$E = I + T$
The internal extent formula for compacted tries

Paolo Boldi Sebastiano Vigna
Università degli Studi di Milano, Italy

Abstract

It is well known [Knu97, pages 399–400] that in a binary tree the external path length minus the internal path length is exactly $2n - 2$, where n is the number of external nodes. We show that a generalization of the formula holds for compacted tries, replacing the role of *paths* with the notion of *extent*, and the value $2n-2$ with the *trie measure*, an estimation of the number of bits that are necessary to describe the trie.

1 Introduction

The well-known formula [Knu97, pages 399–400]

$$E = I + 2n - 2,$$

where n is the number of external nodes, relates the *external path length* E of a binary tree (the sum of the lengths of the paths leading to external nodes) with the *internal path length* I (the sum of the lengths of the paths leading to internal nodes).[1]

A *compacted (binary) trie* is a binary tree where each node (both internal and external) is endowed with a (binary) string (possibly empty) called *compacted path*. For a compacted trie, if we extend in the natural way the values of E and I the formula is no longer valid. In this note we provide a suitable generalization of the formula, using the definition of *extent* of a node (which collapses to the definition of path when all compacted paths are empty). We show that $E = I + T$, where E is the sum of the lengths of external extents, I is the sum of the lengths of internal extents, and T is the *trie measure*, which approximates the number of bits that are necessary to describe the trie. If all compacted paths are empty the trie measure is $2n - 2$, so our equation is a generalization of the classical result. We also provide a generalization to the case of non-binary tries.

2 Definitions

We work out our definitions from scratch closely following Knuth's, as the notation that can be found in the literature is not always consistent.

[1] The formula actually reported by Knuth is slightly different ($E = I + 2n$) because in his notation n is the number of *internal* nodes, which is equal to the number of external nodes minus one. As we will see, for compacted tries the number of external nodes is equal to the size of the set of strings represented by the trie, and so it is a more natural candidate for the letter n.

Binary trees. A *binary tree* is either the empty binary tree or a pair of binary trees (called the *left subtree* and the *right subtree*) [Knu97, page 312].[2]

A binary tree can be represented as a rooted tree[3] in which nodes are either *internal* or *external*. The empty binary tree is represented by a single external root node. Otherwise, a binary tree is represented by an internal root node connected to the representations of the left and right subtree by two edges labelled 0 and 1. Note that external nodes have no children, whereas internal nodes have always exactly two children.[4]

Compacted binary tries. A *compacted binary trie* is either a binary string, called a *compacted path*, or a binary string endowed with a pair of binary tries (called the *left subtrie* and the *right subtrie*). Equivalently, a compacted binary trie can be seen as a labelling of the nodes of a binary tree with compacted paths.

Given a nonempty prefix-free set of strings $S \subseteq 2^*$, the associated compacted binary trie is:

- the only string in the set, if $|S| = 1$;
- otherwise, let p be the longest common prefix of the strings in S; then, the trie associated to S is given by the string p and by the pair of tries associated with the sets $\{ x \in 2^* \mid pbx \in S \}$, for $b = 0, 1$.

A compacted binary trie can be represented as a rooted tree in which, as in the case of binary trees, nodes are either *internal* or *external*. A single string is represented by a single external root node labelled by the string. Otherwise, a string and a pair of subtries are represented by an internal root node labelled by the string, connected to the representations of the first and second subtries by two edges labelled 0 and 1 (see Figure 2). From this representation, the set S can be recovered by looking at the labelled paths going from the root to *external* nodes.

Given a node α of the trie (see again Figure 2):

- the *extent* of α is the longest common prefix of the strings represented by the external nodes that are descendants of α;
- the *compacted path* of α, denoted by c_α, is the string labelling α;
- the *name* of α is the extent of α deprived of its suffix c_α.

We will use the name *internal extent* (*external extent*, resp.) for the extent of an internal (external, resp.) node.

