Evaluating the Performance of Navigation Prediction Model Based on Varied Session Length

Honey Jindal(✉) and Neetu Sardana

Computer Science and Engineering, Jaypee Institute of Information Technology,
Noida, India
honey.cs0990@gmail.com, neetu.sardana@jiit.com

Abstract. Web navigation prediction plays a vital role in web, due to its broad research applications. It can be used for personalization, improvise website design, and business intelligence. Main aim of these applications is to enhance user's satisfaction levels who are visiting the website. Web navigation prediction model tries to predict the future set of the webpage from their historical navigations. The past navigations are collected in the web server log file. Navigations form the sessions of varied length which are used for building the navigation model. Selecting very long sessions or very small sessions degrades the model performance. Thus, selecting an optimal session length is mandate as it would impact the model performance positively. This paper presents pre-investigation measures like page loss, branching factor and session length. We investigate the performance of prediction model based on two different ranges of session length. First range that has been considered is three to seven (3 to 7) and second range is two to ten (2 to 10). The Model has been evaluated on three real datasets. The experimental results show that selecting session of length ranging from 2 to 10 gives better learning hence intensifies accuracy of navigation prediction model. The model accuracy of Set B showed improvement from 0.27 to 8.73% in MSWEB, 0.62 to 2.8% in BMS and 10.81 to 14.23% in Wikispeedia dataset.

Keywords: Prediction · Session length · N-gram · Navigation · Web · Markov

1 Introduction

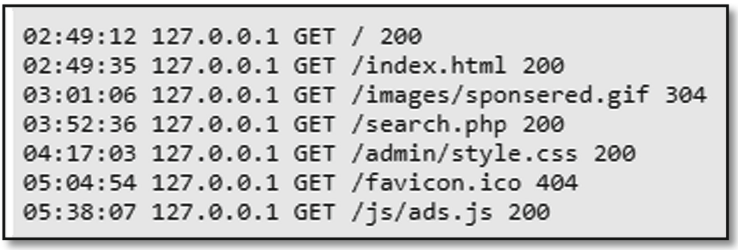
The continuous growth of web is resulting in enormous websites. The structure of websites is also becoming complex. Often users face difficulty in locating the desired information while navigating through the website. With the website designer perspective, the main challenge is to analyze the user behavior and personalize them. This will not only help them in locating required information but also improve user's satisfaction level.

Web Navigation Prediction (WNP) is an emerging research area to address these issues. In WNP, a model is trained such that it predicts the next web page(s) from the visited web pages. WNP can be generalized and applied on different applications [14] like search engines [16], caching systems and latency reduction [17], anomaly detection

P. K. Singh et al. (Eds.): FTNCT 2019, CCIS 1206, pp. 431–443, 2020.
https://doi.org/10.1007/978-981-15-4451-4_34

[8], personalization [5], website design [18], detecting malicious web pages [26, 27], recommendation systems [1], event detection [32] and location prediction [9, 28].

User navigation history is captured in the log file through cookies or web servers. A snapshot of weblog file shown in Fig. 1. The fields of web logs are user IP address, user authentication, date/time, action, return code, size, referral, browser/platform. Each row in the log file [10] represents single web page request. It consist important information about the client and the requested web page. The information is recorded by the server to understand user behavior.



```
02:49:12 127.0.0.1 GET / 200
02:49:35 127.0.0.1 GET /index.html 200
03:01:06 127.0.0.1 GET /images/sponsered.gif 304
03:52:36 127.0.0.1 GET /search.php 200
04:17:03 127.0.0.1 GET /admin/style.css 200
05:04:54 127.0.0.1 GET /favicon.ico 404
05:38:07 127.0.0.1 GET /js/ads.js 200
```

Fig. 1. Web log file [10]

Web logs are preprocessed and sessions are constructed from the log file which is used for making prediction model. Session consisting set of pages can be of varied length. Longer sessions often have noise as they may be repetition of pages or user is following longer path to reach desired page. This results in poor browsing experience of the user and may harm the popularity of the website. According to Janrain [15], about 74% of the online users get frustrated with website when they do not get their required web content. According to Forrester research [11], a good website design can attract more user's and vice versa. Half of the sales will get negative impact, if user is unable to locate his desired information. Due to the negative experience faced by the users on their first visit, 40% users may not return to the website. In 2013, a Monetate/eConsultancy study [15] found that in-house marketers who are personalizing their browsing experience observed 19% uplift in their sales. Smaller sessions will dilute the learning of prediction model so it is important to use the session length that can help in building the prediction model optimally.

