*International Journal of Engineering & Technology*, 7 (4) (2018) 2356-2359

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*
*doi: 10.14419/ijet.v7i4.13901*
***Research paper***

# A novel approach based on sequence prediction for webpage access

**Nguyen Thon Da [1] *, Tan Hanh [2]**

*[1] Faculty of Information Systems, University of Economics and Law, VNU-HCM, Vietnam*
*[2] Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Vietnam*
**Corresponding author E-mail: dant@uel.edu.vn*

**Abstract**

Predicting the next item of a sequence over a finite alphabet is highly important in Web Mining. This paper presents a solution to improve the performance of sequence prediction; first and foremost, predicting what is the next Web page that will be visited by that user for prefetching the Web page. The proposed approach is how to decrease the complexity of the prediction space. Experimental results on a few real-life datasets show that the time execution of this novel approach is better than that of traditional approaches.

***Keywords***: *CPT; CPT+; Sequence Prediction; Web Mining.*

## 1. Introduction

This Sequence Prediction is one of very important and popular in Data Mining, especially Web Mining. Typical works relating to this domain are text compression [1], energy consumption reduction in mobile systems [2], forecast prediction [3].
Many works using different approaches have been proposed such as Machine Learning [4-7], Association Rules [8-12], Sequential Rules [13-15], Markov [16-22] and so on. Besides, a few hybrid methods have been applied such as the combination of Markov and Clustering [23-27], that of Markov, Clustering, and Association Rules [26, 28, 29] and so on.
One of the significant limitations of aforementioned approaches is that they are outmoded. Moreover, they have also a few limitations as follows:
Approaches using Machine Learning build lossy models, which may thus ignore relevant information from training sequences while performing predictions [30].
According to the paper [31], above-mentioned models suffer from some major drawbacks: prediction is not exact due to Markov models. Therefore, these models skipped nearly information contained in training sequences for predicting, and this leads significantly reduce their accuracy. For instance, Markov models typically consider only the last k items of training sequences for performing a prediction, where k is the order of the model. Increasing the order of Markov models is a solution to this problem. Nevertheless, increasing the order of Markov models often leads to a very high state complexity, thus making them may be impractical. Currently, in terms of sequence prediction, the CPT+ is up to 98 times more compact and 4.5 times faster than CPT and has the best overall accuracy when compared to six state-of-the-art models from the literature [31].
In the next section, we formally introduce fundamental knowledge relating to sequence database and sequence prediction, in particular, the CPT+. Section 3 respectively describes sequence prediction for Web page access. In section 4, we present our proposed approach to improve the performance of sequence prediction for Web page access. In section 3, we describe an experimental study. The final section depicts our conclusion.

## 2. Sequence and sequence prediction

A few basic definitions relating to sequence and sequence prediction is presented as follows.

### 2.1. Item set

Itemset I = {$i_1$, $i_2$, $i_3$ …, $i_m$) is an unordered set of distinct items.
Example 1: $I_c$ = {a, b, c, d} is a set of items. These items are different in every pair. Thus, itemsets {a, c, d, b}, {a, c, b, d} and {a, d, b, c} are the same.
Example 2: In the Table 1, itemsets listed include {a, b, x, e}, {c, f, g}, {a, c}, {e, b, f}, {h}…

### 2.2. Sequence

A sequence S is a list of itemsets where S = {$I_1$, $I_2$, $I_3$, $I_n$}, and $I_1$, $I_2$, $I_3$, In are itemsets. Figure 1 illustrates sequences with the following information:
Sequence S1: {a, b, x, e}, {c, f, g}, {k}
Sequence S2: {a, c}, {k}, {a, g, f
Sequence S5: {k}, {e, h, f, c}
In reality, there are some common types of sequences: A sequence of Web pages visited by a user, ordered by the time of access; a sequence of words or characters typed on a laptop by a user, or in a text such as a book; a sequence of products bought by a customer in a retail store; a sequence of proteins in bioinformatics; a sequence of symptoms observed on a patient at a hospital and so on.

### 2.2. Sequence database

A sequence database is a set of sequences SD = ⟨$S_1$, $S_2$, $S_3$,…, $S_k$⟩ having sequence identifiers 1, 2, 3, …, k.
Example 1: A sequence database is shown in Table 1.

