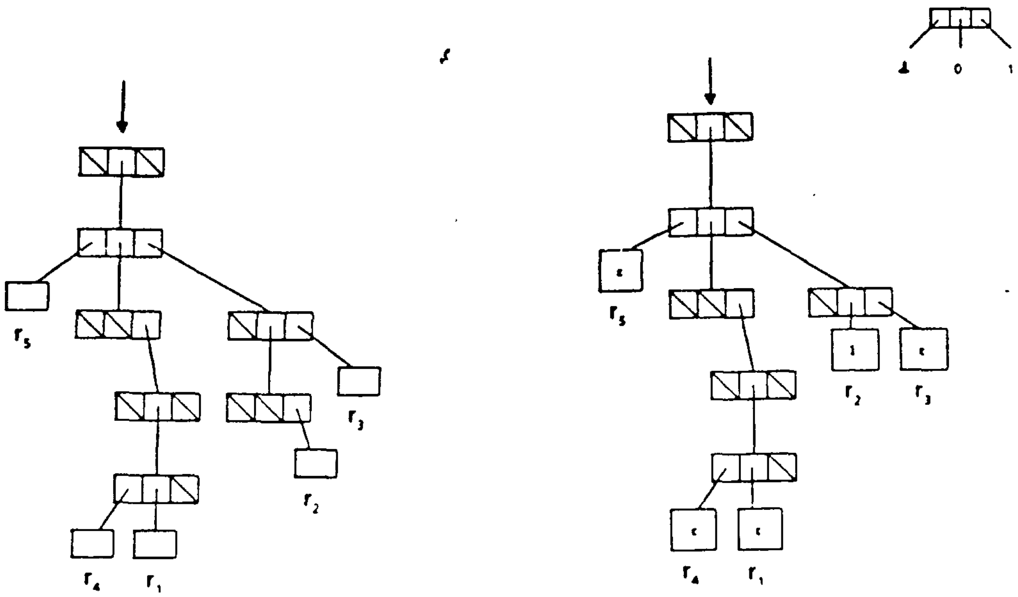


Figure 2. Endmarker prefix tree built from $s = \{00100, 0101, 011, 0010, 0\}$.

Let $r_1, \ldots, r_n$ be a collection of items where each item consists of a *key* part (which uniquely identifies the item) and a *data* part. For the remainder of this section, let ξ be the set of keys of these items and let us assume that the keys are strings composed from the ordered alphabet $\mathcal{A} = \{1, \ldots, m\}$. We shall consider two data structures for the storage of such collections of items; both structures are implementations of the endmarker prefix tree $t(\xi[\perp])$.



(a) full

(b) compact

Figure 3. Endmarker tries for items $r_1,\ldots,r_5$ with respective keys $k_1 = 00100$, $k_2 = 0101$, $k_3 = 011$, $k_4 = 0010$, $k_5 = 0$, and alphabet $\{\perp, 0, 1\}$ with $\perp < 0 < 1$.

3.1 *Full endmarker tries*

The construction of the full endmarker trie of ξ requires that all the keys in ξ have finite length, i.e. $\xi \subseteq \mathcal{A}^*$. In this case, the maximal paths of the endmarker prefix tree $t(\xi[\perp])$ end at terminal nodes. We thus have a bijective correspondence between the terminal nodes of $t(\xi[\perp])$ and the keys of ξ. The following implementation of $t(\xi[\perp])$ will be called the full endmarker trie built from ξ, and will be denoted by $t^{fe}(\xi)$. In $t^{fe}(\xi)$, a nonterminal node of $t(\xi[\perp])$ is represented by an array of pointers to its children; a terminal node v of $t(\xi[\perp])$, corresponding to a key $k \in \xi$, is represented by a pointer to the data of the item whose key is k (compare Figure 3(a))

In order to search for a key k, we traverse $t^{fe}(\xi)$ starting at the root and proceed recursively as follows. If the root is a nonterminal, we search for k in the first subtree when $k = \varepsilon$ and, when $k = iz$ with $i \in A$, we search for z in the $(i+1)$–th subtree. Otherwise, the search ends: it succeeds precisely when the root is a terminal node and $k = \varepsilon$.

The space required by this algorithm is thus proportional to the number $Sf(\xi)$ of nonterminal nodes in $t^{fe}(\xi)$ and its running time is proportional to the *total leaf node path length* $Tf(\xi) = tpl\big(t^{fe}(\xi)\big)$, where

$$tpl(g) := \sum_{all\ leaf\ nodes\ l\ of\ g} depth(l)$$

and $depth(l)$ denotes the edge length of the path that connects the root and l.

For computing the expectations of Sf and Tf with respect to the prefix model, it is convenient to cast the idea of full endmarker trie into a recursive definition.

DEFINITION : The *full endmarker trie* built with a finite set of keys $\xi \subset A^*$ is the $(m+1)$–ary tree, denoted by $t^{fe}(\xi)$, which is recursively defined as follows:

(i) If ξ is empty, $t^{fe}(\xi)$ is the empty tree.

(ii) If $\xi = \{\varepsilon\}$, $t^{fe}(\xi)$ is the tree whose root is a leaf node (i.e. all its subtrees are empty).

(iii) Otherwise, $t^{fe}(\xi)$ is the $(m+1)$–ary tree having an 'internal' root node whose subtrees are $t^{fe}(\xi \cap \{\varepsilon\})$, $t^{fe}(\xi_1), \dots, t^{fe}(\xi_m)$ in order.

This definition yields recursive expressions for $Sf(\xi)$, the number of internal nodes of $t^{fe}(\xi)$, and for its the total leaf node path length $Tf(\xi)$:

$$Sf(\xi) = 1 - \delta_{|\xi|,0} - \delta_{\xi,\{\varepsilon\}} + \sum_{i \in A} Sf(\xi_i), \tag{4}$$

$$Tf(\xi) = |\xi|(1 - \delta_{\xi,\{\varepsilon\}}) + \sum_{i \in A} Tf(\xi_i). \tag{5}$$

THEOREM 6. *The expectations of Sf and Tf over the n element subsets of $\mathcal{A}^{[h]}$ are*

$$E[Sf] = \sum_{1 \le j \le h} m^{h-j} \left[1 - \tau\left(m^{[h]}, m^{[j]}, n, 0\right) - \tau\left(m^{[h]}, m^{[j]}, n, 1\right)\right],$$

$$E[Tf] = \sum_{1 \le j \le h} m^{h-j} \left[\frac{n}{m^{[h]}} m^{[j]} - \tau\left(m^{[h]}, m^{[j]}, n, 1\right)\right],$$

where $\tau(a,b,c,d) = \frac{\binom{a-b}{c-d}}{\binom{a}{b}}$.

