

A Latent Usage Approach for Clustering Web Transaction and Building User Profile

Yanchu Zhang¹, Guandong Xu¹, Xiaofang Zhou²

¹School of Computer Science and Mathematics
Victoria University, PO Box 14428, VIC 8001, Australia
{xu,yzhang}@csm.vu.edu.au

²School of Information Technology & Electrical Engineering
University of Queensland, Brisbane QLD 4072, Australia
zxf@itee.uq.edu.au

Abstract. Web transaction data between web visitors and web functionalities usually convey users' task-oriented behavior patterns. Clustering web transactions, thus, may capture such informative knowledge, in turn, build user profiles, which are associated with different navigational patterns. For some advanced web applications, such as web recommendation or personalization, the aforementioned work is crucial to make web users get their preferred information accurately. On the other hand, the conventional web usage mining techniques for clustering web objects often perform clustering on usage data directly rather than take the underlying semantic relationships among the web objects into account. *Latent Semantic Analysis* (LSA) model is a commonly used approach for capturing semantic associations among co-occurrence observations. In this paper, we propose a LSA-based approach for such purpose. We demonstrated usability and scalability of the proposed approach through performing experiments on two real world datasets. The experimental results have validated the method's effectiveness in comparison with some previous studies.

1 Introduction

With the popularizing and spreading of Internet application, Web has recently become a powerful platform for, not only retrieving information, but also discovering knowledge, from web data repository. Generally, web users may exhibit various types of behaviors associated with their information needs and intended tasks when they are traversing the Web. These task-oriented behaviors are explicitly characterized by sequences of clicks on different web items performed by users. As a result, these tasks are implicitly captured by inducing the underlying relationships among the click-stream data. For example, image a web site designed for information about automobiles; there will be a variety of customer groups with various access interests during their visiting such an E-commerce website. One type of customers intends to make comparison prior to shopping,; a visitor planning to purchase particular type car of wagon, for example, would have to browse the web pages of each manufacturer, compare their offering,, whereas another one will just be more interested in one specific brand car, such as "Ford", rather than one specific car category. In such circum-

stance, these two visitors with different interests may follow distinct access tracks to accomplish their goals and corresponding clickstream data are recorded in web sever log file as well. As a result, mining web log information may reveal user access patterns. Moreover, the discovered informative knowledge (or pattern) will be utilized for providing better web applications or web services, such as web recommendation or personalization.

Generally, web mining techniques can be defined as those methods to extract so-called “nuggets” (or knowledge) from web data repository, such as content, linkage, usage information, by utilizing data mining tools. Among such web data, user clickstream, i.e. usage data, can be mainly utilized to capture users’ navigational patterns and identify user intended tasks. Once the user navigational behaviors are effectively characterized, they will provide benefits for further web applications, in turn, facilitate and improve web service quality for both web-based organizations and for end users. As a result, web usage mining recently has become one more active and hotter topic, and a variety of research communities from database management, artificial intelligence and information systems etc., have addressed this topic and achieved great success as well [1-7]. Meanwhile, with the benefits of great progress in data mining research, many data mining techniques, such as clustering[3, 8, 9] association rule mining [10, 11] and sequential pattern mining [12] are adopted widely to improve the usability and scalability of web mining.

Related work: In general, there are two types of clustering methods performed on the usage data: user transaction clustering and web page clustering [13]. One successful application of web page clustering is adaptive web site. For example, the algorithm called PageGather [3, 14] is proposed to synthesize index pages that are not existing initially, based on partitioning web pages into various groups. The generated index page is conceptually representing the various access interests of users according to their navigational history. Another example is that clustering user rating results has been successfully adopted in collaborative filtering application as a data preparing step to improve the scalability of recommendation using *K-Nearest-Neighbor* (KNN) algorithm [15]. Mobasher et al. [9] utilize user transaction and pageview clustering techniques, which is employing traditional *k*-means clustering algorithm to characterize user access pattern for web personalization based on mining web usage data. These proposed clustering-based techniques have been proven to be efficient from their experimental results since they are really capable of identifying the intrinsic common attributes revealed from their recently historic clickstream data. Generally, these usage patterns are explicitly captured at the level of user transaction or pageview. They, however, do not reveal the underlying characteristics of user navigational activities as well as web pageview. For example, such discovered usage patterns provide little information of why such web transactions or web pages are partitioning together, and latent relationships among the co-occurrence observation data have not been incorporate into the mining process as well. Thus, it is needed to develop LSA-based approaches that can reveal not only common trends explicitly, but also take the latent information into account implicitly during mining. In [16], an algorithm based on *Principal Factor Analysis* (PFA) model derived from statistical analysis, is proposed to generate user access pattern and uncover latent factor by clustering user transactions and analyzing principal factor involved in web usage mining. Analogous, some works [17-19] are addressed to derive user access patterns and web page

segments from various types of web data, by utilizing a so-called *Probabilistic Semantic Latent Analysis* (PLSA) model, which is based on maximum likelihood principle from statistics.