The first sequence ⟨{a, b, x, e}, {c, f, g}, {k}⟩ contains three itemsets. This shows that items a, b, x, and e occurred at the same time; items c, f,g occurred at the same time; were followed by k.

**Table 1:** An Example of Sequence Database

| Sid | Sequences |
|---|---|
| Seq1 | {a, b, x, e}, {c, f, g}, {k} |
| Seq2 | {a, c}, {k}, {a, g, f} |
| Seq3 | {k}, {a, f}, {e} |
| Seq4 | {e, b, f}, {a}, {h}, {k} |
| Seq5 | {k}, {e, h, f, c} |

Table 2 presents a special case of the sequence database. In this case, every itemset only has only one item. These sequence databases are usually used for making predictions. For example, they are utilized for predicting the next symbol of a sequence based on the previously observed symbols.

**Table 2:** A Special Case of Sequence Database

| Sid | Sequences |
|---|---|
| Seq1 | {1}, {2}, {3} |
| Seq2 | {2}, {1} |
| Seq3 | {4}, {1}, {5} |
| Seq4 | {2}, {3}, {5}, {4} |
| Seq5 | {3}, {4}, {2} |
| Seq6 | {2}, {5}, {4}, {5}, {1} |

According to [31], a prediction model is trained with a set of training sequences (known as sequence database).
There are numerous popular applications relating to sequence predictions such as weather forecasting, web page prefetching, stock market prediction, consumer product recommendation and so on. Figure 1 illustrates a sequence prediction model.

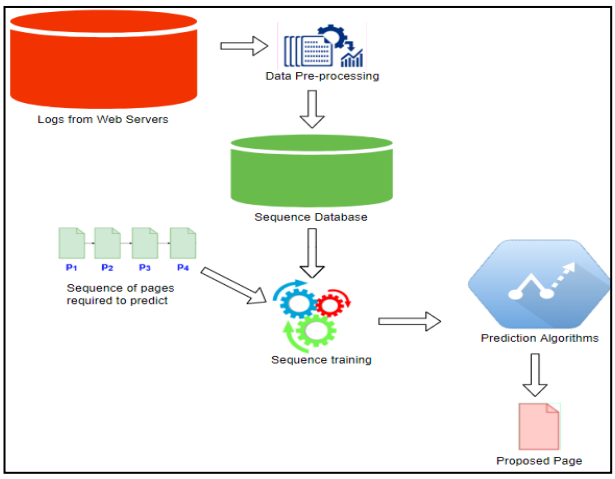


**Fig. 1:** A Sequence Prediction Model.

According to Figure 1, in this model, there are three main phases have been proposed. Firstly, the cleaning data phase is done. Secondly, training a sequence prediction model using some previously seen sequences called the training sequences. The final phase is to use a trained sequence prediction model to perform prediction for new sequences. To sum up, given a set of training sequences, the task of sequence prediction is to find the next element of a target sequence by only observing its previous items [30].
Many works relating to sequence prediction have been proposed. They are mostly based on the Markov models such as Dependency Graph (DG) [32], All-k-order Markov (AKOM) [33], Transition Directed Acyclic Graph (TDAG) [34], CPT [30], and CPT+ [31]. Besides, sequential rule mining and neural networks have been used to predict sequences [31].
The work [31] indicates that on various real-time sequence databases the resulting models CPT [30] and CPT+ [31] are much more exact than other models (For more information, see Figure 2).

| Datasets | CPT+ | CPT | AKOM | DG | LZ78 | PPM | TDAG |
|---|---|---|---|---|---|---|---|
| BMS | **38.25** | 37.90 | 31.26 | 36.46 | 33.46 | 31.06 | 6.95 |
| SIGN | **33.01** | 32.33 | 8.63 | 3.01 | 4.79 | 4.25 | 0.00 |
| MSNBC | 61.50 | **61.64** | 47.88 | 55.68 | 43.64 | 38.06 | 31.14 |
| Bible word | 27.52 | 22.05 | **38.68** | 24.92 | 27.39 | 27.06 | 23.33 |
| Bible char | **73.52** | 69.14 | 7.96 | 0.00 | 3.02 | 0.10 | 9.90 |
| Kosarak | **37.64** | 33.82 | 20.52 | 30.82 | 20.50 | 23.86 | 1.06 |
| FIFA | **35.94** | 34.56 | 25.88 | 24.78 | 24.64 | 22.84 | 7.14 |

**Fig. 2:** Comparison of Accuracy and Model Size [31].

According to the benchmark in Figure 2, they compare their proposed CPT+ sequence prediction model with a few up-to-the-minute models on many different types of data. FIFA, MSNBC, BMS, and Kosarak are sequence databases of Web pages, for instance. It is easy to realize that CPT+ much better than other models in terms of accuracy.

