Web usage prediction and recommendation using web session clustering

Vinh-Trung Luu*, Germain Forestier*, Mathis Ripken*,
Frédéric Fondement*, Pierre-Alain Muller*
*MIPS, Université de Haute Alsace,
12, rue des frères Lumière - 68093 Mulhouse cedex - France
Email: {trung.luu-vinh, germain.forestier, mathis.ripken
frederic.fondement, pierre-alain.muller}@uha.fr

***Abstract*—In recent years, a strong interest has been given to web usage prediction and recommendation methods to improve e-commerce, search engines and other online applications. There have been various efforts carried out in this field, particularly focused on using recordings of web user interactions with websites. In this context, our research focuses on developing a novel approach for web prediction and recommendation. The proposed method relies on hierarchical session clustering by sequence similarity measure and takes advantage of access activity time and access position in prediction session to make a recommendation. The performed experiments reveal that hierarchical parameter and prediction accuracy are relevant. In addition, the paper introduces cost estimation to adapt web visitor behavior to web business purposes using prediction and recommendation results.**

I. Introduction

The Internet is today the richest information source in the world. However, with this extremely large amount of information, the problem web visitors are facing is how to reach relevant content and to discard irrelevant resources or unrelated content. Web visitors exhibit various types of behavior through their browsing activities which are captured during their visit. Consequently, it becomes more and more essential to present proper recommendation content adapted to visitor interests. In other words, visitor needs should be understood and taken into account correctly. Although the raise of individual privacy concern can receive negative comments as the privacy violation, visitors' interest is currently the backbone of e-marketing campaigns. Correspondingly, thousands of different web visitors, for example on an online shopping website, may see thousands of distinct versions of the homepage, which is called content personalization. In order to effectively use collected visitor browsing data for such behavioral targeting, web usage prediction and recommendation (WPR) have been adopted and plays a vital role in behavioral targeting strategy of search engines, entertainment and e-commerce websites. Taking advantage of multiple techniques to target visitor, the process predicts the upcoming request of visitors and sends related promotional contents information with specific recommendation. By pre-fetching, pre-sending or caching such recommendations, network latency effect can be also reduced. Since WPR helps to increase revenue growth, it has turned into a fundamental feature of commercial websites, or even helps to improve search engines performance.

Predictive model construction, which indicates the chances of next accesses of visitor browsing, is the earlier phase of WPR process. In order to build this kind of model, Predictive Analytics (PA) [1] is among the most appropriate methodologies. Consisting of technologies which assist users in predicting web visitor action, PA appliances are widely known to be efficient in e-commerce marketing, search engines or other big data systems by instantly analyzing and discovering web usage patterns [2]. Alternatively, recommender system [3] assists web visitors in making real-time choice and even transform them into customers. There have been multiple presentation techniques to guide personalized navigation such as images, text, hyperlinks, etc. with the support of font size, color, etc. to lead visitors to tailored content as a recommendation, and these features have changed the way of interaction between websites and visitors. Later, under the influence of online campaigns, client behavior is altered and hence advanced web data is produced. Overall, the prediction and recommendation work with web data as a mutual reciprocity is illustrated in Figure 1.

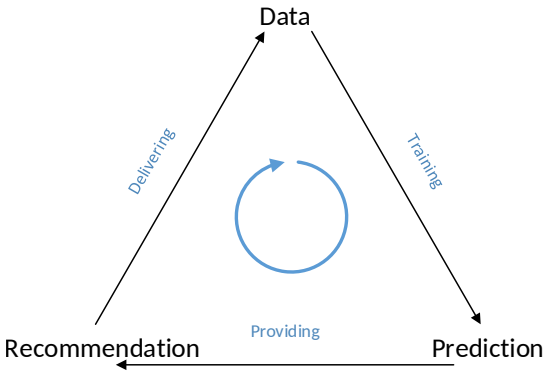

Fig. 1: Round process of prediction, recommendation and web data

Nonetheless, several works have shown an ambiguity of prediction and recommendation by stirring them together. Obviously, correct prediction of visitor action itself does not reflect what they prefer or is suitable for them. The visitor

online behavior should be inferred not necessarily from their custom but according to the web sites or search engines business. Also, a further prediction and recommendation related part is customization cost estimation between web document design and visitor need. In this paper, in order to bridge those research gaps, we introduce a prediction model of web visitor behavior which takes advantage of clustering techniques based on session alignment. The proposed technique produces recommendation patterns without needing web structure but using activity time visitors spend on pages and page indexes, and a measure for web site structure to visitor interest adaptation cost.

