

sCPT : A succinct representation of a Compact Prediction Tree

Rafael Ktistakis, Rajeev Raman

Department of Informatics, University of Leicester, United Kingdom

Abstract

Predicting the next item of a sequence over a finite alphabet has important approaches in various domains. *emph*Lossless approach is a new approach recently introduced by Gueniche et al. 2013 with CPT. CPT (Compact Prediction Tree) keeps all the training data in order for all the relevant information to be available for each prediction and yield more accurate predictions. Later, CPT+ [9] stood as an improved version over CPT. CPT+ improved the high space and time complexity of CPT and increased its accuracy. Compared to other state-of-the-art models from the literature: TDAG, PPM, Lz78, DG, PPM, All-K-order, CPT, CPT+ has the best overall accuracy under evaluation on real life datasets. Recently, CPT+ was used in SPiCe [1] competition where it ranked among the best 10 accurate predictors. The competition used a different way of evaluating the predictors from the way that CPT/+ originally evaluated. Hence, its accuracy success is not based on only one particular evaluation, highlighting its importance. However, even the improved CPT+ utilises a fairly big amount of memory in order to losslessly keep all the training data. This puts some constraints on the size of the training data which will feed CPT+. Tested under real life datasets, CPT+ utilises space from 15 to 175 times more than a *compact representation* of training data (Γ). Below, it is presented a *succinct* representation of CPT+ (sCPT) where it consistently utilises from 1.2 to 3 times the space of Γ , without any significant performance loss in regards to speed. Therefore, sCPT is able to hold undoubtedly more training data, leading to an improved accuracy.

1 Introduction

Let a finite set of items (symbols) $\Sigma = \{e_1, e_2, \dots, e_\sigma\}$ be called *alphabet*, then a *sequence* S is a list of ordered items $S = \langle i_1, i_2, \dots, i_k \rangle$ where $i_m \in \Sigma$ for $1 \leq m \leq k$. A multiset of sequences $D = \langle S_1, S_2, \dots, S_l \rangle$ also called *dataset* is constituted by sequences that created under a similar way. Given a dataset, a prediction algorithm is *trained* on this dataset. Once the prediction algorithm is trained, it will repeatedly accept a *context* which is another sequence of symbols, and outputs a *prediction*. The context is obtained from an unknown *query sequence* which does not belong to the dataset but shares common characteristics with the dataset. Loosely speaking, the context is derived from a part of the query sequence, and the prediction is about what comes *after* the context in the query. Evaluating the prediction result can be based on different tasks (Appendix C) Each task can be correlated with numerous applications such as, products recommendation, web page prefetching, compression algorithms etc. As noted, once the predictor is trained, we are given an unknown query sequence $Q = \langle q_1, \dots, q_j, \dots \rangle$. The prediction algorithm is given a context $\langle q_1, q_2, \dots, q_j, \rangle$ and the predictor aims to say something about what is in $\langle q_{j+1}, q_{j+2}, \dots \rangle$, as it gets explained in Appendix C.

Compact Representation of Training Data

Compact Representation of training data (Γ) will play an important role as a comparison measure of memory utilisations, during this report. It captures the total size of training data that is given as an input to a predictor. It is defined as $\Gamma = \log(\sigma) \cdot k \cdot l$ bits and constitutes a base for comparison among different predictors and their memory utilisations.

Lossless

Lossless sequence prediction is a relatively new approach and therefore appears to have quite a few challenges that someone that studies it has to cope with. It is clear that due to the fact that a whole dataset is utilised, a

lossless approach achieves a higher space utilisation over a lossy approach. This fact leads to two results; firstly, it is somewhat intensive for an algorithm to go through all the data and extract the appropriate information needed for predictions, and secondly, keeping all the relative data *in memory*, through the corresponding data structures, appears to be a considerably hard challenge because the represented structure's space utilisation often reaches an order of magnitude of the initial gathered data. As a result, a lossless sequence prediction approach needs sophisticated tools to search for patterns and sequences in *Big Data* when simultaneously the aforementioned challenges are coped. CPT/CPT+ is currently the only implementations in lossless sequence prediction.

On contrary, a *lossy* approach can be less memory hungry by representing the given information of a dataset through models (like *dependency graphs*) [13]. This leads to implementation that sometimes utilises less space and can be potentially faster. However, a lossy approach lacks accuracy especially in hard cases when information in a dataset is not repeating as other parts of informations or a repetition is either generalised or ignored by a model having a partly loss of information that is not reversible. Often such hard cases lead to more complex models that along with Big data is hard to make algorithms that produce predictions with efficiency and performance.

2 CPT and CPT+

2.1 Data structures of CPT

CPT is constituted by a Trie (*Prediction Tree PT*), an Array of bit-vectors (*Inverted Index II*) and an array of pointers to the last items of sequences inserted to the prediction tree (*Look-up Table LT*). As a lossless approach, CPT utilises the whole dataset provided for building its structures.

Inverted Index

The II plays an important role in finding which sequences should be retrieved from the prediction tree. Every bit-vector of the II corresponds to an alphabet item and matches (with 1/0) whether this alphabet item appears to a sequence or not. Every index of a bit-vector belongs to a sequence with the 0th index to belong to the 1st sequence of the dataset and the last index to the last sequence of the dataset (figure 1). The main operation that is needed is a bit-wise AND among different bit-vectors of the II. Depending on the query sequence items, different bit-vectors are chosen from the II to apply the AND operation. The result is a bit-vector which shows which sequences contain all the query items. According to the above description, it is easy for someone to oversee that the size of the II is increasing along with the number of sequences. In a real-life application, it is usually the case that the alphabet size remains relatively stable. However, the dataset size can be increased as time passes. An example can be the fact of users' weblogs for a specific website that uses a predictor like CPT for caching web-pages. The website has a determined number of pages (constitutes the alphabet), however, the order that a user visits the pages and the amount of the pages that the user visits per session varies. Hence, different transactions are being created (constituting the sequences). Since the application lies on a lossless approach, it means that the bit-vectors of the II would take more and more space. The II is undeniably the less scalable structure to the CPT which takes $\sigma \times l$ bits of space.

