Predicting Users' Revisitation Behaviour Based on Web Access Contextual Clusters

by Hapnes Toba, Christopher Starry Jomei, Lotanto Setiawan, Oscar Karnalim, Hui Li

Submission date: 27-Jul-2023 03:38PM (UTC+0700)
Submission ID: 2137494615
File name: visitation_Behaviour_Based_on_Web_Access_Contextual_Clusters.pdf (327.27K)
Word count: 6069
Character count: 33505

Predicting Users' Revisitation Behaviour Based on Web Access Contextual Clusters

Hapnes Toba
Faculty of Information Technology
Maranatha Christian University
Bandung 40164, Indonesia
hapnestoba@it.maranatha.edu

Christopher Starry Jomei
Faculty of Information Technology
Maranatha Christian University
Bandung 40164, Indonesia
christopherstarry@gmail.com

Lotanto Setiawan
Faculty of Information Technology
Maranatha Christian University
Bandung 40164, Indonesia
lo.tanto@yahoo.com

Oscar Karnalim
Faculty of Engineering and Built Environment
University of Newcastle
Callaghan, Australia
oscar.karnalim@uon.edu.au

Hui Li
College of Computer Science and Technology
Guizhou University
Guiyang, China
cse.huili@gzu.edu.cn

***Abstract*—Most modern browsers record all previously visited web pages for future revisitation. However, not all users utilise such feature. One of the reasons is that the records are displayed at once as a single list, which may overwhelm the users. This paper proposes a predictive model to decide whether a web page will be revisited in the future based on a particular visit. The model can be used to filter web records so that only web pages that may be re-visited are presented. According to our evaluation, the model is considerably effective. It can generate 53.195% accuracy when measured with 10-fold cross validation and 95% meaningful topic identification. Further, attributes rooted from the same website' access frequency are the most salient ones for prediction. In addition, contextual similarities based on k-means clustering and cosine similarity (which are used for defining some attributes) are considerably effective.**

Keywords—browsing history, cosine similarity, k-means clustering, multinomial naïve bayes, user log

I. INTRODUCTION

The World Wide Web (WWW) is an ecosystem where individuals can satisfy their information and communication needs [1]. The importance of this ecosystem is growing rapidly as the number of users is extremely high. In 2017 alone, about 3,578 million people accessed the ecosystem [2]. Consequently, there are a lot of research related to the ecosystem, which mainly focuses on Information Retrieval [3]. As the users become more familiar with WWW, they tend to access web pages more frequently [4]. Most modern browsers (such as Google Chrome, Mozilla Firefox, and Microsoft Edge) record all previously visited web pages so that the users can re-visit them easier. However, the records are rarely used based on two reasons [5]. First, the records are not directly visible from the navigation panel. Second, the records are displayed in a way that does not engage the users; numerous web pages are presented at once as a single list, which may overwhelm the users.

For the latter issue, some works have been proposed to solve it. They commonly re-visualise or rank the records so that the users can be engaged easily [4], [6]-[10]. However, to our knowledge, most existing works do not try to directly filter the web pages. Only web pages which may be accessed further are kept on the list. We would argue such filtering mechanism may help the users to find some web pages for revisitation as the number of displayed web pages is reduced. It is important to note that the filtering mechanism can also be used together with the existing works (i.e., revisualising or ranking the records). The web pages are filtered first before passed to the existing works as the inputs. In order to develop the filtering mechanism, this paper proposes a predictive model to decide whether a web page may be re-visited in the future based on a particular visit.

The model is formed using Multinomial Naïve Bayes [11], with twenty-four attributes on board. The first eight attributes are based on the access frequency of the same website while the remaining attributes are based on the access frequency of websites with the same context, *i.e.* the contextual similarity is defined with either K-Means Clustering [11] or cosine similarity [3]. Naïve Bayes is known for its good performance in dealing with many problems, including the new ones like ours. For some cases, it is even more effective than some advanced algorithms such as Artificial Neural Network. We chose the multinomial one as some of the attributes are numeric. Other algorithms are selected as they are often used as the baselines and we plan to set this work as a baseline for other further research.