This paper includes five sections as follows: Section 2 presents the proposed WPR model. Section 3 describes the experimental results. Related works are discussed in Section 4. Section 5 shows the conclusion and reveals some research plans for the future.

II. Proposed method

A. Prediction

This paper proposes to perform web prediction on top of web session clusters. In order to build relevant session cluster, a similarity measure is adopted to calculate the similarity between train queries data, then clusters are hierarchically constructed. As the resulting clusters contain comparable historical sessions with identical initial pages, it facilitates the prediction process of next accesses from the first ones.

1) Modified combination measure: The original proposed combination measure [4] was experimentally proved to be appropriate to evaluate the global and local similarity of sequences [5]. Nonetheless, clusters of similar sessions are not sufficient to make prediction if session lengths are very short and entry pages are not unified. Firstly, sessions with less than two accesses should be eliminated since they are definitely not suitable for prediction. Second, as an illustration, to forecast next pages that visitor is likely to hit after page `A`, clusters with sessions initialized by `A` like (`ABCDE`, `ABCD`) should be considered if they are available. Similarly, for more-than-one-page inputs such as `AB`, clusters with `A` as the first page will be analyzed to see if they include `AB` to make prediction. As a consequence, it is inappropriate to use the set (`ABCDE`, `BCDE`, `CDE`) as a corresponding prediction cluster of any specific input access.

Consequently, the sequence similarity measure is required to take this issue into consideration to avoid such situations. In other words, order related sessions but including different first accesses should not be considered as similar and in the same cluster. Accordingly, we modified similarity combination measure by assigning a very low similarity score to session pairs with different prefix, hence only pairs of sessions which have the duplicate initials are considered. This modification does not impact the association of similar and duplicate initial sessions. In addition, such a *glocal* (*i.e.* global and local) similarity measure performs somehow similarly to association rules or Markov model in "learning" order and relation between elements, which are pages in session.

2) Clustering: As a tree diagram commonly adopted to build a hierarchical form of clusters, dendrogram does not merely illustrate a cluster set but its multilevel set [6]. The approach consists in merging two most similar clusters at a level into one at the upper level. Correspondingly, cutting the dendrogram at variable levels outputs different sets of clusters and the context-dependent appropriate level can be selected. Furthermore, for the purpose of making elements in the same cluster similar and different clusters dissimilar, it is not trivial to decide which level fits the best the data. An inappropriate level to stop clustering could create indefinite or over-particular clusters. For instance, combining two clusters containing sequences such as (`ABCD`, `ABC`) and (`ABCDEFG`, `ABCDEF`) may make the cluster result not as united as the original two. On the contrary, clusters like (`ABC`) and (`ABC`) should be joined together as they are identical. We set a heuristic threshold to collect clusters based on their intra and inter distances, as the formula below:

$$\textit{diff} = max(inter\_distance) - min(intra\_distance) \quad (1)$$

Given $inter\_distance$ the distance of elements between single clusters and merged cluster, $intra\_distance$ is the distance of elements inside merged cluster and *diff* is the difference of inter and intra distances.

Diff is used as a threshold value to decide to continue or stop to merge clusters. The larger *diff* is, the more distant clusters are. As we use single-linkage as hierarchical strategy due to its experimental advantage with similarity combination measure, clusters are merged if they are nearest neighbors. In order to observe if two clusters are worth being merged, the minimum distance inside merged cluster is compared to the maximum distance from it to single clusters. For example, if we decide to stop cluster at $\textit{diff} > 0.1$, then there are more clusters created than $\textit{diff} > 0.9$ although these clusters are less separated. This threshold is flexible to find optimal clusters depending on the context, and there is even no cluster created if this threshold value is very high.

Additionally, there might be couples of clusters which include different entry page sessions due to linkage strategy. In order to flush prediction clusters, they will be eliminated from the cluster set. If the session similarity measure is not improved as described above, the number of applicable prediction clusters will be significantly less by this removal, and thus the prediction efficiency will be correspondingly reduced.

3) Prediction implementation: If historical sessions are not clustered, time and space consuming investigations through every element of train set must be implemented to consider if a prediction can be made. Accompanying prediction clusters, corresponding session groups of an input can be called and require considerable less time and space. Based on one or more initial prefix of input sessions from the test set, if there exist identical sessions in the matching clusters, the prediction is correct. Figure 2 shows an example of page sequence inputs to anticipate next pages, and a complete session to confirm the accuracy of the prediction. Figure 3 illustrates three examples

of corresponding clusters of inputs in Figure 2, with an equal session to the complete session of Figure 2 in Cluster 1. In this case, the prediction is then accurate.

