Modelling User Behaviour for Web Recommendation Using LDA Model

Conference Paper · December 2008

DOI: 10.1109/WIIAT.2008.313 · Source: DBLP

Guandong Xu
University of Technology Sydney
158 PUBLICATIONS **573** CITATIONS
SEE PROFILE

Yanchun Zhang
Victoria University Melbourne
222 PUBLICATIONS **2,978** CITATIONS
SEE PROFILE

Xun Yi
RMIT University
138 PUBLICATIONS **1,224** CITATIONS
SEE PROFILE

Modelling User Behaviour for Web Recommendation Using LDA Model

Guandong Xu, Yanchun Zhang, and Xun Yi
School of Computer Science & Mathematics
Victoria University, PO Box 14428, VIC 8001, Australia
{Guandong.Xu, Yanchun.Zhang, Xun.Yi}@vu.edu.aul

Abstract

Web users exhibit a variety of navigational interests through clicking a sequence of Web pages. Analysis of Web usage data will lead to discover Web user access pattern and facilitate users locate more preferable Web pages via collaborative recommending technique. Meanwhile, latent semantic analysis techniques provide a powerful means to capture user access pattern and associated task space. In this paper, we propose a collaborative Web recommendation framework, which employs Latent Dirichlet Allocation (LDA) to model underlying topic-simplex space and discover the associations between user sessions and multiple topics via probability inference. Experiments conducted on real Website usage dataset show that this approach can achieve better recommendation accuracy in comparison to existing techniques. The discovered topic-simplex expression can also provide a better interpretation of user navigational preference.

1. Introduction

Due to explosive growth of information over the Internet in last several decades, information overload is becoming a big challenge. In the context of Web search, how to locate desirable Web information Web users are actually interested or needed, is emerging as a very popular topic which attracts much attention not only from researcher community, but also industrial developers [1]. Web recommendation system approach is probably a promising way that is capable of addressing above difficulties.

To-date, there are two kinds of approaches and techniques commonly used in the domain of recommendation system, namely content-based (CB) filtering and collaborative filtering (CF) systems [2]. Content-based filtering systems, such as WebWatcher [3] and client-side agent Letizia [4], 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.

WUM is an application of Web data mining to discover usage pattern from Web log file and identify underlying user visit interest exhibited from user's navigational activity. To discover the user access pattern, machine learning methods, e.g. clustering algorithm, are often used to learn the underlying associations from Web user navigational behaviours, and construct the aggregates of users to represent user access patterns. However, the conventional clustering-based techniques are lack of ability of uncovering the latent semantics associated with the access patterns. To address this, Latent Semantic Indexing (LSI) [1] and Probabilistic Latent Semantic Analysis (PLSA) [6] are proposed to deal with latent semantic analysis. Although these methods have achieved considerable success, however, they still suffer from a number of drawbacks, such as overfitting and inappropriate generative semantics [7].

In this paper, we propose a collaborative Web recommendation scheme based on Latent Dirichlet Allocation (LDA) model [7], which is originally proposed as a probabilistic document-topic model in text mining domain. With LDA, the latent associations between Web user sessions and multiple tasks/topics, and associations between tasks/topics and Web pages are estimated via a variation inference algorithm.

The rest of the paper is organized as follows. In Section 2, we first discuss the Web usage data model used in this study. Then in Section 3, the LDA model is introduced and we present the algorithms for Web usage mining and collaborative Web recommendation based on LDA in Section 4. Experiments and results are demonstrated to evaluate the efficiency and effectiveness of the proposed scheme in Section 5. Finally, we conclude and outline future research directions in Section 6.

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.313

2. Web Usage Data Model

Before we start introducing the statistical model for Web usage mining and Web recommendation, it is essential to briefly discuss Web usage data model used in this study. With the introduction of Web session-page expression, we further formulate the Web usage data model as follows:

- $S = \{s_1, s_2, \cdots s_m\}$: 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: $s_i = \{a_{i1}, a_{i2}, \cdots a_{in}\}$, 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.

