

A Web Recommendation Technique Based on Probabilistic Latent Semantic Analysis

Guandong Xu¹, Yanchun Zhang¹, and 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 user task-oriented behavior pattern. Mining such type of clickstream data will lead to capture usage pattern information. Nowadays Web usage mining technique has become one of most widely used methods for Web recommendation, which customizes Web content to user-preferred style. Traditional techniques of Web usage mining, such as Web user session or Web page clustering, association rule and frequent navigational path mining can only discover usage pattern explicitly. They, however, cannot reveal the underlying navigational activities and identify the latent relationships that are associated with the patterns among Web users as well as Web pages. In this work, we propose a Web recommendation framework incorporating Web usage mining technique based on *Probabilistic Latent Semantic Analysis* (PLSA) model. The main advantages of this method are, not only to discover usage-based access pattern, but also to reveal the underlying latent factor as well. With the discovered user access pattern, we then present user more interested content via collaborative recommendation. To validate the effectiveness of proposed approach, we conduct experiments on real world datasets and make comparisons with some existing traditional techniques. The preliminary experimental results demonstrate the usability of the proposed approach.

1 Introduction

With the popularizing and spreading of Internet applications, Web has recently become a powerful data repository for, not only retrieving information, but also discovering knowledge. 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 objects (i.e. Web pages) performed by users. As a result, extracting the underlying usage pattern among the clickstream data is able to capture these interest-oriented tasks implicitly. 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 one 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 scenario, these two visitors who exhibit different interests may follow distinct access tracks to accomplish their goals, and thus, corresponding click-stream data are recorded in Web sever log file accordingly. Mining Web log information, thus, will lead to reveal user access pattern. Moreover, the discovered informative knowledge (or patterns) will be utilized for providing better Web application, such as Web recommendation or personalization. Generally, Web recommendation can be viewed as a process that recommends customized Web presentation or predicts tailored Web content to users according to their specific tastes or preferences.

Related Work: With the significant development in data mining domain, many advanced techniques, such as k-Nearest Neighbor (*kNN*) algorithm [1-3], Web clustering [4-6], association rule mining [7, 8] and sequential pattern mining [9] are widely utilized to address Web usage mining recently. The successful progress shows that it will, not only, benefit Web structure and presentation design, e.g. Adaptive Web Design, but also, improve the quality of Web applications, such as practical Web personalization and recommendation systems [10-13].

To-date, there are two kinds of approaches commonly used in Web recommendation, namely content-based filtering and collaborative filtering systems [14, 15]. Content-based filtering systems such as WebWatcher [16] and client-side agent Letizia [11] generally generate recommendation based on the pre-constructed user profiles by measuring the similarity of Web content to these profiles, while collaborative filtering systems make recommendation by utilizing the rating of current user for objects via referring other users' preference that is closely similar to current one. Today collaborative filtering systems have been widely adopted in Web recommendation applications and have achieved great success as well [1-3]. In addition, Web usage mining has been proposed as an alternative method for Web recommendation recently [5]. The discovered usage pattern is utilized to determine user access interest, in turn, make collaborative recommendation efficiently.

On the other hand, *Latent Semantic Analysis* (LSA) is an approach to capture the latent or hidden semantic relationships among co-occurrence activities [17]. In practical applications, Single Value Decomposition (SVD) or Primary Component Analysis (PCA) algorithm is employed to generate a reduced latent space, which is the best approximation of original vector space and reserves the main latent information among the co-occurrence activities. [17-19]. LSA has been widely used in information indexing and retrieval applications [18, 20], Web linkage analysis [21, 22] and Web page clustering [23]. Although LSA has achieved great success in some applications, it still has some shortcomings [24]. *Probabilistic Latent Semantic Analysis* (PLSA) is a probabilistic variant of LSA based on maximum likelihood principle. Recently, approaches based on PLSA has been successfully applied in collaborative filtering [25] Web usage mining [26], text learning and mining [27, 28], co-citation analysis [28, 29] and related topics.

Our Approach: In this paper, we propose a Web recommendation framework based on PLSA model. The Web recommendation process exploits the usage pat-

tern derived from Web usage mining to predict user preferred content and customize the presentation. By employing PLSA model on usage data, which is expressed as a page weight matrix, we can not only characterize the underlying relationships among Web access observation but also identify the latent semantic factors that are considered to represent the navigational tasks of users during their browsing period. Such relationships are determined by probability inference, and then are utilized to discover the usage-based access pattern. Furthermore, we make use of these discovered access pattern knowledge for Web recommendation by finding the most similar user access pattern to active user and predicting the preferred content based on the matched pattern.