Our proposed predictive model may also benefit the website owners. If an owner knows that a user may not re-visit their website anymore, they can attract the user to engage again with a kind of discount (for e-commerce website) or Easter egg (*i.e.* unexpected thing that attracts the user's interest). It is true that this kind of event can also be discovered with server log, checking which users have not accessed the website for a long time. However, we would argue that our proposed mechanism is less time-consuming, as the event can be predicted right after the user has visited the website for the last time. Search engines may also benefit from our predictive model. They may exclude some web pages from their suggestions, which make their computation for web page recommendation faster and time-efficient.

II. RELATED WORK

As the need of internet grows rapidly, web user behaviour has been researched in many works while some of them rely on search engine queries. For example, a work in [12] analyses user queries from Excite, a search engine, to gain insight about web user behaviour. Another example is a work in [13] which discovers how user searches images through their queries. Several works are more focused on queries created by a

specific group of people. The group can be either young users [14] or K-12 students [15]. The former group is analysed to quantify any difficulties experienced by young users when interacting with the internet while the latter is analysed to gain an insight about student learning behaviour.

Instead of search engine queries, several research works try to contribute on other technical aspects. First, a work in [16] proposes a combination of web application models with runtime navigation logs for understanding user behaviour further. Second, a work in [17] introduces SMAP-Mine, a data mining algorithm to efficiently predict user behaviour in a mobile web system. Third, a work in [18] develops visual snippet, a compact representation that combines text snippet's and thumbnail's benefits for finding relevant web pages. Fourth, a work in [19] establishes an algorithm for a sparse non-negative matrix factorisation, that is specifically designed to discover web user behaviour. The research on web user behaviour contributes to the growth of other fields. It fasten the research about search engine ranking [20], spam identification [21], user navigation problem [1], the dwell time of web pages [22], and web revisitation.

Web revisitation occurs when an user accesses a previously-visited web page for various reasons [9]. This task can be 15manding as the user may not remember the page's URL and the number of previously-visited web pages (accessed through web browsing history) can be extremely high. Hence, several works try to alleviate the burden by either re-visualising or ranking the content of the browsing history. Revisualising is applied by replacing the conventional list-based browsing history with a more intuitive representation. A work in [4], for instance, provides an interactive visualisation. Another work in [10], proposes a graph-based visualisation of the browsing history based on its sessions.

Ranking is applied with the help of Information Retrieval. A work in [9] combines ranking and clustering mechanism to predict what web pages may be revisited. A work in [8] utilises context and content keywords to suggest web revisitation with the help of relevance feedback. Not all works about web revisitation are about providing better web browsing history. Several works act as a support to the task. A work [23] discovers four web revisitation patterns through a combination of observing web access log and survey. A work in [24] introduces a new metric for web revisitation as a result of considering tabbed browsing.

To the best of our knowledge, most works have not applied a filtering mechanism to the browsing history (keeping only the most potential pages that will b3evisited). We believe that the mechanism may be helpful as the number of web pages is reduced, resulting better revisualising and ranking the content of the browsing history.

III. Methodology

Our work proposes a predictive model to determine the potentiality of a web page to be re-visited. The model can be used to filter user browsing history so that only web pages that would be re-visited are kept. Multinomial Naïve Bayes [11] is used as the predictive model's learning algorithm; Naive Bayes (the core model of Multinomial Naive Bayes) is considerably effective in most dataset due to its independence assumption [25]. The algorithm has been applied on many fields such as pharmaceutical industry [26], nuclear power plant [27] and text analysis (either general as in [25] or academic as in [28]). The Multinomial Naïve Bayes is implemented with by utilizing Weka [29]).

The predictive model relies on twenty-four attributes, which are classified further to three groups. The first group (A1-A8) is derived from the access frequency of the same website. The second group (B1-B8) is derived from the access frequency of websites with the same cluster-based context; wherein the cluster-based context is determined through K-Means Clustering [11]. The third group (C1-C8) is derived from the access frequency of websites with the same cosine-based context; wherein the cosine-based context is defined using cosine similarity [3].