A	AB	ABCDEF
(a) Input 1	(b) Input 2	(c) Complete session

Fig. 2: Possible inputs and complete session to predict, and investigate the prediction accuracy.

ABCDE ABCDEF ABCDF	AMNO AMNOP AMNOPQ AMNOPQ	ABCDC ACDCEC
(a) Cluster 1	(b) Cluster 2	(c) Cluster 3

Fig. 3: Three prediction clusters corresponding to Input 1, and Cluster 2 will be eliminated to predict Input 2 in Figure 2. Besides, complete session of Figure 2 matches the second session of Cluster 1.

Nevertheless, prediction cluster set is generally not able to cover every input to make the entire prediction. Due to the limited size of the train set, removal of unclean clusters, unique or rare queries, etc., there may be access that cannot be predicted in real-time. On the other hand, a proper train should eliminate invalid patterns like input errors, incomplete visitor traces etc., to be best suited for the application context.

B. Recommendation

The recommendation, in any form, should make visitors more convenient in their browsing. Since recommendation is regularly based on prediction, the prediction should be effective in defining targets. Nevertheless, prediction information is basically not relevant to use in recommendation. In other words, browsing behavior and site proposition are two correlated but distinct concepts. In order to improve the system usability, prediction models are required to be integrated into people aspects. For example, if a visitor has been querying about hotel deals, it may be helpful to show them airline promotions to calculate how much they can save totally, instead of more hotel options they are likely to search, that makes them confused. Otherwise, when three information pages of an online course have been browsed and users are predicted to visit the fourth page, it is completely not essential to suggest the fourth one. Alternatively, it is probably the right time to show subscription benefit or schedule advise before they leave. Consequently, one of the prospective approaches to make a recommendation by taking advantage of prediction is to dynamically recommend new information to the visitor. This kind of approach is context-dependent and based on the categories that visitors are predicted to belong to. Namely, prediction model may recognize the matching categories of a visit through its corresponding prediction clusters, then appropriate suggestions can be performed to make some specific content more accessible. In order to effectively comply with visitor demand, such recommendation should be directly started from the first access and active during the visitor session.

Alternatively, another considerable feature to support visitors is browsing time reduction. As visitors have their targets while accessing a website, one essential thing we can focus on is time saving to quickly reach those targets. For this purpose, visitors' preferences like browsing order should not be considered as it may provoke a waste of time. It is reasonable to assume that visitors are prone to spend more activity time on their interesting content [7], [8] (even opening them in a new tab, keep searching in other sites and then going back to them). The difference between activity time and the duration from start time to end time of a web page is that activity time does not include interruption time caused by other irrelevant activities. Particularly, the total time visitors spent on scrolling, highlighting, hovering, etc. on a web page without an idle time is activity time [9]. Furthermore, visitors are likely to leave after reaching these target information [10] as the notion of *maximal forward reference* by Chen et al. [11]. Following that, a session group or cluster of similar navigational page sequences reveals some interests in visiting order of pages that is good for predicting accesses, but not for interest prediction. For instance, the visit sequences $\mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E}$ and $\mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E} \rightarrow \mathtt{F}$ may be in the same cluster of navigation but the destination page of sessions could possibly be E or F. Also, B may be kind of a "bridge" such as category or search page to jump to real content, so it does not reflect the common interest of visitors and they are likely to spend a limited time on it. Therefore, their corresponding navigation cluster probably does not make sense in the recommendation as introduced by [12] or [13]. Briefly, access activity time and index should be two indicators of visitors' interest that should be considered in the recommendation.

According to the previous assumption about the relation between time spent and last visited page to user interest, we assume that two main factors affect the interest prediction of a specific page are: (1) activity time of visitor and (2) the visit order. There may be more than one target for each visitor and these targets are probably of different priorities for the visitor. In this context, a metric of visit destination probability which takes into account position of the page in visit session and time spent on it is proposed. Consequently, this metric shows the probabilities of each page of a session to be interesting for a vistor, by the formula below:

$$Pr(i) = Pos(i) \times T(i) \tag{2}$$

Given $Pr(i)$ the interest probability of page i in a session, $Pos(i)$ the position of page i in that session (1st page of session get the position of 1, the second one's position is 2, and so forth), and $T(i)$ the time visitor spend on page i by their activities. Repeatedly, $T(i)$ is different from the total access duration from start time to end time on a page, it does not include page loading time, browsing interruption, etc.