The main contributions in this work are as follows: firstly, we present a Web recommendation unified framework incorporating Web usage mining technique based on PLSA model. Secondly, we investigate the discovery of user access patterns and latent factors related to these patterns via employing probability inference process, in turn, make use of the discovered usage knowledge for Web recommendation. Particularly, we develop a modified *k-means* clustering algorithm on the transformed session vectors and build up user access patterns in terms of centroids of generated session clusters, which reflect common navigational interests in same user category. Finally, we demonstrate the usability and effectiveness of the proposed model by conducting experiments on two real world datasets. The evaluation results show that usage-based approach is capable of predicting user preferred content more accurately and efficiently in comparison with some traditional techniques.

The rest of the paper is organized as follows. In section 2, we introduce the PLSA model. We present the algorithms for discovering latent factors, Web page categories in section 3. In section 4, we concentrate on how to construct usage-based user access pattern and Web recommendation model upon the discovered usage knowledge as well. To validate the proposed approach, we demonstrate experiment and comparison results conducted on two real world datasets in section 5, and conclude the paper in section 6.

2 Probabilistic Latent Semantic Analysis (PLSA) Model

2.1 Data Sessionization Process

Prior to introduce the principle of PLSA model, we discuss briefly the issue with respect to sessionization process of usage data. In general, the user access interests exhibited may be reflected by the varying degrees of visits in different Web pages during one session. Thus, we can represent a user session as a weighted page vector visited by user during a period. After data preprocessing, we can built up a page set of size n as $P = \{p_1, p_2, \dots, p_n\}$ and user session set of size m as $S = \{s_1, s_2, \dots, s_m\}$. The whole procedures are called page identification and user sessionization respectively. By simplifying user session in the form of page vector, each session can be considered as an n -dimensional page vector $s_i = \{a_{i1}, a_{i2}, \dots, a_{in}\}$, where a_{ij} denotes the weight for page p_j in s_i session.

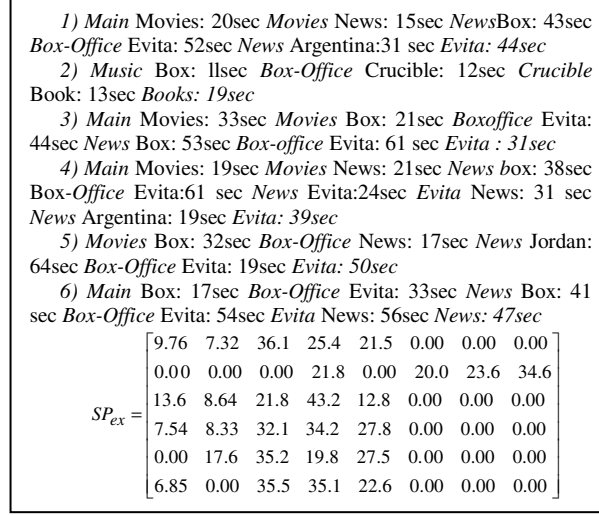


Fig. 1. A usage snapshot and its normalized session-page matrix expression

As a result, the user session data can be formed as Web usage data represented by a session-page matrix $SP = \{a_{ij}\}$. The entry in the session-page matrix, a_{ij} , is the weight associated with the page p_j in the user session s_i , which is usually determined by the number of hit or the amount time spent on the specific page. Generally, the weight a_{ij} associated with page p_j in the session s_i should be normalized across pages in same user session in order to eliminate the influence caused by the amount difference of visiting time durations or hit numbers. For example, Figure 1 depicts an usage data snapshot and its corresponding session-page matrix in the form of normalized weight matrix from [30, 31]. The element in the matrix is determined by the ratio of the visiting time on corresponding page to total visiting time, e.g. $a_{11} = 15/(15 + 43 + 52 + 31 + 44) * 100 = 9.7 \dots$ and so on.

