Discovering Task-Oriented Usage Pattern for Web Recommendation

Guandong Xu[1], Yanchun Zhang[1], Xiaofang Zhou[2]

[1]School of Computer Science and Mathematics
Victoria University, PO Box 14428, VIC 8001, Australia
[2] School of Information Technology & Electrical Engineering
University of Queensland, Brisbane QLD 4072, Australia

{xu,yzhang}@csm.vu.edu.au
zxf@itee.uq.edu.au

Abstract

Web transaction data usually convey user task-oriented behaviour pattern. Web usage mining technique is able to capture such informative knowledge about user task pattern from usage data. With the discovered usage pattern information, it is possible to recommend Web user more preferred content or customized presentation according to the derived task preference. In this paper, we propose a Web recommendation framework based on discovering task-oriented usage pattern with *Probabilistic Latent Semantic Analysis* (PLSA) model. The user intended tasks are characterized by the latent factors through probabilistic inference, to represent the user navigational interests. Moreover, the active user's intuitive task-oriented preference is quantized by the probabilities, by which pages visited in current user session are associated with various tasks as well. Combining the identified task preference of current user with the discovered usage-based Web page categories, we can present user more potentially interested or preferred Web content. The preliminary experiments performed on real world data sets demonstrate the usability and effectiveness of the proposed approach.

Keywords: Task-Oriented Usage Pattern, Web Usage mining, Web Recommendation.

1 Introduction

With the rapid development of a variety of Internet applications, Web has recently become not only a powerful platform and tool for retrieving information, but also a large repository for discovering knowledge. However, how to find needed and related information from the Web is a big challenge that Web information search domain is facing. Among much work addressed to such so-called information overload problem, Web recommendation is one of the instrumental means to help users locate more preferred information. Basically, Web recommendation is considered as the process of identifying user's preference and adapting service to satisfy user's need based on referring the historical behaviour of current user or others who share similar interest to this user.

To-date, there are two kinds of approaches and techniques commonly used in Web recommendation, namely content-based filtering agents and collaborative filtering systems (Dunja 1996; Herlocker, Konstan et al. 2004). Content-based filtering systems, such as WebWatcher (Joachims, Freitag et al. 1997) and client-side agent Letizia (Lieberman 1995), generally generate recommendation based on the pre-constructed user profiles by measuring the similarity of Web content to these profiles. In contrast, 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. Since collaborative filtering technique refers common interest of user group instead of individual's and is capable of presenting more preferable content to users, it has recently been widely adopted in Web recommendation applications and have achieved great success as well (Shardanand and Maes 1995; Konstan, Miller et al. 1997; Herlocker, KONSTAN et al. 1999). In addition, Web usage mining (WUM) has been proposed as an alternative method for not only revealing user access pattern, but also making Web recommendation in recent year (Mobasher, Dai et al. 2002). WUM is an application of data mining to discover usage pattern from Web log files and identify the underlying user functional interests that lead to common navigational activity, and has become an active topic of research and commercialization. Existing WUM techniques, which are well studied and developed in data mining domain, include collaborative filtering based on the k-Nearest Neighbor algorithm (*kNN*) (Shardanand and Maes 1995; Konstan, Miller et al. 1997; Herlocker, KONSTAN et al. 1999), Web clustering (Han, Karypis et al. 1998; Perkowitz and Etzioni 1998; Mobasher, Dai et al. 2002), association rule mining (Agrawal and Srikant 1994; Agarwal, Aggarwal et al. 1999) and sequential pattern mining technique (Agrawal and Srikant 1995). Amongst these methods, Web clustering is an important topic that engages in clustering not only Web users but also pages - discovering clusters of users that exhibit similar access pattern and categories of pages that share close functionality to users. By making use of the discovered knowledge from user clusters or page categories, Web designer may understand the users better, capture the unobservable relationships among pages from user's view ponit deeply, thus, can improve Web structure design and provide more preferable and customized service to the users.