For example, according to (2), the interest probability of page C, with $T(\mathtt{C})$ equals to 12 secs, in session $\mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{C} \rightarrow$

$\mathtt{D} \rightarrow \mathtt{E}$, is $3 \times 12 = 36$. In the same session, $Pr(\mathtt{E})$ with 5 secs of time and $Pos(\mathtt{E}) = 5$ would be 25. It is noticeable that even an exit visit is not likely to have a high probability if the time spent on it is not significant. In this case, the visitor probably left the site not because they find what they needed. Assuming that $Pr(\mathtt{C}) > Pr(\mathtt{D}) > Pr(\mathtt{E}) > Pr(\mathtt{A}) > Pr(\mathtt{B})$ for the session, the probabilities of recommendation should be similarly, in descending order, such as $\mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E} \rightarrow \mathtt{A} \rightarrow \mathtt{B}$ with $\mathtt{C}$ and $\mathtt{B}$ having the highest and lowest values of all following (2), respectively. This kind of sequences is a form of recommendation that converted from the prediction session, and we aim at building recommendation clusters from them since such clusters are convenient for the dynamic recommendation process. From a cluster of prediction as described in Figure 4, a corresponding recommendation cluster like Figure 5 can be derived for example using previously mentioned computation.

$\mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E}$
$\mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E}$
$\mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{D} \rightarrow \mathtt{E}$
$\mathtt{A} \rightarrow \mathtt{B} \rightarrow \mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E}$
$\mathtt{A} \rightarrow \mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E}$

Fig. 4: Cluster of prediction.

$\mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E} \rightarrow \mathtt{A} \rightarrow \mathtt{B}$
$\mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E} \rightarrow \mathtt{A} \rightarrow \mathtt{B}$
$\mathtt{D} \rightarrow \mathtt{E} \rightarrow \mathtt{A} \rightarrow \mathtt{B}$
$\mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E} \rightarrow \mathtt{A} \rightarrow \mathtt{B}$
$\mathtt{C} \rightarrow \mathtt{D} \rightarrow \mathtt{E} \rightarrow \mathtt{A}$

Fig. 5: Cluster for recommendation.

In order to be in one recommendation cluster, these sequences certainly have to be related in order and length because it reflects a group of common browsing behaviour together with the correlation of time spent on particular pages, which is the subject of this approach. As combination measure clusters sequences based on *glocal* similarity, it makes clustered sessions similar in length. Also, combination measure makes the position of a specific page in one session not very different from its position in other sessions in the same cluster, if existing. Therefore, a threshold of session correspondence computed by combination measure is mandatory to validate recommendation clusters. If target contents of visitors in prediction cluster are not similar, such visitors are not targeted for the recommendation. Consequently, there may be no recommendation presented as no equivalent recommendation cluster built on those prediction clusters.

Concerning the frequencies of page appearance in recommendation cluster, the recommendation should start with the most frequent page from first to last index, as long as it has not been visited. For example, such frequencies of $\mathtt{C} = 4$ and of $\mathtt{D} = 1$ in 1st index of recommendation cluster, visitors who belong to the prediction cluster in Figure 4 are interested mostly in $\mathtt{C}$, then $\mathtt{D}$, $\mathtt{E}$, $\mathtt{A}$ and $\mathtt{B}$. Therefore, $\mathtt{C}$ should be on the top of recommendation list if it has not been accessed, and so forth. In other words, the page with the highest probability of interest should be recommended first, then other pages recommended in descending probability (or priority).

The dynamic recommendation process indeed exploits the prediction results when monitoring visitor run-time behavior. If this kind of behavior matches prefix of prediction patterns, their recommendation content will appear in multiple forms. Accordingly, when visitor accesses page $\mathtt{A}$ or $\mathtt{A}$ then $\mathtt{B}$ as described in Figure 6 for example, prediction cluster in Figure 4 may be among the used clusters. Correspondingly, recommendation cluster in Figure 5 is then prepared for navigation suggestion. The process works as shortest paths instruction for the visitor so that they can instantly end up with their expected information.

$\mathtt{A}$	$\mathtt{AB}$
(a) Input 1	(b) Input 2

Fig. 6: Possible inputs for prediction using navigation cluster in Figure 4 and then recommended by recommendation cluster in Figure 5.

The development of recommendation clusters from prediction clusters and prediction ones from collected sessions are presented in Figure 7. Likewise, the prediction and recommendation steps from visitor opening entries are shown in Figure 8.