2.2 PLSA Model

The PLSA model is based on a statistic model called aspect model, which can be utilized to identify the hidden semantic relationships among general co-occurrence activities. Similarly, we can conceptually view the user sessions over Web page space as co-occurrence activities in the context of Web usage mining to discover the latent usage pattern. For the given aspect model, suppose that there is a latent factor space $Z = \{z_1, z_2, \dots, z_k\}$ and each co-occurrence observation data (s_i, p_j) is associated with the factor $z_k \in Z$ by varying degree to z_k .

According to the viewpoint of aspect model, thus, it can be inferred that there are existing different relationships among Web users or pages related to different factors, and the factors can be considered to represent the user access patterns. In this manner,

each observation data (s_i, p_j) can convey the user navigational interests over the k -dimensional latent factor space. The degrees to which such relationships are “explained” by each factor derived from the factor-conditional probabilities. Our goal is to discover the underlying factors and characterize associated factor-conditional probabilities accordingly.

By combining probability definition and Bayesian rule, we can model the probability of an observation data (s_i, p_j) by adopting the latent factor variable z_k as:

$$P(s_i, p_j) = \sum_{z_k \in Z} P(z_k) \bullet P(s_i | z_k) \bullet P(p_j | z_k) \quad (1)$$

Furthermore, the total likelihood of the observation is determined as

$$L_i = \sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet \log P(s_i, p_j) \quad (2)$$

where $m(s_i, p_j)$ is the element of the session-page matrix corresponding to session s_i and page p_j .

In order to estimate the desired probabilities, we utilize Expectation Maximization (EM) algorithm to perform maximum likelihood estimation in latent variable model [32]. Generally, two steps are needed to implement in this algorithm alternately: (1) Expectation (E) step where posterior probabilities are calculated for the latent factors based on the current estimates of conditional probability; and (2) Maximization (M) step, where the estimated conditional probabilities are updated and used to maximize the likelihood based on the posterior probabilities computed in the previous E-step. Iterating of E-step and M-step will result in the monotonically increasing of total likelihood L_i until a local optimal limit is reaching, which means the estimated results can represent the final probabilities of observation data. More details regarding EM algorithm is referred in [33]

It is easily found that the computational complexity of this algorithm is $O(mnk)$, where m is the number of user session, n is the number of page, and k is the number of factors.

3 Discovery of Latent Factors, Usage-Based Web Page Categories

Those probabilities generated in section 2 quantitatively measure the underlying relationships among Web users, pages as well as latent factors. We, thus, utilize the class-conditional probability estimates and clustering algorithm to identify user access interests, partition Web pages and user sessions into various usage-based categories.

3.1 Characterizing Latent Factor

First, we discuss how to capture the latent factor associated with user navigational behavior by characterizing the “dominant” pages. Note that $P(p_j | z_k)$ represents the conditional occurrence probability over the page space corresponding to a specific

factor, whereas $P(z_k | p_j)$ represents the conditional probability distribution over the factor space corresponding to a specific page. In such case, we may consider that the pages whose conditional probabilities $P(z_k | p_j)$ and $P(p_j | z_k)$ are both greater than a predefined threshold μ can be viewed to contribute to one particular functionality related to the latent factor. By exploring and interpreting the content of these pages satisfying aforementioned condition, we may characterize the semantic meaning of each factor. The algorithm to characterize the task-oriented semantic latent factor is described as follows:

Algorithm 1. Characterizing Latent Factor

Input: $P(z_k | p_j)$ and $P(p_j | z_k)$, predefined threshold μ

Output: A set of characteristic page base sets $LF = (LF_1, LF_2, \dots, LF_k)$

1. $LF_1 = LF_2 = \dots = LF_k = \Phi$
2. For each z_k , choose all pages $p_j \in P$
 - If $P(p_j | z_k) \geq \mu$ and $P(z_k | p_j) \geq \mu$ then

$$LF_k = LF_k \cup p_j$$
 - Else go back to step 2
3. If there are still pages to be classified, go back to step 2
4. Output: $LF = \{LF_k\}$

