Personalized Web Content Recommendation based on LDA Profile

Hiroshi FUJIMOTO† Minoru ETOH‡ Akira KINNO† and Yoshikazu AKINAGA†

† NTT DOCOMO R&D Center, 3-6 Hikarino-oka, Yokosuka-shi, Kanagawa, 239-8536 Japan

‡ Osaka University Cybermedia Center, 1-32 Machikaneyama, Toyonaka, Osaka, 560-0043 Japan

E-mail: †{fujimoto, kinno, akinaga}@nttdocomo.co.jp, ‡ etoh@ieee.org

Abstract We propose a web content recommendation method based on latent topic modeling such as LDA. The main technical challenge is how to symbolize web access actions, by words, which are monitored through a web proxy log. We have developed a hierarchical URL dictionary and a cross-hierarchical directory matching method which provides automatic abstraction functionality. We also propose a recommendation scheme based on LDA model which recommends unseen contents as well as seen ones in the past. We show recommendation effectiveness of our method by applying to proxy data of 7500 students in Osaka University.

Keyword Latent Topic Model, Latent Dirichlet Allocation, Web Mining, Content Recommendation

LDA モデルを利用した Web ユーザプロファイリング方式のコンテンツ推薦への適用に関する研究

藤本 拓† 栄藤 稔‡ 金野 晃† 秋永 和計†

†株式会社 NTT ドコモ R&D センタ 〒239-8536 神奈川県横須賀市光の丘 3-6

‡大阪大学サイバーメディアセンター 〒560-0043 大阪府豊中市待兼山町 1-32

E-mail: †{fujimoto, kinno, akinaga}@nttdocomo.co.jp, ‡ etoh@ieee.org

あらまし 文書分類技術で広く利用される潜在トピックモデルを Web 閲覧ログに適用することで，Web 閲覧ユーザのプロファイリングを行う．高精度なプロファイリングを実現するために，階層型辞書を利用した Web 閲覧行動の抽象化を行うことで，大量の閲覧ログから効果的に潜在トピックを抽出する方式を提案する．さらに，LDA モデルを利用した，既知，及び未知コンテンツの推薦方式について述べる．また，大阪大学の学生 7500 人のプロキシログを利用し，提案方式で得られたプロファイリング結果のコンテンツ推薦への応用について評価する．

キーワード Latent Topic Mode, Latent Dirichlet Allocation, Web マイニング，コンテンツ推薦

1. Introduction

Web access user behavior analysis in general occupies the first crucial step of personalized web applications such as advertizing, recommendation, and web search. To realize the analysis needed, the application system monitors web access behavior at sites, which are categorized into clients, servers and proxies. Depending on applications, the monitoring site category and modeling of user web access may differ. This paper focuses on “topic modeling” which means that documents (i.e., users) are represented as mixtures of topics (i.e., abstracted user profile components), where a topic is a probability distribution over words (i.e., user web access actions). There have been comprehensive contributions regarding the topic modeling of user web access behavior. Most successful topic modeling targets domain-specific and application-oriented web analysis. By narrowing user actions to viewed contents, it offers excellent performance for recommendation and targeted advertisement [[1],[14],[5],[2],[15],[9]]. The extracted topics, in other words, abstracted user intentions, enable the system to infer the user’s next action. Please note that they used SVD[4], LSI [12] or pLSA [13] for probabilistic modes, since their contributions appeared in the early 2000’s. As an update, LDA [[3]] or more sophisticated models could be used instead.

The motivation for this paper lies in the authors’ belief that proxy data with a better topic and action model will yield more extensive user analysis, where the term ‘extensive’ means that the results are not domain-specific nor application-oriented, but rather broadened to social group descriptions. The research scope of this paper seems to be similar to [6], which compared LDA to pLSA for probabilistic modeling, and associated user sessions with multiple topics to describe the user sessions in terms of viewed web pages. This paper, however, focuses on the association between words (i.e., user web accesses) and the observed click streams rather than probabilistic modeling. We also use an LDA model for topic modeling though, simply taking viewed pages as words doesn’t work, since a click stream contains many meaningless page Given a lot of proxy data, the key issue is how to select the proper words to symbolize sessions.