Proof: Let $r_{Sf}(\xi) := Sf(\xi) - \sum_{j \in \mathcal{A}} Sf(\xi_j)$. By (4), we can write $r_{Sf}(\xi) = I(\xi) - Z(\xi) - W(\xi)$, where $I(\xi) = 1$, $Z(\xi) = \delta_{|\xi|,0}$ and $W(\xi) = \delta_{|\xi \cap \{\epsilon\}|,|\xi|}$. To compute the generating function $Sf^{(h)}(x)$, we will first compute the generating function $r_{Sf}^{(h)}(x)$ and then apply Theorem 5. By properties (i) and (iii) of Lemma 2 we have $\mathbf{F}_h[I] = (1+x)^{m^{[h]}}$ and $\mathbf{F}_h[Z] = 1$. By direct counting we find

$$\mathbf{F}_h[W] = \sum_{\substack{\xi \in \mathcal{R}(\mathcal{A}^{[h]}) \\ \xi = \{\epsilon\}}} x^{|\xi|} = x.$$

Thus, $r_{Sf}^{(h)}(x) = (1+x)^{m^{[h]}} - 1 - x$. Since $r_{Sf}^{(0)}(x) = 0$, Theorem 5 applied to Sf gives

$$\begin{aligned}
\mathbf{F}_h[Sf] &= \sum_{0 \le j \le h} m^{h-j} (1+x)^{m^{[h]} - m^{[j]}}\, r_{Sf}^{(j)}(x). \\
&= \sum_{1 \le j \le h} m^{h-j} (1+x)^{m^{[h]}} - \sum_{1 \le j \le h} m^{h-j} (1+x)^{m^{[h]} - m^{[j]} + 1}. \\
&= \sum_{1 \le j \le h} m^{h-j} \sum_{n \ge 0} x^n \left[\binom{m^{[h]}}{n} - \binom{m^{[h]} - m^{[j]} + 1}{n}\right].
\end{aligned}$$

Extracting the coefficient of x^n from the above expression yields

$$\binom{m^{[h]}}{n} E[Sf] = \sum_{1 \le j \le h} m^{h-j} \left[\binom{m^{[h]}}{n} - \binom{m^{[h]} - m^{[j]} + 1}{n}\right].$$

The value of $E[Tf]$ can computed in a similar manner. □

3.2 *Compact endmarker tries*

For an arbitrary finite subset $\xi \subset \mathcal{A}^{\circledast}$ (ξ may contain infinitely long keys) the following implementation of the endmarker prefix tree $t(\xi[\perp])$ gives rise to the ***compact endmarker trie***, which will be denoted by $t^{ce}(\xi)$. The construction of $t^{ce}(\xi)$ is analogous as for full endmarker tries, except that the branching of $t(s[\perp])$ is stopped at subtrees T consisting of nodes of out-degree less than 2 (i.e. T is a path). Such a subtree T is collapsed into a single terminal node and the string corresponding to the unique maximal path of T becomes the label of this terminal node.

The search algorithm for $t^{ce}(\xi)$ is analogous as for full endmarkers tries except that, on reaching a terminal node, the label of this terminal node must be compared with the unexamined part of the search key; only when these two are equal is the search successful.

The storage required by this algorithm is proportional to the the number $Sc(\xi)$ of internal nodes of $t^{ce}(\xi)$; its running time is proportional to the total leaf node path length $Tc(\xi) := tpl(t^{ce}(\xi))$.

A recursive definition of the compact endmarker trie $t^{ce}(\xi)$ can be obtained from the recursive definition of the full endmarker trie given in §3.1 by considering finite sets $\xi \subset \mathcal{A}^{\circledast}$ and replacing condition (ii) by the following 'compaction' condition:

$(ii*)$ If $\xi = \{x\}$, $t^{ce}(\xi)$ is equal to a single leaf node with a label equal to x.

This recursive definition of $t^{ce}(\xi)$ yields the following recursive expressions of $Sc(\xi)$ and $Tc(\xi)$:

$$Sc(\xi) = r_{Sc}(\xi) + \sum_{i \in \mathcal{A}} Sc(\xi_i), \tag{6}$$

$$Tc(\xi) = r_{Tc}(\xi) + \sum_{i \in \mathcal{A}} Tc(\xi_i), \tag{7}$$

with $r_{Sc}(\xi) = 1 - \delta_{|\xi|,0} - \delta_{|\xi|,1}$ and $r_{Tc}(\xi) = |\xi|(1 - \delta_{|\xi|,1})$.

THEOREM 7. *The expectations of Sc and Tc over the n element subsets of $\mathcal{A}^{[h]}$ are*

$$E[Sc] = \sum_{1 \le j \le h} m^{h-j}\big[1 - \tau(m^{[h]}, m^{[j]}, n, 0) - m^{[j]}\tau(m^{[h]}, m^{[j]}, n, 1)\big],$$

$$E[Tc] = \sum_{1 \le j \le h} m^{h-j} m^{[j]} \Big[\frac{n}{m^{[h]}} - \tau(m^{[h]}, m^{[j]}, n, 1)\Big],$$

where $\tau(a,b,c,d) = \frac{\binom{a-b}{c-d}}{\binom{a}{b}}$.