Our approach: In this paper, we address these issues by proposing another alternative LSA-based approach for clustering web transaction and generating user profile. After data preprocessing, we produce a user transaction collection and a pageview corpus via user and pageview identification process respectively, in turn, construct the session-pageview matrix as usage data, in which each cell is expressed by a weight representing the contribution made by a specific pageview during one user transaction. In this manner, we could map the relationships among the co-occurrence observations (i.e. user transactions) into a high-dimensional space. Moreover, an improved LSA-based clustering algorithm, named latent usage information (LUI), is proposed to find out user segments with similar behaviors effectively and precisely from aforementioned usage data by using linear algebra theory, especially single value decomposition of matrix due to revealing deeper relationships among web transactions. The discovered user clusters are exploited to generate a variety of goal-oriented user profiles by calculating the centroid of corresponding cluster in the form of weighted pageview set. Experiments are conducted on two real world datasets to validate the usability and scalability of usage mining. Meanwhile, an evaluation metric is adopted to assess the quality of discovered clusters, and comparisons are made with some previous work as well. The experimental results have shown that the proposed approach is capable of effectively discovering user access pattern and revealing the underlying relationships among user visiting records.

The remainder of paper is organized as follows: in section 2, we briefly discuss how to construct session-pageview matrix during data preparation. Section 3 gives the latent usage information (LUI) algorithm. Since the LUI algorithm is based on linear algebra theory, especially the Single Value Decomposition (SVD) of a matrix, some basic background knowledge of SVD is provided in this section as well. In section 4, some experimental results derived on real world datasets are presented and comparisons with previous study are discussed as well. Finally, we conclude and give future works in section 5.

2 Latent Usage Information (LUI) Model

We start with collecting the raw web sever logs of the site and perform data cleaning, pageview identification, and user identification such data preparation measures to construct the co-occurrence observation. More detailed introduction of data preparation steps could be found in [20]. At this stage, we briefly introduce how to build up the session-pageview matrix for web usage mining

2.1 Usage Data Identification

Basically, according to W3C definition, a web pageview can be viewed as a visual rendering of a web page. In this way, the user access interest exhibited may be reflected by the varying degree of visits in different web pages during one session.

Thus, we can represent a user session as a collection of transactions, which includes a series of weighted pageviews, during the visiting period. In other words, the user session can be expressed in the form of pageview vectors. From such viewing point, we generate the following user session expression. Given n web pages in a web site and m web users visiting the web site during a period of time, after appropriate data preprocessing such as page identification and user sessionization, we built up the pageview corpus as $P = \{p_1, p_2, \dots, p_n\}$, and user session collection as $S = \{s_1, s_2, \dots, s_m\}$. Conceptually, modeling of user session in a collection of pages defined by the so-called web page corpus, which consists of all web items visited by whole users, is similar to modeling a document in terms of word frequencies by using a word dictionary in text IR. In short, each user session can be, in turn, expressed as a set of weight-pageview pairs, $s_i = \{ \langle p_1, a_{i1} \rangle, \langle p_2, a_{i2} \rangle, \dots, \langle p_n, a_{in} \rangle \}$. By simplifying the above expression in the form of pageview vector, each user session can be considered as an n -dimensional vector $s_i = \{a_{i1}, a_{i2}, \dots, a_{in}\}$, where a_{ij} denotes the weight for pageview p_j in s_i user session. As a result, the whole user session data can be utilized to form web usage data represented by a session-pageview matrix $SP_{m \times n} = \{a_{ij}\}$ (Figure 1 illustrates the skeletal structure of session-page matrix).