In our previous work (Xu, Zhang et al. 2004; Xu, Zhang et al. 2005), Web user clustering and page grouping

techniques are well investigated to reveal such informative knowledge with regard to user behaviour and page functionality based on mining usage data. Especially, a so-called *Probabilistic Latent Semantic Analysis* (PLSA) model is proposed to address the topic of Web clustering. Different from other existing *Latent Semantic Analysis* (LSA) method, PLSA is to capture not only the underlying relationships among Web users as well as pages, but also reveal the hidden task-oriented pattern derived form WUM with probability inference approach. The main idea of this paper is to extend the above work to Web recommendation by identifying user task-oriented access pattern and integrating usage-based Web page category into Web recommendation process to improve the efficiency of recommendation. Moreover, approaches based on PLSA has been successfully applied in collaborative filtering (Hofmann 2004) Web usage mining (Jin, Zhou et al. 2004; Xu, Zhang et al. 2004; Xu, Zhang et al. 2005), text learning and mining (Cohn and Chang 2000; Hofmann 2001), co-citation analysis (Cohn and Hofmann 2001; Hofmann 2001) and related topics.

In this paper, we propose a Web recommendation framework based on discovering task-oriented usage pattern with PLSA model. The Web recommendation process exploits the usage pattern derived from Web usage mining to predict user preferred content or customized presentation. At data preparation stage, we collect Web transaction data from Web server log files, and construct user session collection and Web page corpus respectively. Conceptually, each user session can be expressed as a weighted Web page vector, in which the element reflects the relative significance contributed by the corresponding Web page in same user session. After integrating all user sessions, the Web access observation (i.e. usage data), on which the mining task is performed, is ultimately constructed in the form of page-based weight matrix. By employing PLSA model, 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 the estimated probabilities, and then are utilized to discover the task-oriented usage patterns in the form of a dominant task sequence. Furthermore, we make use of this discovered knowledge of usage pattern for Web recommendation by combining the task-oriented usage pattern and Web page category into Web recommendation to predict the more potentially interested or preferred content to user.

The main contributions we have done in this work are described as follows: firstly, we present a Web usage mining and Web recommendation integrated framework based on PLSA model. Secondly, we investigate the discovery of user access pattern and latent factor related to these patterns via employing probability inference process, in turn, make use of the discovered usage knowledge for Web recommendation. Particularly, we develop an algorithm for identifying task-oriented usage pattern and predicting user potentially visited pages based on Bayesian updating approach and incorporating Web page category into Web recommendation. Finally, we demonstrate the usability and effectiveness of the proposed model by conducting experiments on two real world datasets.

The rest of the paper is organized as follows. In section 2, we introduce Web usage mining technique with PLSA model, especially we discuss how to identify user access session and achieve probability estimations via Expectation-Maximization (EM) algorithm. We present the algorithms for discovering Web page categories and identifying task-oriented access pattern in section 3. In section 4, we concentrate on how to develop Web recommendation framework upon the discovered usage knowledge. To validate the proposed approach, we conduct preliminary experiments on two real world datasets, present evaluation results in section 5, and conclude the paper and outline future work in section 6.

2 Web Usage Mining with PLSA

Web usage mining usually consists of three steps, i.e. data collection and pre-processing, pattern mining as well as knowledge application. As a result, Web recommendation is actually the ultimate stage of the Web usage mining, i.e. application stage. The overview of Web usage mining and Web recommendation is depicted in Figure 1.

2.1 Usage Data Model

Prior to introducing Web usage mining technique, we briefly discuss the issue with respect to construction of usage data. In general, the exhibited user access interests may be reflected by the varying degrees of visits on different Web pages during one session. Thus, we can represent a user session as a weighted page vector visited by the user during a period. In this paper, we use the following notations to model the co-occurrence activities of Web users and pages:

- $S = \{s_1, s_2, \cdots s_m\}$: a set of m user sessions.
- $P = \{p_1, p_2, \cdots p_n\}$: a set of n Web pages.
- For each user, the navigational session is represented as a sequence of visited pages with corresponding weights: $s_i = \{a_{i,1}, a_{i,2}, \cdots a_{i,n}\}$, where a_{ij} denotes the weight for page p_j visited in s_i user session. The corresponding weight is usually determined by the number of hit or the amount time spent on the specific page. Here, we use both of them to construct usage data from two real world data sets.
- $SP_{m \times n} = \{a_{i,j}\}$: the ultimate usage data in the form of weight matrix with dimensionality of $m \times n$.

