Investigating the PageRank and sequence prediction based approaches for next page prediction

Nguyen Thon Da[1], Tan Hanh[2]
[1]Faculty of Information Systems, University of Economics and Law, VNU-HCM, Vietnam
[2]Department of Information Technology, Posts and Telecommunications Institute of Technology, Vietnam

Article Info

Article history:

Received Jul 25, 2020
Revised Dec 10, 2020
Accepted Dec 17, 2020

Keywords:

CPT+
Markov
PageRank
Sequence prediction

ABSTRACT

Discovering unseen patterns from web clickstream is an upcoming research area. One of the meaningful approaches for making predictions is using sequence prediction that is typically the improved compact prediction tree (CPT+). However, to increase this method's effectiveness, combining it with at least other methods is necessary. This work investigates such PageRank-based methods related to sequence prediction as All-K-Markov, DG, Markov 1st, CPT, CPT+. The experimental results proved that the integration of CPT+ and PageRank is the right solution for next page prediction in terms of accuracy, which is more than a standard method of approximately 0.0621%. Still, the size of the newly created sequence database is reduced up to 35%. Furthermore, our proposed solution has an accuracy that is much higher than other ones. It is intriguing for the next phase (testing one) to make the next page prediction in terms of time performance.

Corresponding Author:

Nguyen Thon Da
Faculty of Information Systems
University of Economics and Law, VNU-HCM
Quarter 3, Linh Xuan Ward, Thu Duc District, Ho Chi Minh City, Vietnam
Email: dant@uel.edu.vn

1. INTRODUCTION

Sequence learning is a significant component of learning in numerous areas of intelligent systems, DNA sequencing [1] and web page prediction. With the increase in information and communication, the study of web page prediction becomes one of the meaningful and challenging tasks. Moreover, Sequential pattern mining is a very active research topic, where hundreds of papers present new algorithms and applications each year [2]. They have numerous real-life applications since data is naturally encoded as sequences of symbols in many fields such as bioinformatics, market basket analysis, text analysis, energy reduction in smart homes, web page clickstream analysis, and e-learning [2]. One of the crucial branches of sequential pattern mining is sequence prediction. Given a set of training sequences, the problem of sequence prediction involves how to find the next item of a target sequence by only observing its previous items [3]. Numerous models have been proposed by researchers to address this sequence prediction issue. Some of them use, for example, neural networks, pattern mining, and a probabilistic approach. However, Markov Chains are also prevailingly used for this.

As an essential part of data mining, next page prediction has become more common in the real world. An example of this is clickstream analysis that one of its significant tasks is to find essential items and is also one of the hot topics these days. The principal aim of this article is to investigate various methods using the PageRank and sequence for next page prediction on real clickstream datasets. In this paper's scope, we focus on techniques related to sequence prediction for next page prediction. This paper is organized as follows. Section 2 introduces the background. Section 3 describes related work. In the next section, we present our proposed approach. In section 5, we offer results and discussion. Finally, section 6 concludes.

2. BACKGROUND

2.1. Next page prediction

The next page prediction task consists of predicting the next page of a sequence based on the previously observed pages. For instance, if a user has visited some webpages X, Y, Z, in that order, one may want to predict what is the next webpage that will be visited by that user. There are two major steps for making the next page prediction shown below.
Step 1: Training sequences → building a sequence prediction model → prediction model
Step 2: (prediction model, a specific sequence) → prediction algorithm → prediction.

2.2. Common sequence prediction models

One of the typical models are proposed for Webpage prediction is a rules-based model [4-9]. The dependency graph (DG) sequence prediction model was proposed in work [10]. The authors used a prediction algorithm patterned after that proposed by James Griffioen and Randy Appleton [11]. They have presented a prefetching scheme for the World Wide Web aimed at reducing the latency perceived by users. Also, the prediction algorithm constructs a dependency graph that depicts the pattern of accesses to different files stored at the server [10]-using the first-order Markov prediction method described in [11] for file prediction, Padmanabhan and Mogul [12] constructed a dependency graph containing nodes for all files ever accessed at a particular WWW server. To effectively predict, [13] estimated from n-grams to yield the conditional probabilities for sequence prediction.

The paper [14] describes how the conflict can be resolved with partial string matching and reports experimental results showing that mixed-case English text can be coded in as little as 2.2 bits/characters with no prior knowledge of the source. The work used the concept of compressibility shown to play a role analogous to that of entropy in classical information theory, where one deals with probabilistic ensembles of sequences. [15] is a model for making sequence prediction using the text-compression method; it limits the growth of storage by retaining the most likely prediction contexts and discarding (forgetting) less likely ones. A robust model is CPT that is described in [3] as below.

The CPT's advantage is that it could compress the training data so that all relevant information is available for each prediction. It also offers a low time complexity for its training phase and is easily adaptable for different applications and contexts [3]. An improved model of CPT is CPT+. The CPT+ address this issue by proposing three novel strategies to reduce CPT's size and prediction time, and increase its accuracy. Experimental results of CPT+ on seven real-life datasets show that the resulting model (CPT+) is up to nearly 100 times more compact and nearly five times faster than CPT, and has the better accuracy than other models such as AKOM [13], CPT [3], DG [10], Lz78 [12], PPM [14], and TDAG [15]. This work aims to improve the effectiveness of next page prediction by using the integration of PageRank with a start-of-the-art approach called CPT+ and providing a comparison of this proposed approach with other ones.