A data-aware measure. Consider the compacted trie associated with a nonempty set $S \subseteq 2^*$. We define the *trie measure* of S [GHSV07] as

$$T(S) = \sum_\alpha (|c_\alpha| + 1) - 1 = O(n\ell)$$

[2] We remark that the definition we use (a slight abstraction on Knuth's) is the simplest and most correct from a combinatorial viewpoint, but might sound unfamiliar. An alternative commonly found description says that a binary tree is given by a node with a left and a right subtree, either of which might be empty; the latter definition, however, does not account for the empty binary tree, which is essential in making the left-child-right-sibling isomorphism with ordered forests work (see again [Knu97, pages 334–335]).

[3] An acyclic connected graph with a chosen node (the root). As observed by Knuth [Knu97, page 312], a tree (in the graph-theoretical sense) and a binary tree are two completely different combinatorial objects.

[4] We remark that it is common to forget about external nodes altogether and consider only internal nodes as "true" nodes of the binary tree. In this setting, there are nodes with no children, nodes with a single child (left or right), and nodes with two children. As noted by Knuth, handling external nodes explicitly makes the structure "more convenient to deal with". In our case, external nodes are essential in the very definition of E.

s_0	001001010
s_1	00100110100100010
s_2	001001101001001

Figure 1: A toy example set S.

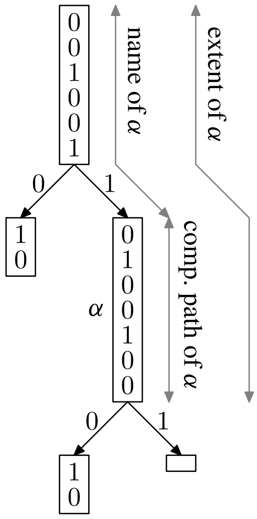

Figure 2: The rooted-tree representation of the compacted trie associated with the set S of Figure 1, and the related names. Arrows display the direction from the root to the external nodes.

where the summation ranges over all nodes of the trie, $n = |S|$ and ℓ is the average length of the elements of S. Actually, $T(S)$ is the number of edges of the standard (non-compacted) trie associated with S.

This measure is directly related to the number of bits required to encode the compacted trie associated with S explicitly: indeed, to do this we just need to encode the trie structure (as a binary tree) and to write down in preorder all the c_α's. Since there are n external nodes (hence $n - 1$ internal nodes), writing a concatenation of the c_α's requires $T(S) - 2n + 2$ bits; then we need $\log \binom{T(s)}{2n-2}$ additional bits to store the starting point of each c_α, whereas the trie structure needs just $2n - 2$ bits (e.g., using Jacobson's representation for binary trees [Jac89]). All in all, the space required to store the trie is

$$T(S) + \log \binom{T(S)}{2n-2}.$$

More precisely, the above number of bits is sufficient to write every trie with n external nodes and measure $T(S)$, and it is necessary for at least one such a trie (whichever representation is used) [FGG$^+$08].

3 $E = I + T$

We start by generalizing the internal path formula for binary trees to an *internal extent formula* for compacted binary tries:

Theorem 1 *Let S be a nonempty prefix-free set of n binary strings with average length ℓ, and consider the compacted binary trie associated with S. Let E be the sum of the lengths of the external extents* *(equivalently: $E = n\ell$, the sum of the lengths of the strings in S), I the sum of the lengths of the internal extents, and T the trie measure of S. Then,*

$$E = I + T.$$

Proof. We prove the theorem by induction on n. The theorem is obviously true for $n = 1$, as in this case $E = |c_\alpha|$, $I = 0$ and $T = |c_\alpha| + 1 - 1 = |c_\alpha|$. Consider now the case of a trie with root α and subtries with their values n_0, n_1, E_0, E_1, I_0, I_1, T_0, and T_1. Then, using the definitions, we have

$$\begin{aligned}
E &= (E_0 + E_1) + (|c_\alpha| + 1)(n_0 + n_1)\\
I &= (I_0 + I_1) + (|c_\alpha| + 1)(n_0 - 1 + n_1 - 1) + |c_\alpha|\\
&= (I_0 + I_1) + (|c_\alpha| + 1)(n_0 + n_1 - 1) - 1\\
T &= (T_0 + 1) + (T_1 + 1) + (|c_\alpha| + 1) - 1 = T_0 + T_1 + |c_\alpha| + 2\\
n &= n_0 + n_1.
\end{aligned}$$