Generally, the element in the session-page matrix, a_{ij}, is the normalized 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. The session normalization is able to capture the relative significance of a page within one user session with respect to others pages accessed by same user. For example, Figure 2 depicts an usage snapshot from log file (Shahabi, Zarkesh et al. 1997; Xiao, Zhang et al. 2001).

The element in the normalized session-page matrix is determined by the ratio of the visiting time on corresponding page to total visiting time, e.g. $a_{11} = 15/(15+43+52+31+44)*100 = 9.7\ldots$ and so on.

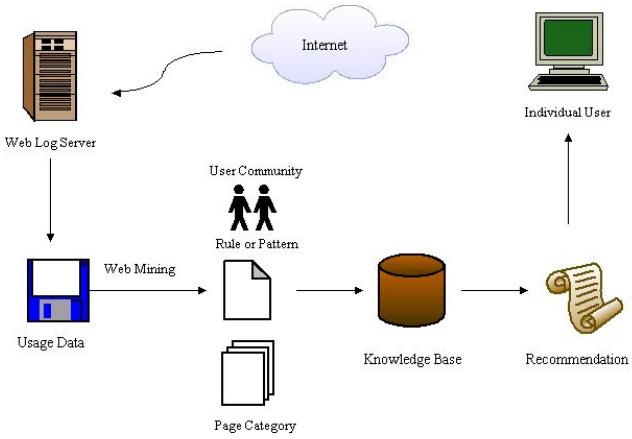

Fig. 1. The overview of Web Mining and Web Recommendation system

1) Main Movies: 20sec *Movies* News: 15sec *News*Box: 43sec *Box-Office* Evita: 52sec *News* Argentina:31 sec *Evita: 44sec*

2) Music Box: llsec *Box-Office* Crucible: 12sec *Crucible* Book: 13sec *Books: 19sec*

3) Main Movies: 33sec *Movies* Box: 21sec *Boxoffice* Evita: 44sec *News* Box: 53sec *Box-office* Evita: 61 sec *Evita : 31sec*

4) Main Movies: 19sec *Movies* News: 21sec *News b*ox: 38sec Box-*Office* Evita:61 sec *News* Evita:24sec *Evita* News: 31 sec *News* Argentina: 19sec *Evita: 39sec*

5) Movies Box: 32sec *Box-Office* News: 17sec *News* Jordan: 64sec *Box-Office* Evita: 19sec *Evita: 50sec*

6) Main Box: 17sec *Box-Office* Evita: 33sec *News* Box: 41 sec *Box-Office* Evita: 54sec *Evita* News: 56sec *News: 47sec*

$$SP_{ex} = \begin{bmatrix} 9.76 & 7.32 & 36.1 & 25.4 & 21.5 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 21.8 & 0.00 & 20.0 & 23.6 & 34.6 \\ 13.6 & 8.64 & 21.8 & 43.2 & 12.8 & 0.00 & 0.00 & 0.00 \\ 7.54 & 8.33 & 32.1 & 34.2 & 27.8 & 0.00 & 0.00 & 0.00 \\ 0.00 & 17.6 & 35.2 & 19.8 & 27.5 & 0.00 & 0.00 & 0.00 \\ 6.85 & 0.00 & 35.5 & 35.1 & 22.6 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

Fig. 2. A usage snapshot and its normalized session-page matrix expression

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 (Hofmann 1999). Similarly, we can conceptually view the user sessions over Web pages 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 $Z = \{z_1, z_2, \cdots z_k\}$ 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. In this manner, each usage data (s_i, p_j) can convey the user navigational interest by mapping the observation data into the *k*-dimensional latent factor space. The degrees to which such relationships are "explained" by each factor are represented by the factor-conditional probabilities. Below is some background of PLSA. We use following probability definitions to model usage data:

- $P(s_i)$ denotes the probability that a particular user session s_i will be observed in the occurrences data,
- $P(z_k \mid s_i)$ denotes a user session-specific probability distribution on the unobserved class factor z_k explained above,
- $P(p_j \mid z_k)$ denotes the class-conditional probability distribution of pages over the latent variable z_k.