2.3. Using PageRank for sequence prediction

The research [16] indicates that PageRank has gone from being used to evaluate the importance of web pages to a much broader set of applications. The PageRank algorithm supporting the sequence prediction is described as below. Let SDfull be the original sequence database. After the PageRank computation is processed, the sequence database is arranged into two parts. Part 1: SDhigh is a set of sequential data series with a high average PageRank index and the second part: SDlow is a set of sequential data series with a low average PageRank. The relationships of these data sets are determined by the formula (1):

$$SD_{full} = SD_{high} \cup SD_{low} \tag{1}$$

Consider SD_{high} is a data set containing strings of the form *P_{PR}, where* is any sequence of data and predictable PPR page (P_{PR} always follows strings *). It is considering the websites that visit the P_{PR} page, when the PageRank index of the P_{PR} page is high, the more pages that will directly visit the P_{PR} page (according to PageRank's nature). This means that the number of sequences that successfully predict the P_{PR} page will increase. Conversely, when the PageRank of a P_{PR} page is low, the fewer pages will go directly to the P_{PR} page and is the number of strings that predict the P_{PR} page's success will drop. When calculating the average of the PageRank indices on the data sequence, the higher the average number of sequential series, the more strings of form * P_{PR} will appear. This also means that the number of predicted success sequences will be more and will be added to the number of predicted success sequences according to the CPT+ algorithm. Therefore, the integration of PageRank calculation into CPT+ is significant for predicting Web access.

3. LITERATURE REVIEW

This research [17] aims to predict the user's behaviour using the Apriori prefix tree (PT) algorithm. Using the popularity value of pages, the authors of the work [18] bias conventional PageRank algorithm and model a next page prediction system that produces page predictions under given top-n value. The work [19] introduced an approach for personalized page ranking and recommendation by integrating association mining and PageRank to meet user's search goals. The effectiveness of their proposed method was verified through a few experimental evaluations.

The work [20] proposed a PageRank-like algorithm is proposed for conducting web page access prediction, and they extended the use of PageRank algorithm for next page prediction with several navigational attributes. In the research [21] provides a solution to web page access prediction aiming to increase accuracy and efficiency by reducing the sequence space with the integration of PageRank into CPT+. Moreover, the work [22] proposed a method in which an ambiguous prediction problem can be resolved using web PageRank and Markov model. Its experimental result shows a reduced number of vague predictions after applying the PageRank method. Assuming a set of successive past top-k rankings, the paper [23] introduced a method for predicting the ranking position of a web page by ranking trend sequences used for Markov models training. Due to the accuracy of the low order Markov model usually is not satisfactory, the article [24] utilized popularity and similarity-based PageRank algorithm to make predictions when the ambiguous results are found. The work [25] proposed the use of a PageRank-based algorithm for the web site's graph, and they proved, through experimentation, that their approach results in more accurate and representative predictions than the ones produced from the pure usage-based approaches.

However, these Markov based models suffer from some significant drawbacks, most of them assume the Markovian hypothesis that each event solely depends on the previous circumstances, and thus, prediction accuracy using these models can decrease [26]. Furthermore, both rules-based models and Markov based models do not use all the information contained in training sequences to perform predictions, and this can severely reduce their accuracy [26]. Also, the research [27] investigated related work for web page access prediction. The research indicates that the combination of the CPT+ [26] with the PageRank is a meaningful choice for next page prediction. The scope of this research is that it deals with issues regarding predicting the next items effectively in clickstream or web access context that is being used in various areas in real life. For instance, the result of this work can be applied to predicting behaviour in the e-commerce context or predicting users' trends while visiting various websites.

4. RESEARCH METHOD

In this section, we first utilize K-fold cross validation to divide each real dataset into ten equal parts and perform the work [21] on training datasets. Secondly, we evaluate various sequence prediction based methods by using [28] for smaller datasets with their size shorten by the approach [21].

4.1. Integrating K-fold cross validation method to improve data mining accuracy for web access prediction

The objective of cross-validation is to test the model's ability to predict new data [29]. In particular, the K-fold cross validation method [30] divides the set of observations into K groups, approximately with equal size [31]. K is usually chosen as 5 or 10, and as K becomes extensive, the size difference between the training set and the subsamples will be smaller again, as this difference decreases, the deviation of the technique the lower the [32]. The data is trained and tested K times, each time t, trained on the set D\Dt and tested on Dt (D is the original data set, and Dt is the set test data) [30]. The estimate of cross-validation accuracy is the sum of the correct classifications divided by the number of entities in the original dataset. The purpose of K-fold cross validation is mainly used to estimate the ability of the machine learning model on invisible data.

4.2. Develop training data sets and improve accuracy

4.2.1. Data

We changed the dataset KOSARAK (collected from http://fimi.ua.ac.be/data) into a sequence database (including 100,000 sequences) in a format that is defined as follows. There is a text file which represents a sequence from a sequence database. A single space and a -1 separate each item from a sequence. The value "-2" shows the end of a sequence.