This paper analyses the performance of prediction model based on two different ranges of session length. The two set of ranges are Set A (3 to 7) and Set B (2 to 10). Generally Set A has been used in past studies [4]. We will compare this range with longer session length range two to ten (2–10) to find suitable session length for model building. This study analyzes the impact of varied session length on prediction model.

1.1 Research Objectives

1. This paper highlights pre-investigations measures which are required to inject good quality inputs to the training model.

2. Web navigations have been analysed and detail summary of how pre-investigations will affect the model is discussed.
3. We have evaluated model performance using varied session length on three real datasets(MSWEB, BMS and Wikispeedia)

The rest of the paper is organized as follows. Section 2 gives preliminaries and model representation. Related work is presented in the Sect. 3. Experimental details are described in Sect. 4 and conclude the paper in Sect. 5.

2 Preliminaries

This section describes the basic terminology, representation and modeling of a session.

- ***Sessions:*** A Session represents the web page(s) visit order of the user during the website navigation. A session, S is represented as {P1, P2, …, Pn} where n denotes the number of pages. Each user browsing history is stored in a session.
- ***N-grams:*** In WNP, N-gram is prominently used to represent the training model. The N-gram can be represented as <p_1, p_2, ….,p_N>. This depicts sequences of web page(s) navigation of the user's. Each web page is represented with unique page id. For example, if we consider session consisting six pages having session length as six, S = <P11, P22, P5, P13, P20, P8>. In the given example, 1-gram will contain five sessions <P11, P22>, <P22, P5>, <P5, P13>, <P13, P20>, <P20, P8> and 2-gram will contain four sessions <P11, P22, P5>, <P22, P5, P13>, <P5, P13, P20>, <P13, P20, P8>. N-gram is a fixed length representation of sessions. Due to the fixed length representation of the training set, the model complexity, state-space complexity, computational complexity required to build the model can be easily determined.
- ***Markov Model (MM):*** Markov model [2, 3, 12, 13] is the well known representation used for the WNP. User navigation behavior is captured in the log file and analyzed to predict the next desired information. The log file is pre-processed to find the sessions. These sessions are used as the input for modeling the Markov model. MM is the graphical representation of sessions. Each node is represented by the pages and links between them represents the transition probability to move from one state to another. Markov models can be formed in varied order. In first-order MM, each state is represented with single page. For instance, a link between state A and B is formed using the transition probability. The transition probability is defined as the ratio of number of times <A, B> occurs to the number of times <A> occurs.

Transition probability to move from A to B is given by,

$$P(A \rightarrow B) = \frac{\mu(\mathrm{A,B})}{\mu(\mathrm{A})} \quad \text{where, } \mu \text{ denotes frequency}$$

In the second-order MM, each state is represented with two pages. For instance, a link between state <A, B> and <C> is formed using the transition probability.

The transition probability is defined as the ratio of number of times <A, B, C> occurs to the number of times <A, B> occurs.

Transition probability to move from <A, B> to C is given by,

$$P((A,B) \rightarrow C) = \frac{\mu(\mathrm{A,B,C})}{\mu(\mathrm{A,B})} \quad \text{where}\ \mu\ \text{denotes frequency}$$

Similarly, higher order MM can be formed. In a Kth-order MM, each state is represented by K web pages. Since, the accuracy of Kth-order MM is low, All-Kth Markov model (KMM) comes into existence. In KMM, all lower order models are nested inside the higher order model. If a higher order KMM fail to predict then the search begins in the next subsequent lower order model.

- ***All-Kth Modified Markov Model (KMMM):*** The accuracy of MM is very low. Therefore, Modified Markov model (MMM) is proposed by Mamoun et al. [2]. In this model order of the pages does not matter. For example, if the sessions have same set of pages then they are represented in the same state. In order to further enhance the performance of MMM, all-Kth model are embedded with it. This model is known as All-Kth Modified Markov Model (KMMM). Jindal et al. [7] and Mamoun et al. [2] analyzed that All-Kth Modified Markov Model (KMMM) is proved to the compressed and effective prediction model. Therefore, in this work we choose KMMM as a prediction model to evaluate the performance over varied session length.