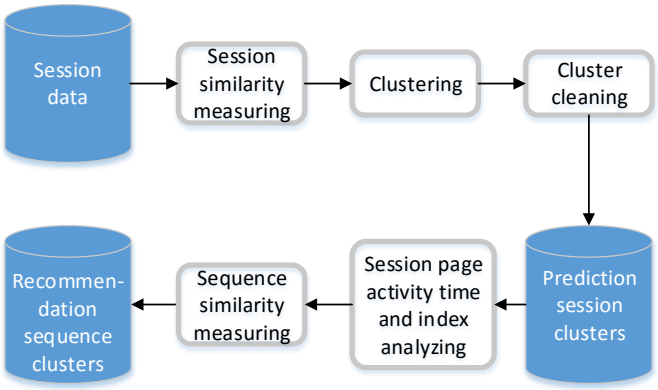

Fig. 7: Visitor sessions grow into prediction session clusters, and prediction session clusters turn into recommendation sequence clusters.

C. Cost to adapt web site structure to recommender system

As the visitor navigation is supposed to be lead by web site structure, it may cost site owners or content providers if they require enhancing site structure to benefit visitors. Accordingly, a cost metric taking advantages of prediction and recommendation cluster should be considered. Since it is difficult to make an assumption about the specific structure of site or recommender system, an effort of estimation to reconstruct visitor tendency to the endorsement of the website is reasonable to study. Particularly, a conversion cost to migrate a prediction cluster to its corresponding recommendation cluster can be regarded as an initial solution. In detail, this conversion needs to be based on sub-conversion

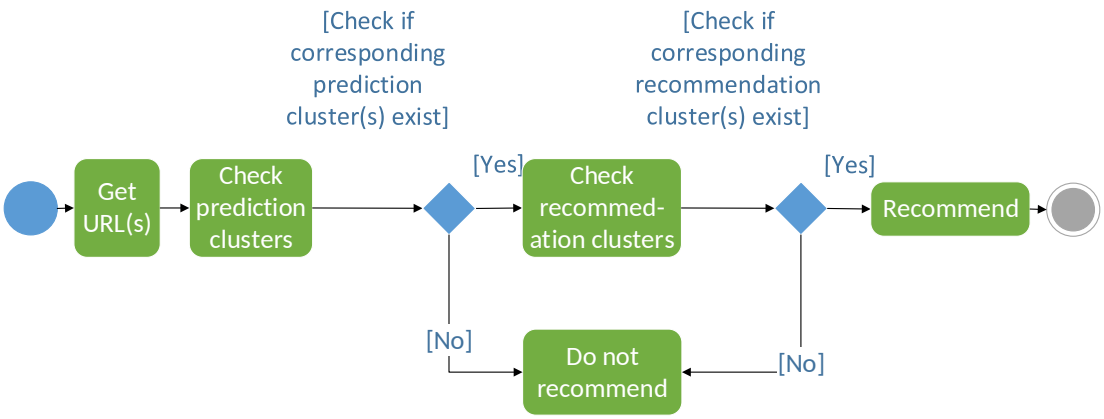

Fig. 8: The representation of prediction and recommendation workflow.

of each prediction sequence to correlative recommendation one. In addition, the cost of sub-conversion should be globally and locally computed through combination measure since the measure correctly reflects the similarity between them and the more similar sessions are, the less conversion cost is supposed to be. Namely, sequence conversion cost and similarity are in inverse ratio. For example, the two prediction and recommendation sequences of web access in Figure 9 have the similarity of 0.4 by combination measure. Consequently, their conversion cost will then be -0.4. This conversion cost is also appropriate to apply in web usability evaluation.

$$\begin{array}{l} \mathrm{A} \rightarrow \mathrm{B} \rightarrow \mathrm{C} \rightarrow \mathrm{D} \rightarrow \mathrm{E} \\ \mathrm{C} \rightarrow \mathrm{D} \rightarrow \mathrm{E} \rightarrow \mathrm{A} \rightarrow \mathrm{B} \end{array}$$

Fig. 9: First prediction and recommendation sequences of clusters in Figure 4 and 5.

The adaptation of visitor manner to business strategy is more beneficial if it meets more end-users, hence cluster size plays an important role in conversion cost calculation. Concerning two equal prediction-to-recommendation conversion costs, which one consists of greater number of sessions is the preferred one. Alternatively, the one which costs less for conversion among same size pairs of prediction and recommendation clusters is most productive. The formula to compute total conversion cost, in order to switch prediction cluster sequences to recommendation ones, should take cluster size into consideration as follows:

$$C(P,R) = \frac{Sum(sim(i,j) \times -1)}{N} \tag{3}$$

where $C(P,R)$ is given as total conversion cost between clusters P(prediction) and R(recommendation), $sim(i,j)$ is the computed similarity score between 2 corresponding prediction and recommendation sequences i and j and N is the sequence number of each cluster.