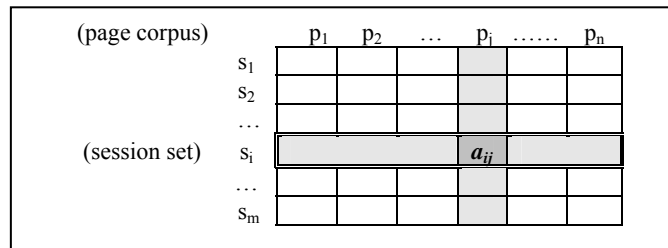


Fig. 1. Skeletal structure of session-pageview matrix

```

202.161.108.167 - - [01/Feb/2003:00:00:03 +1100]
"GET/timetables/city/2003s1/cc 4logo.gif HTTP/1.1" 206
14102 "http://www.cs.rmit.edu.au/timetables/city/2003s1/
cover.html "Mozilla/4.0 (compatible; MSIE 5.5; Windows 98)"

213.183.13.65 - - [01/Feb/2003:00:00:16 +1100]
"GET/~winikoff/palm/dev.html HTTP/1.1" 302 244
"http://www.google.de/search?q=sources+onboardc+examples&ie
=UTF-8&oe=UTF-8&hl=de&meta=" Scooter/3.3"

```

Fig. 2. Sample of web access log

The cell value in the session-page matrix, a_{ij} , can be represented by a weight associated with the contribution of page p_j in the user session s_i , which is usually determined by the number of hit or the amount time spent by specific user on the corresponding page. Generally, in order to eliminate the influence caused by the relative amount difference of visiting time duration or hit number, the normalization manipulation across pageviews space in same user session is performed. Figure 2 illustrates two items retrieved from a web access log, in which each field is separated by space.

Particularly, note that the first and fourth fields are identified as the visitor IP address and the requested URL respectively, and utilized to assist usage data collection.

2.2 Latent Usage Information Algorithm

Once usage matrix is constructed, we may applying conventional clustering on user transaction data to classify user sessions into various groups, within which the classified sessions share both common access interest exhibited from their visiting records. It is intuitive to perform clustering algorithm directly on each row vector of usage matrix to determine the relative “close” session cluster by using similarity-based measure, such as commonly adopted cosine similarity from Information Retrieval. In [9], an algorithm named PACT is proposed based on the above discussed technique. However, this kind of clustering technique only capture the mutual relationships between session data explicitly, it is incapable of revealing the “deeper” underlying characteristics of usage pattern. In this work, we propose the latent usage information (LUI) algorithm to group user sessions semantically through taking latent information into account. For better understanding the LUI algorithm, we first discuss some theoretical background of the SVD algorithm.

• Single Value Decomposition Algorithm

The SVD definition of a matrix is illustrated as follows[21]: For a real matrix $A=[a_{ij}]_{m \times n}$, without loss of generality, suppose $m \geq n$ and there exists SVD of A:

$$A = U_{m \times m} \sum_{m \times n} V_{n \times n}$$

where U and V are orthogonal matrices. Matrices U and V can be respectively denoted as $U_{m \times m}=[u_1, u_2, \dots, u_m]_{m \times m}$ and $V_{n \times n}=[v_1, v_2, \dots, v_n]_{n \times n}$, where u_i ($i=1, \dots, m$) is a m -dimensional vector $u_i=(u_{1i}, u_{2i}, \dots, u_{mi})^T$ and v_j ($j=1, \dots, n$) is a n -dimensional matrix $v_j=(v_{1j}, v_{2j}, \dots, v_{nj})^T$. Suppose $\text{rank}(A)=r$ and single values of A are diagonal elements of \sum as follows:

$$\sigma_1 \geq \sigma_2 \geq \dots \sigma_r \geq \sigma_{r+1} = \dots = \sigma_n = 0.$$

For a given threshold ε ($0 < \varepsilon \leq 1$), we choose a parameter k such that $(\sigma_k - \sigma_{k+1})/\sigma_k \geq \varepsilon$. Then, we denote $U_k=[u_1, u_2, \dots, u_k]_{m \times k}$, $V_k=[v_1, v_2, \dots, v_k]_{n \times k}$, $\sum_k=\text{diag}(\sigma_1, \sigma_2, \dots, \sigma_k)$, and $A_k=U_k \sum_k V_k$

As known from the theorem in algebra [21], A_k is the best approximation matrix to A and conveys main and latent information among the usage data. This property makes it possible to find out relative “close” user session at the semantic latent level based on their mutual similarity.