Look-Up Table

The Look-Up Table is the structure that adds the necessary "glue" between the II and the Prediction Tree (figure 1). After that the II calculates which sequences contain the query items then the LT will simply point out to those sequences in the PT. The LT is an array with pointers to every end of every sequence in the PT. The pointer for the 1st sequence is at the 0th position of the LT, the pointer for the 2nd sequence at the 1st position etc. As the II, the size of the LT depends on the number of the sequences that the CPT is trained on. For every sequence, the LT uses 64 bits. In section 3 it will be presented a more elegant solution where the LT can be omitted completely.

Prediction Tree

The prediction Tree of the CPT Predictor is constituted by a trie structure (figure 1). After the build phase where the prediction tree has been built, the main operation that the predictor needs to execute, as a part of a prediction task, is to retrieve back sequences. Which sequences will be retrieved is a matter that is decided by the II. Hence, the prediction tree is important to support instructions that give back the relative sequences. Since the LT points to the end of every sequence, it makes sense to retrieve sequences beginning from root and going up till the root. The retrieved sequence can be reversed in a later time. Therefore, a *getparent()* instruction should be supported in a node of the prediction tree. Currently, the trie offers such functionality which takes $\mathcal{O}(k)$ time for the sequence retrieval to get completed. The space required for the PT is a factor of the total number of nodes. Every node needs 32 bytes of stored information which includes a pointer to the parent node, pointers to all children nodes and the label of the node. Regardless of the space usage limitation that it is already mentioned, there is an extra limitation that can be studied in a near future. The fact that the pointers of all the children of a node are being contained in a simple array makes the build phase more and more less efficient for bigger and bigger datasets. That is because for every sequence insertion in the PT, all the children of a node has to be parsed in order to determine whether a new node insertion is required or not. A *Ternary Tree* [2] could solve such a problem during the building phase.

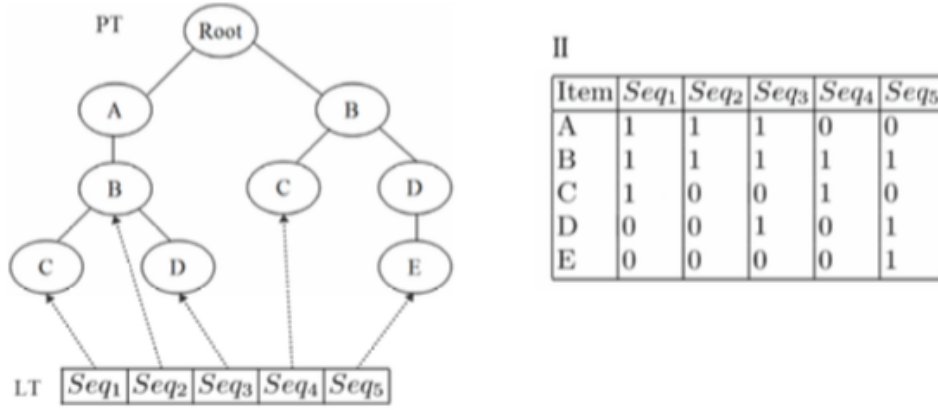


Figure 1: A Prediction Tree (PT), Inverted Index (II) and Lookup Table (LT)

In Figure 1 it is shown an example of CPT during multiple sequence insertions.

2.2 Data structures of CPT+

CPT+ keeps the data structure of CPT, however it applies two main new strategies for improving the memory utilisation of CPT [9] (Figure 2). The first one, called *Frequent subsequence compression (FSC)*, identifies the frequencies of subsequences that appear within a dataset. Each frequent subsequence can be assigned a unique id. This unique id can replace consecutive nodes of the PT that constitute the same subsequence. Of course, a new structure has to be introduced in order to hold the correlation between a frequent subsequence and an id. A *subsequence Dictionary* takes this role and translates a subsequence to an id (or vice-versa) in $\mathcal{O}(1)$ time.

The second strategy is the *Simple Branches Compression (SBC)*. A *simple branch* is a branch that leads to a single leaf. As a result, each node of a simple branch has between 0 and 1 children. This strategy identifies the simple branches and replaces them with a single node which represents the whole simple branch.

2.3 CPT and CPT+ Predictor

In order for the CPT to be used for predictions, the aforementioned data structures have to work together providing the necessary information when it is needed. A bitwise AND shows which sequences contain the query items and then these sequences can be located by using the Look-up Table which points to the end of each corresponding sequence in the PT. Then, PT can be traversed backwards (going to parent until root is reached). Having retrieved a requested sequence from the PT and along with the query sequence, a *consequent* can be resulted as a part of the retrieved sequence. The query items are used as a set of items and each one of these items is matched in the retrieved sequence and removed from it. Having removed all these items from