4.2.2. Method

We propose a combination among numerous techniques such as K-folder-validation check, PageRank algorithm on sequence database, analysing sequences to predict next pages effectively. The proposed procedure includes 4 main phases introduced below.

Phase 1 : we run randomly all sequences inside the considered datasets (also the input sequence database). Then we split the randomized dataset into 10 equal parts. The first part is used as a testing dataset called DBTest1, 9 remaining parts become another dataset for a training dataset called DBTrain1.

Phase 2 : We calculate the PageRank value for all sequences in the dataset DBTrain1. Thanks to the [21] we reduced redundant sequences.

Phase 3 : We analyse sequences and use an effective sequence prediction to predict next pages on the testing dataset (DBTest1).

We repeat the mentioned phases for 9 remaining parts according to the theory of K-Fold Cross Validation.

Phase 4 : We evaluate the accuracy of sequence-prediction-based models, analyse and draw conclusions.

4.2.3. Evaluation framework

We used the evaluation Framework introduced in the article [26] to evaluate prediction models. A prediction can be either a success if the prediction is accurate, a failure if the prediction is inaccurate, or a none-match if the model is unable to perform a prediction. Besides, coverage is the ratio of sequences without prediction against the total number of test sequences, and accuracy is the number of successful predictions against the total number of test sequences [26].

5. RESULTS AND DISCUSSIONS

After creating 10 data sets according to the above method, we proceeded to take ten training sets (with the size of 90,000 lines) of these 10 data sets to implement the solution to shorten the data series. By using the PageRank algorithm, sequential databases with corresponding precision are created, as illustrated in Table 1. In which R_i is the accuracy of the collapsed sequential databases during the ith K-Fold Check Validation. According to Table 1, the values 100, 98, 96 down to 58, 56 are the size (in per cent) of the collapsed database, respectively, compared to the training database.

Experimental results show that when applying the PageRank solution to reduce the training data, gradually set size from 2%, 4%, 6%, up to 34% (corresponding to the compact data set is 98%, 96%, 94%, down to 66%), accuracy is higher than that of the initial training database. The compact training sequence database's construction took a long time due to the large data set (100,000 lines), and the number of nodes in the directed graph is many (23,496 nodes).

According to the test results illustrated in Figure 1, the average predictive accuracy of the initial training database (sized 90,000) is 99.936%, when removing the redundant data series to the database If the collapsed data reaches a size of 66% (59,400 lines), the average predicted accuracy is 100% (an increase of 0.0621%). Figure 1 illustrates a chart comparing the average of the predicted accuracy on the collapsed datasets in size without losing the predictive accuracy by PageRank and CPT+.

Note that, when the size is reduced to 66%, the peak accuracy is 100%, and a process of degradation of accuracy is reached when the size is 62% or less. From the above experimental results, we have the basis to confirm that when using the reduced training data set of size, 66% (59,400) to continue for the next stage is the test phase (predictive) is very feasible. Comparison of web access prediction models by integrating PageRank: The empirical results detailed in Table 1 and Figure 2 show that the solution of integrating PageRank with CPT + and DG is suitable with the predicted accuracy of web access is approximately 100% for CPT+ and over 80 % for DG. In contrast, the solution of integrating PageRank with CPT (an old version of CPT+) is not suitable because the accuracy of web access prediction, in this case, has not reached 50%.

Furthermore, Figure 2 also shows that integrating PageRank with CPT+ is more effective than all the other methods (DG, Markov1, AKOM, LZ78, CPT) (see Appendix from 1 to 6). It can be easily understood that the pattern of DG stands in second place with accuracy from 80% to 93%. There is a steady increase from 80% (accuracy) at a size of 100% to 93% at 56% (reduced sequence database). Besides, Markov1 is the third place with accuracy from 65% to 83%. The figures for AKOM and LZ78 show similar trends. In contrast, the accuracy of CPT is below 50% in most cases. The figure for CPT indicates a steady fall from about 48% to about 38% (accuracy) when reducing the size of the sequence database using the PageRank algorithm. The pattern of CPT+ is relatively stable, with the accuracy reached approximately 100% in most cases. Therefore, our proposed approach to integrating PageRank with CPT+ is an effective solution for predicting web access.

Table 1. The comparison of the accuracy of sequence prediction models

Size (%)	Average of accuracy					
	DG	CPT	Markov1	AKOM	LZ78	CPT+
100	80.116	48.088	65.932	59.451	59.945	99.936
98	80.585	48.031	66.338	59.773	60.319	99.941
96	81.060	48.007	66.799	60.171	60.630	99.948
94	81.486	47.946	67.312	60.591	60.922	99.955
92	81.996	47.955	67.986	61.145	61.415	99.960
90	82.499	47.924	68.580	61.577	61.811	99.964
88	83.044	47.631	70.282	62.123	62.267	99.968
86	83.517	46.994	70.076	62.702	62.707	99.971
84	84.087	46.332	70.925	63.353	63.265	99.972
82	84.678	45.741	71.728	64.083	64.002	99.976
80	85.292	44.877	72.592	64.829	64.502	99.966
78	85.931	44.091	73.495	65.780	65.095	99.982
76	86.828	43.295	74.501	66.478	65.897	99.985
74	87.834	42.397	75.466	67.283	66.486	99.990
72	88.497	41.700	76.362	68.146	67.260	99.993
70	89.311	41.105	77.278	69.733	68.037	99.997
68	89.931	40.611	78.080	68.887	68.773	99.998
66	90.307	40.281	78.918	70.293	69.443	99.998
64	90.781	39.932	79.299	70.986	70.407	99.985
62	89.613	39.124	80.771	71.879	71.195	99.979
60	90.731	39.084	81.868	72.873	72.534	99.862
58	92.086	38.125	83.127	73.959	73.317	99.690
56	93.304	37.970	82.887	74.931	73.958	99.683