Proof: For a set $\xi \subset \mathcal{A}^*$, let $M(\xi)$ be the number of nodes of the full end-marker trie $t^{fe}(\xi)$ having exactly one terminal node among its descendants. Relations (4), (5), (6), (7) imply

$$M(\xi) = Sf(\xi) - Sc(\xi) = Tf(\xi) - Tc(\xi). \tag{8}$$

Also, $r_M(\xi) := M(\xi) - M(\xi_1) - \ldots - M(\xi_m) = (1 - |\xi \cap \{\varepsilon\}|)\delta_{|\xi|,1}$. By direct counting we find $\mathbf{F}_0[r_M] = 0$,

$$\mathbf{F}_h[r_M] = \sum_{\substack{\xi \in \mathcal{R}(\mathcal{A}^{[h]}) \\ |\xi| = 1, \xi \neq \{\epsilon\}}} x^{|\xi|} = (m^{[h]} - 1)x,$$

and from Theorem 5 we deduce

$$\begin{aligned}\mathbf{F}_h[M] &= \sum_{0 \le j \le h} m^{h-j}(1+x)^{m^{[h]} - m^{[j]}}\, r_M^{(j)}(x) \\ &= \sum_{1 \le j \le h} m^{h-j}(m^{[j]} - 1)\, x\, (1+x)^{m^{[h]} - m^{[j]}}.\end{aligned}$$

Extracting the coefficient x^n from this expression yields

$$\binom{m^{[h]}}{n} E[M] = \sum_{1 \le j \le h} m^{h-j}\left[(m^{[j]} - 1)\binom{m^{[h]} - m^{[j]}}{n-1}\right].$$

The claimed expectations $E[Sc]$ and $E[Tc]$ now follow from (8) and Theorem 6. □

Remark. This generating function approach can be also used for calculating the expectations of other tries cost functions of interest considered in [delaT87b].

4. The binary set–intersection prefix model

The sample space of the *binary set–intersection prefix model* consists of a class of ordered pairs (ξ,η) of sets of binary string keys. This class depends on four parameters: the size l of the first component ξ, the size n of the second component η, the size k of the intersection $\xi \cap \eta$, and the maximum length h of the of the binary string keys. For nonnegative integers h, l, n, and k, the probability space of the binary set–intersection prefix model is

$$\mathcal{I}_{h,l,n,k} := \{\, (\xi,\eta) \mid \xi,\eta \subset \{0,1\}^{[h]}\ ,\ |\xi| = l,\ |\eta| = n,\ |\xi \cap \eta| = k \,\},$$

where $\{0,1\}^{[h]} = \{0,1\}^{[0]} \cup \{0,1\}^{[1]} \cup \ldots \{0,1\}^{[h]}$, and all set pairs (ξ,η) are assumed to be equally probable. The expectation of a real–valued mapping $X(\xi,\eta)$ over the pairs $(\xi,\eta) \in \mathcal{I}_{h,l,n,k}$ will be denoted by $E[X]$. The sum

$$N_{h,l,n,k}[X] := \sum_{\substack{\xi,\eta \subset \{0,1\}^{[h]} \\ |\xi|=l,\ |\eta|=n,\ |\xi\cap\eta|=k}} X(\xi,\eta)$$

is related to the expectation of X by $N_{h,l,n,k}[X] = |\mathcal{I}_{h,l,n,k}| E[X]$ and will be called the *normalized expectation* of X.

4.1 *Translation rules*

Throughout this section X will denote a real–valued mapping of ordered pairs (ξ,η) of sets $\xi,\eta \subseteq \{0,1\}^{[h]}$. To each such mapping X we associate the *generating function of normalized expectations* $X^{(h)}(x,y,t)$,

$$\begin{aligned} X^{(h)}(x,y,t) &:= \sum_{\xi,\eta \subset \{0,1\}^{[h]}} X(\xi,\eta)\, x^{|\xi|}\, y^{|\eta|}\, t^{|\xi\cap\eta|} \\ &= \sum_{l,n,k \geq 0} N_{h,l,n,k}[X]\, x^m\, y^n\, t^k. \end{aligned}$$

We shall now establish translation rules, between X and its generating function $X^{(h)}(x,y,t)$, similar to those derived for the prefix model in §2.1. As earlier, the translation rules will be formulated as properties of the operator $\mathbf{F}_h$ which maps a functions X into its generating function $\mathbf{F}_h[X] := X^{(h)}(x,y,t)$. Some of these properties can be conveniently expressed in terms of the mappings $\overline{\mathbf{P}}_c(\xi,\eta) := (\xi_c,\eta_c)$ (where $\xi_c = \{x \mid cx \in \xi\}$) for each $c \in \{0,1\}$, and also $\overline{\mathbf{P}}_\perp(\xi,\eta) := (\xi \cap \{\varepsilon\}, \eta \cap \{\varepsilon\})$.

LEMMA 8. [*Additive-multiplicative rule*] *Let X, Y, and Z be real-valued mappings of ordered pairs (ξ,η) of sets $\xi,\eta \subset \{0,1\}^{[h]}$.*

(*i*) $\mathbf{F}_h[\lambda.X] = \lambda \mathbf{F}_h[X]$;

(*ii*) $\mathbf{F}_h[X+Y] = \mathbf{F}_h[X] + \mathbf{F}_h[Y]$;

(*iii*) $\mathbf{F}_h[(Y \circ \overline{\mathbf{P}}_0).(Z \circ \overline{\mathbf{P}}_1)] = (1+x+y+xyt)\,\mathbf{F}_{h-1}[Y]\,\mathbf{F}_{h-1}[Z],\ h \geq 1$;

(*iv*) $\mathbf{F}_h[(X \circ \overline{\mathbf{P}}_\perp).(Y \circ \overline{\mathbf{P}}_0).(Z \circ \overline{\mathbf{P}}_1)] = \mathbf{F}_0[X]\,\mathbf{F}_{h-1}[Y]\,\mathbf{F}_{h-1}[Z], h \geq 1$.

Proof: Properties (*i*) and (*ii*) follow immediately from the definitions of $\mathbf{F}_h$ and $\mathbf{P}_c$. Property (*iii*) can be deduced from (*iv*) (taking $X(s) := I(s) = 1$) and

$$\begin{aligned} X^{(0)}(x,y,t) &= \sum_{\xi,\eta \subset \{\varepsilon\}} x^{|\xi|} y^{|\eta|} t^{|\xi \cap \eta|} \\ &= 1 + x + y + xyt. \end{aligned} \tag{9}$$

In order to prove (*iv*), we first observe the mapping $\xi \to (\xi \cap \{\varepsilon\}, \xi_0, \xi_1)$ is a bijection between $\mathcal{R}(\{\varepsilon\}) \times \{0,1\}^{[h-1]} \times \{0,1\}^{[h-1]}$ and $\{0,1\}^{[h]}$. Taking