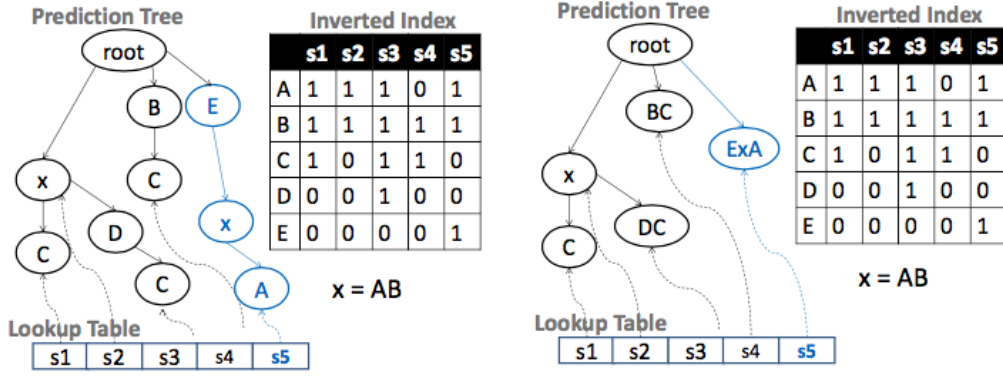


Figure 2: FSC and SBC compression strategies

the retrieved sequence, a *consequent* is created. Consequent is eventually the part of the retrieved sequence that specifies what comes next after a query sequence. For example, if the sequence $S_1 = \langle a, b, x, z, c, d, e \rangle$ was retrieved from the prediction tree and the query sequence is $S_q = \langle a, b, a, c \rangle$ then the consequent will be $S_c = \langle d, e \rangle$. CPT+ follows the same process for retrieved the matched sequences from PT through LT and II, however as a final step the retrieved sequences has to get *decoded* through the subsequence Dictionary.

CPT uses different techniques [8] to make use of the consequents. Such techniques could facilitate different prediction tasks (see Appendix C). Recursively shrinking the query size (creating sub-queries) and producing multiple consequents is one of them. All the consequents then are inserted into a table that counts the occurrences of the items. For each consequent a different weight is applied to its occurrences. This weight depends on the length of the corresponding sub-query. Finally, the item with the highest count in the table constitutes the prediction. The process for CPT+ is slightly more efficient since it calculates the sub-queries via a *noise* ratio. This noise ratio is simply a measure to filter the items of the query that have a low frequency in the training data. This low frequency is defined by this ratio. Then, the identified “noisy” items can be removed one by one, creating the sub-queries and following the same process as CPT.

2.4 Strengths & Weaknesses

CPT and CPT+ (as lossless approaches) utilise all the available sequences in a trie where they are easily accessible through the II and the LT. Therefore, making predictions can be potentially fast and accurate as described in [8] and [9]. Locating similar sequences can be a relatively easy task that makes CPT and CPT+ work as a good tool for predictors that exploit such sequences. Moreover, CPT can implement all the prediction tasks described in Appendix C, offering a bigger diversity for implementations on several applications. It can be considered undoubtedly as a good state-of-the-art predictor since it is placed among the best 10 best accurate predictors in SPiCe [1] competition.

However, even though with the improvement of CPT+, where size of the structures were reduced and speed was enhanced, some performance issues can be still addressed. Bit-vectors are the main element of II which is used for locating CPT+’s sequences. When more and more sequences are being added, the bitwise AND operations are becoming slower and slower, leading to performance issues and overall scalability issues since the vectors are becoming much larger. It is common for real life applications to deal with datasets that contain an enormous number of sequences. Therefore, Bit-vectors can be CPT+’s Achilles’ heel regarding its performance. Also, space usage of II in such cases can be wasteful since with *Succinct Data Structures* (see Section A) a better and more memory efficient structure can be proposed (section 3). In addition, CPT+ suffers from memory consistency making it more hard to predict its behaviour in cases large datasets are fed the predictor, since its overall size depends upon the number of sequences.

It is interesting to point out that one of the two compression strategies of CPT+, partially fails to save space from the PT. The SBC strategy depends on simple branches that lead to a single leaf. However, sequences that share same prefixes will be part of the same single branch. LT will point to all those different sequences that share the same single branch. However, replacing a simple branch with a node will result in conflicts with sequences that who share the prefix. LT will still point to nodes that should have been freed. If the memory is indeed freed, then LT points to a “bad” memory address otherwise memory saved does not correspond to the

numbers of nodes replaced during SBC strategy. Looking at Figure 2, let us assume that LT s4 sequence points to x . Thus, s4 is a prefix of s5. If the branch of sequence s5 is identified as a simple branch then node x cannot be freed. LT should still point to x and x would point to node E. In this example, the memory saved is zero. Of course there are workarounds for making SBC strategy work on such cases but it is not studied in “CPT+: Decreasing the Time/Space Complexity of the Compact Prediction Tree”. The memory evaluation made upon the number of nodes saved and not on the actual memory utilisation as reported by the system. Details can be found on CPT and CPT+ implementation on <http://goo.gl/LE4uYO>.

In Figure 3 it is presented the memory usage of CPT+ and its data structures as it was reported by Java Virtual Machine (JVM) in respect of each data structure object. For this evaluation, synthetic datasets were used in order to illustrate the linear pattern of the memory increment according to sequence number. Datasets were generated through IBM Quest Dataset Generator [6]. It is an interesting fact that CPT+ utilises almost 40 times the space of Γ . Even the smallest structure like the LT, is around 5 times of Γ . If it is considered that in a lossless approach where all available training data are kept and relatively big quantities of training data might be given, then CPT+ cannot be scalable in applications which depend on *Big Data*.