Based on these definitions, we calculate probability of an observed pair (s_i, p_j) by adopting the latent factor variable z_k as:

$$P(s_i, p_j) = P(s_i) \bullet P(p_j \mid s_i) \quad (1)$$

$$P(p_j \mid s_i) = \sum_{z \in Z} P(P_j \mid z) \bullet P(z \mid s_i) \quad (2)$$

By applying Bayesian formula, we obtain the probability of an observation data associated with the latent factor as:

$$P(s_i, p_j) = \sum_{z \in Z} P(z) \bullet P(s_i \mid z) \bullet P(P_j \mid z) \quad (3)$$

Following the likelihood principle, the total likelihood *Li* is determined as

$$L_i = \sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet \log P(s_i, p_j) \quad (4)$$

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

From knowledge of statistics, Expectation Maximization (EM) algorithm is an effective way to perform maximum likelihood estimation in latent variable model (Dempster, Laird et al. 1977). Usually, two steps namely Expectation (E) and Maximization (M) step are iterating in this algorithm, i.e. E step leads to calculate the posterior probabilities for the latent factors based on the current estimates of conditional probability; whereas M step results in updating the estimated conditional probabilities and maximizing the likelihood based on the posterior probabilities computed in the previous E-step, i.e.

(1) In the E-step, we can simply apply Bayesian formula to generate following variable based on usage observation:

$$P(z_k \mid s_i, p_j) = \frac{P(z_k) \bullet P(s_i \mid z_k) \bullet P(p_j \mid z_k)}{\sum_{z_k \in Z} P(z_k) \bullet P(s_i \mid z_k) \bullet P(p_j \mid z_k)} \quad (5)$$

(2) In M-step, we can compute:

$$P(p_j \mid z_k) = \frac{\sum_{s_i \in S} m(s_i, p_j) \bullet P(z_k \mid s_i, p_j)}{\sum_{s_i \in S, p_j' \in P} m(s_i, p_j') \bullet P(z_k \mid s_i, p_j')} \quad (6)$$

$$P(s_i \mid z_k) = \frac{\sum_{p_j \in P} m(s_i, p_j) \bullet P(z_k \mid s_i, p_j)}{\sum_{s_i' \in S, p_j \in P} m(s_i', p_j) \bullet P(z_k \mid s_i', p_j)} \quad (7)$$

$$P(z_k) = \frac{1}{R} \sum_{s_i \in S, p_j \in P} m(s_i, p_j) \bullet P(z_k \mid s_i, p_j) \quad (8)$$

where

$$R = \sum_{s_i \in S, p_j \in P} m(s_{i,}, p_j) \quad (9)$$

Substituting equation (6)-(8) into (3) and (4) will result in the monotonically increasing of total likelihood Li of the observation data. The executing of E-step and M-step is repeating until Li is converging to a local optimal limit, which means the estimated results can represent the final probabilities of observation data.

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 Identifying Web Page Category and Task-Oriented Access Pattern

As we discussed in section 2, note that each latent factor z_k do really represent specific aspect associated with co-occurrence in nature. For each factor, the degree related to the co-occurrence is expressed by the factor-based probability estimate. From this viewing point, we, thus, can utilize the class-conditional probability estimates generated by the PLSA model and clustering algorithm to partition Web pages into various usage-based groups. Meanwhile, we can infer the latent factors by interpreting the meaning of "dominant" Web pages whose probabilities are exceeding the predefined threshold.