3 Related Work

During website browsing, user navigation history is captured in the web log file. The web log file cannot be used directly for analysis sand prediction as it consists of lot of noisy information like image, video, audio and robotics files. Thus, these log files are cleaned and pre-processed. During this phase the noisy information is filtered and user's as well as sessions are identified. Sessions are the sequence of the navigation trails of the users. Users' are identified using their IP address.

In past several sessions generation techniques were found which attempts to obtain relevant patterns from the web log file. Broadly, three session generation techniques have been used in the past namely, time-based, navigation-based and integer programming.

- Time-based: Catledge et al. [19] and Cooley et al. [21] have used page-stay time and session duration thresholds. Zhang et al. [20] proposed dynamic time-oriented method. The sub sessions are formed from the session when their time exceeds from the respective thresholds. Time-oriented heuristics do not consider website structure, thus most of the useful navigation patterns are missed in the session generation. Session generated may have duplicate web pages in the same session. For example {P2, P1, P1, P7, P3} or {P2, P1, P7, P1, P3} are allowed in the time-based heuristics. Here, {P1, P1} or {P1, P7, P1} causes unnecessary duplication of web page P1 that makes sessions longer. Moreover, these heuristics are not reliable as user(s) might get involved in some other activities during web page navigation. Other factors like web page content, content size, web page components, busy communication line may impact the session formation.

- Navigation-based: Cooley et al. [22, 23] have proposed navigation-based graphical structure of web sessions. In this network, nodes are represented by the web pages and edges are represented by the direct link between the web pages. For each navigated session, if there is no connection found between the two consecutive web pages then backward browsed webpage is inserted. This artificial insertion generates longer sessions.
- Integer programming: Dell et al. [24, 25] proposed integer programming based session generation techniques. Herein, web sessions are partitioned into the chunks using IP and agent information through logarithmic objective function. This objective function assigns the each web page to the chunk of the particular session such that there is no duplicate web page found in the session. For example, the given session is {P1, P3, P6, P3, P6, P8, P7, P6, P6, P8, P10} there is actually no link present between page P7 and P10. In this approach the session will split into two subsessions as {P1, P3, P6, P8, P7} and {P10}. However, according to website topology, the correct subsessions should consist of {P1, P3, P6, P8, P7} and {P1, P3, P6, P8, P10. In addition, the obtained subsession with web page P10 have no correlation with other web pages which is not correct.

Session generation techniques presented varied session identification methods but they do not focused on deriving optimal session length. West et al. [29] observed that session length defines the user navigation behavior. Shorter path means user step towards the right direction and longer path means user did not get the right path. He might be circling around the desired page. In addition, longer path requires more state-space complexity and high computational cost [30]. It makes the prediction model development cumbersome [30] and degrades model performance. Since, the success of pattern discovery depends on the quality of input session injected to it [31], we have evaluated the impact of varied session lengths on web navigation prediction model. The paper discusses the pre-investigation measures that required to be performed before generating a prediction model. The pre-investigations are required mainly to choose the optimal session length for web navigation prediction as the quality of prediction accuracy depends upon the input sessions. To the best of our knowledge, no work has been done in past that inquires the optimal session length for web navigation prediction. Although logs are generated, cleaned and later used for prediction in so many application areas that we have mentioned in the paper but none of them have discussed the session length to be important component which need attention.

4 Experimental Details

Selecting an optimal Session length is a major concern before developing the prediction model. This is because the accuracy of the model depends on the sessions taken as an input. This study main focus is to analyze the effect of session length over prediction model. This section presents the experimental details like the dataset used, pre-investigation measures, evaluative parameters and the results obtained. Therefore, we analyses the performance of prediction model based on two different ranges of session length. We perform experiment on two sets. Set A consists sessions whose length lies in between 3 to 7. Set B consists sessions whose length lies in between 2 to 10. Generally Set A has been used in past studies [4]. We will compare this range with Set B. The training and testing for both sets is divided in the ratio of 0.7 and 0.3.