$W := (X \circ \overline{\mathbf{P}}_{\perp})(Y \circ \overline{\mathbf{P}}_0)(Z \circ \overline{\mathbf{P}}_1)$, we have

$$\begin{aligned}
\mathbf{F}_h[W] &= \sum_{\xi,\eta \in \mathcal{R}(\{0,1\}^{[h]})} W(\xi,\eta)\, x^{|\xi|} y^{|\eta|}\, t^{|\xi\cap\eta|} \\
&= \sum_{\xi,\eta \in \mathcal{R}(\{0,1\}^{[h]})} X(\xi \cap \{\varepsilon\}\,,\, \eta \cap \{\varepsilon\})\, Y(\xi_0,\eta_0) Z(\xi_1,\eta_1) \\
&\qquad\qquad x^{|\xi_0|+|\xi_1|+|\xi\cap\{\varepsilon\}|}\, y^{|\eta_0|+|\eta_1|+|\eta\cap\{\varepsilon\}|} \\
&\qquad\qquad t^{|\xi_0\cap\eta_0|+|\xi_1\cap\eta_1|+|\xi\cap\eta\cap\{\varepsilon\}|} \\
&= Y^{(h-1)}(x,y,t)\, Z^{(h-1)}(x,y,t) \sum_{\mu,\nu \in \mathcal{R}(\{\varepsilon\})} X(\mu,\nu)\, x^{|\mu|}\, y^{|\nu|}\, t^{|\mu\cap\nu|} \\
&= Y^{(h-1)}(x,y,t)\, Z^{(h-1)}(x,y,t)\, X^{(0)}(x,y,t),
\end{aligned}$$

which is as claimed in (iv). □

LEMMA 9. [*Initialization rule*]. *If* $I(\xi,\eta) := 1$ *then*

$$I^{(h)}(x,y,t) = (1 + x + y + xyt)^{2^{|h|}}.$$

Proof: The case $h = 0$ follows from (9). For $h > 1$ we write $I = (I \circ \overline{\mathbf{P}}_0).(I \circ \overline{\mathbf{P}}_1)$ and applying (iii) of Lemma 8 deduce $I^{(h)}(x,y,t) = (1 + x + y + xyt)[I^{(h-1)}(x,y,t)]^2$, $h \geq 1$. Solving this recurrence yields the desired expression of $I^{(h)}(x,y,t)$. □

THEOREM 10. *Let* X *and* Y *be real-valued functions of pairs* (ξ,η) *of subsets* $\xi,\eta \subseteq \{0,1\}^{[h]}$, *and let us assume that* $h \geq 1$.

(i) *If* $X(\xi,\eta) = Y(\xi\cap\{\varepsilon\}\,,\,\eta\cap\{\varepsilon\})$ *then* $X^{(h)}(x,y,t) = (1+x)^{2^{|h|}-1}\, Y^{(0)}(x,y,t)$.

(ii) *If* $X = Y \circ \overline{\mathbf{P}}_c$, *with* $c \in \{0,1\}$, *then*

$$X^{(h)}(x,y,t) = (1 + x + y + xyt)^{2^h}\, Y^{(h-1)}(x,y,t).$$

(*iii*) *If* $r_X(\xi,\eta) := X(\xi,\eta) - X(\xi_0,\eta_0) - X(\xi_1,\eta_1))$ *then*

$$X^{(h)}(x) = r_X^{(h)}(x) + 2\,(1 + x + y + xyt)^{2^h}\, X^{(h-1)}(x). \tag{10}$$

Proof: These properties can be proved with the help of Lemma 8 and Lemma 9 proceeding in a similar manner as in the proof of Theorem 3. □

Note that recurrence (10) can be solved by means of Lemma 4, which also works as the iteration rule for the set–intersection prefix model.

The following Theorem 11 reduces the calculation of the generating function $X^{(h)}(x,y,t)$ to the determination of the generating function $r_X^{(h)}(\xi,\eta)$ of $r_X(\xi,\eta) := X(\xi,\eta) - X(\xi_0,\eta_0) - X(\xi_1,\eta_1)$. In §5, we will apply this theorem to compute the average running time of the set intersection algorithms.

THEOREM 11. *If Let X be a real-valued function of pairs* (ξ,η) *of subsets* $\xi,\eta \subseteq \{0,1\}^{[h]}$, *and let* $r_X(\xi,\eta) = X(\xi,\eta) - X(\xi_0,\eta_0) - X(\xi_1,\eta_1)$. *Then,*

$$X^{(h)}(x,y,t) = \sum_{0\le j\le h} 2^{h-j}\,(1 + x + y + xyt)^{2^{[h]} - 2^{[j]}}\, r_X^{(j)}(x,y,t).$$

Proof: This expression of the generating function $X^{(h)}(x,y,t)$ can be obtained by solving the recurrence equation (10) with the help of Lemma 4. □

5. Analysis of algorithms for set intersection

We now present two algorithms for computing the intersection of sets of binary string keys. For each of them we will compute the exact average running time with respect to the binary set–intersection prefix model.

5.1 *Average set–intersection time using full endmarker tries*

The set intersection $\mathit{Intersectf}(\xi,\eta) := \xi \cap \eta$, with $\xi, \eta \subseteq \{0,1\}^{[h]}$, can be computed by the following algorithm:

1. If $|\xi| = 0$ or $|\eta| = 0$ then $\mathit{Intersectf}(\xi,\eta) \leftarrow \emptyset$;
2. If $\xi = \{\varepsilon\}$ then $\mathit{Intersectf}(\xi,\eta) \leftarrow \xi$;
3. If $\eta = \{\varepsilon\}$ then $\mathit{Intersectf}(\xi,\eta) \leftarrow \eta$;
4. Otherwise,

$$\mathit{Intersectf}(\xi,\eta) \leftarrow (\xi \cap \eta \cap \{\varepsilon\}) \cup 0\mathit{Intersectf}(\xi_0,\eta_0) \cup 1\mathit{Intersectf}(\xi_1,\eta_1).$$

Let $t^{fe}(\xi)$ and $t^{fe}(\eta)$ be the full endmarker tries built from ξ and η respectively. The following observations lead to an implementation of the above algorithm. The root node of $t^{fe}(\xi)$ (respectively $t^{fe}(\eta)$) is a terminal node if and only if $\xi = \{\varepsilon\}$ (respectively $\eta = \{\varepsilon\}$). If $\xi, \eta \not\subseteq \{\varepsilon\}$, the zero length string $\varepsilon \in \xi \cap \eta$ exactly when both of the first sons of $t^{fe}(\xi)$ and $t^{fe}(\eta)$ are nonempty. The sets ξ_0 and ξ_1 are represented by the second and third subtrees of $t^{fe}(\xi)$; η_0 and η_1 are represented by the second and third subtrees of $t^{fe}(\eta)$.