3.1 Discovering Web Page Category

Note that the set of $P(z_k \mid p_j)$ 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:

$$vp_j = (c_{j,1}, c_{j,2}, ..., c_{j,k}) \quad (10)$$

Where $c_{j,s}$ is the occurrence probability of page p_j on factor z_s. In this way, 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:

$$sim(p_i, p_j) = \left(vp_i, vp_j\right) \Big/ \left(\left\|vp_i\right\|_2 \bullet \left\|vp_j\right\|_2\right) \quad (11)$$

where $\left(vp_i, vp_j\right) = \sum_{m=1}^{k} c_{i,m} c_{j,m}$, $\left\|vp_i\right\|_2 = \sqrt{\sum_{l=1}^{k} C_{i,l}^2}$

With the page similarity measurement (11), we propose a modified k-means clustering algorithm to partition Web pages into corresponding categories. The detail of the clustering algorithm is described as follows:

Algorithm 1 Clustering Web Page

Input: the set of $P(z_k \mid p_j)$, predefined threshold μ

Output: A set of Web page categories and centroids $PCL = \{PCL_1, \cdots, PCL_p\}$

1. Select the first page p_1 as the initial cluster PCL_1 and the centroid of this cluster: $PCL_1 = \{p_1\}$ and $Cid_1 = p_1$.
2. For each page p_j, measure the similarity between p_j and the centroid of each existing cluster $sim(p_j, Cid_i)$
3. If $sim\left(p_j, Cid_t\right) = \max_i (sim(p_j, Cid_i)) > \mu$, then insert p_j into the cluster PCL_t and update the centroid of PCL_t as

$$Cid_t = 1/\left|PCL_t\right| \bullet \sum_{j \in SCL_t} vp_j \quad (12)$$

where $\left|PCL_t\right|$ is the number of sessions in the cluster

Otherwise, p_j will create a new cluster itself 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 $PCL = \{PCL_p\}$

In addition, note that $P(p_j \mid z_k)$ represents the conditional occurrence probability over the page space corresponding to a specific factor, whereas $P(z_k \mid p_j)$ represents the conditional probability distribution over the factor space corresponding to a specific page, which is expressed in the form of:

$$P(z_k \mid p_j) = \frac{P(p_j \mid z_k) \bullet P(z_k)}{\sum_{z_k \in Z} P(p_j \mid z_k) \bullet P(z_k)} \quad (13)$$

In such expression, we may consider that the pages whose conditional probabilities $P(p_j \mid z_k)$ and $P(z_k \mid p_j)$ are both greater than a predefined threshold μ can be viewed to contribute to one particular functionality related to the latent factor. Furthermore, we choose all pages satisfying aforementioned condition to form "dominant" page sets to characterize the latent factor.

3.2 Identifying Task-Oriented Access Pattern

Suppose that the conditional probability estimates are derived from the PLSA model as described above, we, in turn, utilize them to identify the user's underlying access task and to predict the potentially interested Web content to user in recommendation process.

Since the user session is represented as a sequence of visited pages, we can capture the task sequence derived from clicked pages within the session accordingly. This

aim is accomplished by computing the posterior probability of each task based on Bayesian updating approach, given that pages are independent on tasks. These posterior probabilities associated with the various tasks indicate the likelihood of user's underlying intention. The usage pattern, therefore, is characterized as a sequence of tasks with corresponding probabilities. By presetting an appropriate threshold, we choose all tasks whose posterior probabilities are greater than the preset value as a dominant task collection to reflect the user's initial intention. In section 5, we present two examples of task-oriented access pattern to illustrate how such task sequences are represented in terms of probability weights.

4 Web Recommendation Based on Task-Oriented access pattern

The discovered task-oriented access pattern can actually reveal the user's intrinsic access intend associated with latent task factors. As a result, incorporating the identified sequence of dominant tasks with the task-based page categories derived from previous section will lead to discover the potential pages more likely to be visited or interested by the user in following period. The detailed algorithm of Web recommendation is described as follows:

Algorithm 2 Web Recommendation

Input: the active user session $s_i =< p_1^i, p_2^i, \cdots, p_t^i >, p_j^i \in P$, a set of estimated conditional probabilities $P(p_j \mid z_k)$ and threshold.