4.1 Dataset Description

We have conducted experiments on three datasets: MSWEB, BMS and Wikispeedia. The detail characteristics of each dataset have been presented in Table 1.

Table 1. Dataset summary

Dataset & Year	Source	Application	#Sessions	#Unique pages	Avg. session length
MSWEB (1999)	www.microsoft.com	Microsoft website	38000	294	3.01
BMS (2000)	www.gazelle.com	E-Commerce	59601	497	2.42
Wikispeedia (2009)	www.snap.stanford.edu	Wiki Pages	51000	3326	5.5

- *Dataset 1: MSWEB*

This dataset was collected from the Microsoft logs. The data consists of 38000 sessions from random users in February, 1998. Each row represents sequence of areas of the website that the user visits in a period of one week.

- *Dataset 2: BMS*

This dataset was collected from e-commerce web server logs (Gazelle.com) and used as a part of the KDD Cup 2000 competition. It contains 59,601 web sessions of items and 497 distinct items. The average length of the sessions is 2.42 items.

- *Dataset 3: Wikispeedia*

This is a popular online web page game. In this, each player has given a task to find a shortest path from source to destination web page. The player navigates from source web page to destination web page using the hyperlinks. The player has no knowledge of the global network structure. Therefore, he uses local information provided on the webpages. The player's navigations were collected in the web log file which consists 4606 articles and 3326 distinct articles. It comprises 51 K navigations collected over 2009.

The details of training and testing sessions are summarized in Table 2. After 0.7 (training) and 0.3 (testing) split, sessions are further divided categorized into N-gram using sliding window concept. It has been clearly observed that training and testing sessions of Set B is more as compared to Set A sessions. This is because Set B is a superset of Set A.

Table 2. Training and testing dataset

N-Gram	MSWEB (Set A)		MSWEB (Set B)		BMS (Set A)		BMS (Set B)		Wikispeedia (Set A)		Wikispeedia (Set B)	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
1	22670	1918	58278	3351	19211	4046	43114	7592	92388	14612	152104	13818
2	27670	1918	36172	2111	19211	4046	25402	4503	69468	14612	119879	13584
3	15258	1167	22499	1360	10187	2423	15462	2880	47152	13277	88218	12590
4	7585	680	13565	873	4919	1268	9278	1725	26925	9895	58988	10009
5	3176	362	7895	555	2087	667	5530	1124	11874	5615	35735	6675
6	896	154	4089	347	626	282	2946	739	3484	2384	20089	4254

4.2 Pre-investigation Measures

Pre-investigation measures are the metrics which is used to measure the effectiveness of input data. Measuring quality of data is very important before developing a model. A good quality input data injected to the model will produce better results. This section presents two pre-investigation measures: Page Loss and Branching Factor.

(a) **Page Loss**

Page loss determines the missing percentage of the pages in the training model. It defines as the ratio of number of web pages missing in the dataset to the total number of web pages of the website. The page loss would yield unseen pages and will not generate predictions for such pages. It also impacts models negatively. For example, if a web page P9 occurs in a test dataset which was not available in the training model; then training model will fail to generate predictions for web page 9. This measure is important to understand model incapability before model development phase. Table 3 depicts page loss of set A and set B training model. While investigating the datasets, we have found that some web pages were lost while dividing the dataset into training and testing. It has been observed that page loss is less in set B as compared to set A. Since, Set B has long session range; it produces more subset of sessions in the dataset with large combination of pages. Addressing this page loss is important, because it will give rise to more cold-start pages and cold-start sessions.

Table 3. Training page loss

Page loss	MSWEB	BMS	Wikispeedia
Set A	37 (12.58%)	112 (22.53%)	1057 (22.94%)
Set B	16 (5.44%)	110 (22.13%)	438 (9.50%)

(b) **Branching Factor**

Branching factor measures the network characteristics of the model. It is defined as the average number of outlinks present in the model corresponding to each state. Branching factor determines model prediction capability. It gives network structure insights which is helpful to understand "how much predictions a model may generate corresponding to its current state". This pre-investigation measure is important to compute average outlink percentage of the network states. Table 4 presents branching factor of Set A and B on varied N-grams. It has been found that branching factor of Set B is higher in all the datasets. This is because Set B injects more sessions in the training model which will have more outlinks corresponding to each state.