Adding the equations $E_j = I_j + T_j$ for $j = 0, 1$ (which hold by inductive hypothesis) we have

$$E_0 + E_1 = I_0 + I_1 + T_0 + T_1.$$

We add $(|c_\alpha| + 1)(n_0 + n_1)$ to both sides, getting

$$\begin{aligned}
E_0 + E_1 + (|c_\alpha| + 1)(n_0 + n_1) &= I_0 + I_1 + (|c_\alpha| + 1)(n_0 + n_1 - 1) + T_0 + T_1 + |c_\alpha| + 1\\
E_0 + E_1 + (|c_\alpha| + 1)(n_0 + n_1) &= I + 1 + T - 1,
\end{aligned}$$

which entails the thesis. ▮

As noted in the introduction, when all compacted paths are empty E is equal to the external path length, I is equal to the internal path length, and the trie measure is exactly $\left(\sum_\alpha 1\right) - 1 = 2n - 2$. Thus, the internal extent formula is truly a generalization of the internal path formula.

4 A simple application

We were lead to the equation $E = I + T$ by the problem of bounding the average length of an internal extent in terms of the average length of an external extent, that is, in terms of ℓ, the average length of the strings in S. This bound can now be easily obtained:

Corollary 1 *Let S, with $|S| \geq 2$, be a set of binary strings. With the notation of Theorem 1,*

$$I/(n-1) \leq \ell - 3/2 + 1/n.$$

Proof. We just divide both members of the internal extent equation by n:

$$\begin{aligned}
\frac{E}{n} &= \frac{I}{n} + \frac{T}{n}\\
&= \frac{I}{n-1} + \frac{I}{n} - \frac{I}{n-1} + \frac{2n - 2 + \sum_\alpha |c_\alpha|}{n}\\
&= \frac{I}{n-1} + \frac{3}{2} - \frac{1}{n} - \frac{I - (n-1)(n/2 - 1 + \sum_\alpha |c_\alpha|)}{n(n-1)}\\
&\geq \frac{I}{n-1} + \frac{3}{2} - \frac{1}{n}.
\end{aligned}$$

To see why the last bound is true, note that in a trie with $n-1$ internal nodes the contribution to I of the edges (i.e., excluding the compacted paths) is at most $(n-1)(n-2)/2$ (the worst case is a linear trie). On the other hand, the contribution of compacted paths to each internal path cannot be more than $\sum_\alpha |c_\alpha|$, so the overall contribution cannot be more than $(n-1)\sum_\alpha |c_\alpha|$. We conclude that

$$I \leq (n-1)\Big((n-2)/2 + \sum_\alpha |c_\alpha|\Big). \blacksquare$$

Note that the bound is essentially tight, as in a linear trie with empty compacted paths $E = n(n+1)/2 - 1$ and $I = (n-2)(n-1)/2$, so $E/n - I/(n-1) = 3/2 - 1/n$.

5 A generalization to non-binary tries

Given an alphabet Σ, a *compacted trie over* Σ is defined as follows: it is either a single string $x \in \Sigma^*$, or a string $x \in \Sigma^*$ together with a subset $X \subseteq \Sigma$ with $|X| > 1$ endowed with a function ζ that assigns a compacted trie over Σ to each element of X.

Given a nonempty prefix-free set of strings $S \subseteq \Sigma^*$, the associated compacted trie over Σ is:

- the only string in the set, if $|S| = 1$;
- otherwise, let p be the longest common prefix of the strings in S; then, the trie associated with S is given by p, the set $X \subseteq \Sigma$ of all $a \in \Sigma$ such that pa is the prefix of some string in S, and by the function ζ mapping a to the compacted trie associated with the set $\{\, x \in \Sigma^* \mid pax \in S \,\}$.