The function $\mathit{Intersectf}(\xi,\eta)$ can be realized as the following simultaneous traversal of the trees $t^{fe}(\xi)$ and $t^{fe}(\eta)$. We start at the roots of the tries. Step 1 is implemented by testing whether one of the two trees is empty, and Step 2 and Step 3 by testing whether the root node of the appropriate trie is a terminal node. In Step 4, we can compute $\xi \cap \eta \cap \{\varepsilon\}$ by examining the first subtrees $t^{fe}(s)$ and $t^{fe}(s)$; the recursive call to $\mathit{Intersectf}(\xi_0,\eta_0)$ (respectively $\mathit{Intersectf}(\xi_1,\eta_1)$) is then realized by simultaneously visiting the second (respectively third) subtrees of $t^{fe}(\xi)$ and $t^{fe}(\eta)$; these subtrees represent the sets ξ_0 and η_0 (respectively ξ_1 and η_1).

The total time necessary to compute the intersection is thus proportional to the number, $\mathit{Ife}(\xi,\eta)$, of pairs of nodes that are simultaneously visited in $t^{fe}(\xi)$ and $t^{fe}(\eta)$ (i.e., $\mathit{Ife}(\xi,\eta)$ equals the total number of times that Step 4 is executed). The function $r_{Ife}(\xi,\eta) := \mathit{Ife}(\xi,\eta) - \mathit{Ife}(\xi_0,\eta_0) - \mathit{Ife}(\xi_1,\eta_1)$ can be written as $r_{Ife}(\xi,\eta) = 1 - \delta_{b(\xi,\eta)}$ where

$$b(\xi,\eta) = (\ \xi = \{\varepsilon\})\ \mathit{or}\ (|\xi| = 0)\ \mathit{or}\ (\eta = \{\varepsilon\})\ \mathit{or}\ (|\eta| = 0). \qquad (11)$$

The results of the following lemma will be helpful in extracting coefficients from the generating functions that will emerge from our computations. The coefficient of the term $x^l y^n t^k$ in a polynomial $P(x,y,t)$ will be denoted by $[l,n,k]P(x,y,t)$.

LEMMA 12. The coefficient

$$K_{l,n,k}[\alpha,\beta] := [l,n,k]\left\{\left[\,(1+x)^\alpha + (1+y)^\alpha - 1\right](1+x+y+xyt)^\beta\right\}$$

equals

$$K_{l,n,k}[\alpha,\beta] = I_{l,n,k}[\alpha,\beta] + I_{n,l,k}[\alpha,\beta] - I_{l,n,k}[0,\beta]. \qquad (12)$$

where $I_{l,n,k}[\alpha,\beta] := \binom{\beta}{k}\binom{\beta-k}{n-k}\binom{\beta+\alpha-n}{l-k}$. Also, $K_{l,n,k}[0,2^{[h]}] = I_{l,n,k}[0,2^{[h]}] = |\mathcal{I}_{h,l,n,k}|$.

Proof: The identity

$$\begin{aligned} I_{l,n,k}[\alpha,\beta] &= [x^l y^n t^k]\{(1+x)^\alpha\,(1+x+y+xyt)^\beta\} \\ &= \binom{\beta}{k}\binom{\beta-k}{n-k}\binom{\beta+\alpha-n}{l-k}, \end{aligned}$$

was established in [FRS85] by expanding $(1+x)^\alpha\,(1+x+y+xyt)^\beta$ first in t, and then in x and in y. By the symmetry of $(1+x)^\alpha\,(1+x+y+xyt)^\beta$ with respect to x and y, we deduce $[x^l y^n t^k](1+y)^\alpha\,(1+x+y+xyt)^\beta = I_{n,l,k}[\alpha,\beta]$, and $I_{n,l,k}[0,\beta] = I_{l,n,k}[0,\beta]$. Thus relation (12) follows, and also $K_{l,n,k}[0,\beta] = I_{l,n,k}[0,\beta] = |\mathcal{I}_{h,l,n,k}|$. □

With the help of Theorem 11, we shall now calculate the average time necessary to compute the intersection of a pair of sets $(\xi,\eta) \in \mathcal{I}_{h,l,n,k}$ by means of the implementation of $\mathit{Intersectf}(\xi,\eta)$ described above.

THEOREM 13. *The expected value of $Ife(\xi,\eta)$ over the pairs of sets $(\xi,\eta) \in \mathcal{I}_{h,l,n,k}$ is*

$$E[Ife] = (2^{[h]}-1) - \frac{1}{|\mathcal{I}_{h,l,n,k}|} \sum_{1\le j\le h} 2^{h-j} K_{l,n,k}[2^{[j]}-1\,,\, 2^{[h]}-2^{[j]}+1],$$

where $|\mathcal{I}_{h,l,n,k}| = K_{l,n,k}[0\,,\,2^{[h]}]$.

Proof: From Lemma 8 and Lemma 9 we deduce $r^{(h)}_{Ife}(x,y,t) = (1+x+y+xyz)^{2^{[h]}} - Z^{(h)}(x,y,t)$, with $Z(s) = \delta_{b(\xi,\eta)}$ and $b(\xi,\eta)$ as in (11). By direct counting we find

$$Z^{(h)}(x,y,t) = \sum_{\substack{\xi,\eta\subset\{0,1\}^{[h]} \\ (|\xi|=0)\ or\ (\xi=\{\epsilon\})\ or\ (|\eta|=0)\ or\ (\eta=\{\epsilon\})}} x^{|\xi|}\,y^{|\eta|}\,t^{|\xi\cap\eta|}$$

$$= (1+x+y+xyz)\left[(1+x)^{2^{[h]}-1} + (1+y)^{2^{[h]}-1} - 1\right].$$

Since $r^{(0)}_{Ife}(x,y,t) = 0$, Theorem 11 yields

$$Ife^{(h)}(x,y,t) = (2^{[h]}-1)(1+x+y+xyz)^{2^{[h]}} - \sum_{1\le j\le h} 2^{h-j}\,(1+x+y+xyz)^{2^{[h]}-2^{[j]}}\,Z^{(j)}(x,y,t).$$

Extracting the coefficient of $x^l y^n t^k$ from this expression with the help of Lemma 12, we arrive at the desired expectation of $E[Ife]$. □

Remark. It may be noted that $l = n = k$ implies $Ife(\xi,\xi) = S^{fe}(\xi)$, the number of internal nodes in the full endmarker trie of ξ. This is reflected by our calculations. When we set $l = n = k$, the expression of the expectation $E[Ife]$ given in Theorem 13 reduces (after some algebraic cancellations) to the expectation $E_{hn}[Sf]$ with respect to the prefix model computed earlier in Theorem 6 (with the alphabet size $m = 2$).