Table 4. Branching factor

N-Gram	MSWEB (Set A)	MSWEB (Set B)	BMS (Set A)	BMS (Set B)	Wikispeedia (Set A)	Wikispeedia (Set B)
1	17.23	19.99	16.36	26.57	9.40	10.05
2	3.76	4.205	2.93	3.549	3.25	3.47
3	2.32	2.430	2.02	2.221	2.01	2.04
4	1.92	1.927	1.78	1.858	1.73	1.79
5	1.79	1.731	1.70	1.716	1.64	1.67
6	1.76	1.655	1.68	1.658	1.62	1.51

4.3 Evaluation Parameters

In this section, we will define some prediction parameters used to evaluate model performance [6, 7]. The definitions of the predicting parameters are given below:

Definition 1: Prediction Accuracy

Prediction accuracy is defined as the ratio of correct predictions to the total number of test cases.

$$Prediction\ Accuracy = \frac{Correctly\ predicted\ test\ cases}{Total\ test\ cases}$$

Definition 2: Model Accuracy

Model accuracy is defined as the ratio of correct predictions to the total predictions.

$$Model\ Accuracy = \frac{Correctly\ predicted\ test\ cases}{Total\ test\ cases\ matched\ with\ the\ training\ model}$$

Definition 3: Coverage
Coverage is defined as the ratio of total number of predictions to the number of total test cases.

$$Coverage = \frac{Total\ Predictions}{Total\ test\ cases}$$

4.4 Experimental Results

(1) *Coverage*
Coverage is the evaluative measure which defines percentage of outlinks (prediction paths) covered by the test state. The value of coverage is depended on network structure. Table 5 presents coverage of the Set A and B over varied N-grams. It has been found that coverage of set B is more in all the datasets. Since, the branching factor of training models of Set B is higher; the model with Set B covers more outlinks during prediction as compared to model with Set A.

Table 5. Coverage

N-Gram	MSWEB (Set A)	MSWEB (Set B)	BMS (Set A)	BMS (Set B)	Wikispeedia (Set A)	Wikispeedia (Set B)
1	5.40	5.46	7.04	7.04	7.32	7.33
2	6.15	6.37	5.24	5.68	8.32	12.87
3	4.62	5.11	3.72	4.12	6.50	9.21
4	3.70	4.06	3.34	3.87	5.54	8.29
5	3.29	3.55	3.26	3.70	5.06	6.88
6	3.23	3.51	3.07	3.57	4.94	6.58

(2) *Prediction Accuracy*
Table 6 presents the effect of varying the session length on the prediction accuracy of the model. It has been seen clearly that the prediction accuracy decreases as N-gram increases. This is because the number of training examples becomes less as session length increases (N) (see Table 2). We have observed that the prediction accuracy of set B is higher as compared to set A on both datasets. This is because set B has less page loss while having high coverage corresponding to each test example session. Due to more availability of sessions, Set B has more chances to make correct predictions than Set A.

Table 6. Prediction accuracy of Set A and B

N-Gram	MSWEB (Set A)	MSWEB (Set B)	BMS (Set A)	BMS (Set B)	Wikispeedia (Set A)	Wikispeedia (Set B)
1	65.01	73.85	51.26	54.06	43.96	55.93
2	74.03	74.56	41.89	44.60	38.57	52.80
3	63.00	63.09	38.05	40.16	33.54	46.42
4	54.35	54.98	37.77	39.45	33.54	45.98
5	49.10	49.87	38.23	39.73	34.60	45.41
6	47.40	47.83	35.46	36.08	33.97	45.79

(3) Model Accuracy
The difference between model and prediction accuracy is that during the evaluation phase, model accuracy removes unseen test sessions from the total test set. Unseen sessions are those which are not known to the training model.

Model accuracy with respect to varied session length is presented in Table 7. It shows model prediction ability with respect to the test sessions which are available in the training model. It has been observed that Set B has more correct prediction ability than Set A on all datasets. Since, Set B has more outlinks for each state as compared to Set A. It generates more predictions and has more chances to make correct predictions.