TABLE I. The Attributes of Proposed Predictive Model

ID	Definition
A1	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website are the same
A2	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website are different
A3	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites are the same
A4	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites are different
A5	The proportion of accessed websites with the same name as the currently-accessed one from 24 hours before
A6	The proportion of accessed websites with different name as the currently-accessed one from 24 hours before
A7	The proportion of accessed websites with the same name as the currently-accessed one from 7 days before
A8	The proportion of accessed websites with different name as the currently-accessed one from 7 days before
B1	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share the same cluster-based context
B2	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share different cluster-based context
B3	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share the same cluster-based context
B4	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share different cluster-based context
B5	The proportion of accessed websites with the same cluster-based context as the currently-accessed one from 24 hours before
B6	The proportion of accessed websites with different cluster-based context as the currently-accessed one from 24 hours before
B7	The proportion of accessed websites with the same cluster-based context as the currently-accessed one from 7 days before
B8	The proportion of accessed websites with different cluster-based context as the currently-accessed one from 7 days before
C1	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share the same cosine-based context
C2	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share different cosine-based context
C3	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share the same cosine-based context
C4	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share different cosine-based context
C5	The proportion of accessed websites with the same cosine-based context as the currently-accessed one from 24 hours before
C6	The proportion of accessed websites with different cosine-based context as the currently-accessed one from 24 hours before
C7	The proportion of accessed websites with the same cosine-based context as the currently-accessed one from 7 days before

ID	Definition
C8	The proportion of accessed websites with different cosine-based context as the currently-accessed one from 7 days before

Cluster-based contextual similarity is defined based on web page clusters provided beforehand. Fig. 1 shows how the clusters are formed with web page links as the input. Each link's web page is crawled with JSoup [30], and tokenised (with all stop words are removed). Since the methodology will be evaluated on Chinese users, only Chinese characters are considered as terms and the stop words are taken from [31]. These terms are therefore indexed as term-frequency tuples (where the frequency refers to the term's occurrence frequency on a particular web page) and used to form similarity vectors on clustering. The clustering itself is performed with K-Means Clustering [11] (where K is assigned with 10 and the implementation is based on Weka [29]).

Two websites are considered to have the same cluster-based contextual similarity if, given a number of web page clusters, their links are more frequently occurred on web pages which fall on the same clusters than different clusters. Fig. 2 shows how the similarity is calculated.

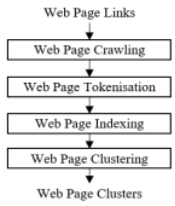

Fig. 1. How web page clusters are formed

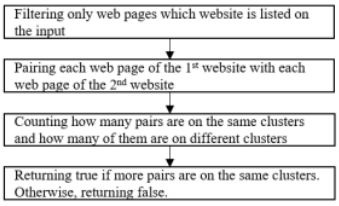

Fig. 2. Defining cluster-based contextual similarity between two websites

To illustrate this, let us assume we have two websites (*dom1.com* and *dom2.com*) and their cluster-based contextual similarity will be calculated based on two web page clusters (*cluster-1* and *cluster-2*); *cluster-1* has three web pages (*dom1.com/a1, dom2.com/b1,* & *dom3.com/c1*) and *cluster-2* has four web pages (*dom1.com/d2*, *dom1.com/e2, dom3.com/f2,* and *dom4.com/g2*). At first, all unrelated web pages are removed. Consequently, cluster-1 and cluster-2 have only two web pages each (dom1.com/a1 & dom2.com/b1 for cluster-1 and dom1.com/d2 & dom1.com/e2 for cluster-2). Secondly, each web page of the 1st website (dom1) will be paired with each web page of the 2nd website (dom2). It results in three pairs: dom1.com/a1 with dom2.com/b1, dom1.com/d2 with dom2.com/b1, and dom1.com/e2 with dom2.com/b1.

Thirdly, the number of pairs which member fall on the same cluster and different clusters are calculated. In our case, two pairs have their members on different clusters (dom1.com/d2 with dom2.com/b1 and dom1.com/e2 with dom2.com/b1) and a pair has their members on the same cluster (dom1.com/a1 with dom2.com/b1). Fourthly, since the number of pairs which members fall on different clusters is higher, dom1.com is considered as different to dom2.com based on cluster-based contextual similarity. This similarity function would return false.

Considering a large dataset may not be able to be clustered at once due to memory limitation, our cluster-based contextual similarity measurement can be tuned to accept multiple cluster groups at once (where each group refers to a result from an execution of clustering algorithm). In such manner, the similarity result will be defined through voting mechanism: a similarity result with higher occurrence frequency among groups will be considered as the final result. For example, if there are five cluster groups and three of them consider given two websites are the same, it can be concluded that those two websites share the same cluster-based contextual similarity according to those groups.