## 2.3. Using CPT for sequence prediction

The CPT is a promising proposed prediction model [31]. This is a model that provides more accurate predictions, have reasonable size and is noise tolerant. Its idea is to build a lossless model (or a model where the loss of information can be controlled) and to use all relevant information to perform each sequence prediction. Challenges that would tackle: (1) Define an efficient structure in terms of space to store sequences, (2) The structure must be incrementally updatable to add new sequences, (3) The prediction algorithm offers accurate predictions, and if possible, is also effective in terms of time [31]. Proposal of authors is to build the Compact Prediction Tree (CPT) with a tree-structure for storing training sequences, an indexing mechanism, and each sequence is inserted one after the other in the CPT.
The training process of CPT receive a set of training sequences and generates three different structures: (1) a PT (Prediction Tree), (2) an LT (Lookup Table) and (3) an Inverted Index. Sequences are considered one by one to build gradually these structures during the training process.

## 2.4. Using CPT+ for sequence prediction

Despite the CPT is considered as one of the most accurate sequence prediction models, it is still unsuitable for applications where the number of sequences is very large. Therefore, an important backward of CPT is its high time and space complexity. Authors of the work [34] tackled this issue by proposing three novel strategies to reduce CPT's size and prediction time and increase its accuracy [31].
The task of sequence prediction is composed of predicting the next symbol of a sequence based on the previously observed symbols. For example, if a user has visited pages p1, p2, p3, in that order, one may want to predict next pages that will be visited by that user.
To predict Web page access, two following steps would be done. First, one must train a sequence prediction model for Web page prediction using the sequences of Web pages visited by several users. In the second step, a prediction model trained with the sequences of Web pages visited by several users will be used to perform Web page access predictions.
We utilized CPT+ [31] for Web page access prediction. Our novel proposal is to decrease the predictive space reserving the accuracy. Further, the time execution of the new proposal is also better than that of the ordinary.
Let S = pageA, pageB, pageE be the sequence we need predict the next page. In the sequence S, PageA is followed by pageB and pageB is followed by pageE.
Suppose that we have a sequence database as follows.
Seq1: pageA, pageB, pageC, pageX, pageY, pageU
Seq2: pageH, pageA, pageB, pageE, pageZ, pageG, pageU, pageF
Seq3: pageK, pageA, pageB, pageE
Seq4: pageA, pageB, pageN, pageX, pageV

Seq5: pageN, pageA, pageC, pageX, pageZ, pageB, pageU, pageX
Seq6: pageM, pageJ, pageA, pageB, pageE, pageR, pageO
Seq7: pageN, pageA, pageB, pageE, pageI, pageZ, pageF, pageK
Seq8: pageN, pageA, pageB, pageH, pageI, pageZ, pageS, pageV
Seq9: pageM, pageJ, pageA, pageB, pageE, pageR, pageO
Seq10: pageH, pageZ, pageB, pageE, pageM
In this context, we need predict the next page of above sequence S. To address this, we apply the CPT+ and propose two strategies for decreasing the prediction space as follows.
Strategy 1: Remove sequences in the sequence database that contain only one sequence S and S is at the last position of each of sequences because removed these sequences are redundant for predicting the next page. Let $T_1$ be the time for executing this task.
Then, the strategy 2 will be done.
Strategy 2: Remove sequences belong to the sequence database without sequence S due to redundancy. Let $N_1$, $T_2$ be the number of remaining sequences after removing these sequence and the time to execute this task, respectively.
After doing the above strategy, we continue using the CPT+ to making Web page access prediction. At this time, the prediction space has been shortened with respect to the ordinary one.
Let reduced-SDB, $T_3$ be the new sequence database and the time execution, respectively. In this case, the size of reduced-SDB is $N_1$.
Let N be the number of sequences of ordinary sequence database and T be the time execution for predicting the next page of the sequence S.
It is not difficult to realize that $N \geq N_1$.