Table 7. Model accuracy of Set A and B

N-Gram	MSWEB (Set A)	MSWEB (Set B)	BMS (Set A)	BMS (Set B)	Wikispeedia (Set A)	Wikispeedia (Set B)
1	65.15	73.88	53.31	57.02	44.24	55.93
2	74.28	74.60	44.88	48.08	38.82	52.80
3	62.56	63.13	41.16	43.94	33.76	46.42
4	54.77	55.04	41.87	44.57	33.79	45.98
5	49.67	50.00	42.85	45.94	34.91	45.41
6	47.34	47.98	40.48	42.93	34.35	45.80

4.5 Discussion

From the experiment results, we have inferred that early investigation of the input would yield better predictions. Before making predictions, the optimal split of training and testing dataset and optimal session length should be consider. To investigate the performance of prediction model, two investigation parameters have been used. Page loss indicates the amount of page loss in the training and testing dataset. It is important to consider because it provide insight of cold-start web pages and cold-start sessions or unseen sessions. Presence of unseen sessions makes model difficult to learn and causes prediction failure. Second investigation parameter is the branching factor. This measure is important as it provides insight of the number of predictions possible from the training state. The Set B has less page loss and high branching factor as compared to

Set A which indicates Set B is more preferable. Our experimental results revealed that model trained with Set B attains better coverage, prediction and model accuracies. The experimental results confirm the inference drawn from the pre-investigations measures.

5 Conclusion and Future Work

In this paper, we conduct experiments to evaluate the performance of prediction model over varied session length. For this, we select two set of session length. In set A, session with length 3 to 7 are selected and in Set B sessions with length 2 to 10 are selected. We evaluate the effectiveness of the input sessions injected to the model using two pre-investigation measures: page loss and branching factor. A set which has less page loss and high branching factor should be considered for the predictions.

In addition, we evaluate the performance of the model using evaluative measures over varied N-grams. The measures used in the study are: coverage, prediction accuracy and model accuracy. More crucially, it has been observed that set B has high coverage and high accuracy as compared to set A. It has been found that the session length do impacts the coverage and accuracy of the prediction model. Session length ranging from 2 to 10 is found to be best for development of prediction model. The model accuracy of Set B showed improvement from 0.27 to 8.73% in MSWEB, 0.62 to 2.8% in BMS and 10.81 to 14.23% in Wikispeedia dataset.

In the near future, we plan to do focus domain-centric session evaluation as user browsing behaviour varies with domains. Moreover, other pre-investigations measures can be explored which are required to develop high quality sessions.

References

1. Abrishami, S., Naghibzadeh, M., Jalali, M.: Web page recommendation based on semantic web usage mining. In: Aberer, K., et al. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 393–405. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35386-4_29
2. Awad, M.A., Khalil, I.: Prediction of user's web browsing behavior: application of markov model. IEEE Trans. Syst. Man Cybern. - Part B: Syst. Hum. **42**(4), 1131–1142 (2012)
3. Awad, M.A., Khan, L.R.: Web navigation prediction using multiple evidence combination and domain knowledge. IEEE Trans. Syst. Man Cybern. - Part B: Syst. Hum. **37**(6), 1054–1062 (2007)
4. Madhurai, B.C., Anand, C.J., Ramya, K., Phanidra, M.: Analysis of users' web navigation behaviour using GRPA with variable length Markov chains. Int. J. Data Min. Knowl. Manag. Process **1**(2), 1–20 (2011)
5. Pierrakos, D., Paliouras, G.: Personalizing web directories with the aid of web usage data. IEEE Trans. Knowl. Data Eng. **22**(9), 1331–1344 (2010)
6. Pirolli, P.L.T., Pitkow, J.E.: Distributions of surfers' paths through the World Wide Web: empirical characterizations. World Wide Web **1**(2), 29–34 (1999). https://doi.org/10.1023/A:1019288403823
7. Jindal, H., Sardana, N.: Web navigation prediction using Markov based models: an experimental study. Int. J. Web Eng. Technol. **11**(4), 310–334 (2016)