To realize the symbolization, we have proposed a word association scheme called "CHDM: cross-hierarchical directory matching method" which extracts multiple words from each user session by matching against a directory database [7]. We have also extracted interest profiles of University students, while shown the optimality of the method by employing perplexity analysis.

In this paper, we define recommendation scheme of the LDA model. Given proxy log as training data set, it can predict probabilities of accesses to both seen and unseen contents in the leaning set for each user using LDA outputs. We also show the recommendation effectiveness of CHDM using real proxy data of 7500 students in Osaka University. Precision/recall analysis is employed to confirm the optimality of CHDM.

2. Proxy based Web User Profiling

2.1. LDA-based topic modeling

We assume topic modeling where the user accesses Web pages under certain topics (i.e., abstracted user intentions or tasks). For example, the user accesses a certain SNS site under his latent topic "SNS-addict", or accesses a certain job site under her latent topic "Job Hunting". In this case, by applying concepts of LDA, a Web user should correspond to a document, accessed web contents correspond to words, and their latent topics correspond to topics of documents. The observed accesses of each user are input to the LDA model, which then outputs the association between users and topics. In detail, the input and the outputs are as follows:

Inputs: a matrix N where each element $\nu(m,v)$ denotes the counts of contents v each user m accessed.

Output1: a matrix θ where each element $\vartheta(m,k)$ denotes the topic k distribution of each user m.

Output2: a matrix Φ where each element $\varphi(k,v)$ denotes the contents v distribution of each topic k.

2.2. Cross-Hierarchical Directory Matching

The goal of *topic modeling* is to derive the optimal outputs Θ and Φ. To realize this, optimal input N is needed. The simplest way that takes all the accessed URLs as words (i.e. the approach of [[6]]) doesn’t work, since many of them are not related to users' intention. Moreover, it is said in the text mining domain that word set should be abstracted by dictionaries for a proper model [10].

Cross-Hierarchical Directory matching (CHDM) is a method that uses a hierarchical dictionary to get a set of abstracted URLs that are broader in concept than the originally accessed URLs. The dictionary C should have an ontology structure, a category hierarchy that supports path abstraction. Categories c_h are numbered $\{1,2,\ldots,|C|\}$ in order of breadth-first search, so h is smaller than h' if c_h is an ancestor and broader in concept than $c_{h'}$. For example, if c_h is "newspaper" and $c_{h'}$ is "local newspaper", $c_{h'}$ is subordinate to c_h.

Moreover one or more URLs of Web site are registered to each category c_h. (To distinguish these URLs from proxy log entries, we call the former SURL.) If two SURLs are registered to different two categories and one category is subordinate to the other, the two sites have the same relationship with regard to conceptual hierarchy. For example, 'The New York Times' registered to c_h is a broader in concept than ‘China - The New York Times' which is registered to $c_{h'}$.

A basic idea of CHDM is to get a set of abstracted URLs by getting the hierarchical relationships of all URLs and discarding URLs of subordinate concepts. To know the hierarchy of URLs, we get a set of SURLs that the URLs belong to (matching step). Since we know their hierarchical relationship, we can discard all the SURLs of subordinate concepts and so create set of abstracted SURLS (abstraction step). This is the word set assigned to the session.

Figure 1 shows a simple example. URLs accessed at t_1, t_3, t_4, t_5, and t_6 belongs to SURLs respectively in the dictionary that are shown in the column 'Matched SURL'. Corresponding categories of the matched SURLs are also obtained straightforwardly in the column 'Matched Category'. Then we can get pairs (c_2, 'http://x2.y.z/'), (c_3, 'http://x.z.y/'), (c_4, http://x4.y.z/), and (c_5, 'http://x.y.z/w/') assigned to the session.

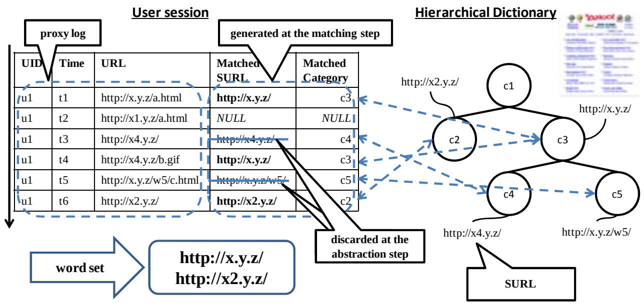

Figure 1. Example of Cross-Hierarchical Directory matching.