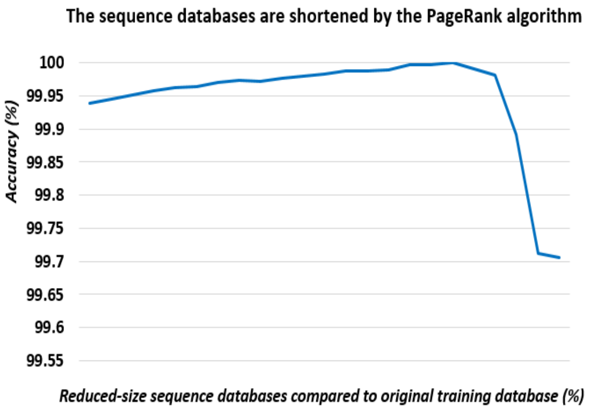

Figure 1. Sequence databases are shortened by PageRank algorithm

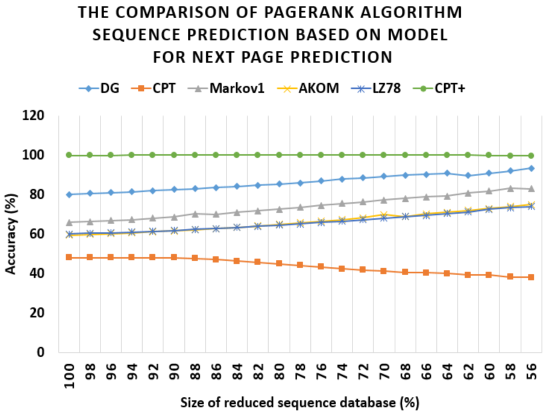

Figure 2. The comparison of PageRank algorithm sequence prediction based model for next page prediction

6. CONCLUSION

In this paper, we presented an investigation about the sequence prediction used to predict the next page. Experimental results on the real dataset Kosarak shows that the integration of CPT+ with PageRank is a bit higher than any sequence prediction based models. Besides, the size of the shortening sequence database is reduced down to nearly 35%. This cutting in terms of the size of sequence databases takes advantage of the testing phase (prediction phase). Thus, the combination of PageRank with sequence processing support to the CPT+ for next page prediction. In particular, redundant sequences are removed from datasets (sequence databases). The prediction space is shorten depending on the number of removed sequences. Besides, after removing neccessary data, the accuracy is still not changed even better. From present time continuing forwards in time, we are goint to develop a novel algorithms or continue improving the available algorithms such as CPT+, PageRank to solve the next page prediction issue more effective. A part from that, a solution using big data tools is also considered to escalate the performance in terms of time execution for deal with this issue.

APPENDIX

Appendix 1. The accuracy of model DG on reduced datasets

No. l	%	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	**100**	**80.065**	**79.995**	**80.076**	**80.085**	**80.237**	**80.157**	**79.92**	**80.07**	**80.212**	**80.344**
2	**98**	80.63	80.429	80.59	80.52	80.747	80.52	80.251	80.743	80.707	80.713
3	**96**	81.18	80.934	81.078	80.973	81.267	80.882	80.838	81.197	81.028	81.218
4	**94**	81.674	81.201	81.53	81.374	81.508	81.36	81.365	81.689	81.508	81.654
5	**92**	82.132	81.779	82.018	81.909	82.032	81.91	81.953	82.059	82.091	82.075
6	**90**	82.74	82.092	82.549	82.453	82.537	82.525	82.43	82.594	82.433	82.639
7	**88**	83.248	82.855	82.857	83.08	83.072	83.142	83.038	83.026	83.005	83.119
8	**86**	83.555	83.389	83.383	83.525	83.331	83.565	83.558	83.62	83.649	83.594
9	**84**	84.165	83.819	83.979	84.268	83.879	84.151	84.272	84.161	83.941	84.236
10	**82**	84.649	84.701	84.554	84.749	84.466	84.739	84.738	84.751	84.487	84.944
11	**80**	85.208	85.404	85.147	85.509	85.254	85.303	85.194	85.119	85.25	85.528
12	**78**	85.811	85.996	85.92	86.03	85.849	85.9	85.766	85.915	85.879	86.245
13	**76**	86.289	86.899	86.861	86.86	86.724	86.913	86.806	86.809	86.92	87.198
14	**74**	87.138	87.803	88.146	88.039	87.732	87.926	87.766	87.896	87.871	88.027
15	**72**	88.003	88.527	88.658	88.496	88.574	88.417	88.553	88.705	88.574	88.463
16	**70**	88.978	89.299	89.096	89.757	89.531	89.069	89.655	89.066	89.676	88.98
17	**68**	89.878	90.225	89.525	90.213	90.196	89.558	90.23	89.736	90.287	89.458
18	**66**	90.481	90.268	89.934	90.267	90.324	90.526	90.317	90.581	90.353	90.022
19	**64**	90.458	90.619	91	90.597	90.199	91.262	90.689	91.42	89.876	91.223
20	**62**	88.913	89.501	90.19	89.244	89.281	90.143	89.301	90.331	89.29	90.006
21	**60**	90.161	90.345	91.406	90.259	90.218	91.56	90.449	91.447	90.214	91.37
22	**58**	90.896	91.491	93.213	91.581	91.286	93.448	91.661	93.109	91.598	93.278
23	**56**	92.504	92.633	94.186	92.907	92.992	94.048	93.006	94.153	92.887	93.903