Output: the dominant task sequence corresponding to the user session $TL = \{z_1^i, \cdots, z_t^i\}$ and the top-N recommendation pages $RS = \{p_j^r\}$.

1. For each task $z_k \in Z$, which is independent on the pages, calculate the posterior probability of z_s given all pages in s_i by employing Bayesian updating method (Russell and Norvig 1995):

$$P(z_k \mid s_i) = \alpha P(z_k) \prod_{p_j^i \in S_i} P(p_j^i \mid z_k)$$

   where α is a constant.

2. Choose all tasks whose conditional probabilities are greater than a preset threshold as the dominant task sequence corresponding to the user session.

$$TL = \{z_k \mid z_k \in Z, P(z_k \mid s_i) > \mu\}$$

3. For each z_k in *TL*, incorporate it with the corresponding task-based page category, then compute the recommendation score for each page p_j as

$$rs(p_j) = \sqrt{\sum_{z_k, p_j} P(z_k \mid s_i) \cdot P(p_j \mid z_k)}, p_j \in P, z_k \in TL$$

   Note that the recommendation score will be 0 if the page is already visited in the current session

4. Sort the computed recommendation scores from step 3 in a descending order, i.e. $rs = (rs(p_1^r), \cdots, rs(p_n^r))$, and choose the N pages with the highest scores to construct the top-N recommendation set.

$$RS = \{p_j^r \mid rs(p_j^r) > rs(p_{J+1}^r), j = 1, 2, \cdots, N-1\}$$

5 Experiments and Evaluations

In order to evaluate the effectiveness of the proposed method based on PLSA model and explore the discovered latent semantic factor, 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 (www.ecn.purdue.edu/KDDCUP/). 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 (Mobasher 2004). The data is based on a 2-week Web log file during April of 2002. After data pre-processing 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".

Discovery of latent task factor with PLSA model has been investigated in our previous work (Xu, Zhang et al. 2004). In this part, we just present the experimental results in terms of Web page category, task-oriented access pattern as well as the evaluation of recommendation.

5.2 Examples of Web Page Categories

At this stage, we utilize aforementioned clustering algorithm to partition the Web pages into various clusters. By analysing 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 intuitive latent factors, and the latter may reveal the "overlapping" relationships in content among Web pages.

In Table 1, we list three Web page groups out of total generated groups from KDDCUP data set, which is expressed by top ranked page information such as page numbers and their relative URLs as well. It shows that each of these three page groups reflects sole usage task, which is consistent with the corresponding factor depicted in Table 1 of (Xu, Zhang et al. 2005). Table 2 illustrates two Web page groups from CTI data set correspondingly. In this table, the upper row lists the top ranked pages and their corresponding content 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 of (Xu, Zhang et al. 2005). On the other hand, the listed significative 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 of (Xu, Zhang et al. 2005).

Table 1. Examples of Web page groups from KDDCUP

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
12	dpt_about	20	dpt_affiliate
13	dpt_about_mgmtteam	21	new_security
14	dpt_about_boarddirectors	22	new_shipping
15	dpt_about_healthwellness	23	new_returns
16	dpt_about_careers	24	dpt_terms
17	dpt_about_investor	57	dpt_about_the_press
18	dpt_about_pressrelease	58	dpt_about_advisoryboard
19	dpt_refer		

Table 2. Examples of Web page groups from CTI

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		

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. For example, the instrumental and suggestive task list based on the discovered page groups can be added into the original Web page as the means of *Adaptive Web Site Design*, to provide better service to users.

5.3 Examples of Task-Oriented Usage Patterns

As described in section 4, we exploit the posterior probability derived from PLSA model to identify the task-oriented usage pattern and predict the user's potentially visited Web pages by combining the task-oriented page categories into recommendation process. In the following table, we demonstrate two examples of derived task-based usage patterns through employing algorithm 2 on two real user sessions from KDDCUP and CTI dataset respectively. We list the active user sessions as well as tasks model derived from their sessions.