Let $T_4 = T_1 + T_2 + T_3$.

We will check whether if $T_4 > T$ or $T_4 < T$.
In Section 3, we will present our experimental results.

# 3. Experimental results

## 3.1. Experimental environment

With regard to the hardware platform, we used a laptop computer using the configuration: Intel 8-core processor-i7-4800M, CPU@2.70 GHz, 32GB RAM; 256 GB hard drive (SSD 256 MB). The software environment uses the following configuration: the operating system is Ubuntu 14.04 LTS 64 bit and the Java development platform is the JDK 8u131.

## 3.2. Data

We have collected data and built sequence databases from Web Log data of two Websites http://villazest.co.za (offering travel services) and http://palmviewsanibel.com (offering resort services). See http://bit.ly/2NrttAF and http://bit.ly/2Lxc3ki respectively for detailed data collected from two these Web log files.
The Web Log Data file of the Website villazest.co.za contains 334,723 accesses visited by users from 07-Feb-2015 to 27-Sep-2015 (approximately 77 MB). Also, the Web Log Data file of the Website pamviewsanibel.com contains 4,217,576 accesses visited by users from 05-Apr-2013 to 19-May-2013 (approximately 1.0 GB).
Besides, we also built a tool to remove redundant data to remain meaningful data by using the C Sharp (C#) programming language and Log Parser Studio (tinyurl.com/logparser-studio). Then items (visited links) of sequences were encoded into integer numbers to access faster and save the space for presenting.

```
8 -1 2 -1 19 -1 5 -1 -2
7 -1 23 -1 28 -1 15 -1 -2
14 -1 3 -1 2 -1 27 -1 19 -1 -2
4 -1 16 -2 23 -1 8 -1 25 -1 51 -1 -2
18 -1 2 -1 42 - 1 9 -1 -2
```

**Fig. 3:** A Formatted Sequence Database.

Figure 3 shows a formatted sequence database. Each item from a sequence is an integer number having positive value and items from the same itemset within a sequence are separated by unique spaces. In this case, the value "-1" indicates the end of an itemset. The value "-2" depicts the end of a certain sequence. We could consider every item as a page visited by users.
Information relating to datasets is illustrated in Table 3.

**Table 3:** Real-Time Datasets

| Datasets | Number of sequences |
|---|---|
| villazest.co.za | 966 |
| pamviewsanibel.com | 2749 |

## 3.3. Experimental results

After predicting the next pages of a few of sequences in sequence databases from Website villazest.co.za, we obtain results presenting in Figure 4.
Results of sequence prediction in Figure 4 show that the time execution of our approach is from 1.5 to approximately 2 times than that of the ordinary approach.
Similarly, with the Website pamviewsanibel.com, results of sequence prediction in Figure 5 show that the time execution of our approach is from 2.6 to 4 times than that of ordinary. Thus, T is much slower than $T_4$.
Rather, in order to check the accuracy of our approach, we also obtain results described in Figure 4 and Figure 5.

| Sequences | Original approach | | Our approach | |
|---|---|---|---|---|
| | Time execution (milliseconds) | Accuracy | Time execution (milliseconds) | Accuracy |
| S1: <9, 4, 8> | 216 | 99.99 | 114 | 100 |
| S2: <9, 4, 8, 0> | 150 | | 89 | |
| S3: <9, 4, 8, 0, 1> | 161 | | 89 | |

**Fig. 4:** Dataset from Villazest.Co.Za.

| Sequences | Original approach | | Our approach | |
|---|---|---|---|---|
| | Time execution (milliseconds) | Accuracy | Time execution (milliseconds) | Accuracy |
| S1: <2, 1, 7> | 287 | 98.76 | 106 | 100 |
| S2: <7, 4, 3> | 288 | | 110 | |
| S3: <1, 4, 3> | 291 | | 100 | |
| S4: <2, 3, 1> | 403 | | 102 | |

**Fig. 5:** Dataset from pamviewsanibel.com.

The above results led the recognition that the time execution is significantly decreased when comparing our approach with the original approach for sequence prediction. Moreover, the accuracy of our approach is better than that of the original approach.
Next, we present the complexity of our approach.
As mentioned above, N is the number of sequences of original sequence database and $N_1$ is the number of remaining sequences after removing redundant sequences. As we knew, N1 is much less than N.
The complexity of our approach is similar to that of CPT+ presented in the work [31]. The different thing here is the size of prediction space. This means that the number of sequences used in our approach is much less than that of sequences used ordinary approach.