Similarly to what happens for compacted binary tries, a compacted trie over Σ can be represented as a rooted tree where each node is labelled by a (possibly empty) string over Σ and internal nodes have at most $|\Sigma|$ (but not less than two) children, each associated with a distinct symbol of Σ. The notation of Figure 2 carries on easily, and the definition of trie measure is extended in the natural way.

We now want to generalize the internal extent formula (Theorem 1) to non-binary tries.

Theorem 2 *Let S be a nonempty prefix-free set of n strings over an alphabet with σ symbols, and consider the compacted trie associated with S. For each $d = 0, \ldots, \sigma$, let $Y(d)$ be the sum of the lengths of the extents of nodes with d children, $n(d)$ be the number of such nodes, and T be the trie measure. Then,*

$$E = \sum_{d=2}^{\sigma} (d-1)Y(d) + T.$$

Proof. By induction on the number of nodes. This is true for a one-node trie; for the induction step, suppose that the root of a trie has a compacted path of length c, and h subtries ($2 \leq h \leq \sigma$); for the i-th subtrie, by induction hypothesis, since $E = Y(0)$ we have

$$Y_i(0) = \sum_{d=2}^{\sigma} (d-1)Y_i(d) + T_i. \tag{1}$$

Observe that, for every $d = 0, 2, 3, \ldots, \sigma$,

$$Y(d) = \sum_{i=1}^{h} Y_i(d) + (c+1)\left([d = h] + \sum_{i=1}^{h} n_i(d)\right) - [d = h]$$

where we used Iverson's notation.[5] Moreover

$$n(d) = [d = h] + \sum_{i=1}^{h} n_i(d)$$

so

$$Y(d) = \sum_{i=1}^{h} Y_i(d) + (c+1)n(d) - [d = h].$$

Further

$$T = \sum_{i=1}^{h} T_i + h + c.$$

Summing (1) memberwise, we obtain

$$\sum_{i=1}^{h} Y_i(0) = \sum_{d=2}^{\sigma}(d-1)\left(\sum_{i=1}^{h} Y_i(d)\right) + \sum_{i=1}^{h} T_i$$

that is equivalent to

$$Y(0) - (c+1)n(0) + [0 = h] = \sum_{d=2}^{\sigma}(d-1)\left(Y(d) - (c+1)n(d) + [d = h]\right) + T - h - c,$$

hence

$$Y(0) = \sum_{d=2}^{\sigma}(d-1)Y(d) - (c+1)\left(\sum_{d=2}^{\sigma}(d-1)n(d) - n(0)\right) + h - 1 + T - h - c.$$

Since $\sum_{d=2}^{\sigma}(d-1)n(d) = n(0) - 1$, we have

$$Y(0) = \sum_{d=2}^{\sigma}(d-1)Y(d) + T. \blacksquare$$

6 Acknowledgments

We would like to thank the anonymous referee for spotting subtle inconsistencies in the first version of this paper.

References

[FGG$^+$08] Paolo Ferragina, Roberto Grossi, Ankur Gupta, Rahul Shah, and Jeffrey S. Vitter. On searching compressed string collections cache-obliviously. In *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 181–190. ACM, 2008.

[5] For a given Boolean predicate ϕ, we let $[\phi]$ be 0 if ϕ is false, 1 if ϕ is true [Knu92].

[GHSV07] Ankur Gupta, Wing-Kai Hon, Rahul Shah, and Jeffrey Scott Vitter. Compressed data structures: Dictionaries and data-aware measures. *Theoretical Computer Science*, 387(3):313–331, 2007.

[Jac89] Guy Jacobson. Space-efficient static trees and graphs. In *30th Annual Symposium on Foundations of Computer Science (FOCS '89)*, pages 549–554, Research Triangle Park, North Carolina, 1989. IEEE Computer Society Press.

[Knu92] Donald E. Knuth. Two notes on notation. *American Mathematical Monthly*, 99(5):403–422, May 1992.

[Knu97] Donald E. Knuth. *The Art of Computer Programming, Volume 1, Fundamental Algorithms*. Addison-Wesley, Reading, MA, USA, third edition, 1997.