• Representation of User Transaction in Latent Space

Once SVD implementation is completed, we may rewrite user sessions with the obtained approximation matrix U_k , \sum_k and V_k and map them into another k -dimensional latent space. For a given session, it is represented as a coordinate vector with respect to pageviews: $s_i = \{a_{i1}, a_{i2}, \dots, a_{in}\}$. The projection of coordinate vector s_i in the k -dimensional latent subspace is reparameterize as

$$s'_i = s_i V_k \sum_k = (t_{i1}, t_{i2}, \dots, t_{ik}) \quad (1)$$

where $t_{ij} = \sum_{k=1}^n a_{ik} v_{kj} \sigma_j$, $j = 1, 2, \dots, k$.

- **Similarity Measure**

We adopt traditional Cosine similarity to capture common interests shared by user sessions, i.e. for two vectors $x=(x_1, x_2, \dots, x_k)$ and $y=(y_1, y_2, \dots, y_k)$ in k -dimensional space, the similarity between them is defined as

$$\text{sim}(x, y) = (x \bullet y) / (\|x\|_2 \|y\|_2), \text{ where } x \bullet y = \sum_{i=1}^k x_i y_i, \|x\|_2 = \sqrt{\sum_{i=1}^k x_i^2}.$$

In this manner, the similarity between two user sessions is defined as:

$$\text{sim}(s'_i, s'_j) = \frac{(s'_i \bullet s'_j)}{\|s'_i\|_2 \|s'_j\|_2} \quad (2)$$

3 Constructing User Profile based on Latent Usage Information

In this section, we present the algorithms for clustering web transaction and generating user profile based on the discovered clusters as well.

3.1 Clustering Web Transaction

Here we adopt a modified standard K -means clustering algorithm, named MK -means clustering, to classify user session based on the transformed SP matrix over the latent k -dimensional space. This algorithm does not need to predefine value k and k initial centroids, whereas the standard k -means has to do so to start clustering. The algorithm is described as follows:

Algorithm: MK -means clustering

Input: usage data SP' and similarity threshold ε

1. Choose the first user session s_i' as the initial cluster C_1 and centroid of this cluster, i.e. $C_1 = \{s_i'\}$ and $Cid_1 = s_i'$.
2. For each session s_i' , calculate the similarity between s_i' and the centroids of other existing cluster $\text{sim}(s_i', Cid_j)$.
3. if $\text{sim}(s_i', Cid_k) = \max_j (\text{sim}(s_i', Cid_j)) > \varepsilon$, then allocate s_i' into C_k and recalculate the centroid of cluster C_k as $Cid_k = 1/|C_k| \sum_{j \in C_k} s_j'$;
4. Otherwise, let s_i' itself construct a new cluster and be the centroid of this cluster.
5. Repeat step 2 to 4 until all user sessions are processed and all centroids do not update any more.

Output: cluster set $CS=\{C_k\}$

3.2 Building User Profile

As we mentioned above, each user session is represented as a weight-based pageview vector. In this way, it is reasonable to derive the centroid of cluster obtained by aforementioned algorithm as the user profile. In this work, we compute the mean vector to represent the centroid. For each session cluster $C_k \in CS$, the value for each pageview in the mean vector is determined by the ratio of the sum of pageview weights in C_k to the number of sessions in the cluster. In order to eliminate the diversity in visiting quantity of each session, the weights are normalized in calculating the centroid of cluster. Thus, the maximum weight in user profile is updated to be 1, whereas other weights are divided by the maximum weight across session. Meanwhile, some low-contribution pageviews (i.e. those with mean weights below one certain limit) are filtered out. The algorithm for constructing user profile is as follows:

1. For each pageview in cluster, we compute the mean value of pageview as

$$wt(p, pf) = 1/|C_k| \sum_{s \in C_k} w(p, s) \quad (3)$$

where $w(p, s)$ is the weight of pageview p in session $s \in C_k$.

2. For each cluster, furthermore, we construct its mean vector (i.e. centroid) as

$$mv_C = \{ \langle p, wt(p, pf) \rangle \mid p \in P \} \quad (4)$$

3. For each pageview weight within user profile, if the value $<$ threshold μ , the corresponding item will be removed from the vector with its weight, otherwise keep it leave.
4. Sort the pageviews with their weights in descending order and output the mean vector as user profile.

$$pf_{c_k} = \{ \langle p_{1k}, wt(p_{1k}, pf) \rangle, \langle p_{2k}, wt(p_{2k}, pf) \rangle, \dots, \langle p_{tk}, wt(p_{tk}, pf) \rangle \} \quad (5)$$

where

$$wt(p_{1k}, pf) > wt(p_{2k}, pf) > \dots > wt(p_{tk}, pf) > \mu, PF = \{ pf_{c_k} \}, k = 1, 2, \dots, t.$$

4 Experiment and Evaluation

In order to evaluate the effectiveness of the proposed LUI-based clustering algorithm and user profile generating algorithm, and explore the discovered user access pattern, we conducted preliminary experiments on two real world data sets. Some comparisons with previous work are made as well.