Appendix 2. The accuracy of model CPT on reduced datasets

No.	%	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	**100**	**47.971**	**48.06**	**48.004**	**47.911**	**47.807**	**48.033**	**47.873**	**48.165**	**49.009**	**48.052**
2	**98**	48.136	48.086	47.95	47.889	47.862	48.119	48.007	48.155	48.061	48.049
3	**96**	48.103	48.037	47.988	47.819	47.857	48.082	47.983	48.147	47.972	48.081
4	**94**	48.017	47.979	47.915	47.808	47.807	47.976	47.933	48.131	47.884	48.01
5	**92**	48.05	47.873	47.885	47.816	47.798	48.054	47.915	48.116	47.998	48.04
6	**90**	48.007	47.897	47.913	47.783	47.784	48.005	47.813	48.071	47.965	47.999
7	**88**	47.941	47.656	47.577	47.569	47.477	47.625	47.508	47.765	47.677	47.511
8	**86**	47.401	47.143	46.774	46.937	46.938	46.853	46.964	46.986	47.141	46.806
9	**84**	46.971	46.594	45.902	46.444	46.282	46.032	46.43	46.148	46.602	45.914
10	**82**	46.26	45.992	45.065	45.812	45.64	45.244	45.813	45.395	47.141	45.048
11	**80**	45.743	45.313	44.204	45.198	44.924	44.291	45.146	44.441	45.328	44.179
12	**78**	45.168	44.53	43.167	44.377	44.325	43.3	44.581	43.46	44.651	43.352
13	**76**	44.455	43.961	42.202	43.729	43.607	42.426	43.826	42.427	43.93	42.383
14	**74**	43.757	43.184	41.243	42.981	42.659	41.358	43.111	41.276	43.107	41.289
15	**72**	42.939	42.448	40.576	42.205	42.114	40.676	42.237	40.619	42.571	40.614
16	**70**	42.217	41.645	40.219	41.442	41.497	40.425	41.367	40.244	41.643	40.354
17	**68**	41.281	41.075	39.792	40.999	40.958	40.051	40.804	39.857	41.236	40.057
18	**66**	40.913	40.885	39.47	40.628	40.634	39.772	40.437	39.554	40.795	39.726
19	**64**	40.479	40.365	39.18	40.266	40.276	39.467	40.119	39.302	40.498	39.225
20	**62**	39.95	39.913	38.789	39.994	39.731	39.088	36.628	38.902	40.013	38.883
21	**60**	39.543	39.56	38.385	39.404	39.279	38.645	39.343	38.513	39.553	38.622
22	**58**	39.156	38.942	37.944	38.947	34.848	38.133	38.833	38.194	39.014	38.045
23	**56**	38.337	38.503	37.113	38.397	38.363	37.373	38.359	37.313	38.561	37.395

Appendix 3. The accuracy of model Markov1 on reduced datasets

No.	%	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	**100**	**65.924**	**65.934**	**66.004**	**65.737**	**66.04**	**66.115**	**65.77**	**65.864**	**65.997**	**65.93**
2	**98**	66.189	66.243	66.538	66.263	66.417	66.502	66.229	66.236	66.428	66.335
3	**96**	66.841	66.744	67.061	66.714	66.803	67.031	66.587	66.671	66.811	66.722
4	**94**	67.282	67.037	67.442	67.248	67.262	67.463	67.252	67.292	67.379	67.466
5	**92**	67.905	67.892	68.002	68.007	67.941	68.082	67.903	67.794	68.283	68.046
6	**90**	68.623	68.381	68.704	68.556	68.582	68.587	68.736	68.387	68.67	68.576
7	**88**	69.23	69.18	69.161	69.264	69.244	69.379	79.289	69.704	69.198	69.167
8	**86**	69.888	69.952	69.849	70.278	69.804	70.228	70.093	70.276	70.619	69.777
9	**84**	70.68	70.931	72.02	70.966	70.324	71.034	70.969	70.836	70.884	70.609
10	**82**	71.848	71.885	71.8	71.911	71.704	71.659	71.729	71.625	71.752	71.367
11	**80**	72.822	72.696	72.598	72.754	72.448	72.674	72.499	72.483	72.635	72.307