Cosine-based contextual similarity is defined based on predefined contextual groups. Two websites are considered similar if their links fall on the same contextual group when measured using cosine similarity [3]. It is important to note that the links are compared instead of the web pages since comparing web pages is time-inefficient.The contextual groups should depict various topics. Hence, in our case, these groups are defined from various software descriptions on Wikipedia [32]. One hundred most frequently-accessed software are selected for this task (where, in our case, the frequencies are calculated based on user software access log from our evaluation dataset). Since given website links are commonly written with English characters, the software descriptions are limited to English-written descriptions.

Collected software descriptions are then tokenised, indexed, and clustered to contextual groups. The tokenisation works by splitting the descriptions with non-alphanumeric characters and removing all stop words [33]. The indexing converts resulted terms to term-frequency tuples where the frequency is the term's occurrence frequency toward a particular description. The clustering categorises these descriptions to five contextual groups with the help of K-Means Clustering [11]. The contextual group for each given website link is determined in twofold. At first, the link is tokenised and indexed in the same way as software descriptions are tokenised and indexed. Later, the link's index will be compared to each contextual group's index and a group which leads the highest cosine similarity will be considered as the contextual group of the link [3].

IV. Evaluation and Discussion

This section is divided to seven sub-section. The first sub-section will discuss about evaluation dataset. Remaining sub-sections will discuss about specific evaluation mechanisms and their results. One of them is about cluster-based contextual similarity; one of them is about cosine-based contextual similarity; and the last four are about our prediction model for web revisitation.

A. Evaluation Dataset

Our evaluation dataset was extracted from software and web access logs of 980 Chinese users for four weeks. The access logs were recorded using a dedicated monitoring application installed on their laptop / computer. In total, there are 5,912,189 software access entries and 3,260,710 web access entries. Each entry has three fields: user ID, access timestamp, and accessed software / web page link. For referencing purpose, this dataset will be labelled as main dataset. In addition to main dataset, another dataset called cosine dataset is also used in this evaluation. As its name states, it is dedicated for evaluating cosine-based contextual similarity. Thirty web page pairs were selected by the second author of this paper; two thirds of them share the same topic on their respective web page contents while the others do not.

B. The Cluster Descriptiveness of Cluster-based Contextual Similarity

Clusters resulted from our proposed technique for determining cluster-based contextual similarity are considered descriptive if their respective terms either have a meaning or lead to a particular topic. To determine how many cluster terms are meaningful, all web page links from the main dataset are grouped per one thousand links wherein each group are categorised further to ten clusters with our proposed clustering technique. In total, there are 118 groups with 1180 clusters. A cluster will be selected randomly for each group, and the meaning of its most frequently-occurred term will be rated by two examiners. The examiners have three options to choose on: meaningful, not meaningful, and undecidable. If the examiners share different perspectives toward some cluster terms, a discussion will be performed till a consensus is met.

Our experiment reveals that the proportion of meaningful cluster terms (50.85%) is higher than that proportion for meaningless cluster terms (22.88%). Hence, it can be stated that our proposed technique may lead to descriptive clusters. Several cluster terms (26.27%) are still on "undecidable" status since these terms are either overly general or have multiple interpretations.To determine whether terms on a cluster may lead to a more-general topic, the first three thousand web pages from the main dataset are divided to three groups (with one thousand pages per group) wherein each group's pages are categorised to ten clusters with our proposed technique. For each cluster, we define a general term based on its top-10 frequently-occurred words, and we could also able to define the topic of 95% clusters in Chinese. In other words, our proposed technique may lead to descriptive clusters; each cluster has its own specific topic.

C. The Accuracy of Cosine-based Contextual Similarity

Cosine-based contextual similarity is accurate if it can distinguish web page topic as human does. In this context, cosine dataset is used where the accuracy is defined based on how many web page pairs are detected to have the same result as human judgement. Considering cosine similarity returns a proportion degree while human judgement returns a boolean value, cosine similarity's result will be discretised to the boolean one. The proportion degree that is higher or equal to 75% will be converted to true (which means given two web pages share the same topic). Otherwise, it will be converted to false. Our experiment shows that cosine similarity can correctly predict the topics of 54% web page pairs (27% for true positive and 27% for true negative). In other words, cosine-based contextual similarity may be considerably accurate; its prediction is higher than 50%.