4.1 Data sets

The first dataset used is downloaded from KDDCUP (www.ecn.purdue.edu/kddcup/). The data set is common-used data resource provided to test and compare methods (prediction algorithm, clustering approaches, etc.) for data mining purpose. Data preprocessing is needed to perform on the raw data set since there are some short user sessions existing in the data set, which mean they are of less contribution for data mining. Support filtering technique is used to eliminate these user sessions, leaving only sessions with at least four pages. After data preparation, we have setup an data set including 9308 user sessions and 69 pages, where every session consists of 11.88 pages in average. We refer this data set to “KDDCUP data”. In this data set, the entries in session-page matrix associated with the specific page in the given session are determined by the numbers of web page hits by the given user.

The second data set is from a university website log files and was made available by the author of [13]. The data is based on a random collection of users visiting this site for a 2-week period during April of 2002. After data preprocessing, the filtered data contains 13745 sessions and 683 pages. This data file is expressed as a session-page matrix where each column is a page and each row is a session represented as a vector. The entries in the table correspond to the amount of time (in seconds) spent on pages during a given session. For convenience, we refer this data as “CTI data”. For each dataset, we randomly choose 1000 transaction as the evaluation set, whereas the remainder part is selected as the training set for constructing user profiles.

4.2 Results of generated user profiles

We utilize aforementioned LUI method to classify user transactions. For comparison purpose, we also perform PACT approach based on standard K-means used in [9] to generate user profiles. From the results, it is found that generated profiles are “overlapping” of pageviews since some pageviews are listed in more than one user clusters. Table 1 depicts 2 user profiles generated from KDD dataset using LUI approach. Each user profile is listed in a ordered pageviews’ sequence with weights, which means the greater weight of a pageview contribute, the more likely it is to be visited. The first profile in Table 1 represents the activities involved in online-shopping circumstance such as login, shopping_cart, checkout etc., especially occurring in purchasing leg-wear products, whereas second user profile reflects customers’ concern focused on the interests with regard to the department store itself.

Analogously, some informative finding can be obtained in Table 2, which is derived from CTI dataset. In this table, three profiles are generated: the first one reflects the main topic of international student concerning issues regarding applying for admission, and second one involves in the online applying process for graduation, whereas the final one indicates the most common activities happened during students browsing the university website, especially while they are determining course selection, i.e. selecting course, searching syllabus list, and then going through specific syllabus.

Table 1. Examples of generated user profiles from KDD dataset

Pageview #	Pageview content	weight
29	Main-shopping_cart	1.00
4	Products-productDetailleagwear	0.86
27	Main-Login2	0.67
8	Main-home	0.53
44	Check-expressCheckout	0.38
65	Main-welcome	0.33
32	Main-registration	0.32
45	Checkout-confirm_order	0.26

Pageview #	Pageview content	weight
11	Main-vendor2	1.00
8	Main-home	0.40
12	Articles-dpt_about	0.34
13	Articles-dpt_about_mgmtteam	0.15
14	Articles-dpt_about_broadofdirectors	0.11

Table 2. Examples of generated user profiles from CTI dataset

Pageview #	Pageview content	weight
19	Admissions-requirement	1.00
3	Admissions-costs	0.41
15	Admissions-intrnational	0.24
13	Admissions-I20visa	0.21
387	Homepage	0.11
0	Admission	0.11

Pageview #	Pageview content	weight
349	Gradapp-tologin	1.00
20	Admissions-statuscheck	0.35
340	Gradapp-login	0.32
333	Gradapp-appstat_shell	0.13
0	Admissions	0.11

Pageview #	Pageview content	weight
387	Homepage	1.00
59	Courses	0.78
71	Course-syllabilist	0.40
661	Program-course	0.17
72	Course-syllabisearch	0.12