## 4. Conclusion

This paper introduced an approach to improve the performance of sequence prediction, in particular, predict what is the next Web page that will be visited by that user. The way the paper proposed is removing redundant sequences, decreasing the prediction space of sequence database in advance. Then, CPT+ was utilized.
We used two real-time datasets and obtained the results better than the ordinary approach in terms of time execution and accuracy. Experimental results show that our approach is from 1.5 times to 4 times faster than the ordinary approach. Rather, the accuracy of our approach is from 0.3% to 1% better than that of the ordinary approach.
In the future, we have a plan to improve the performance of our approach in terms of the time execution and accuracy, particularly, the accuracy for sequence prediction. Besides, another work brings many research opportunities is the combination of different methods to improve the performance of sequence prediction for Web page access.

## Acknowledgement

The authors thank Prof. Philippe Fournier-Viger (http://philippe-fournier-viger.com/) for his support in terms of sequence prediction and its applications. Currently, he is the director of Center of Innovative Industrial Design, Harbin Institute of Technology (Shenzhen, China)).

## References

[1] T. C. Bell, J. G. Cleary, and I. H. Witten, Text compression: Prentice-Hall, Inc., 1990.
[2] C. Draa, J. Tayeb, S. Niar, and E. Grislin, "Application sequence prediction for energy consumption reduction in mobile systems." pp. 23-30.
[3] A. J. Majda, I. Timofeyev, and E. V. Eijnden, “Models for stochastic climate prediction,” Proceedings of the National Academy of Sciences, vol. 96, no. 26, pp. 14687-14691, 1999. https://doi.org/10.1073/pnas.96.26.14687.
[4] G. Suchacka, and S. Stemplewski, "Application of Neural Network to Predict Purchases in Online Store." pp. 221-231.
[5] S. Bahram, D. Sen, and R. S. Amant, "Prediction of web page accessibility based on structural and textual features." p. 31.
[6] M. T. Hassan, K. N. Junejo, and A. Karim, "Learning and predicting key Web navigation patterns using Bayesian models." pp. 877-887.
[7] D. Bonino, F. Corno, and G. Squillero, "A real-time evolutionary algorithm for web prediction." pp. 139-145.
[8] M. Li, X. Yu, and K. H. Ryu, “MapReduce-based web mining for prediction of web-user navigation,” Journal of Information Science, vol. 40, no. 5, pp. 557-567, 2014. https://doi.org/10.1177/0165551514544096.
[9] N. Labroche, N. Monmarché, and G. Venturini, "A new clustering algorithm based on the chemical recognition system of ants." pp. 345-349.
[10] L. Jianhui, and Z. Bingjie, "A Web Prediction Pattern Recommendation Algorithm." pp. 263-266.
[11] Q. Yang, T. Li, and K. Wang, “Building association-rule based sequential classifiers for web-document prediction,” Data mining and knowledge discovery, vol. 8, no. 3, pp. 253-273, 2004. https://doi.org/10.1023/B:DAMI.0000023675.04946.f1.
[12] R. Geetharamani, P. Revathy, and S. G. Jacob, “Prediction of users webpage access behaviour using association rule mining,” Sadhana, vol. 40, no. 8, pp. 2353-2365, 2015. https://doi.org/10.1007/s12046-015-0424-0.
[13] P. Fournier-Viger, T. Gueniche, and V. S. Tseng, "Using Partially-Ordered Sequential Rules to Generate More Accurate Sequence Prediction." pp. 431-442.
[14] E. Frias-Martinez, and V. Karamcheti, "A prediction model for user access sequences."
[15] M. Géry, and H. Haddad, "Evaluation of web usage mining approaches for user's next request prediction." pp. 74-81.