Appendix 3. The accuracy of model Markov1 on reduced datasets *(continue)*

No.	%	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
12	**78**	73.399	73.603	73.32	73.654	73.565	73.478	73.328	73.625	73.654	73.322
13	**76**	74.146	74.548	74.331	74.641	74.577	74.39	74.463	74.648	74.799	74.468
14	**74**	75.188	75.694	75.243	75.521	75.472	75.361	75.562	75.462	75.794	75.364
15	**72**	76.124	76.84	75.873	76.477	76.593	76.156	76.574	76.165	76.811	76.002
16	**70**	77.104	77.65	76.747	77.771	77.536	76.998	77.476	76.887	77.77	76.844
17	**68**	77.848	78.579	77.554	78.375	78.329	77.718	78.292	77.856	78.646	77.599
18	**66**	78.727	79.258	78.457	79.112	78.991	78.94	78.924	78.804	79.346	78.62
19	**64**	75.587	79.831	79.69	79.743	79.784	80.016	79.861	79.876	71.241	79.608
20	**62**	80.523	80.561	81.052	80.637	80.548	81.185	80.619	81.042	80.782	80.834
21	**60**	81.457	81.467	82.372	81.656	81.447	82.408	81.606	82.53	81.645	82.4
22	**58**	82.588	82.506	83.921	82.596	82.486	84.188	82.669	84.065	73.708	84.019
23	**56**	83.534	83.485	75.681	83.639	83.844	84.483	83.848	84.578	83.774	84.461

Appendix 4. The accuracy of model AKOM on reduced datasets

No.	%	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	**100**	**59.185**	**59.346**	**59.538**	**59.357**	**59.612**	**59.542**	**59.397**	**59.546**	**59.415**	**59.568**
2	**98**	59.604	59.648	59.882	59.753	59.922	59.754	59.729	59.962	59.739	59.733
3	**96**	60.071	60.094	60.399	60.097	60.205	60.346	60.048	60.151	60.076	60.22
4	**94**	60.375	60.299	60.707	60.583	60.665	60.592	60.513	60.808	60.689	60.678
5	**92**	61.064	61.089	61.123	61.167	61.122	61.157	61.092	61.199	61.274	61.158
6	**90**	61.6	61.442	61.716	61.553	61.686	61.571	61.568	61.514	61.617	61.502
7	**88**	62.082	62.066	62.108	62.181	62.165	62.153	61.974	62.317	61.944	62.239
8	**86**	62.485	62.649	62.642	62.815	62.722	62.832	62.822	62.977	62.58	62.5
9	**84**	63.413	63.4	63.37	63.429	63.157	63.411	63.441	63.529	63.114	63.266
10	**82**	64.092	64.206	64.081	63.969	64.147	63.968	64.07	64.267	64.184	63.842
11	**80**	64.808	65.013	64.754	64.708	64.711	64.882	64.862	64.885	64.942	64.729
12	**78**	65.414	67.713	65.591	65.507	65.489	65.565	65.567	65.837	65.535	65.586
13	**76**	66.011	66.489	66.295	66.528	66.356	66.43	66.553	66.733	66.777	66.61
14	**74**	66.895	67.211	67.311	67.306	67.234	67.202	67.374	67.391	67.486	67.418
15	**72**	67.778	68.477	68.021	68.213	68.278	67.823	68.316	68.056	68.456	68.046
16	**70**	68.465	69.222	68.609	69.352	69.242	68.323	69.132	76.887	69.37	68.725
17	**68**	69.304	69.86	62.262	69.891	69.885	69.047	69.852	69.397	70.082	69.293
18	**66**	70.042	70.392	69.943	70.603	70.376	70.187	70.423	70.179	70.705	70.075
19	**64**	70.727	70.964	71.004	70.916	71.238	70.869	70.942	71.225	71.214	71.106
20	**62**	71.496	71.544	72.261	71.91	71.719	72.022	71.751	72.33	72.001	72.06
21	**60**	72.363	72.379	73.412	72.724	72.624	73.459	72.464	73.557	72.691	73.53
22	**58**	73.331	73.767	74.001	73.658	73.515	75.005	73.495	74.901	73.708	75.039
23	**56**	74.334	74.289	75.502	74.808	74.761	75.579	74.601	75.572	74.713	75.699