3. Latent Dirichlet Allocation Model

The basic idea of the LDA is that usage data is modelled as a random mixture over latent topics with a probability distribution, where each topic is represented by a distribution over page space. In this sense, any random mixing distribution, $\mathrm{T}(z)$, is determined by an underlying distribution thereby representing uncertainty over a particular $\pi(\cdot)$ as $P_k(\pi(\cdot))$, where P_k is defined over all $\pi \in \mathrm{P}_k$. The generative model is, therefore, expressed as follows:

1. Pick a mixing distribution $\pi(\cdot)$ from P_k with a probability $P_k(\pi)$

2. For each usage session

(a) Choose a topic z with probability $\pi(z)$.

(b) Choose a page p_i from the topic z with $T_z(p_i)$.

The probability of observing a sequence of pages, $p = p_1, p_2 \cdots p_n$ in this model is:

$$P_{LDA}(p) = \int_{P_k} \left\{ \prod_{i=1}^{n} \sum_{z=1}^{k} \pi(z) T_z(p_i) \right\} P_k(\pi) d\pi \qquad (1)$$

where $P_k(\pi) = \Gamma\left(\sum_{z=1}^{k} \alpha_z\right) \prod_{z=1}^{k} \frac{\pi(z)^{\alpha_z - 1}}{\Gamma(\alpha_z)}$ is the Dirichlet distribution with parameters $\alpha_1, \alpha_2 \cdots \alpha_k$. In this model, the ultimate aim is to estimate the parameters for Dirichlet distribution and the parameters for each of the k topic models. Although the integral in this expression is intractable for exact inference, $T_z(p_i)$ is actually estimated by using a wide range of approximation inference algorithms, such as a variational inference algorithm {Blei, 2003 #64}.

4. Modelling User Behaviour for Web Recommendation Using LDA Model

4.1 Discovering User Access Pattern Based on LDA

Given m Web user sessions containing z topics expressed over n distinctive pages, we can represent $P(p|z)$ with a set of z multinomial distributions ϕ over the n pages, such that $P(p|z=j) = \phi_p^{(j)}$, and $P(z)$ with a set of m multinomial distributions θ over the z topics, such that for a page in a Web session s, $P(z=j) = \theta_j^{(s)}$. Here we use the LDA model described above to generate ϕ and θ that result in maximum likelihood estimates in the context of WUM. The complete probability model is as follows:

$$\begin{aligned} &\theta \sim Dirichlet(\alpha) \\ &z_i \left| \theta^{s_i} \right. \sim Discrete(\theta^{s_i}) \\ &\phi \sim Dirichlet(\beta) \\ &p_j \left| z_i, \phi^{z_i} \right. \sim Discrete(\phi^{z_i}) \end{aligned} \qquad (2)$$

Here, z stands for a set of hidden topics, θ^{s_i} denotes a Web session (s_i) preference distribution over the topics and ϕ^{z_i} represents the specific topic z_i 's association distribution over page space. α and β are hyper-parameters of the prior of θ and ϕ.

We use the variational inference algorithm to estimate each Web session's correlation with multiple topics (θ), and the associations between the topics and Web pages (β), with which we can capture user visit preference distributions exhibited by each Web session and identify the semantics of the topic space.

Given this representation, for each latent topic, we can consider user sessions with θ^{s_i} exceeding a threshold as the "prototypical" user sessions associated with that topic. In other words, these top user sessions contribute significantly to this topic via their navigational behaviours, in turn, are used to construct this topic-specific user access pattern.

Algorithm 1 Building Topic-Oriented User Access Pattern

Input: The session-topic preference distribution θ, usage data and a predefined threshold μ

1. For each latent topic z_j, choose all user sessions with $\theta_{z_j}^{s} \geq \mu$ to construct a user session aggregation R.

2. For each latent topic z_j, compute the topic-specific aggregated user access pattern by taking into account the discovered sessions' associations with z_j

$$\overrightarrow{ap_j} = \frac{\sum_{s \in R} \theta_{z_j}^{s} \bullet \vec{s}}{|R|} \tag{3}$$

where $|R|$ is the number of the selected sessions in R.

3. Output a set of topic-oriented user access patterns over the k multiple topics, $TOAP = \left\{ \overrightarrow{ap_1}, \overrightarrow{ap_2}, \ldots \overrightarrow{ap_k} \right\}$.