3.2 Identifying Web Page Categories

Note that the set of $P(z_k | p_j)$ is conceptually representing the probability distribution over the latent factor space for a specific Web page p_j , we, thus, construct the page-factor matrix based on the calculated probability estimates, to reflect the relationship between Web pages and latent factors, which is expressed as follows:

$$pr_j = (c_{j,1}, c_{j,2}, \dots, c_{j,k}) \quad (3)$$

Where $c_{j,s}$ is the occurrence probability of page p_j on factor z_s . In this manner, the distance between two page vectors may reflect the functionality similarity exhibited by them. We, therefore, define their similarity by applying well-known cosine similarity as:

$$sim(p_i, p_j) = (pr_i, pr_j) / (\|pr_i\|_2 \cdot \|pr_j\|_2) \quad (4)$$

where $(pr_i, pr_j) = \sum_{m=1}^k c_{i,m} c_{j,m}$, $\|pr_i\|_2 = \sqrt{\sum_{l=1}^k C_{i,l}^2}$

With the page similarity measurement (4), we propose a modified k-means clustering algorithm to partition Web pages into corresponding categories. The detail of the clustering algorithm is described in [34]. The discovered Web page categories reflect, either user “sole” access interest or cross-interest navigational intention.

4 Clustering User Session and Making Web Recommendation

Similarly, we employ clustering algorithm on the probabilistic variable set of $P(z_k | s_i)$, which represents the probability of a latent class factor z_k exhibited by a given user session s_i to capture user access pattern. The clustering user session via modified k-means clustering algorithm is described as follows:

Algorithm 2. Clustering User Session

Input: the set of $P(z_k | s_i)$, predefined threshold μ

Output: A set of user session clusters $SCL = \{SCL_1, SCL_2, \dots, SCL_p\}$ and corresponding centroids $Cid = \{Cid_1, Cid_2, \dots, Cid_p\}$

1. Select the first session s_1 as the initial cluster SCL_1 and the centroid of this cluster: $SCL_1 = \{s_1\}$ and $Cid_1 = s_1$.
2. For each session s_i , measure the similarity between s_i and the centroid of each existing cluster $sim(s_i, Cid_j)$
3. If $sim(s_i, Cid_t) = \max_j(sim(s_i, Cid_j)) > \mu$, then insert s_i into the cluster SCL_t and update the centroid of SCL_t as

$$Cid_t = 1/|SCL_t| \bullet \sum_{j \in SCL_t} sr_j \quad (5)$$

where sr_j is the transformed user session over factor space, $|SCL_t|$ is the number of sessions in the cluster;

Otherwise, s_i will create a new cluster and is the centroid of the new cluster.

4. If there are still sessions to be classified into one of existing clusters or a session that itself is a cluster, go back to step 2 iteratively until it converges (i.e. all clusters' centroid are no longer changed)
5. Output $SCL = \{SCL_p\}$, $Cid = \{Cid_p\}$

As we mentioned above, each user session is represented as a weighted page vector. In this manner, it is reasonable to derive the centroid of cluster obtained by aforementioned algorithm as the user access pattern (i.e. user profile).

Generally, Web recommendation process is usually carried out in two ways. On the one hand, we can take the current active user's historic behavior or pattern into consideration, and predict the preferable information to the specific user. On the other hand, by finding the most similar access pattern to the current active user from the learned models of other users, we can recommend the tailored Web content. The former one is sometime called memory-based approach, whereas the latter one is called model-based approach respectively. In this work, we adopt the model-based technique in our Web recommendation framework. We consider the usage-based access patterns generated in section 3 as the aggregated representations of common navigational behaviors, and utilize them as a basis for recommending potentially visited Web pages to current user.

Similar to the method proposed in [5], we utilize the commonly used *cosine* function to measure the similarity between the current active user session and discovered usage pattern. We, then, choose the best suitable pattern, which shares the highest similarity with the current session, as the matched pattern for current user. Finally, we generate the top- N recommendation pages based on the historically visited probabilities of pages visited by other users in the selected profile. The procedure is as follows:

Algorithm 3. Web Recommendation

Input: An active user session and a set of user profiles

Output: The top- N recommendation pages

1. The active session and the patterns are to be treated as n -dimensional vectors over the page space within a site, i.e. $s_p = Cid_p = [w_1^p, w_2^p, \dots, w_n^p]$, where w_i^p is the significance contributed by page p_i in this pattern, and $s_a = [w_1^a, w_2^a, \dots, w_n^a]$, where $w_i^a = 1$, if page p_i is already accessed, and otherwise $w_i^a = 0$.
2. Measure the similarities between the active session and all derived usage patterns, and choose the maximum one out of the calculated similarities as the most matched pattern:

$$sim(s_a, s_p) = (s_a \cdot s_p) / (\|s_a\|_2 \|s_p\|_2) \quad sim(s_a, s_p^{mat}) = \max_j (sim(s_a, s_p^j)) \quad (6)$$

3. Incorporate the selected pattern s_p^{mat} with the active session s_a , then calculate the recommendation score $rs(p_i)$ for each page p_i :

$$rs(p_i) = \sqrt{w_i^{mat} \times sim(s_a, s_p^{mat})} \quad (7)$$

Thus, each page in the profile will be assigned a recommendation score between 0 and 1. Note that the recommendation score will be 0 if the page is already visited in the current session.