Appendix 5. The accuracy of model LZ78 on reduced datasets

No.	%	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	**100**	**59.805**	**59.89**	**60.14**	**59.744**	**60.027**	**59.745**	**60.134**	**59.739**	**59.962**	**60.266**
2	**98**	60.166	60.35	60.424	60.259	60.243	60.466	60.386	60.235	60.431	60.23
3	**96**	60.386	60.491	60.72	60.546	60.665	60.834	60.403	60.502	60.949	60.799
4	**94**	60.989	60.618	60.682	61.02	60.901	61.151	60.959	61.06	60.812	61.026
5	**92**	61.306	61.364	61.566	61.474	61.201	61.472	61.272	61.433	61.619	61.439
6	**90**	61.777	61.854	61.71	61.803	61.931	61.692	61.873	61.878	61.617	61.972
7	**88**	62.424	61.931	62.412	62.308	62.291	62.599	61.783	62.201	62.293	62.425
8	**86**	62.646	62.649	62.609	62.463	62.824	62.889	62.702	62.856	62.881	62.554
9	**84**	63.674	63.371	63.101	63.687	63.413	62.962	63.295	62.815	63.114	63.22
10	**82**	64.439	64.256	63.48	64.117	64.043	63.774	64.218	63.947	63.98	63.763
11	**80**	64.724	64.89	64.332	64.634	64.6	64.356	64.475	63.947	64.811	64.255
12	**78**	65.239	65.556	64.836	65.281	65.295	64.459	65.197	64.77	65.567	64.745
13	**76**	66.156	66.354	65.634	66.083	66.184	65.179	66.25	65.367	66.457	65.301
14	**74**	66.566	66.986	66.011	66.843	66.735	66.115	66.785	65.936	67.056	65.826
15	**72**	67.526	68.008	66.409	67.668	67.685	66.534	67.806	66.408	67.771	66.785
16	**70**	68.134	68.272	67.42	68.7	68.74	67.533	68.113	67.221	68.796	67.436
17	**68**	68.78	69.168	68.132	69.241	69.363	68.263	69.278	68.061	69.292	68.149
18	**66**	69.342	69.941	68.896	69.821	69.558	69.263	69.463	68.948	70.05	69.143
19	**64**	70.191	70.649	70.27	70.63	70.524	70.349	70.388	70.255	70.478	70.376
20	**62**	70.988	71.286	71.259	71.288	70.83	71.367	71.208	71.33	71.334	71.297
21	**60**	71.678	71.992	72.705	72.168	71.722	75.572	71.814	72.623	72.301	72.771
22	**58**	72.449	72.767	74.832	73.048	72.645	74.104	72.746	73.943	73.025	74.16
23	**56**	73.571	73.454	74.68	73.808	73.952	74.245	73.729	74.221	73.757	74.573

Appendix 6. The accuracy of model CPT+ on reduced datasets

No.	%	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	**100**	**99.927**	**99.941**	**99.936**	**99.924**	**99.938**	**99.938**	**99.941**	**99.934**	**99.936**	**99.947**
2	**98**	99.921	99.958	99.937	99.925	99.941	99.951	99.951	99.942	99.935	99.946
3	**96**	99.931	99.97	99.944	99.933	99.946	99.954	99.952	99.956	99.949	99.947
4	**94**	99.937	99.964	99.949	99.943	99.954	99.966	99.959	99.957	99.959	99.957
5	**92**	99.94	99.977	99.955	99.959	99.966	99.966	99.959	99.966	99.957	99.959
6	**90**	99.956	99.976	99.967	99.962	99.964	99.967	99.956	99.964	99.962	99.964
7	**88**	99.954	99.978	99.969	99.969	99.963	99.974	99.969	99.971	99.965	99.972
8	**86**	99.96	99.979	99.974	99.97	99.966	99.979	99.981	99.962	99.97	99.966
9	**84**	99.967	99.979	99.979	99.975	99.969	99.977	99.979	99.961	99.967	99.971
10	**82**	99.976	99.98	99.974	99.982	99.974	99.978	99.986	99.966	99.972	99.97
11	**80**	99.973	99.984	99.982	99.977	99.997	99.978	99.988	99.969	99.973	99.84
12	**78**	99.977	99.989	99.985	99.981	99.981	99.99	99.985	99.983	99.968	99.981
13	**76**	99.981	99.985	99.987	99.987	99.974	99.996	99.983	99.987	99.978	99.987
14	**74**	99.982	99.998	99.998	99.991	99.984	99.998	99.989	99.991	99.978	99.993
15	**72**	99.986	99.993	100	99.991	99.991	99.998	99.988	100	99.984	100
16	**70**	99.988	99.998	99.998	99.998	99.998	99.998	99.998	99.998	99.995	100
17	**68**	99.998	100	99.998	99.998	100	99.995	99.998	99.993	99.998	100
18	**66**	100	100	99.995	99.997	100	99.997	100	99.997	99.997	100
19	**64**	99.997	100	99.954	99.997	100	99.938	100	99.997	99.997	99.957
20	**62**	100	99.997	99.941	99.997	99.997	99.955	100	99.947	99.997	99.93
21	**60**	99.997	99.866	99.748	99.994	99.824	99.751	99.997	99.718	99.675	99.76
22	**58**	**99.636**	99.613	99.84	99.581	99.543	99.892	99.693	99.718	99.515	99.844
23	**56**	99.634	99.61	99.81	99.571	99.542	99.89	99.693	99.716	99.511	99.835

ACKNOWLEDGEMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2019-34-06.

REFERENCES

[1] R. Sun and C. L. Giles, "Sequence learning: From recognition and prediction to sequential decision making," *IEEE Intelligent Systems*, vol. 16, no. 4, pp. 67-70, 2001.

[2] P. Fournier-Viger, J. C.-W. Lin, R. U. Kiran, Y. S. Koh, and R. Thomas, "A survey of sequential pattern mining," *Data Science and Pattern Recognition*, vol. 1, no. 1, pp. 54-77, 2017.

[3] T. Gueniche, P. Fournier-Viger, and V. S. Tseng, "Compact prediction tree: A lossless model for accurate sequence prediction," in *International Conference on Advanced Data Mining and Applications*, vol. 8347, pp. 177-188, 2013

[4] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, vol. 1215, 1994, pp. 487-499.

[5] P. Fournier-Viger, T. Gueniche, S. Zida, and V. S. Tseng, "ERMiner: sequential rule mining using equivalence classes," in *International Symposium on Intelligent Data Analysis*, vol. 8819, pp. 108-119, 2014.