As the total conversion cost between clusters in Figure 4 and 5 is -1.9 by combination measure, this cost should be divided by the number of sequences in the cluster to find the representative cost, as $-1.9/5 = -0.38$, in accordance with (3). For the sake of optimization, that minimizes this cost but makes web site more convenient to visitors, pairs of clusters with lower total conversion cost should be in higher implementation priority to advance web site reconstruction.

III. Experimental Result

The prediction accuracy performed on test sets should take into account both matching number of sessions and set size. Nigam et al. [14] measured this accuracy as the ratio between correct prediction number and test session number. The formula is defined as follow:

$$Prediction\_Accuracy = \frac{Correct\_prediction\_number}{Test\_session\_number} \tag{4}$$

The experiments were conducted on three datasets of 2000 individual sessions each from a web site of University campus with more than $20,000$ visits monthly. The three datasets were recorded at the different time in the month and day in order to have representative visits. Dataset collection contains information of sessionID, accessed URLs, activity time etc. of sessions was implemented by Beampulse company, which provides such services written in Java and Javascript. Each dataset is randomly split into train and test sets, with ratio 80% and 20% respectfully. Experimental results show the significant impact of hierarchical criteria on prediction clusters number and thus prediction accuracy. As previously named in (1) as *diff*, this parameter is based on intra and inter distances between created clusters. Consequently, it is a threshold to decide whether the merging process of sub-clusters should carry on in the dendrogram. The sooner this process stops, the more clusters are generated in the result and vice versa. Apparently, more clusters imply less elements contained in a cluster, that makes clusters more specific. This provokes fewer eliminations to make the cluster set pure, and thus more visit patterns remain. Although a lower *diff* value improves the accuracy of prediction besides time saving in bottom-up hierarchical clustering, it may trade off time cost of seeking and loading appropriate clusters for visitor access inputs. Figure 10 and 11

illustrate the experimental correlation between the number of clusters, prediction accuracy and the hierarchical parameter. As ones can see, the number of matches between test set and prediction clusters increases if there are more prediction clusters created (*i.e.* smaller values of hierarchical parameter). For example, in Figure 10 the cluster number reaches 174 when *diff* = 0.1 in Dataset 1. Similarly, the prediction accuracy rises 27.5% with Dataset 2 at the same *diff* value, as presented in Figure 11. Also, in accordance with the experimental results, the number of output prediction clusters is consistently higher than correct prediction number at every hierarchical parameter. Furthermore, the prediction accuracy is impacted not only by prediction and recommendation nature but also the coverage of the train set to test set.

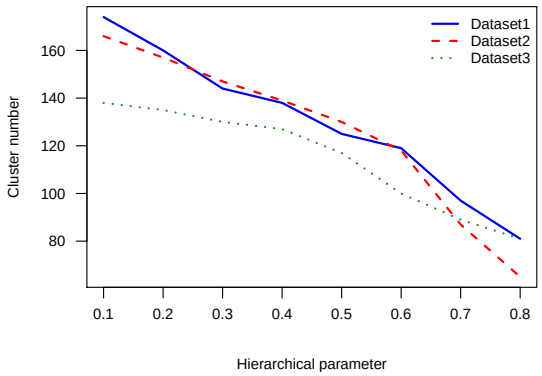

Fig. 10: The hierarchical parameter is inversely proportional to the number of clusters.

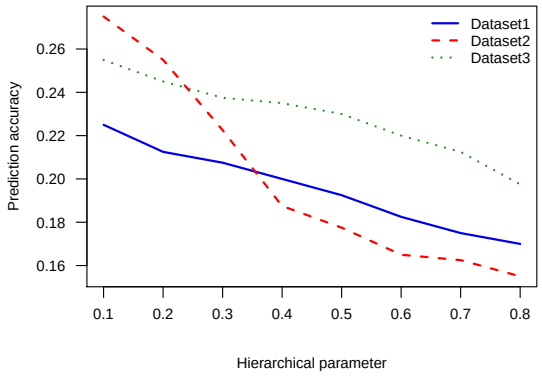

Fig. 11: The hierarchical parameter is inversely proportional to the prediction accuracy.

In order to conduct the experiments of recommendation that take advantages of prediction experimental results, a recommender system should be implemented and tested under real user conditions. Accordingly, the recommendation performance can be measured by the ignorance of visitor on non-target pages and their activity time on target ones.