4. Sort the calculated recommendation scores in step 3 in a descending order, i.e. $rs = (w_1^{mat}, w_2^{mat}, \dots, w_n^{mat})$, and select the N pages with the highest recommendation score to construct the top- N recommendation set:

$$REC(S) = \{p_j^{mat} \mid rs(p_j^{mat}) > rs(p_{j+1}^{mat}), j = 1, 2, \dots, N-1\} \quad (8)$$

5 Experiments and Evaluations

In order to evaluate the effectiveness of the proposed method based on PLSA model and efficiency of Web recommendation, we have conducted preliminary experiments on two real world data sets.

5.1 Data Sets

The first data set we used is downloaded from KDDCUP website. After data preparation, we have setup an evaluation 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 numbers of Web page hits by the given user determines the elements in session-page matrix associated with the specific page in the given session.

The second data set is from a academic Website log files [35]. The data is based on a 2-week Web log file during April of 2002. After data preprocessing stage, the filtered data contains 13745 sessions and 683 pages. 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”.

5.2 Latent Factors Based on PLSA Model

We conduct the experiments on the two data sets to characterize the latent factors and group usage-based Web pages. Firstly, we present the experimental results of the derived latent factors from two real data sets based on PLSA model respectively. Table 1 illustrates the results extracted from the KDDCUP data set, whereas Table 2 presents the results of CTI data set. From these tables, it is shown that the descriptive labels of latent factors are characterized by some “prominent” pages whose probabilistic weights are exceeding one predefined threshold. This work is done by interpreting the contents of corresponding pages since these “dominant” pages contribute greatly to the latent factors. With the derived characteristic factor, we may semantically discover usage-based access pattern.

Table 1. Labels of factors from KDDCUP

Factor	Label	Dominant Page
1	Department search	6, 7
2	Product information of Legwear	4
3	Vendor service info	10,36,37,39
4	Freegift, especially legcare	1,9,33
5	Product information of Legcare	5
6	Online shopping process	27,29,32,42,44,45,60
7	Assortment of various lifestyle	3,26
8	Vendor2’s Assortment	11,34
9	Boutique	2
10	Replenishment info of Department	6,25,26,30
11	Article regarding Department	12,13,22,23
12	Home page	8,35

5.3 Examples of Web Page Categories

At this stage, we utilize aforementioned clustering algorithm to partition Web pages into various clusters. By analyzing the discovered clusters, we may conclude that many of groups do really reflect the single user access task; whereas others may cover

two or more tasks, which may be relevant in nature. As indicated above, the former can be considered to correspond to the primary latent factor, and the latter may reveal the “overlapping” of functionality in content.

In Table 3, we list two Web page categories out of total generated categories from KDDCUP data set, which is expressed by top ranked page information such as page numbers and their relative URLs as well. It is seen that each of these two page groups reflects sole usage task, which is consistent with the corresponding factor depicted in Table 1. Table 4 illustrates two Web page groups from CTI dataset accordingly. In this table, the upper row lists the top ranked pages and their corresponding contents from one of the generated page clusters, which reflect the task regarding searching postgraduate program information, and it is easily to conclude that these pages are all contributed to factor #13 displayed in Table 2. On the other hand, the listed significant pages in lower row in the table involve in the “overlapping” of two dominant tasks, which are corresponding to factor #3 and #15 depicted in Table 2.

Note that with these generated Web page categories, we may make use of these intrinsic relationships among Web pages to reinforce the improvement of Web organization or functionality design, e.g. *Adaptive Web Site Design*.