5.2 *Average set–intersection time using compact endmarker tries*

We shall now consider another algorithm for set intersection, which is based on compact endmarker tries. Let $Part(\alpha,\beta)$ be the function of $\alpha,\beta \subseteq \{0,1\}^{[h]}$ that has the value α when $\alpha \subset \beta$, and the value $\emptyset$ otherwise. The set intersection $Intersectc(\xi,\eta) := \xi \cap \eta$, with $\xi,\eta \subseteq \{0,1\}^{[h]}$, can be computed by the following algorithm:

1. If $|\xi| = 0$ or $|\eta| = 0$ then $Intersectc(\xi,\eta) \leftarrow \emptyset$;
2. If $|\xi| = 1$ then $Intersectc(\xi,\eta) \leftarrow Part(\xi,\eta)$;
3. If $|\eta| = 1$ then $Intersectc(\xi,\eta) \leftarrow Part(\eta,\xi)$;
4. Otherwise,

$$Intersectc(\xi,\eta) \leftarrow (\xi \cap \eta \cap \{\varepsilon\}) \cup 0Intersectc(\xi_0,\eta_0) \cup 1Intersectc(\xi_1,\eta_1).$$

Let $t^{ce}(\xi)$ and $t^{ce}(\xi)$ be the respective compact endmarker tries of ξ and η, and let us assume that $|\xi|,|\eta| \geq 2$. Then, $\varepsilon \in \xi \cap \eta$ precisely when the first sons of $t^{ce}(\xi)$ and $t^{ce}(\eta)$ are nonempty. The sets ξ_0 and η_0 are represented by the respective second subtrees of $t^{ce}(\xi)$ and $t^{ce}(\eta)$; ξ_1 and η_1 are represented by the third subtrees of $t^{ce}(\xi)$ and $t^{ce}(\eta)$.

The algorithm $Intersectc(\xi,\eta)$ can thus be implemented by the simultaneous traversal of the compact endmarker tries $t^{ce}(\xi)$ and $t^{ce}(\eta)$. We start at the root nodes of the tries, and implement Step 1 by testing whether one of the trees is empty. Step 2 (respectively Step 3) is realized by testing whether the root node of $t^{ce}(\xi)$ (respectively $t^{ce}(\eta)$) is a terminal node. If it is, i.e. $\xi = \{x\}$ (repectively $\eta = \{y\}$), $Part(\{x\},\eta)$ (respectively $Part(\xi,\{y\})$) is implemented by searching for the key x in $t^{ce}(\eta)$ (respectively searching for y in $t^{ce}(\xi)$). If this search is successful, we return the value $\{x\}$ (respectively $\{y\}$); otherwise, we return the value $\emptyset$. Since Step 4 is executed precisely when $|\xi|,|\eta| \geq 2$, we can then compute $\xi \cap \eta \cap \{\varepsilon\}$ by simply examining the first subtrees of $t^{ce}(\xi)$ and $t^{ce}(\eta)$ (these subtrees are terminal nodes precisely when $\varepsilon \in \xi \cap \eta$). The recursive call $Intersectc(\xi_0, \eta_0)$ (respectively $Intersectc(\xi_1, \eta_1)$) can be implemented by simultaneously visiting the second sons (respectively third sons) of $t^{ce}(\xi)$ and $t^{ce}(\eta)$, which are the root nodes of compact endmarker tries representing the sets ξ_0 and η_0 (respectively ξ_1 and η_1).

The time required to compute $\xi \cap \eta$ by the above algorithm is proportional to $Ice(\xi, \eta)$, which is defined as the number of pairs of internal nodes simultaneously visited in tries $t^{ce}(\xi)$ and $t^{ce}(\eta)$ (i.e., the number of times that Step 4 is executed) plus the number of internal nodes visited in only one of the tries after a terminal node has been reached in the other (i.e., the number of nodes visited while executing the calls to *Part*).

We shall calculate the expectation of *Ice* in two ways. Our first calculation makes use of the relation between full and compact endmarker tries. That is, the compact endmarker trie $t^{ce}(\xi)$ results from the full endmarker trie $t^{fe}(\xi)$ by pruning every internal node that has only one terminal node among its descendants. Hence, $M(\xi, \eta) := Ife(\xi, \eta) - Ice(\xi, \eta)$ is equal to the number of pairs of internal nodes of $t^{fe}(\xi)$ and $t^{fe}(\eta)$ simultaneously visited, in the implementation of $Intersectf(\xi, \eta)$ given in §5.1, such that each internal node in the pair has only one terminal among its descendants. Thus the function $r_M(\xi, \eta) := M(\xi, \eta) - M(\xi_0, \eta_0) - M(\xi_1, \eta_1)$ has the expresion $r_M(\xi, \eta) = \delta_{(|\xi|=1)\, and\, (\xi \neq \{\varepsilon\})}\, \delta_{(|\eta|=1)\, and\, (\eta \neq \{\varepsilon\})}$.

THEOREM 14. *The expectation of $M(\xi, \eta)$ over the pairs $(\xi, \eta) \in I_{h,l,n,k}$ is*

$$E[M] = \frac{1}{|I_{h,l,n,k}|} \sum_{1 \le j \le h} 2^{h-j} \left(2^{[j]} - 1\right) \Big\{ K_{l-1,n-1,k-1}[0\,,\, 2^{[h]} - 2^{[j]}] + \left(2^{[j]} - 2\right) K_{l-1,n-1,k}[0\,,\, 2^{[h]} - 2^{[j]}] \Big\},$$

where $|I_{h,l,n,k}| = K_{l,n,k}[0\,,\, 2^{[h]}]$.