4.2 Collaborative Web Recommendation Algorithm

Given a set of user access models and current active user session, the algorithm of generating the top-*N* most weighted pages recommendation is outlined as follows:

Algorithm 2 Collaborative Web Recommendation

Input: An active user session and a set of learned user access patterns
Output: The top-*N* most weighted recommendation pages

1. Let the active session and the learned access patterns be as n-dimensional vectors over the page set, i.e. $\overrightarrow{ap_j} = [w_1^j, w_2^j, \cdots, w_n^j]$, where w_i^j is the relative weight on page p_i in this j^{th} user access model; the active user session $\overrightarrow{as_a} = [w_1^a, w_2^a, \cdots w_n^a]$, where $w_i^a = 1$, if page p_i is already clicked, and otherwise $w_i^a = 0$.

2. Measure the similarities between the active session and all learned user access models, and choose the maximum one out of the calculated similarities as the most closely matched access pattern:

$$sim(\overrightarrow{as_a}, \overrightarrow{ap_j}) = (\overrightarrow{as_a} \cdot \overrightarrow{ap_j}) \Big/ \left\| \overrightarrow{as_a} \right\|_2 \left\| \overrightarrow{ap_j} \right\|_2 \tag{4}$$

where $\overrightarrow{as_a} \cdot \overrightarrow{ap_j} = \sum_{i=1}^{n} w_i^j w_i^a$, $\left\| \overrightarrow{as_a} \right\|_2 = \sqrt{\sum_{i=1}^{n} (w_i^a)^2}$

$$sim(\overrightarrow{as_a}, \overrightarrow{ap_{mat}}) = \max_j (sim(\overrightarrow{as_a}, \overrightarrow{ap_j}))$$

3. Refer to the page weight distribution in the most closely matched access pattern $\overrightarrow{ap_{mat}}$, then calculate the recommendation score $RS(p_i)$ for each page p_i:

$$RS(p_i) = \sqrt{w_i^{mat} \times sim(\overrightarrow{as_a}, \overrightarrow{ap_{mat}})} \tag{5}$$

4. Sort the calculated recommendation scores in step 3 in a descending order, and select the N pages with the top-N highest recommendation:

$$WebR(S) = \{ p_j^{mat} \mid RS(p_j^{mat}) > RS(p_{j+1}^{mat}), j = 1, 2, \cdots N-1 \} \tag{6}$$

5. Experiments and Results

5.1 Dataset

The Web log dataset is from a academic Website log files [8]. 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". After performing analysis on the sensitivity of involved Web content, we eventually choose 30 topics as the initial input parameter used in LDA model. In the context of Web recommendation, the effectiveness of the proposed recommendation scheme is normally evaluated by the precision of recommendation. Here, we exploit a metric called hit precision [5] to measure the effectiveness in terms of top-N recommendation.

5.2. Samples of Topics and User Navigational Preference Distribution

As stated in Section 4, we can use LDA to identify the semantics of latent topics from the contents of prominent pages contributing significantly to each topic. We first present two examples out of 30 discovered topics in Table 1. To illustrate theses topics, we also list the URLs of prominent pages as well as their corresponding probabilities (based on ϕ), respectively. Meanwhile, the estimate of each user session's association with multiple topics (θ) could be used to model each user's navigational preference over topic space.

Table 1. Examples of topics discovered from CTI dataset

Topic #1		
Page #	URL	Probability
1	/admissions/	0.332
590	/programs/2002/gradis2002.asp	0.227
591	/programs/2002/gradmis2002.asp	0.115
259	/courses/syllabus.asp?course=584-97-301&q=3&y=2002&id=653	0.084
390	/news/news.asp	0.075
414	/pdf/promos/2002/is2002.pdf	0.029
594	/programs/2002/maat2002.asp	0.026
575	/programs/	0.018
666	/programs/masters.asp	0.017
Topic #18		
4	/admissions/costs.asp	0.287
593	/programs/2002/gradtc2002.asp	0.146
592	/programs/2002/gradse2002.asp	0.134
355	/cti/gradassist/assistsubmit.asp	0.110
196	/courses/syllabus.asp?course=450-96-305&q=3&y=2002&id=273	0.037
464	/people/facultyinfo.asp?id=216	0.036
577	/programs/2001/phdincs2001.asp	0.032
605	/programs/courses.asp?depcode=21&deptmne=csc&courseid=211	0.027
248	/courses/syllabus.asp?course=566-98-301&q=3&y=2002&id=302	0.025
354	/cti/gradassist/assistantship_form.asp?section=news	0.024
247	/courses/syllabus.asp?course=566-98-301&q=3&y=2002&id=302	0.024
597	/programs/bulletin.asp	0.019
385	/hyperlink/hyperspring2002/lobby.asp	0.016
249	/courses/syllabus.asp?course=567-98-302&q=3&y=2002&id=423	0.015
641	/programs/courses.asp?depcode=98&deptmne=tdc&courseid=463	0.015