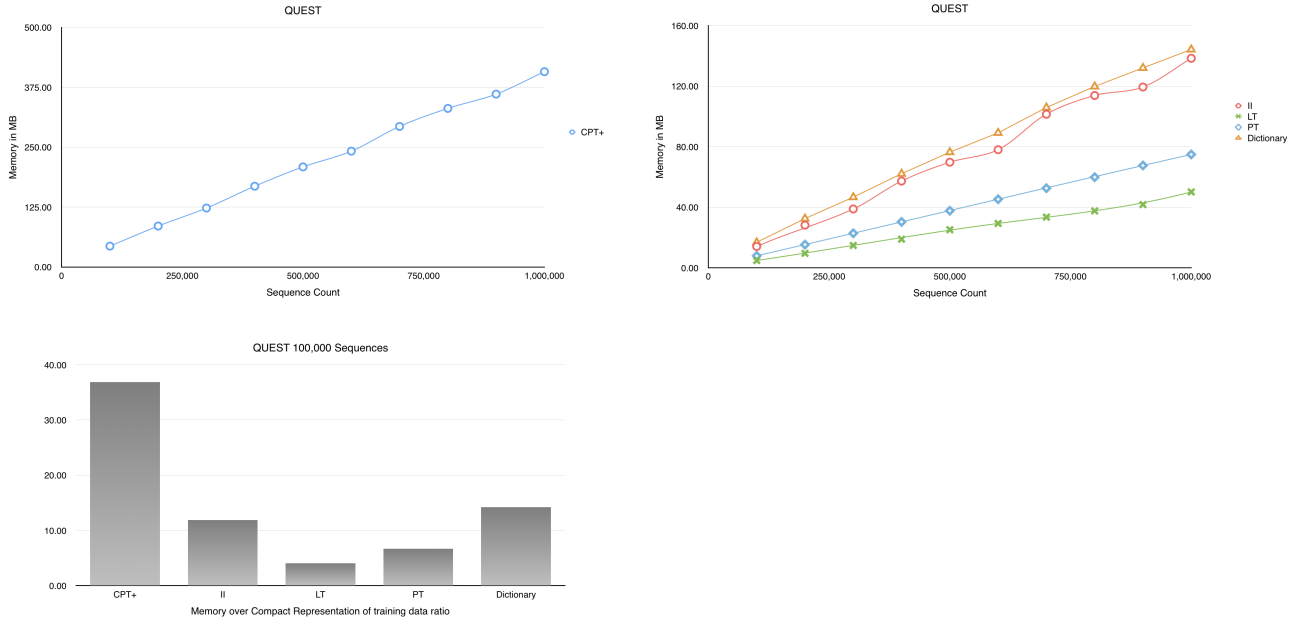


Figure 3: CPT+ memory usage and its structures for datasets of different count number (top); Ratio of CPT+/structures over Γ

3 Succinct CPT

Succinct CPT copes with all the memory disadvantages of CPT by introducing different succinct data structures for the roles of PT, II and LT. By using succinct data structures, fast operations like *rank/select* (Appendix B) are offered. Thus, a significant memory reduction is achieved and the performance of CPT+ is relatively retained. The predictor of CPT+ remains the same. In order for this to be achieved, the combined succinct data structures offer the same functionalities and instructions as the former CPT+ structures (II, PT, LT).

3.1 Elias-Fano for Inverted indices

The II as shown in Figure 1 contains which alphabet item is contained in which sequence. However, every bit-vector can be treated as an inverted list of the positions of the set bits. Hence, the bit-vector for item A of the Figure 1 can be written as $\langle 1, 2, 3 \rangle$. Since, every position in the bit-vector is correlated with a sequence serial number (1^{st} , 2^{nd} , \dots , l^{th} sequence) it is obvious to conclude that the aforementioned inverted list is monotone sequence of numbers. This is important because the compression technique which is mainly being used and described next, is based on *gap encoding*. As it will be proved later, gap encoding will make possible to store

the II in space smaller or at least same as the space of the initial dataset. The techniques described below are based on a paper by Elias [4] and a paper which demonstrates IIs which similar techniques [15].

Elias Technique

It is clear that every bit-vector of the II can be seen as an inverted list of the sequences that the alphabet item is contained. Because the list is monotone can be seen as

$$0 \leq x_0 \leq x_1 \leq \dots \leq x_{n-2} \leq x_{n-1} \leq u$$

where $u > 0$ as a known upper bound. That bound is always known because during prediction phase the dataset size is known too. Now, this list of numbers can be split in two bit arrays according to the formulas below:

lower bit array: for every item of the list, the array continuously stores its $\lambda = \max\{0, \lfloor \log(u/n) \rfloor\}$ lower bits.

upper bit array: then the upper bits are stored in a sequence of unary-coded gaps (see Figure 4)