From the table, it is easily found that the two users have visited 10 and 11 pages respectively during their browsing period. The task-based usage patterns as well as their corresponding probabilities are depicted in the third and fourth column of the table. The upper part of the table shows that the user's activity actually involves in multiple purposes. However, the user's main intention is to perform online shopping as the probability of task #6 is significantly greater than the occurrence probabilities of other tasks. Therefore, we conclude that the dominant theme of first user's behaviour is actually locating on task #4.

Table 3. Examples of Task-Oriented Usage Pattern

#	Real user session	Task # & title	Prob.
1	1. boutique 2. search-result 3. ProductDetailLegcare 4. shopping_cart 5. login2 6. Welcome 7. expressCheckout 8. your_account 9. confirm_order 10. vendor	Online shopping (#6) Product Legcare (#2) Boutique (#9) Department search (#1) Vendor info (3)	0.94 0.02 0.02 0.01 0.01
2	1. admissions/ 2. admissions/requirements 3. admissions/mailrequest 4. admissions/orientation 5. gradapp/appmain_right 6. /news/default 7. /programs/ 8. programs/gradcs2002 9. programs/gradect2002 10. /programs/gradhci2002 11. /programs/core_guide	Admission (#4) Postgrad Program (#13)	0.63 0.37

For another user, we can find that the user was mainly conducting two tasks, i.e. task #4 and task #13. Incorporating the derived task model in table 1, we can further identify that task #4 represents prospective students searching for admission information, such as requirement, orientation etc., whereas task #13 reflects the activity of those students who are particularly interested the postgraduate programs in IT discipline. Unlike the first user, the second user clearly exhibits the cross-interest as the difference of the two corresponding probabilities is not quite significant.

Once the task model of user is identified, it is further utilized to recommend user preferred content accurately

5.4 Evaluation Metric for Web Recommendation

From the view of the user, the effectiveness of the proposed approach is evaluated by the precision of recommendation. Here, we exploit a metric called *hit precision* (Mobasher, Dai et al. 2002) 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 rth 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.

Table 4 gives the effectiveness of recommendation in terms of hit precision. From the table, it is shown that bigger the N number is, higher the *hitp* value is. In most case, the hit precision parameters are larger than 30%.

Table 4. The Results of Recommendation Hit Precision

N	5	6	7	8	9	10	11	12
hitp	0.17	0.19	0.20	0.22	0.27	0.30	0.32	0.33
N	13	14	15	16	17	18	19	20
hitp	0.34	0.35	0.36	0.36	0.37	0.37	0.37	0.38

6 Conclusion and Future Work

Web transaction data between Web visitors and Web functionalities usually convey user task-oriented behavior pattern. As a result, there is an increasing demand to develop techniques that can not only discover user task-oriented usage patterns, but also provide more benefits for recommend user more interested or preferred content In this paper, we have developed a Web recommendation technique by exploiting the knowledge of usage pattern from Web usage mining process based on PLSA model. With the proposed probabilistic method, we can measure 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 leads to build up task-oriented usage patterns and Web page categories, identify the hidden factors conceptually representing user interests or tasks. The discovered usage patterns can result in improvement of Web recommendation. We demonstrate the usability and effectiveness of our technique through experiments performed on the real world datasets.

Our future work will focus on the following issues: we intend to conduct experimental work on more datasets to validate the scalability of our approach. Meanwhile we plan to develop other machine learning algorithms to improve the accuracy of Web recommendation.

Acknowledgement

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

7 Reference

Agarwal, R., C. Aggarwal, et al. (1999): A Tree Projection Algorithm for Generation of Frequent Itemsets. *Journal of Parallel and Distributed Computing* **61**(3): 350-371.

Agrawal, R. and R. Srikant (1994): Jorge B. Bocca and Matthias Jarke and Carlo Zaniolo. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, Santiago, Chile, 487-499, Morgan Kaufmann.