[6] P. Fournier-Viger, R. Nkambou, and V. S.-M. Tseng, "RuleGrowth: mining sequential rules common to several sequences by pattern-growth," in *Proceedings of the 2011 ACM symposium on applied computing*, 2011, pp. 956-961.

[7] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E. M. Nguifo, "CMRules: Mining sequential rules common to several sequences," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 63-76, 2012.

[8] M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," *Machine learning*, vol. 42, no. 1-2, pp. 31-60, 2001.

[9] P. Fournier-Viger, C.-W. Wu, V. S. Tseng, and R. Nkambou, "Mining sequential rules common to several sequences with the window size constraint," in *Canadian Conference on Artificial Intelligence*, vol. 7310, 2012, pp. 299-304.

[10] V. N. Padmanabhan and J. C. Mogul, "Using predictive prefetching to improve World Wide Web latency," *ACM SIGCOMM Computer Communication Review*, vol. 26, no. 3, pp. 22-36, 1996.

[11] J. Griffioen and R. Appleton, "Reducing File System Latency using a Predictive Approach," *USTC'94: Proceedings of the USENIX Summer 1994 Technical Conference on USENIX Summer 1994 Technical Conference*, vol. 1, 1994, pp. 197-207.

[12] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE transactions on Information Theory*, vol. 24, no. 5, pp. 530-536, 1978.

[13] J. Pitkow and P. Pirolli, "Mininglongestrepeatin g subsequencestopredict worldwidewebsurfing," *Proceedings of USITS' 99: The 2nd USENIX Symposium on Internet Technologies & Systems*, 1999, pp. 1-13.

[14] J. Cleary and I. Witten, "Data compression using adaptive coding and partial string matching," *IEEE transactions on Communications*, vol. 32, no. 4, pp. 396-402, 1984.

[15] P. Laird and R. Saul, "Discrete sequence prediction and its applications," *Machine learning*, vol. 15, no. 1, pp. 43-68, 1994.

[16] D. F. Gleich, "PageRank beyond the Web," *Siam Review*, vol. 57, no. 3, pp. 321-363, 2015.

[17] 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.

[18] B. D. Gunel and P. Senkul, "Investigating the effect of duration page size and frequency on next page recommendation with page rank algorithm," in *Proceedings of the Fifth ACM Web Search and Data Mining Conference*, 2012.

[19] J.-H. Su, B.-W. Wang, and V. S. Tseng, "Effective ranking and recommendation on web page retrieval by integrating association mining and PageRank," *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, NSW, 2008, pp. 455-458.

[20] P. Thwe, "Proposed approach for web page access prediction using popularity and similarity based page rank algorithm," *International Journal of Scientific & Technology Research*, vol. 2, no. 3, pp. 240-246, 2013.

[21] N. T. Da, T. Hanh, and P. H. Duy, "Improving Webpage Access Predictions Based on Sequence Prediction and Pagerank Algorithm," *Interdisciplinary Journal of Information, Knowledge & Management*, vol. 14, 2019.

[22] B. Nigam, S. Tokekar, and S. Jain, "Predicting the next accessed web page using Markov model and PageRank," *International Journal of Data Mining and Emerging Technologies*, vol. 3, no. 2, pp. 73-80, 2013.

[23] M. Vazirgiannis, D. Drosos, P. Senellart, and A. Vlachou, "Web page rank prediction with Markov models," in *WWW '08: Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 1075-1076.

[24] P. Thwe, "Using Markov Model and Popularity and Similarity Based PageRank Algorithm for Web Page Access Prediction," in *International Conference on Advances in Engineering and Technology (ICATE)*, 2014.

[25] M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis, "Web path recommendations based on page ranking and markov models," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*, 2005, pp. 2-9.

[26] T. Gueniche, P. Fournier-Viger, R. Raman, and V. S. Tseng, "CPT+: Decreasing the time/space complexity of the Compact Prediction Tree," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2015, pp. 625-636.

[27] N. T. Da, T. Hanh, and P. H. Duy, "A Survey of Webpage Access Prediction," *2018 International Conference on Advanced Technologies for Communications (ATC)*, Ho Chi Minh City, 2018, pp. 315-320.

[28] S. Das, “Time series analysis. Princeton university press,” *Princeton, NJ*, 1994.

[29] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010.

[30] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, vol. 2, 1995, pp. 1137-1145.

[31] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An introduction to statistical learning,” *Springer*, 2013.

[32] M. Kuhn and K. Johnson, “Applied predictive modeling,” *Springer*, 2013.

BIOGRAPHIES OF AUTHORS

Nguyen Thon Da is a final-year Ph.D. student in Information Systems at Posts and Telecommunications Institute of Technology, Vietnam, in the year 2020. He is currently a lecturer in the Faculty of Information Systems, University of Economics and Law, VNU-HCM, Vietnam. His research interests are Data Mining, Data Analytics, Big Data, and Sequence Prediction. https://orcid.org/0000-0002-2660-5011 (ORCID ID)

Tan Hanh received a PhD degree in Informatics from Grenoble INP, France, in 2009. Currently, he is vice president of the Posts and Telecommunications Institute of Technology. His research interests are Distributed Systems, Machine Learning, Information Retrieval, and Data Mining. https://orcid.org/0000-0002-8244-7046 (ORCID ID)