Table 2. Labels of factors from CTI

Factor	Label	Factor	Label
1	specific syllabi	11	international_study
2	grad_app_process	12	Faculty-search
3	grad_assist_app	13	postgrad_program
4	admission	14	UG_scholarship
5	advising	15	tutoring_gradassist
6	program_bachelor	16	Mycti_stud_profile
7	syllabi list	17	schedule
8	course info	18	CS_PhD_research
9	jobs	19	specific news
10	calendar	20	Home page

Table 3. Examples of Web page categories from KDDCUP dataset

Page	Content	Page	Content
10	main/vendor	38	articles/dpt_payment
28	articles/dpt_privacy	39	articles/dpt_shipping
37	articles/dpt_contact	40	articles/dpt_returns
27	main/login2	50	account/past_orders
32	main/registration	52	account/credit_info
42	account/your_account	60	checkout/thankyou
44	checkout/expresCheckout	64	account/create_credit
45	checkout/confirm_order	65	main/welcome
47	Account/address	66	account/edit_credit

Table 4. Examples of Web page categories from CTI dataset

Page	Content	Page	Content
386	/News	588	/Prog/2002/Gradect2002
575	/Programs	590	/Prog/2002/Gradis2002
586	/Prog/2002/Gradcs2002	591	/Prog/2002/Gradmis2002
587	/Prog/2002/Gradds2002	592	/Prog/2002/Gradse2002
65	/course/internship	406	/pdf/forms/assistantship
70	/course/studyabroad	666	/program/master
352	/cti/.../applicant_login	678	/resource/default
353	/cti/.../assistantship_form	679	/resource/tutoring
355	/cti/.../assistsubmit		

5.4 Evaluation Metric of Web Recommendation

From the view of the user, the efficiency of Web recommendation is evaluated by the precision of recommendation. Here, we exploit a metric called *hit precision* [5] to measure the effectiveness in the context of top- N recommendation. Given a user session in the test set, we extract the first j pages as an active session to generate a top- N recommendation set via the procedure described in section 4. Since the recommendation set is in descending order, we then obtain the rank of $j + 1$ page in the sorted recommendation list. Furthermore, for each rank $r > 0$, we sum the number of test data that exactly rank the r th as $Nb(r)$. Let $S(r) = \sum_{i=1}^r Nb(i)$, and $hitp = S(N)/|T|$, where $|T|$ represents the number of testing data in the whole test set. Thus, *hitp* stands for the hit precision of Web recommendation process.

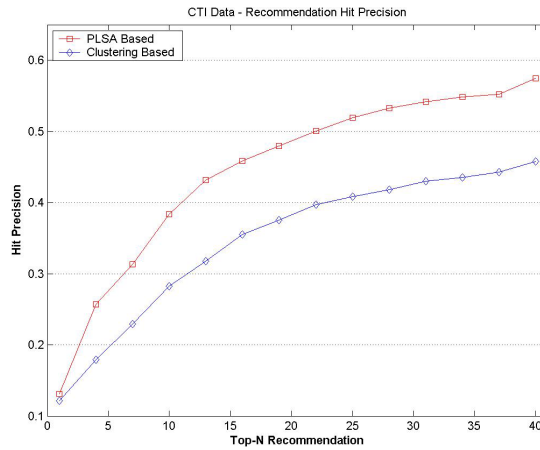


Fig. 2. Web recommendation evaluation upon hitp comparison for CTI dataset

In order to compare our approach with other existing methods, we implement a baseline method that is based on the clustering technique [5]. This method is to generate usage-based session clusters by performing k-means clustering process on usage data explicitly. Then the cluster centroids are derived as the aggregated access patterns.

Figure 2 depicts the comparison results of *hitp* coefficient using the two methods discussed above respectively performed on CTI dataset. The results demonstrate that the proposed PLSA-based technique consistently overweighs standard clustering-based algorithm in terms of hit precision parameter. In this scenario, it can be concluded that our approach is capable of making Web recommendation more accurately and effectively against conventional method. In addition to recommendation, this approach is able to identify the hidden factors why such user sessions or Web pages are grouped together in same category.

6 Conclusion and Future Work

Web usage mining is an emerging technique that can be utilized to, not only reveal Web user access interest, but also improve Web recommendation. This will provide benefits for improvement of Web applications, such as increasing the click-rate of Web site and providing more customized or preferred presentation to users.