[16] D. Dhyani, S. Bhowmick, and W.-K. Ng, "Modelling and predicting a Web page accesses using Markov processes." pp. 332-336.
[17] V. M. Rao, and V. V. Kumari, “An efficient hybrid successive Markov model for predicting web user usage behavior using web usage mining,” International Journal of Data Engineering (IJDE), vol. 1, no. 5, pp. 43-62, 2010.
[18] X. Dongshan, and S. Junyi, “A new markov model for web access prediction,” Computing in Science & Engineering, vol. 4, no. 6, pp. 34-39, 2002. https://doi.org/10.1109/MCISE.2002.1046594.
[19] C. S. Iliopoulos, C. Makris, Y. Panagis, K. Perdikuri, E. Theodoridis, and A. Tsakalidis, “The weighted suffix tree: an efficient data structure for handling molecular weighted sequences and its applications,” Fundamenta Informaticae, vol. 71, no. 2, 3, pp. 259-277, 2006.
[20] V. S. Tseng, K. W. Lin, and J.-C. Chang, “Prediction of user navigation patterns by mining the temporal web usage evolution,” Soft Computing-A Fusion of Foundations, Methodologies and Applications, vol. 12, no. 2, pp. 157-163, 2008.
[21] M. Narvekar, and S. S. Banu, “Predicting user's Web navigation behavior using hybrid approach,” Procedia Computer Science, vol. 45, pp. 3-12, 2015. https://doi.org/10.1016/j.procs.2015.03.073.
[22] B. Nigam, S. Tokekar, and S. Jain, “Evaluation of models for predicting user's next request in web usage mining,” international Journal on Cybernetics & informatics (UCi), vol. 4, pp. 1-13, 2015.
[23] M. A. Awad, and I. Khalil, “Prediction of user's web-browsing behavior: Application of markov model,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1131-1142, 2012. https://doi.org/10.1109/TSMCB.2012.2187441.
[24] M. Awad, L. Khan, and B. Thuraisingham, “Predicting WWW surfing using multiple evidence combination,” The VLDB Journal—The International Journal on Very Large Data Bases, vol. 17, no. 3, pp. 401-417, 2008.
[25] P. Sampath, A. Wahi, and D. Ramya, “A COMPARATIVE ANALYSIS OF MARKOV MODEL WITH CLUSTERING AND ASSOCIATION RULE MINING FOR BETTER WEB PAGE PREDICTION,” Journal of Theoretical & Applied Information Technology, vol. 63, no. 3, 2014.
[26] J. Zhu, J. Hong, and J. G. Hughes, "Using markov chains for link prediction in adaptive web sites," Soft-Ware 2002: Computing in an Imperfect World, pp. 60-73: Springer, 2002. https://doi.org/10.1007/3-540-46019-5_5.
[27] P. Thwe, "Using Markov Model and Popularity and Similarity Based PageRank Algorithm for Web Page Access Prediction."
[28] S. Dubey, and N. Mishra, “Web page prediction using hybrid model,” International Journal on Computer Science and Engineering, vol. 3, no. 5, pp. 2170-2176, 2011.
[29] F. Khalil, J. Li, and H. Wang, "Integrating recommendation models for improved web page prediction accuracy." pp. 91-100.
[30] T. Gueniche, P. Fournier-Viger, and V. S. Tseng, "Compact Prediction Tree: A Lossless Model for Accurate Sequence Prediction." pp. 177-188.
[31] T. Gueniche, P. Fournier-Viger, R. Raman, and V. S. Tseng, "CPT+: Decreasing the time/space complexity of the Compact Prediction Tree." pp. 625-636.
[32] V. Padmanabhan, and J. Mogul, “Using Prefetching to Improve World Wide Web Latency,” Computer Communications, vol. 16, pp. 358-368, 1998.
[33] J. Pitkow, and P. Pirolli, "Mininglongestrepeatin g subsequences-topredict worldwidewebsurfing." p. 1.
[34] P. Laird, and R. Saul, “Discrete sequence prediction and its applications,” Machine learning, vol. 15, no. 1, pp. 43-68, 1994. https://doi.org/10.1007/BF01000408.