From Table 1, it indicates two topic examples out of 30 topic set discovered from CTI dataset based on β. By interpreting the URL contents of the predominant pages with probabilities exceeding 1%, it is shown that topic #1 is in relation to the activities of searching information with respect to Master degrees in disciplines of IS or MIS, such as admission and course syllabus etc, whereas topic #18 is referred to common access interests on browsing associated pages regarding Postgraduate and PhD program, course costs, application of assistantship as well as related faculty and syllabus information.

5.3. Quantitative Analysis

To conduct quantitative analysis with the proposed LDA-based approach, we employ the evaluation metric aforementioned to compute the recommendation accuracy. The results are shown in Figure 1. In order to compare our approach with other existing methods, we also carry out experiments on the CTI dataset with conventional clustering-based and PLSA-based approaches. In a similar manner, the usage-based session clusters by performing k-means clustering and probability inference with PLSA model [5, 9] are constructed to aggregate user sessions with similar access preferences, and the centroids of clusters are derived as the aggregated users access patterns.

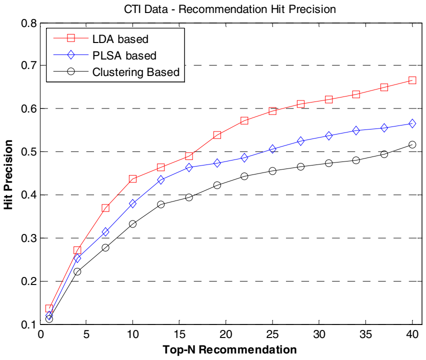

Figure 1. Hit precision comparisons on CTI dataset

The results demonstrate that the proposed LDA-based technique consistently outperforms the standard clustering-based and PLSA-based algorithms in terms of hit precision parameter. From this comparison, it can be concluded that the proposed approach is capable of making Web recommendation more accurate and effective against the conventional methods.

6. Conclusion and Future work

In this paper, we proposed a collaborative Web recommendation scheme by incorporating Web user access pattern based on Latent Dirichlet Allocation (LDA) model. With the LDA model, the associations between user sessions and multiple topics and the associations between topics and Web page space are estimated via a variational probability inference technique. Interpreting the predominant Web pages with significant contribution probabilities results in revealing the semantics of underlying topic space, and examining the association between user session and multiple topics leads to discover user navigational preference distribution over topic space. The experiments on real Website usage log have shown that this approach can achieve better recommendation accuracy in comparison to existing techniques.

In future, we aim to develop a new recommendation scoring schema in Web recommendation based on Markov Chain Learning technique.

References

1. Zhang, Y., J. Yu, and J. Hou, Web Communities, Analysis, Construction and Applications. 2005: Springer.
2. Herlocker, J.L., et al., Evaluating Collaborative Filtering Recommender Systems. ACM Transaction on Information Systems (TOIS), 2004. 22(1): p. 5 - 53
3. 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, p. 770-777, Nagoya, Japan.
4. Lieberman, H. Letizia: An Agent that Assists Web Browsing. in Proc. of the 1995 International Joint Conference on Artificial Intelligence. 1995, p. 924-929, Montreal, Canada: Morgan Kaufmann.
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. Hofmann, T. Probabilistic Latent Semantic Analysis. in Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval. 1999, p. 50-57, Berkeley, California, USA: ACM Press.
7. Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003(3): p. 993-1022.
8. 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.
9. Xu, G., Y. Zhang, and X. Zhou. A Web Recommendation Technique Based on Probabilistic Latent Semantic Analysis. in Proceeding of 6th International Conference of Web Information System Engineering (WISE'2005). 2005, p. 15-28, New York City, USA: LNCS 3806.