In this paper, we have developed a Web recommendation technique by exploiting the pattern knowledge from Web usage mining process based on PLSA. With the proposed probabilistic method, we modeled the co-occurrence activities (i.e. user session) in terms of probability estimations to capture the underlying relationships among users and pages. Analysis of the estimated probabilities could result in building up usage-based Web page categories, discovering usage pattern, and identifying the hidden factors associated with corresponding interests. The discovered usage pattern has been utilized to improve the accuracy of Web recommendation. We demonstrated the effectiveness and efficiency of our technique through experiments performed on the real world datasets and comparison with previous work.

Our future work will focus on the following issues: we intend to conduct more experiments to validate the scalability of our approach. Meanwhile we plan to develop other methods by combining various types of Web data, such as content information into recommendation process to improve the accuracy.

References

1. Herlocker, J., et al. An Algorithmic Framework for Performing Collaborative Filtering. in Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR'99). 1999. Berkeley, CA.
2. Konstan, J., et al., Grouplens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, 1997. **40**: p. 77-87.
3. Shardanand, U. and P. Maes. Social Information Filtering: Algorithms for Automating 'Word of Mouth'. in Proceedings of the Computer-Human Interaction Conference (CHI95). 1995. Denver, CO.
4. 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.

5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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 (IJCAI'97)*. 1997. Nagoya, Japan.
11. 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.
12. 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.
13. 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.
14. Herlocker, J.L., et al., Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 2004. **22**(1): p. 5 - 53.
15. Dunja, M., Personal Web Watcher: design and implementation. 1996, Department of Intelligent Systems, J. Stefan Institute, Slovenia.
16. Joachims, T., D. Freitag, and T. Mitchell. WebWatcher: A Tour Guide for the World Wide Web. in *Proceedings of the International Joint Conference in AI (IJCAI97)*. 1997. Los Angeles.
17. Baeza-Yates, R. and B. Ribeiro-Neto, Modern information retrieval. 1999, Sydney: Addison Wesley.
18. Deerwester, S., et al., Indexing by latent semantic analysis. *Journal American Society for information retrieval*, 1990. **41**(6): p. 391-407.
19. Dumais, S.T. Latent semantic indexing (LSI): Trec-3 report. in *Proceeding of the Text REtrieval Conference (TREC-3)*. 1995.
20. Berry, M.W., S.T. Dumais, and G.W. O' Brie0146-4833n, Using linear algebra for intelligent information retrieval. *SIAM Review*, 1995. **37**(4): p. 573-595.
21. Hou, J. and Y. Zhang. Constructing Good Quality Web Page Communities. in *Proc. of the 13th Australasian Database Conferences (ADC2002)*. 2002. Melbourne, Australia: ACS Inc.
22. Hou, J. and Y. Zhang, Effectively Finding Relevant Web Pages from Linkage Information. *IEEE Trans. Knowl. Data Eng.*, 2003. **15**(4): p. 940-951.
23. Xu, G., Y. Zhang, and X. Zhou. A Latent Usage Approach for Clustering Web Transaction and Building User Profile. in *The First International Conference on Advanced Data Mining and Applications (ADMA 2005)*. 2005. Wuhan, china: Springer.

24. Hofmann, T. Probabilistic Latent Semantic Analysis. in Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval. 1999. Berkeley, California: ACM Press.
25. Hofmann, T., Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems*, 2004. **22**(1): p. 89-115.
26. 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.
27. Cohn, D. and H. Chang. Learning to probabilistically identify authoritative documents. in Proc. of the 17th International Conference on Machine Learning. 2000. San Francisco, CA: Morgan Kaufmann.
28. Hofmann, T., Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal*, 2001. **42**(1): p. 177-196.
29. Cohn, D. and T. Hofmann, The missing link: A probabilistic model of document content and hypertext connectivity, in *Advances in Neural Information Processing Systems*, T.G.D. Todd K. Leen, and Tresp, V., Editor. 2001, MIT Press.
30. Shahabi, C., et al. Knowledge discovery from user web-page navigational. in Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97). 1997: IEEE Computer Society.
31. Xiao, J., et al. Measuring similarity of interests for clustering web-users. in Proceedings of the 12th Australasian Database conference (ADC2001). 2001. Queensland, Australia: ACS Inc.
32. Dempster, A.P., N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statist. Soc. B*, 1977. **39**(2): p. 1-38.
33. 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.
34. 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.
35. Mobasher, B., Web Usage Mining and Personalization, in *Practical Handbook of Internet Computing*, M.P. Singh, Editor. 2004, CRC Press.