Proof: Direct counting yields

$$r_M^{(h)}(x,y,t) = \sum_{\substack{\xi,\eta \subset \{0,1\}^{[h]} \\ (|\xi|=1) \text{ and } (\xi \neq \{\epsilon\}) \text{ and } (|\eta|=1) \text{ and } (\eta \neq \{\epsilon\})}} x^{|\xi|}\, y^{|\eta|}\, t^{|\xi \cap \eta|}$$

$$= (2^{[h]} - 1)\, x\, y\, \left[t + 2^{[h]} - 2 \right].$$

Applying Theorem 11 to M gives

$$M^{(h)}(x,y,t) = \sum_{1 \leq j \leq h} 2^{h-j}\, (2^{[j]} - 1)\, (1 + x + y + xyz)^{2^{[h]} - 2^{[j]}}\, x\, y\, \left[t + 2^{[j]} - 2 \right],$$

and the desired expectation $E[M]$ results by extracting the coefficient of $x^l y^n t^k$ with the help of Lemma 12. □

THEOREM 15. *The average total time $E[Ice]$ required to compute the intersection using compact endmarker tries is*

$$\begin{aligned} E[Ice] = (2^{[h]} - 1) \\ - \frac{1}{|\mathcal{I}_{h,l,n,k}|} \Bigg\{ & \sum_{1 \leq j \leq h} 2^{h-j} K_{l,n,k}[2^{[j]} - 1\,,\, 2^{[h]} - 2^{[j]} + 1] \\ & + \sum_{1 \leq j \leq h} 2^{h-j} (2^{[j]} - 1) \Big[K_{l-1,n-1,k-1}[0\,,\, 2^{[h]} - 2^{[j]}] \\ & \qquad + (2^{[j]} - 2) K_{l-1,n-1,k}[0\,,\, 2^{[h]} - 2^{[j]}] \Big] \Bigg\}, \end{aligned}$$

where $|\mathcal{I}_{h,l,n,k}| = K_{l,n,k}[0\,,\, 2^{[h]}]$.

Proof: This expression of the expectation can be obtained from the relation $E[Ice] = E[Ife] - E[M]$, and the values of $E[Ife]$ and $E[M]$ provided by Theorem 13 and Theorem 14. □

The following alternative way of computing $E[Ice]$ yields additional information of interest to the cost analysis. We break up the values of the function *Ice* into two components,

$$Ice(\xi,\eta) = A(\xi,\eta) + B(\xi,\eta). \qquad (13)$$

The first component, $A(\xi,\eta)$, is the number of pairs internal nodes of $t^{ce}(\xi)$ and $t^{ce}(\eta)$ that are simultaneously visited in the implementation of of the above algorithm for *Intersectc*(ξ,η) (i.e., the number of times Step 4 is is executed). This quantity is of interest in its own right since, as remarked by Trabb Pardo in [Tra78], $A(\xi,\eta)$ measures the risk of computing the intersection $\xi \cap \eta$ to find that it is empty. The second component, $B(\xi,\eta)$, is the number of internal nodes visited in only one of the tries after an internal node has been encountered in the other (i.e., the number of nodes visited in the execution of the calls to *Part*(ξ,η)).

Since Step 4 is executed precisely when $|\xi|, |\eta| \geq 2$, $r_A(\xi,\eta) := A(\xi,\eta) - A(\xi_0,\eta_0) - A(\xi_1,\eta_1)$ can be written as

$$r_A(\xi,\eta) = 1 - Z(\xi,\eta), \tag{14}$$

with $Z(\xi,\eta) = \delta_{(|\xi|\leq 1)\,or\,(|\eta|\leq 1)}$. We further observe that an internal node v of $t^{ce}(\xi)$ (respectively $t^{ce}(\eta)$) is visited in the process of executing the function *Part*(ξ,η) (respectively *Part*(η,ξ)) precisely when the string x, corresponding to the path that connects the root and v, satisfies $|\xi_x| \geq 2$ and $|\eta_x| = 1$ (respectively $|\eta_x| \geq 2$ and $|\xi_x| = 1$). Thus, $r_B(\xi,\eta) := B(\xi,\eta) - B(\xi_0,\eta_0) - B(\xi_1,\eta_1)$ has the expression

$$r_B(\xi,\eta) = \delta_{|\xi|=1}\,\delta_{\xi\neq\{\epsilon\}}\,\delta_{|\eta|\geq 2} + \delta_{|\eta|=1}\,\delta_{\eta\neq\{\epsilon\}}\,\delta_{|\xi|\geq 2}. \tag{15}$$

THEOREM 16. *The expectation of $A(\xi,\eta)$ over the pairs $(\xi,\eta) \in \mathcal{I}_{h,l,n,k}$ is*

$$\begin{aligned} E[A] = (2^{[h]} - 1) \\ - \frac{1}{|\mathcal{I}_{h,l,n,k}|}\Big\{ \sum_{1\leq j\leq h} 2^{h-j} 2^{[j]} \Big[K_{l,n,k}[2^{[j]} - 1\,,\, 2^{[h]} - 2^{[j]} + 1] \\ - (2^{[j]} - 1)\, K_{l-1,n-1,k}[0\,,\, 2^{[h]} - 2^{[j]}]\Big] \\ - \sum_{1\leq j\leq h} 2^{h-j}(2^{[j]} - 1)\, K_{l,n,k}[2^{[j]}\,,\, 2^{[h]} - 2^{[j]}]\Big\}, \end{aligned}$$

with $|\mathcal{I}_{h,l,n,k}| = K_{l,n,k}[0\,,\, 2^{[h]}]$.