At the abstraction step, ' http://x4.y.z/' and 'http://x.y.z/w/' are discarded since corresponding categories (c_4 and c_5) are subordinate concepts of c_3. As a result, the set of remaining SURLS, i.e. ('http://x.y.z/', and 'http://x2.y.z') is an abstracted set of accessed web URLs in the session, and so is assigned as the word set. Details of the algorithm are shown in [7].

3. Recommendation on LDA Model

Given some training period and LDA outputs of the training data, the task is to predict access probabilities of viewed contents during test period after the training. The probabilities $p(v_{test}|m)$ derived as follows:

$$p(v_{test}|m) = \sum_k p(v_{test}|k)p(k|m) \quad (1)$$

where v_{test} is a content accessed by user m at the test period, and k is a latent topic of the access. If the test period is short enough compared to the learning period, $p(k|m)$ of the test set is the same as that of training set under the assumption that topics of each user are not changed suddenly. So $p(k|m)$ can be derived using the LDA outputs as follows:

$$p(k|m) = \frac{\vartheta(m,k)+\alpha}{\sum_k(\vartheta(m,k)+\alpha)} \quad (2)$$

where α is a hyper parameter of LDA model.

On the other hand, $p(v_{train}|k)$ can be derived as the same manner as $p(k|m)$ if a content v_{test} is seen in the training set, i.e.:

$$p(v_{test}|k) = \frac{\varphi(k,v)+\beta}{\sum_k(\varphi(k,v)+\beta)} \quad (3)$$

where β is a hyper parameter of LDA model. Viewed contents during the test period, however, will not seen in the test set. In this case, $p(v_{test}|k)$ cannot be derived straightforwardly.

To predict $p(v_{test}|k)$ for unseen contents, we search v_{learn} in the matrix N such that v_{learn} is broader concept of the v_{test}. Conceptual hierarchy of the two contents can be known by referring a hierarchical dictionary mentioned in Section 3 if v_{learn} are registered in the dictionary. We add some words with top categories in the dictionary to matrix N. Even if v_{learn} with broader concept is not found in the original word set of matrix N, one of added words with top categories will be matched as the broader concept of the unseen v_{test}. Please note that adding of the words does not affects the result of LDA since $v(m,v)$ of the added contents are set very small values.

Once words v_{learn} with broader concept is found, $p(v_{test}|k)$ can be approximated as follows:

$$p(v_{test}|k) \simeq \frac{\varphi(k,v_{learn})+\beta}{\sum_k(\varphi(k,v_{learn})+\beta)} \quad (4)$$

4. Experiments and Results

In this section, we show the recommendation effectiveness of CHDM by applying the scheme mentioned in Section 3 to proxy data of 7500 students in Osaka University.

4.1. Data sets

We use a set of 40 GB proxy log recorded accesses from over 7500 students in Osaka University. The log is recorded four month from April to July 2010. We divided the records into sessions for each user where the session timeout δ was set to 1800 [sec]. This yielded 175831 sessions for 7537 users. We also prepared a dictionary by crawling Yahoo! JAPAN Directory 0 in July 2010 for the hierarchical dictionary that has about 570 thousands distinct SURLs.

We chose proxy log of the first three month for learning set and next 1week for test set. We match the log entries of the learning set against the dictionary in the manner of CHDM. This yielded, as the first result, over 20 thousand distinct words including many very minor words. We eliminated minor words (those with fewer than 5 users) to obtain about 2400 test words.

After these pre-processing, we run LDA and get LDA outputs, i.e., θ and Φ. These outputs indicate very interesting profiles of university students. Details are shown in Appendix and [8].