D. The Effectiveness of The Same Website's Access Frequency for Predicting Web Revisitation

Eight attributes used for predicting web revisitation (A1-A8 from Table 1) are rooted from the same website's access frequency. In this section, the effectiveness of those attributes will be evaluated. A prediction model based only on those attributes will be built and evaluated toward web access log entries from our main dataset (which are 3,260,710 entries in total). An entry indicates that a web page may be re-visited if at least one entry with the same user accesses the same website at later timestamp. Otherwise, it indicates no revisitation. In the dataset, the number of web access entries for revisitation is significantly higher than that number for no revisitation since, per user, a particular website was usually accessed more than once and only one of those indicates no revisitation (the last access). Therefore, to balance the dataset, only $N/2^{th}$ web access entries are considered for revisitation (where N refers to the number of web access entries for a user toward a particular website). This phase results in 18,500 entries in total. Half of them indicates web revisitation while another half indicates no web revisitation. For convenient referencing, these entries will be labelled as prediction dataset in the rest of this paper.

An attribute is considered effective for predicting web revisitation if its inexistence leads to a huge accuracy reduction [35]. The accuracy will be defined through 10-fold cross validation (10 is selected due to its popularity as a threshold for k-fold cross validation [36]). Prior determining the accuracy reduction, an accuracy when A1-A8 are used together should be measured as the baseline. On our dataset, the accuracy is 53.256%, which is adequately acceptable. The accuracy reduction for each attribute can be seen on Table 2. Each value is calculated by subtracting the baseline accuracy with an accuracy where given attribute is excluded. A3 (average delta time in minutes between five latest access times where currently-accessed website and previously-accessed websites are the same) is the most significant one for predicting web revisitation. It gains the largest reduction (1.948%) compared to other attributes. A1 (delta value between the current and previous access time where both currently- and previously-accessed website are the same), on the contrary, should not be used since its inexistence leads to higher accuracy and the largest negative reduction. A4 and A5 should not also be used as they still lead to negative reduction (even though it is not as significant as A1).

TABLE II. ACCURACY REDUCTION FOR ATTRIBUTES FROM THE SAME WEBSITE'S ACCESS FREQUENCY

Attribute ID	Accuracy Reduction (%)
A1	-0.009
A2	0.014
A3	1.948
A4	-0.061
A5	-0.001
A6	-0.001
A7	0.000
A8	0.000

Among those attributes, A7 and A8 are two attributes which inexistence does not affect resulted accuracy. One of the possible reasons is that these attributes need longer data collection period to show their impact (since their value is calculated for 7 days and our whole data collection period is only 28 days).When each of those attributes are correlated with target classes (using correlation-based feature selection from Weka [29]), that is rooted from Pearson correlation [34], only A3 and A1 indicate a non-zero correlation degree (which is 0.0069 and 0.0032 respectively). Hence, from feature selection perspective, it can be stated that both attributes may be effective for predicting web revisitation (even though A1's impact is still need to be observed further as its inexistence leads to higher accuracy).

E. The Effectiveness of The Same Website's Access Frequency for Predicting Web Revisitation

Attributes that are related to the access frequency of websites with the same cluster-based context (B1-B8 from Table 1) are evaluated in this section. The evaluation will be conducted in a similar manner as evaluating A1-A8 (with the

prediction dataset as instances). An attribute is considered effective if its inexistence results in significant accuracy reduction [35] when measured with 10-fold cross validation [36]. When B1-B8 are used altogether to form a predictive model for web revisitation, the model leads to 50.027% accuracy. It is lower than the baseline accuracy of A1-A8 (53.256%). This phenomenon may be caused by at least one of two reasons. First, B1-B8 may be less effective than A1-A8 for predicting web revisitation. Second, the currently-used technique for determining cluster-based contextual similarity may be ineffective when applied on the dataset. Table 3 shows the accuracy reduction for each attribute. B2 (delta value between the current and previous access time where both currently- and previously-accessed website share different cluster-based context) is the only attribute which leads to a positive accuracy reduction. Other attributes lead to either zero impact (B1, B3, and B6) or negative reduction (B4, B5, B7, and B8).