Proof: From the expression of r_A given in (14), and with the help of Lemma 8 and Lemma 9, we deduce

$$r_A^{(h)}(x,y,t) = (1+x+y+xyt)^{2^{[h]}} - Z^{(h)}(x,y,t),$$

and by direct counting we find

$$\begin{aligned} Z^{(h)}(x,y,t) = 2^{[h]}(1+x+y+xyt)\big[(1+x)^{2^{[h]}-1} + (1+y)^{2^{[h]}-1} - 1\big] \\ - (2^{[h]}-1)\big[(1+x)^{2^{[h]}} + (1+y)^{2^{[h]}} - 1 + 2^{[h]}xy\big]. \end{aligned}$$

Substituting the resulting expression of $r_A^{(h)}(x,y,t)$ in the expression of $A^{(h)}(x,y,t)$ provided by Theorem 11, and noting that $r_A^{(0)}(x,y,t) = 0$, we obtain

$$\begin{aligned} A^{(h)}(x,y,t) = (2^{[h]}-1)(1+x+y+xyt)^{2^{[h]}} \\ - \sum_{1\le j\le h} 2^{h-j}(1+x+y+xyt)^{2^{[h]}-2^{[j]}} Z^{(j)}(x,y,t). \end{aligned}$$

Extracting the coefficient of the term $x^l y^n t^k$ with the aid of Lemma 12 yields the desired expression of $E[A]$. □

THEOREM 17. *The expectation of $B(\xi,\eta)$ over the pairs $(\xi,\eta) \in \mathcal{I}_{h,l,n,k}$ is*

$$\begin{aligned} E[B] = \frac{1}{|\mathcal{I}_{h,l,n,k}|}\Big\{ \sum_{1\le j\le h} 2^{h-j}(2^{[j]}-1)\Big[K_{l,n,k}[2^{[j]}-1\,,\,2^{[h]}-2^{[j]}+1] \\ - K_{l,n,k}[2^{[j]}\,,\,2^{[h]}-2^{[j]}] \\ - K_{l-1,n-1,k-1}[0\,,\,2^{[h]}-2^{[j]}] \\ - 2\,(2^{[j]}-1)K_{l-1,n-1,k}[0\,,\,2^{[h]}-2^{[j]}]\Big]\Big\}, \end{aligned}$$

where $|\mathcal{I}_{h,l,n,k}| = K_{l,n,k}[0\,,\,2^{[h]}]$.

Proof: Using the expression of r_B given in (15), and by direct counting, we obtain

$$\begin{aligned} r_B^{(h)}(x,y,t) = (2^{[h]}-1)\,(1+x+y+xyt)\big[(1+x)^{2^{[h]}-1} + (1+y)^{2^{[h]}-1} - 1\big] \\ - (2^{[h]}-1)\,\big[(1+x)^{2^{[h]}} + (1+y)^{2^{[h]}} - 1\big] \\ - (2^{[h]}-1)\,[xyt + 2\,(2^{[h]}-1)\,xy\,]. \end{aligned}$$

Since furthermore $r_B^{(0)}(x,y,t) = 0$, the expression of $B^{(h)}(x,y,t)$ furnished by Theorem 11 is

$$\begin{aligned}B^{(h)}(x,y,t) = \sum_{1\le j\le h} & 2^j\,(2^{[j]}-1)\,(1+x+y+xyt)^{2^{[h]}-2^{[j]}+1} \\ & \qquad \left[(1+x)^{2^{[j]}-1}+(1+y)^{2^{[j]}-1}-1\right] \\ - \sum_{1\le j\le h} & 2^j\,(2^{[j]}-1)\,(1+x+y+xyt)^{2^{[h]}-2^{[j]}} \\ & \qquad \left[(1+x)^{2^{[j]}}+(1+y)^{2^{[j]}}-1+xyt+2\,(2^{[j]}-1)\,xy\right].\end{aligned}$$

Extracting the coefficient of the term $x^l y^n t^k$ by means of Lemma 12 yields the sought value of $E[B]$. □

Since $E[Ice] = E[A] + E[B]$, adding the values of $E[A]$ and $E[B]$ provided by Theorem 16 and Theorem 17 gives an independent derivation of the expression of $E[Ice]$ computed earlier in Theorem 15.

REFERENCES

delaB59 R. de la Brandais, "File Searching Using Variable Length Keys," *Proc. Western Joint Computer Conference* 15, pp. 295–298, 1959.

delaT87a P. de la Torre, "Analysis of Tries," Report CS–TR–1890, Ph. D. Thesis, Department of Computer Science, University of Maryland, 1987.

delaT87b P. de la Torre, "Analysis of Tries that Store Prefixing Keys," Systems Research Center, University of Maryland, 1987.

Dev82 L. Devroye, "A Note on the Average Depth of Tries," *Computing,* vol. 28, no. 4, pp. 367–371, 1982.

Fla83 Ph. Flajolet, "Methods in the Analysis of Algorithms: Evaluation of a Recursive Partitioning Process," *Proceedings of the 1983 International*

FCT–Conference, Borgholm, Sweden, pp. 141–158 in *Lecture Notes in Computer Science 158,* ed. M. Karpinski, 1983.

FRS85 Ph. Flajolet, M. Regnier, and D. Sotteau, "Algebraic Methods for Trie Statistics ," *Annals of Discrete Mathematics*, vol. 25, pp. 145–188, 1985.

FS86 Ph. Flajolet, and R. Sedgewick, "Digital Search Trees Revisited," *SIAM J. Comput.*, vol. 15, no. 3, pp. 748–767, 1986.

Fra77 J. Françon, "On the Analysis of Algorithms for Trees," *Theor. Comp. Sci.*, vol. 4, pp. 155–169, 1977.

Fre60 E. Fredkin, "Trie Memory," *CACM*, vol 3, no. 9, 490–499, 1960.

Gon84 G. H. Gonnet, *Handbook of Algorithms and Data Structures*, Addison–Wesley, Reading, Mass. 1984.

Kno86 G. D. Knott, "Including Prefixes in Doubly–Chained Tries," Report CAR–TR–236, Computer Science Department, University of Maryland, 1986.

Knu73 D. E. Knuth, *The Art of Computer Programming*, Volume 3: *Sorting and Searching*; Addison–Wesley, Reading, Mass. 1973.

Pla84 D. Plateau, "A Pruned Trie to Index a Sorted File and its Evaluation," *Inf. Systems*, vol. 9, no. 2, pp. 157–165, 1984.

Pla83 D. Plateau, "Une Structure Compacte pour Indexer un Fichier Totalment Ordonné: évaluation et mise en ouvre," Thesis, Univ. Paris XI Orsay, 1983.

Reg81 M. Regnier, "On the Average Height of Trees in Digital Search and Dynamic Hashing," *Inf. Proc. Letters*, vol. 13, no. 2, pp. 64–66, 1981.

Tra78 L. I. Trabb Pardo, "Set Representation and Set Intersection," Report STAN–CS–78–681, Ph. D. Thesis, Department of Computer Science, Stanford University, 1978.