4.3 Evaluation of user transaction clusters

In order to evaluate the quality of clusters derived from LUI approach, we adopt one specific metric, named the *Weighted Average Visit Percentage (WAVP)*[9]. This evaluation method is based on assessing each user profile individually according to the likelihood that a user session which contains any pageviews in the transaction cluster will include the rest pageviews in the cluster during the same session. The

principle of WAVP metric is discussed as follows: suppose T is one of transaction set within the evaluation set, and for s specific cluster C , let T_c denote a subset of T whose elements contain at least one pageview from C . Moreover, the weighted average visit percentage of T_c may conceptually be determined by the similarity between T_c and the cluster C if we consider the T_c and C as in the form of pageview vector. As a result, the WAVP is computed as:

$$WAVP = \left(\sum_{t \in T_c} \frac{\vec{t} \cdot \vec{C}}{|T_c|} \right) / \left(\sum_{p \in Pf} wt(p, pf) \right) \quad (6)$$

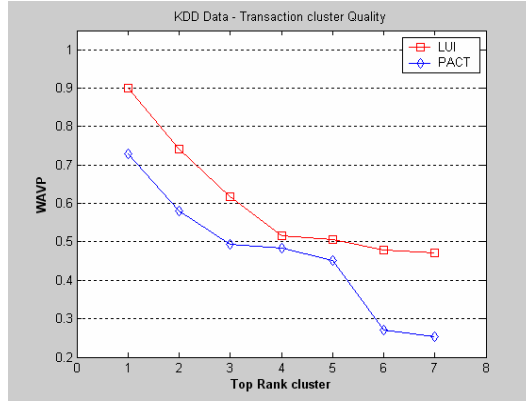


Fig. 3. User cluster quality analysis results upon WAVP comparison for KDD dataset

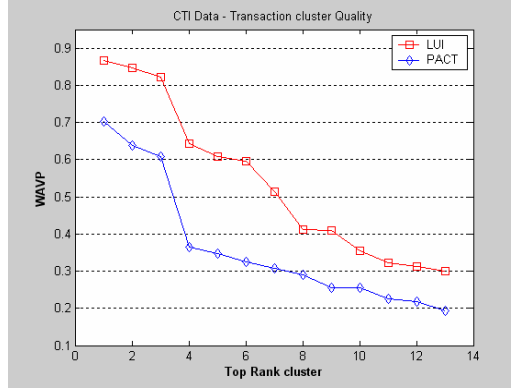


Fig. 4. User cluster quality analysis results upon WAVP comparison for CTI dataset

From the definition of WAVP, it is known that the higher WAVP value is, the better quality of obtained transaction cluster possesses.

As mentioned above, for comparison purpose, we conduct data simulations upon two real world datasets by using two approaches. Figure 3 and Figure 4 depict the comparison results of WAVP values for KDD and CTI datasets with PACT respectively. In each figure, the obtained user profiles are arrayed in the descending rank according to their WAVP values, which reflect the quality of various clustering algo-

rithms. From these two curve lines, it is easily concluded that the proposed LUI-based technique overweighs standard K-means based algorithm in term of WAVP parameter. Moreover, LUI approach is capable of capturing the latent relationships among user transaction and discovering user profiles representing the actual navigational behaviors more effectively and accurately.

5 Conclusion and Future Work

In this paper, we proposed a LSA-based approach, named LUI, for grouping web transaction and generating user profile. Firstly, we mapped the relationships among the co-occurrence observations (i.e. user transactions) into a high-dimensional space to construct the usage data in the form of session-pageview matrix. Then, an dimension reducing algorithm (i.e. single value decomposition) was employed on the usage matrix to capture the latent usage information for partitioning user transaction. Based on the decomposed latent usage information, we proposed a modified k -means clustering algorithm to generate user session clusters. Moreover, the discovered user groups are utilized to construct user profiles expressed in the form of a weighted pageview collection, which represents the common usage pattern associated with one kind of specific visitors' access interests. The constructed user profiles corresponding to various task-oriented behaviors are represented as a set of pageview-weight pairs' collection, in which each weight reflects the significance contributed by the page. Experiments are conducted on two real world datasets to validate the usability and scalability of usage mining. Meanwhile, an evaluation metric is adopted to assess the quality of discovered clusters, and comparisons are made with some previous works as well. The experimental results have shown that the proposed approach is capable of effectively discovering user access pattern and revealing the underlying relationships among user visiting records as well.

The future works will be focused on the research issues, such as performing experiments over more datasets, broadening comparison and make use of discovered user profiles for further web application, for example, web recommendation and personalization.