TABLE III. ACCURACY REDUCTION FOR ATTRIBUTES FROM THE ACCESS FREQUENCY OF WEBSITES WITH THE SAME CLUSTER-BASED CONTEXT

Attribute ID	Accuracy Reduction (%)
B1	0.000
B2	0.171
B3	0.000
B4	-0.011
B5	-0.008
B6	0.000
B7	-0.008
B8	-0.008

When correlated with target classes (using correlation-based feature selection from Weka [29]), all of these features generate zero correlation degree. In other words, from feature perspective, none of these attributes may be effective for predicting web revisitation.

F. *The Effectiveness of The Access Frequency of Websites with the Same Cosine-based Context for Predicting Web Revisitation*

Attributes rooted from the access frequency of websites with the same cosine-based context (C1-C8 from Table 1) will be evaluated in this section. The evaluation will be conducted in a similar manner as evaluating A1-A8 and B1-B8, involving the prediction dataset as instances. An attribute is considered effective if its inexistence results in significant accuracy reduction [35] when measured with 10-fold cross validation [36]. The baseline accuracy is defined by utilizing C1-C8 altogether to form a model and evaluate the model with 10-fold cross validation. The resulted accuracy is 49.755%, which is lower than the baseline accuracies for A1-A8 (53.256%) and B1-B8 (50.027%). In other words, C1-C8 may be less effective than other attribute categories. Further, proposed cosine-based contextual similarity may be not accurate for determining similarity. This finding is supported by the fact that it only leads to 54% accuracy when compared to human judgment.

TABLE IV. ACCURACY REDUCTION FOR ATTRIBUTES FROM THE ACCESS FREQUENCY OF WEBSITES WITH THE SAME COSINE-BASED CONTEXT

Attribute ID	Accuracy Reduction (%)
C1	0.066
C2	-0.115
C3	0.066
C4	-0.1347
C5	0.046
C6	-0.115
C7	0.046
C8	0.012

As seen on Table 4, C1 (delta value between the current and previous access time where both currently- and previously-accessed website share the same cosine-based context) and C3 (average delta time between five latest access times where currently-accessed website and previously-accessed websites share the same cosine-based context) generate the largest positive reduction, followed by C5, C7, and C8 respectively. Only three attributes yield negative reduction: C2, C4, and C6. It can be therefore stated that, according to accuracy reduction, all attributes except C2, C4, and C6 may be effective for predicting web revisitation. It is important to note that the previous finding is not aligned with our finding from correlation-based feature selection from Weka [29]. When correlated with their target classes, all attributes show no correlation.

G. *The Effectiveness of Proposed Model for Predicting Web Revisitation*

This section aims to evaluate the overall effectiveness when all attributes are used together. With 10-fold cross validation on board, these attributes lead to 53.195% accuracy when evaluated toward the prediction dataset. It is higher than the baseline accuracies of B1-B8 (50.027%) and C1-C8 (49.755%). However, it is still slightly lower than the baseline accuracy of A1-A8 (53.256%). To sum up, according to our dataset, only attributes rooted from the same website's access frequency are enough to generate the highest accuracy. Other kinds of attributes may worsen the result even though the reduction is insignificant.

V. CONCLUSION

This research proposes a predictive model to determine whether a website will be revisited by a particular user. The model relies on 24 attributes which are grouped to three categories. The first group is derived from the access frequency of the same website; the second group is derived from the access frequency of websites with the same cluster-based context; and the third group is derived from the access frequency of websites with the same cosine-based context.

According to our evaluation, four findings can be concluded. First, the predictive model is considerably accurate. It generates 53.195% accuracy when all attributes are involved. Second, attributes derived from the access frequency of the same website may be potential for predicting web revisitation. These attributes alone generate 53.256% accuracy, which is slightly higher than the accuracy of all attributes. Third, cluster-based contextual similarity (that is used for defining attributes) generates considerably descriptive clusters. The clusters have 50.85% meaningful terms while the topics of 95% clusters can be defined based on their top-10 frequently-occurred words. Fourth, cosine-based contextual similarity (that is used for defining attributes) is considerably effective. It generates 54% accuracy. These findings confirmed the effectiveness of our approach. As a comparison, some problems, such as keyphrase extraction often have lower accuracy (smaller than 30%) [28, 34-35].