IV. Related work

A considerable amount of mining techniques in web usage prediction and recommendation (WPR) has been proposed. Anitha et al. [15] proposed WPR model by integrating pair-wise nearest neighbor clustering with support pruned in all k-th order Markov model. This approach is similar to [16], that takes advantage of Markov model in processing clusters of non-sequential sequences, but Markov trades prediction accuracy for space and time complexity as well as low coverage. Additionally, pair-wise nearest neighbour clustering may group sequences with common elements in different orders, and eliminates similarly ordered sequences with less common elements. Another related work by Thwe et al. [17] came up with Popularity and Similarity based Page Rank (PSPR) algorithm to solve vague result output by Markov model. The algorithm takes page properties such as frequency, duration, page size into account to rank the popularity of web pages. Nevertheless, duration may not reflect the interest of visitor in web page and page size can possibly be inaccurate at client size due to network traffic. A hybrid prediction model by [18] recommends a combination of Markov and Hidden Markov although there are difficulties handling Hidden Markov such as time and space complexity, or small training set making initial model over-trained. Su et al. [19] proposed a n-gram prediction referred to path-based model. However, they did not take short session sequences into consideration although these sequences can be part of their corresponding cluster. Applying only the maximum occurred frequency of next click in WPR is another weakness of this method as it narrows down appropriate choices of visitor when browsing. One more disadvantage of n-gram process is to compromise between precision and applicability.

In [20], [21], a prediction scheme using Web Access Sequence (WAS) clustering was presented, yet it was based on session similarity measure which is not effective in variable length sessions. Wang et al. [22] combined Jaccard index and k-medoids in the HBM clustering algorithm to group correspondence sessions before applying association rules. As this approach does not deal with web pages order in session when clustering, it may not result the optimal clusters for mining by association rules [23]. In [24], Jalali et al. worked with Longest Common Subsequence (LCS) to classify visitors' navigation pattern to forecast and serve their future requests. Nonetheless, LCS has the disadvantage of finding out the succession of common web pages of sessions which plays an important role in session similarity evaluation. For the reason of usability, accuracy and changeability, visitor-profile based suggestions in WPR like [25], [18] that require visitors' input are impractical, since nowadays even cookies may not be available. Yet another shortcoming of this approach is the nature of prediction and recommendation frequently depends on previous requests of the same session, since even the same visitor may hit the site for particular objectives at different times. Alternatively, this type of model is more appropriate to social network or e-learning system. It is noticeable that most of the mentioned works are ambiguous between prediction and recommendation since they confusedly regard the most likely next access as visitor interest.

V. Conclusion

Our web usage prediction and recommendation (WPR) proposes a personalized modeling of web visitor navigation based on prediction and recommendation clusters. Concerning the improved similarity of sessions in prediction clusters, a sequence similarity combination measure was efficiently modified and applied. The experimental result also revealed the correlation between hierarchical clustering criteria and prediction accuracy. The recommendation process exploits prediction cluster to anticipate future behaviour of corresponding visitors, then considers activity time visitors spend on and position of page in sessions to suggest them shortest path to the supposed desirable content. Besides, we discussed the adaptation cost estimation for converting behaviour patterns of visitors to well-defined rules of business, correspondingly to help client users save browsing time. In order to calculate this kind of estimation cost, we used the WPR result and the combination measure to consider the adaptation effort from website design to the proposition.

The results of our preliminary phase indicate an applicable process of WPR. In future work, besides current visit order and activity time, we want to enhance our WPR model by using other visitor behavioral features such as device, browser, operating system types, etc. Obviously, it is necessary to implement a recommender system on the idea and later find out the correctness of adaptation cost estimation. Additionally, the experiments should be extended to additional web sites, with more advanced usage pattern than the used one, to evaluate the WPR model efficiency. We are also interested in web visitor satisfactory measure to enhance the system performance. Last but not least, a performance comparison between our proposition and others is required to better highlight the method advantages.

Acknowledgement

The authors would like to express gratitude for datasets provided by the Beampulse company to test their methods. They also thank VIED and Campus France for their research funding.

References

[1] J. F. Hair Jr, "Knowledge creation in marketing: the role of predictive analytics," *European Business Review*, vol. 19, no. 4, pp. 303–315, 2007.

[2] W. Lee, B.-W. On, I. Lee, and J. Choi, "A big data management system for energy consumption prediction models," in *Digital Information Management (ICDIM), 2014 Ninth International Conference on*. IEEE, 2014, pp. 156–161.

[3] N. N. Chan, W. Gaaloul, and S. Tata, "A recommender system based on historical usage data for web service discovery," *Service Oriented Computing and Applications*, vol. 6, no. 1, pp. 51–63, 2012.