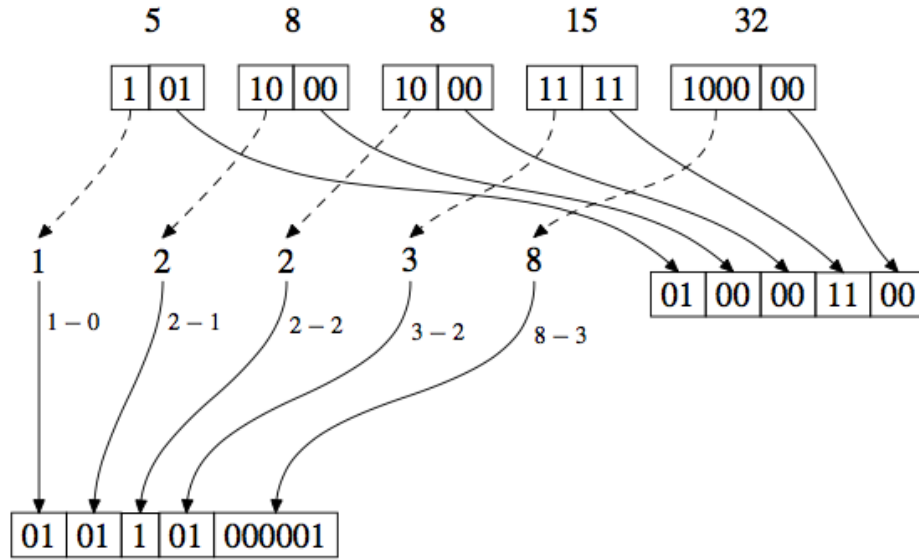


Figure 4: In this example we have a list of numbers 5, 8, 8, 15, 32 with upper bound 36. For the upper bit array $\lambda = \lfloor \log 36/5 \rfloor = 2$. On right it is shown all the λ bits of all elements which are gathered together to create the lower-bits array. On left, the gaps for every number upper bits are calculated and stored sequentially in unary code in the upper-bits array.

Space property

Space wise, it can be proven [15] that for every item of a list it is used at most $2 + \lceil \log(u/n) \rceil$ bits. The dataset containing the sequences, from where the II will be created, can be considered as a memoryless source. Then, every symbol appears with a probability of $p_\kappa = \frac{1}{\sigma}$ in the dataset. The bits needed to code a symbol will be $\sum_{\kappa=1}^{\sigma} p_\kappa \cdot \log(\frac{1}{p_\kappa}) = \log(\sigma)$. An II that uses Elias way of recording the set bit positions for every bit-vector, will take a total space:

$$\sum_{l=1}^{\sigma} n_l \cdot (\log(\frac{l}{n_l}) + 2) \quad (1)$$

We set as M to be the total length of the dataset. Then:

$$M \geq l \quad (2)$$

Using (1), we can oversimplify it as:

$$\sum_{\iota=1}^{\sigma} n_{\iota} \cdot (\log(\frac{M}{n_{\iota}}) + 2) \quad (3)$$

For any bit-vector of the II:

$$p_{\iota} = \frac{n_{\iota}}{M}, \text{ using (2)} \quad (4)$$

Finally, using (3 and (4)), the total memory of an II using Elias technique will be:

$$\sum_{\iota=1}^{\sigma} M \cdot p_{\iota} \cdot (\log(\frac{1}{p_{\iota}}) + 2) \quad (5)$$

In last equation (5) and by claiming that the oversimplification at (3) is usually $M > l$, it can be assumed for most cases that the total memory is less than the memory needed to code the whole dataset. So:

$$\sum_{\iota=1}^{\sigma} M \cdot p_{\iota} \cdot (\log(\frac{1}{p_{\iota}})) \leq M \cdot \log(\sigma) \quad (6)$$

This final statement is important due to the fact that most of the times a usual data structure consumes space that is a factor of the initial dataset size. In the above case, the space is usually less or the same. It is not guaranteed but experiments with real life datasets (presented in Section 4) follow the aforementioned behaviour is followed most of the times.

Accessing Values

Accessing and retrieving back any random value x_i stored with this way can be done by following the steps on the upper bit array and applying rank/select operations (Appendix B):

1. $[\frac{x_i}{2^{\lambda}}] = \theta$
2. $\varrho = \text{select}_0(\theta)$
3. $\zeta = \text{rank}_1(\varrho)$
4. concatenate (the binary representations) of $\varrho - \zeta$ with the value of lower bit array at index ζ

Similarly, we can continue accessing all the values sequentially by keeping track of the number of the 0s parsed in every position of the upper array. Every time that an 1 is met then we have met a new stored value and therefore we should concatenate our bits. A 0 means that we have to increase our variable that counts zeros since the gap to our next number is being increased.

II in sCPT

CPT predictor uses the II in order to determine which sequences contain specific alphabet items. With an II which uses bit-vectors this can be done by applying a bit-wise AND to the relevant bit-vectors. In inverted indices that follow the Elias a gap encoding way, the same functionality can be achieved by *decoding* every list of values for every alphabet item and then by applying an intersection among the relevant lists. However, intersecting lists that might contain a relative big amount of values is inefficient because every intersection will depend on the length of the largest list. With Elias-Fano II we can take advantage of the fact that we can *jump in* a specific value of a list and complete an intersection in relation to the smallest list. Therefore, the overall speed performance is not affected much as shown in experimental evaluation.

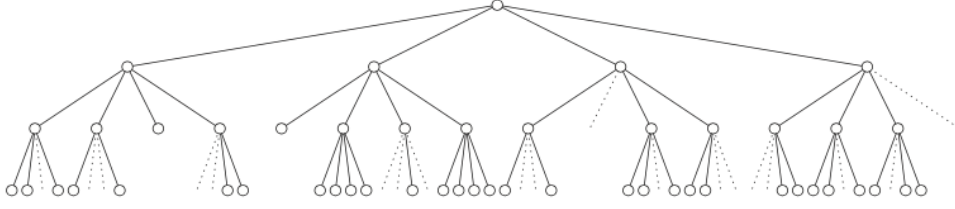


Figure 6: Generalized Jacobson encoding of a 4-ary tree: 1111 1111 1111 1011 1110 1101 1001 0000 0011 0000 1111 0010 1111 1001 1101 1100 0011 1101 1011 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000— 200 bits