For future works, we plan to incorporate more-advanced clustering algorithm on cluster-based contextual similarity for higher accuracy via a comparative study. Further, we also plan to implement this predictive model as a browser plug-in so that it can be used to filter web browsing history.

ACKNOWLEDGMENT

The authors would like to thank Guizhou University, for providing the dataset and Maranatha Christian University, for providing the international student exchange grants.

REFERENCES

[1] M. Vigo and S. Harper, "Real-time detection of navigation problems on the World 'Wild' Web," *International Journal of Human-Computer Studies*, vol. 101, pp. 1–9, 2017.

[2] J. Clement, "Number of internet users worldwide 2005-2018," *Statista*, 07-Jan-2020. [Online]. Available: https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/. [Accessed: 13-Feb-2020].

[3] B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Boston: Pearson, 2010.

[4] W. Du, Z. Qian, P. Parsons, and Y. Chen, "Personal Web Library: Organizing and Visualizing Web Browsing History," *International Journal of Web Information Systems*, pp. 212–232, Feb. 2018.

[5] S. S. Won, J. Jin, and J. I. Hong, "Contextual web history," Proceedings of the 27th international conference on Human factors in computing systems - CHI 09, 2009.

[6] M. V. Kleek, B. Moore, C. Xu, and D. R. Karger, "Eyebrowse," Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA 10, 2010.

[7] M. Mayer, "Web History Tools and Revisitation Support: A Survey of Existing Approaches and Directions," *Foundations and Trends® in Human–Computer Interaction*, vol. 2, no. 3, pp. 173–278, 2007.

[8] L. Jin, L. Feng, G. Liu, and C. Wang, "Personal Web Revisitation by Context and Content Keywords with Relevance Feedback," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1508–1521, Jan. 2017.

[9] G. Papadakis, R. Kawase, E. Herder, and W. Nejdl, "Methods for web revisitation prediction: survey and experimentation," *User Modeling and User-Adapted Interaction*, vol. 25, no. 4, pp. 331–369, 2015.

[10] J. Šimko, M. Tvarozek, and M. Bielikova, "Semantic History Map: Graphs Aiding Web Revisitation Support," *2010 Workshops on Database and Expert Systems Applications*, 2010.

[11] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[12] B. J. Jansen, A. Spink, and T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the web," *Information Processing & Management*, vol. 36, no. 2, pp. 207–227, 2000.

[13] J. Y. Park, N. Ohare, R. Schifanella, A. Jaimes, and C.-W. Chung, "A Large-Scale Study of User Image Search Behavior on the Web," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15*, 2015.

[14] S. D. Torres, I. Weber, and D. Hiemstra, "Analysis of Search and Browsing Behavior of Young Users on the Web," *ACM Transactions on the Web*, vol. 8, no. 2, pp. 1–54, Jan. 2014.

[15] J. Jadav, C. Tappert, M. Kollmer, A. M. Burke, and P. Dhiman, "Using text analysis on web filter data to explore K-12 student learning behavior," *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2016.

[16] C. Bernaschina, M. Brambilla, A. Mauri, and E. Umuhoza, "A Big Data Analysis Framework for Model-Based Web User Behavior Analytics," *Lecture Notes in Computer Science Web Engineering*, pp. 98–114, 2017.

[17] V. S. Tseng and K. W. Lin, "Efficient mining and prediction of user behavior patterns in mobile web systems," *Information and Software Technology*, vol. 48, no. 6, pp. 357–369, 2006.

[18] J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu, "Visual snippets: summarizing web pages for search and revisitation," *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, pp. 2023-2032, 2009.

[19] A. Demachi, S. Matsushima, and K. Yamanishi, "Web Behavior Analysis Using Sparse Non-Negative Matrix Factorization," *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 574-583, 2016.

[20] E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," *ACM SIGIR Forum*, vol. 52, no. 1, pp. 11–18, 2019.

[21] Y. Liu, R. Cen, M. Zhang, S. Ma, and L. Ru, "Identifying web spam with user behavior analysis," *Proceedings of the 4th international workshop on Adversarial information retrieval on the web - AIRWeb 08*, pp. 9–16, 2008.

[22] C. Liu, R. W. White, and S. Dumais, "Understanding web browsing behaviors through Weibull analysis of dwell time," *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR 10*, pp. 379–386, 2010.

[23] E. Adar, J. Teevan, and S. T. Dumais, "Large scale analysis of web revisitation patterns," *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI 08*, pp. 1197–1206, 2008.