6 Acknowledgement

This research has been partly supported through ARC Discovery Project Grant (DP0345710) and National Natural Science Foundation of China (No 60403002).

Reference

1. Joachims, T., D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the *world wide web*. in *The 15th International Joint Conference on Artificial Intelligence (ICJAI'97)*. 1997. Nagoya, Japan.

2. Lieberman, H. *Letizia: An agent that assists web browsing*. in *Proc. of the 1995 International Joint Conference on Artificial Intelligence*. 1995. Montreal, Canada: Morgan Kaufmann.
3. Perkowitz, M. and O. Etzioni. *Adaptive Web Sites: Automatically Synthesizing Web Pages*. in *Proceedings of the 15th National Conference on Artificial Intelligence*. 1998. Madison, WI: AAAI.
4. Ngu, D.S.W. and X. Wu. *Sitehelper: A localized agent that helps incremental exploration of the world wide web*. in *Proceedings of 6th International World Wide Web Conference*. 1997. Santa Clara, CA: ACM Press.
5. Cohen, E., B. Krishnamurthy, and J. Rexford. *Improving end-to-end performance of the web using server volumes and proxy lters*. in *Proc. of the ACM SIGCOMM '98*. 1998. Vancouver, British Columbia, Canada: ACM Press.
6. Büchner, A.G. and M.D. Mulvenna, *Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining*. SIGMOD Record, 1998. **27**(4): p. 54-61.
7. Mobasher, B., R. Cooley, and J. Srivastava. *Creating adaptive web sites through usage-based clustering of URLs*. in *Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*. 1999: IEEE Computer Society.
8. Han, E., et al., *Hypergraph Based Clustering in High-Dimensional Data Sets: A Summary of Results*. IEEE Data Engineering Bulletin, 1998. **21**(1): p. 15-22.
9. Mobasher, B., et al., *Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization*. Data Mining and Knowledge Discovery, 2002. **6**(1): p. 61-82.
10. Agarwal, R., C. Aggarwal, and V. Prasad, *A Tree Projection Algorithm for Generation of Frequent Itemsets*. Journal of Parallel and Distributed Computing, 1999. **61**(3): p. 350-371.
11. Agrawal, R. and R. Srikant. *Jorge B. Bocca and Matthias Jarke and Carlo Zaniolo*. in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*. 1994. Santiago, Chile: Morgan Kaufmann.
12. Agrawal, R. and R. Srikant. *Mining Sequential Patterns*. in *Proceedings of the International Conference on Data Engineering (ICDE)*. 1995. Taipei, Taiwan: IEEE Computer Society Press.
13. Mobasher, B., *Web Usage Mining and Personalization*, in *Practical Handbook of Internet Computing*, M.P. Singh, Editor. 2004, CRC Press.
14. Perkowitz, M. and O. Etzioni, *Adaptive Web sites*. Communications of the ACM, 2000. **43**(8): p. 152 - 158.
15. O'Conner, M. and J. Herlocker. *Clustering Items for Collaborative Filtering*. in *Proceedings of the ACM SIGIR Workshop on Recommender Systems*. 1999. Berkeley, CA: ACM Press.
16. Zhou, Y., X. Jin, and B. Mobasher. *A Recommendation Model Based on Latent Principal Factors in Web Navigation Data*. in *Proceedings of the 3rd International Workshop on Web Dynamics*. 2004. New York: ACM Press.
17. Xu, G., et al. *Discovering User Access Pattern Based on Probabilistic Latent Factor Model*. in *Proceeding of 16th Australasian Database Conference*. 2004. Newcastle, Australia: ACS Inc.
18. Xu, G., Y. Zhang, and X. Zhou. *Using Probabilistic Semantic Latent Analysis for Web Page Grouping*. in *15th International Workshop on Research Issues on Data Engineering: Stream Data Mining and Applications (RIDE-SDMA'2005)*. 2005. Tyoko, Japan.
19. Jin, X., Y. Zhou, and B. Mobasher. *A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content*. in *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04)*. 2004. San Jose.
20. Cooley, R., B. Mobasher, and J. Srivastava, *Data Preparation for Mining World Wide Web Browsing Patterns*. Journal of Knowledge and Information Systems, 1999. **1**(1): p. 5-32.
21. Datta, B.N., *Numerical Linear Algebra and Application*. 1995: Brooks/Cole Publishing Company.