get child $rank_1(i) \cdot K - 1$

get right sibling $rank_1(i) \cdot K + 1$

get parent $select_1(\frac{i+1}{K})$

sibling exist if remaining of $\frac{i}{K}$ is not zero

Storing bitmap

A relative serious drawback of the described approach for representing a tree using a binary bitmap is the fact that the bitmap gets quite long. The overall length is twice as long as the total number of nodes. Usually, rank/select operations have a space and time overhead that depends upon the type of the bit-vector that they support. Deciding how to store the bitmap could affect both the memory utilisation and the speed performance. According to SDSL-lite library [7], a rank which supports a bit-vector has a 25% of its total length bits overhead and an $\mathcal{O}(1)$ time complexity. These stats vary according to the representation of the bit-vector (such other bit-vector representation is RRR-vectors [14]). CPT+ predictor is mainly depended on accessing parents nodes since LT points to the end of sequences. Elias Fano representation of bit-vectors, as described in section 3.1, is implement with sd-vector by SDSL-lite library. Getting the parent node in a bitmap requires only select operation. The select operation on a sd-vector has a space overhead of 64 bits and an $\mathcal{O}(1)$ time complexity. These properties and the way the bitmap is stored (as already explained for inverted indices) make it an ideal option for storing the bitmap of the PT. Accessing the parent node will be achieved quite as fast as in the PT while representing it almost succinctly and saving a big amount of space.

3.3 Look-Up Table Omission

In CPT+ Look-Up table is used for pointing in the last node of a sequence within the trie. However shown in Figure 7, the nodes of the trie can be counted in a level-order way and thus an II upon the node count would be easy to be constructed. The 1_{st} sequence will be the 6_{th} column in the inverted index due to the fact that this sequence ends on the node 6 in the trie. However, the II length now depends on the trie total number of nodes. The total number of nodes is usually much higher than the total number of sequences (in case that is less due to high sequence repetition then this approach is even better for the II memory utilisation). As a result, the II would contain longer bit-vectors for every alphabet item. It is common to assume that no memory is saved but the opposite happens. On the other hand, sCPT uses bit-vectors coded with Elias Fano technique. It can be proven that the extra overhead per bit-vector will be $\log_2(M)$, where M is the ratio of nodes count over sequence number (assuming that node count is higher than sequence number). Also, in most datasets the sequence number is greater than the alphabet size. Thus, this overhead will be less than the 64 bits per sequence which is required by LT.

4 Experimental Results and Performance Comparison

For the experimental results a C++ version of CPT and CPT+ implemented. The design and architecture is the same as CPT and CPT+ original implementation in Java 8 (provided at <http://goo.gl/LE4uYO>). The experiments

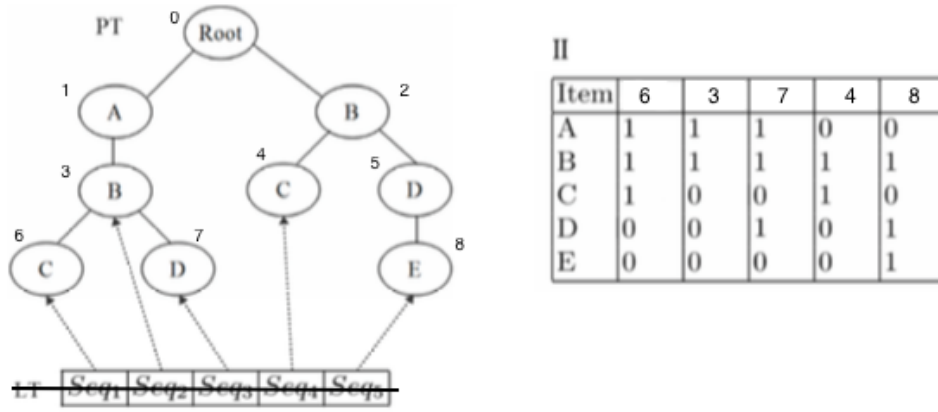


Figure 7: Illustration of how LT can be omitted and have LT take its role

performed on a Intel Xeon CPU E5-2620 v3 at 2.40GHz, Cache 15MB, 16GB Ram DDR4 at 1866MHz. The datasets that were used are available through SPMF library [6] and constitute real life datasets. The BMS, Kosarak, MSNBC Fifa, Nasa_07 and Nasa_08 datasets consist of webpages logs by users on webpages. SIGN dataset is transcribed from videos, in sign language. BIBLE is constituted by the set of sentences divided into words which represents a sequence.

To evaluate memory, both CPT+ and sCPT were trained with the whole dataset and then the memory calculation performed upon the total memory retained in C++ Heap segment after the training is finished. The speed performance evaluated by feeding the predictor with queries of length 3 items which are generated through the same dataset that trained CPT+ and sCPT respectively. Since the accuracy of the prediction is not currently evaluated and the purpose of this evaluation was to perform a full prediction process (to evaluate speed when predictor does all its steps), a K-Fold [12] validation environment was not set up. It should pointed out that both CPT+ and sCPT use exact the same predictor. What changes is the instruction call to their data structures respectively (achieved through Encapsulation).

Space Comparison

For the space comparisons the compact representation of training data usage (Γ) will be used as a metric between CPT+ and sCPT. CPT+ utilises $15 \cdot \Gamma$ to $174 \cdot \Gamma$ (Figure 8). sCPT space utilisation lies between $1.2 \cdot \Gamma$ to $4.7 \cdot \Gamma$. The succinct representation of CPT+ can be considered more consistent and predictable since its variance range for memory is much less than CPT+. The inverted indices that correspond to the biggest structure in regards of memory consumption among the other structures in CPT+, is presented opposite the II representation in sCPT at Figure 9. II in CPT+ takes $0.2 \cdot \Gamma$ to $163 \cdot \Gamma$ while in sCPT $0.5 \cdot \Gamma$ to $2.9 \cdot \Gamma$. These comparisons illustrate the memory benefits of sCPT over CPT. Therefore, the ability for sCPT to scale is much greater than this of CPT+.