8. Xie, Y., Tang, S.: Online anomaly detection based on web usage mining. In: 26th IEEE International Parallel and Distributed Processing Symposium Workshops & PhD Forum, pp. 1177–1182 (2012)
9. Xue, A.Y., et al.: Solving the data sparsity problem in destination prediction. The VLDB Journal **24**(2), 219–243 (2015)
10. Web Server Log File Samples - IIS and Apache. http://www.herongyang.com/Windows/Web-Log-File-IIS-Apache-Sample.html. Accessed 18 May 2019
11. Usability. http://www.usability.gov. Accessed Nov 2015
12. Sunil, K., Sanjeev, G., Abhinav, G.: A survey on Markov model. MIT Int. J. Comput. Sci. Inf. Technol. **4**(1), 29–33 (2014)
13. Borges, J., Levene, M.: Evaluating variable-length Markov chain models for analysis of user web navigation sessions. IEEE Trans. Knowl. Data Eng. **19**(4), 441–452 (2007)
14. Facca, F.M., Lanzi, P.L.: Mining interesting knowledge from weblogs: a survey. Data Knowl. Eng. **53**(3), 225–241 (2005)
15. 94% of businesses say personalisation is critical to their success. https://econsultancy.com/94-of-businesses-say-personalisation-is-critical-to-their-success/. Accessed 25 May 2019
16. Daxin, J., et al.: Mining search and browse logs for web search: a survey. ACM Trans. Comput. Logic **4**(4), 1–42 (2013)
17. Praveen, K., et al.: Pre fetching web pages for improving user access latency using integrated web usage mining. In: International Conference on Communication, Control and Intelligent Systems. IEEE (2015)
18. Ketukumar, B., Patel, A.R: Web data mining in e-commerce. Study, analysis, issues and improving business decision making. Ph.D. thesis, Hemchandracharya North Gujarat University, Patan, India (2014)
19. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide Web. Comput. Netw. ISDN Syst. **27**(6), 1065–1073 (1995)
20. Zhang, J., Ghorbani, A.A.: The reconstruction of user sessions from a server log using improved time-oriented heuristics. In: Second Annual Conference on Communication Networks and Services Research Proceedings, pp. 315–322. IEEE (2004)
21. Cooley, R., et al.: Data preparation for mining World Wide Web browsing patterns. Knowl. Inf. Syst. **1**(1), 5–32 (1999)
22. Cooley, R., et al.: Web usage mining: discovery and application of interesting patterns from web data. Ph.D. thesis, Dept. of Computer Science, University of Minnesota (2000)
23. Cooley, R., Tan, P.-N., Srivastava, J.: Discovery of interesting usage patterns from web data. In: Masand, B., Spiliopoulou, M. (eds.) WebKDD 1999. LNCS (LNAI), vol. 1836, pp. 163–182. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-44934-5_10
24. Dell, R.F., et al.: Web user session reconstruction using integer programming. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 385–388. IEEE Computer Society (2008)
25. Dell, R.F., Román, P.E., Velasquez, J.D.: Fast combinatorial algorithm for web user session reconstruction. In: Proceedings of 24th IFIP TC7 Conference, Buenos Aires, Argentina (2009)
26. Lui, J., et al.: A Markov detection tree-based centralized scheme to automatically identify malicious webpages on cloud platforms. IEEE Access **6**, 74025–74038 (2018)
27. Kazemian, H.B., Ahmed, S.: Comparisons of machine learning techniques for detecting malicious webpages. Expert Syst. Appl. **42**(3), 1166–1177 (2015)

28. Ahmad, S., et al.: A stochastic approach towards travel route optimization and recommendation based on users constraints using Markov chain. IEEE Access **7**, 90760–90776 (2019)
29. West, R., Leskovec, J.: Human way finding in information networks. In: Proceedings of the 21st International Conference on World Wide Web, pp. 619–628. ACM (2012)
30. Jindal, H., et al.: Elimination of backward browsing using decomposition and compression for efficient navigation prediction. Int. J. Web Based Commun. **14**(2), 196–223 (2018)
31. Bayir, M.A., Toroslu, I.H., Demirbas, M., Cosar, A.: Discovering better navigation sequences for the session construction problem. Data Knowl. Eng. **73**, 58–72 (2012)
32. Xu, J., et al.: Automatic generation of social event storyboard from image click-through data. IEEE Trans. Circuits Syst. Video Technol. 1–12 (2015, accepted)