[4] V. Luu, M. Ripken, G. Forestier, F. Fondement, and P. Muller, "Using glocal event alignment for comparing sequences of significantly different lengths," in *International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, 2016, pp. 58–72.

[5] V.-T. Luu, G. Forestier, F. Fondement, and P.-A. Muller, "Web site audience segmentation using hybrid alignment techniques," in *Trends and Applications in Knowledge Discovery and Data Mining*. Springer, 2015, pp. 29–40.

[6] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[7] J. Nielsen, "How long do users stay on web pages," *useit. com: Jakob Nielsens Website*, 2011.

[8] J. Kim, D. W. Oard, and K. Romanik, "User modeling for information filtering based on implicit feedback," 2001.

[9] M. Claypool, P. Le, M. Wased, and D. Brown, "Implicit interest indicators," in *Proceedings of the 6th international conference on Intelligent user interfaces*. ACM, 2001, pp. 33–40.

[10] D. Shen, X. Wang, and H.-L. Chen, "Managing web-based learning resources for k-12 education: lessons learned from web analytics," in *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, vol. 2008, no. 1, 2008, pp. 470–475.

[11] M. S. Chen, J. S. Park, and P. S. Yu, "Data mining for path traversal patterns in a web environment," in *Distributed Computing Systems, 1996., Proceedings of the 16th International Conference on*. IEEE, 1996, pp. 385–392.

[12] D. L. Nkweteyim, "A collaborative filtering approach to predict web pages of interest from navigation patterns of past users within an academic website," Ph.D. dissertation, University of Pittsburgh, 2005.

[13] M. Jalali, N. Mustapha, M. N. Sulaiman, and A. Mamat, "Webpum: A web-based recommendation system to predict user future movements," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6201–6212, 2010.

[14] B. Nigam, S. Tokekar, and S. Jain, "Evaluation of models for predicting user's next request in web usage mining," *international Journal on Cybernetics & informatics (UCi)*, vol. 4, pp. 1–13.

[15] A. Anitha, "A new web usage mining approach for next page access prediction," *International Journal of Computer Applications*, vol. 8, no. 11, pp. 7–10, 2010.

[16] D. Sejal, T. Kamalakant, V. Tejaswi, D. Anvekar, K. Venugopal, S. Iyengar, and L. Patnaik, "Wnpwr: Web navigation prediction framework for webpage recommendation," in *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*. IEEE, 2015, pp. 120–125.

[17] P. Thwe, "Proposed approach for web page access prediction using popularity and similarity based page rank algorithm," *International Journal of Science and Technology Research*, vol. 2, no. 3, 2013.

[18] Q. Liu, Y. Wang, J. Li, Y. Jia, and Y. Ren, "Predicting user likes in online media based on conceptualized social network profiles," in *Web Technologies and Applications*. Springer, 2014, pp. 82–92.

[19] Z. Su, Q. Yang, Y. Lu, and H. Zhang, "Whatnext: A prediction system for web requests using n-gram sequence models," in *Web Information Systems Engineering, 2000. Proceedings of the First International Conference on*, vol. 1. IEEE, 2000, pp. 214–221.

[20] C. Dimopoulos, C. Makris, Y. Panagis, E. Theodoridis, and A. Tsakalidis, "A web page usage prediction scheme using sequence indexing and clustering techniques," *Data & Knowledge Engineering*, vol. 69, no. 4, pp. 371–382, 2010.

[21] B. S. Chordia and K. P. Adhiya, "Grouping web access sequences using sequence alignment method," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 3, pp. 308–314, 2011.

[22] F.-H. Wang and H.-M. Shao, "Effective personalized recommendation based on time-framed navigation clustering and association mining," *Expert systems with applications*, vol. 27, no. 3, pp. 365–377, 2004.

[23] D. Lobo, "Association rules: Normalizing the lift," in *Digital Information Management (ICDIM), 2014 Ninth International Conference on*. IEEE, 2014, pp. 151–155.

[24] M. Jalali, N. Mustapha, M. N. B. Sulaiman, and A. Mamat, "A web usage mining approach based on lcs algorithm in online predicting recommendation systems," in *Information Visualisation, 2008. IV'08. 12th International Conference*. IEEE, 2008, pp. 302–307.

[25] A. Espinosa, M. Regts, J. Tashiro, and M. Vargas-Martin, "Prediction model based on user profile and partial course progress for a digital media learning environment," in *The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications*, 2012, pp. 120–123.