Speed Comparison

Figure 8 depicts the speed performance of sCPT opposite to CPT+. Positive number means that sCPT is slower and negative that it is faster. sCPT through experiments found to be 10% to 60% slower in half of the datasets while its performance considered the same or slightly faster in the rest of the datasets. The performance differences lie only on the data structures instructions which used for each approach and not in any variations of the predictor.

5 Conclusion

CPT+ is proven to be one of the state-of-the-art predictors which ranked among the best in SPiCe competition. However its memory consumption lack of consistency and predictability which make it difficult for CPT+ to scale up as a lossless approach. sCPT overcomes these weaknesses of CPT+ and offers a consistent and predictable memory utilisation without sacrificing any serious amount of performance. With sCPT the ability

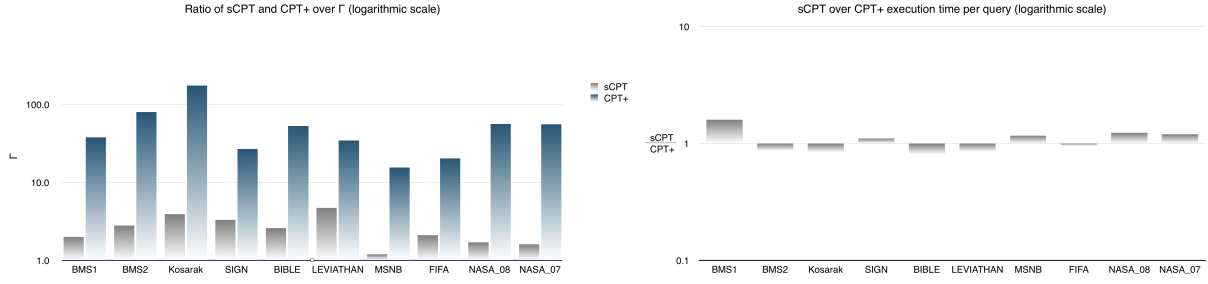


Figure 8: at right: sCPT and CPT+ over Γ , at left: execution time ratio of sCPT over CPT+. The execution time (clock time) measured for 1000 queries, same for both implementations.

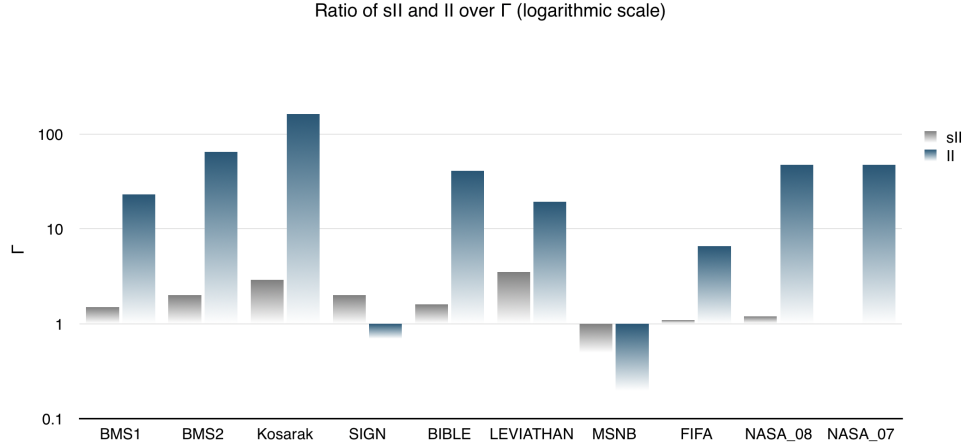


Figure 9: II of sCPT over Γ and II of CPT+ over Γ . Negative number means that the structure takes less space than the compact representation of training data

of greater scalability is offered. As already proven by CPT and CPT+, losslessly keeping all the available training data offers greater accuracy. As a future work for sCPT would be to evaluate the fact that scaling the lossless predictor in a factor where much more training data is available, would improve prediction accuracy even more.

A Succinct Data Structures

This kind of structures are high space efficient data structures (introduced by Jacobson 1989) that support rapid and efficient operations, (look Appendix B), like *rank* & *select* [3, 11]. A data structure in order to be considered *succinct*, it should use space that approaches the information-theoretic lower bound of the space that is required to represent the data. One interesting fact is that in contrast with other compressed representations succinct data structures do not sacrifice performance in order to deliver space efficiency when represent and retrieve back any relative data.

A.1 Why Succinct Data Structures

Using *Succinct data structures* can help reducing the space required by a lossless approach and keep most of the data "in memory" (which is stated as a main challenge in Section 1). This leaves open room to improve an algorithm's scalability. If we consider that compression methods are mostly based in compression and decompression mechanisms, it is almost clear that much time is spent in compressing and decompressing data. Even though, it would have been achieved a space reduction by using a compression method, the overall performance would not have been any good. Even worse, in a scalable level such an algorithm would take huge amounts of time to complete leaving out an important aspect of the algorithm which is scalability. Also, if we consider prediction tools like backward search on FM-Index [5] which their complexity to find a search pattern depends only on the times a rank/select were performed, this leads to algorithms which are more scalable with a robust performance time which is in-dependable of the input training set; aspect crucial for an algorithms performance.