[24] H. Zhang and S. Zhao, "Measuring web page revisitation in tabbed browsing," *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI 11*, pp. 1831–1834, 2011.

[25] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification," *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 26–39, 2016.

[26] H. Zhang, Z.-X. Cao, M. Li, Y.-Z. Li, and C. Peng, "Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals," *Food and Chemical Toxicology*, vol. 97, pp. 141–149, 2016.

[27] F.-C. Chen and M. R. Jahanshahi, "NB-CNN: Deep Learning-Based Crack Detection Using Convolutional Neural Network and Naïve Bayes Data Fusion," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4392–4400, 2018.

[28] F. C. Jonathan and O. Karnalim, "Semi-Supervised Keyphrase Extraction on Scientific Article using Fact-based Sentiment," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 16, no. 4, p. 1771, Jan. 2018.

[29] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining: practical machine learning tools and techniques*. Amsterdam: Morgan Kaufmann, 2017.

[30] "jsoup Java HTML Parser, with best of DOM, CSS, and jquery," *jsoup Java HTML Parser, with best of DOM, CSS, and jquery*. [Online]. Available: https://jsoup.org/. [Accessed: 13-Feb-2020].

[31] Xangis, "Xangis/extra-stopwords," *GitHub*. [Online]. Available: https://github.com/Xangis/extra-stopwords/blob/master/chinese. [Accessed: 13-Feb-2020].

[32] "Main Page," *Wikipedia*, 05-Feb-2020. [Online]. Available: https://en.wikipedia.org/wiki/Main_Page. [Accessed: 13-Feb-2020].

[33] Xangis, "Xangis/extra-stopwords," *GitHub*. [Online]. Available: https://github.com/Xangis/extra-stopwords/blob/master/english. [Accessed: 13-Feb-2020].

[34] "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.

[35] O. Karnalim, "Software Keyphrase Extraction with Domain-Specific Features," *2016 International Conference on Advanced Computing and Applications (ACOMP)*, pp. 43-50, 2016.

[36] G. J. Mclachlan, K.-A. Do, and C. Ambroise, "Analyzing Microarray Gene Expression Data," *Wiley Series in Probability and Statistics*, 2004.

Predicting Users' Revisitation Behaviour Based on Web Access Contextual Clusters

ORIGINALITY REPORT

3%	2%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
2	Submitted to University of New South Wales Student Paper	<1%
3	deepai.org Internet Source	<1%
4	jtec.utem.edu.my Internet Source	<1%
5	Markel Vigo, Simon Harper. "Real-time detection of navigation problems on the World 'Wild' Web", International Journal of Human-Computer Studies, 2017 Publication	<1%
6	Luo, Dixin, Hongteng Xu, Hongyuan Zha, Jun Du, Rong Xie, Xiaokang Yang, and Wenjun Zhang. "You Are What You Watch and When You Watch: Inferring Household Structures	<1%

From IPTV Viewing Data", IEEE Transactions on Broadcasting, 2014.
Publication

7	content.iospress.com Internet Source	<1 %
8	dmc.tamuc.edu Internet Source	<1 %
9	repository.tudelft.nl Internet Source	<1 %
10	"Intelligent Systems in Cybernetics and Automation Control Theory", Springer Science and Business Media LLC, 2019 Publication	<1 %
11	George Papadakis, Ricardo Kawase, Eelco Herder, Wolfgang Nejdl. "Methods for web revisitation prediction: survey and experimentation", User Modeling and User-Adapted Interaction, 2015 Publication	<1 %
12	Muftah Afrizal Pangestu, Oscar Karnalim, Simon. "Mapping Similarity Detectors of Code Clone to Academic Integrity in Programming", 2021 IEEE World Conference on Engineering Education (EDUNINE), 2021 Publication	<1 %
13	escholarship.org Internet Source	<1 %

14	journal.uad.ac.id Internet Source	<1%
15	Liu, Yue-Cen, Wei-Hsun Wen, and Wei-Guang Teng. "Discovering and Exploiting Cumulative Cues for Informational Web Search", 2010 International Conference on Technologies and Applications of Artificial Intelligence, 2010. Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Predicting Users' Revisitation Behaviour Based on Web Access Contextual Clusters

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6