Agrawal, R. and R. Srikant (1995): Mining Sequential Patterns. *Proceedings of the International Conference on Data Engineering (ICDE)*, Taipei, Taiwan, 3-14, IEEE Computer Society Press.

Cohn, D. and H. Chang (2000): Learning to probabilistically identify authoritative documents. *Proc. of the 17th International Conference on Machine Learning*, San Francisco, CA, 167-174, Morgan Kaufmann.

Cohn, D. and T. Hofmann (2001): The missing link: A probabilistic model of document content and hypertext connectivity: an in *Advances in Neural Information Processing Systems*. T. G. D. Todd K. Leen, and Tresp, V.(eds). MIT Press.

Dempster, A. P., N. M. Laird, et al. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statist. Soc. B* **39**(2): 1-38.

Dunja, M. (1996). Personal Web Watcher: design and implementation, Department of Intelligent Systems, J. Stefan Institute, Slovenia.

Han, E., G. Karypis, et al. (1998): Hypergraph Based Clustering in High-Dimensional Data Sets: A Summary of Results. *IEEE Data Engineering Bulletin* **21**(1): 15-22.

Herlocker, J., J. KONSTAN, et al. (1999): An Algorithmic Framework for Performing Collaborative Filtering. *Proceedings of the 22nd ACM Conference on Researchand Development in Information Retrieval (SIGIR'99)*, Berkeley, CA.

Herlocker, J. L., J. A. Konstan, et al. (2004): Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* **22**(1): 5 - 53.

Hofmann, T. (1999): Probabilistic Latent Semantic Analysis. *Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, 50-57, ACM Press.

Hofmann, T. (2001): Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal* **42**(1): 177-196.

Hofmann, T. (2004): Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems* **22**(1): 89-115.

Jin, X., Y. Zhou, et al. (2004): A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content. *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04)*, San Jose.

Joachims, T., D. Freitag, et al. (1997): WebWatcher: A Tour Guide for the World Wide Web. *Proceedings of the International Joint Conference in AI (IJCAI97)*, Los Angeles.

Konstan, J., B. Miller, et al. (1997): Grouplens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM* **40**: 77-87.

Lieberman, H. (1995): Letizia: An agent that assists web browsing. *Proc. of the 1995 International Joint Conference on Artificial Intelligence*, Montreal, Canada, 924-929, Morgan Kaufmann.

Mobasher, B. (2004): Web Usage Mining and Personalization: an in *Practical Handbook of Internet Computing*. M. P. Singh(eds). CRC Press.

Mobasher, B., H. Dai, et al. (2002): Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery* **6**(1): 61-82.

Perkowitz, M. and O. Etzioni (1998): Adaptive Web Sites: Automatically Synthesizing Web Pages. *Proceedings of the 15th National Conference on Artificial Intelligence*, Madison, WI, 727-732, AAAI.

Russell, S. J. and P. Norvig (1995): *Artificial Intelligence, A Modern Approach*, Prentice Hall.

Shahabi, C., A. Zarkesh, et al. (1997): Knowledge discovery from user web-page navigational. *Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97)*, 20-29, IEEE Computer Society.

Shardanand, U. and P. Maes (1995): Social Information Filtering: Algorithms for Automating 'Word of Mouth'. *Proceedings of the Computer-Human Interaction Conference (CHI95)*, Denver, CO.

Xiao, J., Y. Zhang, et al. (2001): Measuring similarity of interests for clustering web-users. *Proceedings of the 12th Australasian Database conference (ADC2001)*, Queensland, Australia, **35:** 107-114, ACS Inc.

Xu, G., Y. Zhang, et al. (2004): Discovering User Access Pattern Based on Probabilistic Latent Factor Model. *Proceeding of 16th Australasian Database Conference*, Newcastle, Australia, **39:** ACS Inc.

Xu, G., Y. Zhang, et al. (2005): Using Probabilistic Semantic Latent Analysis for Web Page Grouping. *15th International Workshop on Research Issues on Data Engineering: Stream Data Mining and Applications (RIDE-SDMA'2005)*, Tokyo, Japan.