B Functionality of Rank & Select

Let B be a sequence of n items. Then **rank(c)** and **select(c)** can be defined as:

rank(o) Counts the number of items which appear in B from its start to its o -th position [11].

select(o) Finds the o -th occurrence of an element in B [11].

The time complexity of rank/select queries depends on the structure where these two operation are built on. For example, it is possible to execute rank/select in constant $\mathcal{O}(1)$ time for some bit-vectors, like rrr-vectors [14], or another binary rank index [3].

C Prediction Tasks

The adoption of which item is predicted after the context and in $\langle q_{j+1}, q_{j+2}, \dots \rangle$ is only specified by the prediction task.

Right next item: The prediction algorithm gives one prediction for the q_{j+1} coming after the given context.

Top K predictions: The prediction algorithm gives K alternative predictions for the q_{j+1} item of the given context.

Distribution of right next items: For each item of the alphabet Σ , the prediction task gives the item's probability to be the q_{j+1} -th item.

Item in a future window For a given value of w , it is predicted an item that appears somewhere among the q_{j+1} -th, q_{j+2} -th, \dots , q_{j+w} -th items. So, w mostly behaves as a *window* value where the predicted item could appear.

References

- [1] Borja Balle, Rémi Eyraud, Franco M Luque, Ariadna Quattoni, and Sicco Verwer. “Results of the Sequence Prediction Challenge (SPiCe): a Competition on Learning the Next Symbol in a Sequence”. In: *13th International Conference in Grammatical Inference*. Vol. 57. Proceedings of the 13th International Conference in Grammatical Inference. Delft, Netherlands: JMLR W&CP, Oct. 2016. URL: <https://hal.archives-ouvertes.fr/hal-01399429> (cit. on pp. 1, 4).
- [2] Jon L. Bentley and Robert Sedgewick. “Fast algorithms for sorting and searching strings”. In: *SODA*. Vol. 97. 1997, pp. 360–369. (Visited on 01/19/2017) (cit. on p. 3).
- [3] Craig Dillabaugh. *Succinct Data Structures*. 2007 (cit. on p. 12).
- [4] Peter Elias. “Efficient Storage and Retrieval by Content and Address of Static Files”. In: *J. ACM* 21.2 (Apr. 1974), pp. 246–260. ISSN: 0004-5411. DOI: 10.1145/321812.321820. (Visited on 01/24/2017) (cit. on p. 6).
- [5] P. Ferragina and G. Manzini. “Opportunistic data structures with applications”. In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. 2000, pp. 390–398. DOI: 10.1109/SFCS.2000.892127 (cit. on p. 12).
- [6] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu., and V. S. Tseng. “SPMF: a Java Open-Source Pattern Mining Library”. In: *Journal of Machine Learning Research (JMLR)* 15 (2014), pp. 3389–3393 (cit. on pp. 5, 10).
- [7] Simon Gog. *simongog/sdsl-lite*. 2015. URL: <https://github.com/simongog/sdsl-lite> (cit. on p. 9).
- [8] Ted Gueniche, Philippe Fournier-Viger, and Vincent S. Tseng. “Compact Prediction Tree: A Lossless Model for Accurate Sequence Prediction”. In: *Advanced Data Mining and Applications* (2013), pp. 177–188. DOI: 10.1007/978-3-642-53917-6_16 (cit. on pp. 1, 4).
- [9] Ted Gueniche, Philippe Fournier-Viger, Rajeev Raman, and Vincent S. Tseng. “CPT+: Decreasing the Time/Space Complexity of the Compact Prediction Tree”. In: *Lecture Notes in Computer Science* (2015), pp. 625–636. DOI: 10.1007/978-3-319-18032-8_49 (cit. on pp. 1, 3–5).
- [10] Guy Jacobson. “Space-efficient static trees and graphs”. In: *Foundations of Computer Science, 1989., 30th Annual Symposium on*. IEEE, 1989, pp. 549–554. (Visited on 02/18/2016) (cit. on p. 8).
- [11] Guy Jacobson. “Space-efficient Static Trees and Graphs”. In: *30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, North Carolina, USA, 30 October - 1 November 1989*. 1989, pp. 549–554. DOI: 10.1109/SFCS.1989.63533 (cit. on p. 12).
- [12] Ron Kohavi. “A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection”. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI’95*. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143. ISBN: 1-55860-363-8. URL: <http://dl.acm.org/citation.cfm?id=1643031.1643047> (cit. on p. 10).
- [13] Venkata N. Padmanabhan and Jeffrey C. Mogul. “Using Predictive Prefetching to Improve World Wide Web Latency”. In: *SIGCOMM Comput. Commun. Rev.* 26.3 (July 1996), pp. 22–36. ISSN: 0146-4833. DOI: 10.1145/235160.235164 (cit. on p. 2).
- [14] Rajeev Raman, Venkatesh Raman, and S. Srinivasa Rao. “Succinct Indexable Dictionaries with Applications to Encoding K-ary Trees and Multisets”. In: *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms. SODA ’02*. San Francisco, California: Society for Industrial and Applied Mathematics, 2002, pp. 233–242. ISBN: 0-89871-513-X. URL: <http://dl.acm.org/citation.cfm?id=545381.545411> (cit. on pp. 9, 12).
- [15] Sebastiano Vigna. “Quasi-succinct Indices”. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. WSDM ’13*. New York, NY, USA: ACM, 2013, pp. 83–92. ISBN: 978-1-4503-1869-3. DOI: 10.1145/2433396.2433409. (Visited on 07/28/2016) (cit. on p. 6).