4.2. Evaluation Metrics

We evaluate recommendation effectiveness of CHDM. First we derive access probabilities of all the contents for each user by the manner mentioned in Section 3 from the LDA outputs, and extract contents with top-N high access probabilities as a recommendation set. We then employ two evaluation metrics as follows:

$$precision(m,N) = \frac{|R_{m,N} \cap V_m|}{|R_{m,N}|}, recall(m,N) = \frac{|R_{m,N} \cap V_m|}{|V_m|} \quad (5)$$

where R_{mN} is the recommendation set for user m and V_m is a set of viewed contents by the user.

4.3. Evaluation Results

We first evaluate number of topics of CHDM to search the best model for recommendation. The results are shown in Figure 2. The figure represents changes of average precision and recall of all the test users by changing number of topics where recommendation set is top-5 or top-10. The results show that 24 stopics is a good choice and yields a better LDA model than the other values. (Note that higher topics yield more computational cost in running LDA.)

Next we show the optimality analysis of CHDM. We prepare three evaluation set. The first generates recommendation set by CHDM with predicting both seen/unseen contents (predict-seen/unseen). The second generates recommendation set by CHDM with predicting only seen contents in the training set (predict-seen). And the third generates recommendation set by simply chosing top-N popular contents among all the users in the training set (choose-popular).

The results are shown in Figure 3. The figure represents precision-recall curve for each recommendation size from top-1 to top-10. CHDM with seen/unseen-contents leads the highest precision/recall compared with the others. Especially precision is 1.5 times larger than that of chose-popular at top-5, so the recommendation by our approach is quite effective even when test set involves unseen contents.

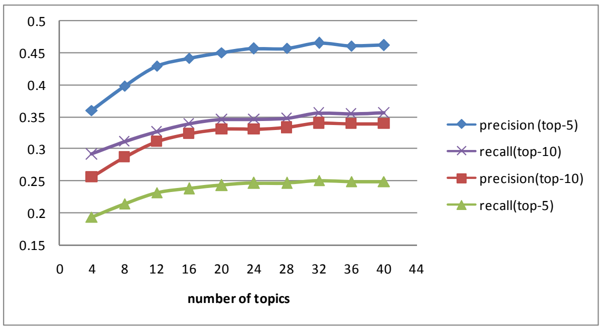

Figure 2. Evaluation of number of topics.

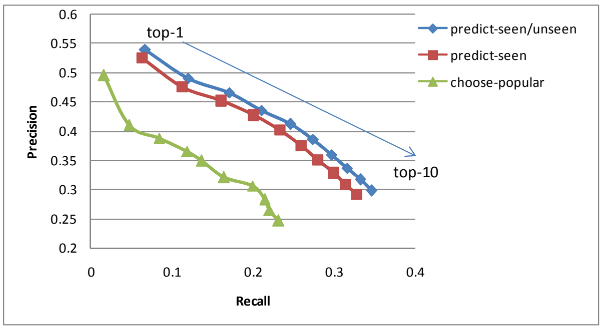

Figure 3. Precision-recall curve for each recommendation size.

5. Conclusion

One of a key application of profiling of Web users is recommendation. In this paper, we define a recommendation scheme for topic modeling with LDA model. It can predict unseen contents as well as seen contents in the training set. In future, we intend to apply our model to Web recommendation system and evaluate the effectiveness on real application.

References

[1] A.S. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. In Proc. of the 16th international conference on World Wide Web, Alberta, Canada ,2007.

[2] C. Lin, G.R. Xue, H.J. Zeng, and Y. Yu. Using Probabilistic Latent Semantic Analysis for Personalized Web Search. In the Lecture Notes in Computer Science, 2005, vol. 3399/2005, pp.707-717.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research archive vol. 3 (March 2003), pp. 993 - 1022.

[4] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. Numerische Mathematik vol.14, num.5, pp.403-420.

[5] G. Xu, Y. Zhang, and X. Zhou. Using probabilistic latent semantic analysis for Web page grouping. In Proc. of the Research Issues in Data Engineering: Stream Data Mining and Applications, 2005.

[6] G. Xu, Y. Zhang, and X. Yi. Modeling User Behavior for Web Recommendation Using LDA Model. In Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technol, Melbourne, 2008.

[7] H. Fujimoto, M.Etoh, A.Kinno, and Y. Akinaga. Web User Profiling on Proxy Logs and its Evaluation in Personalization. In Proc. of the 13th Asia-Pacific Web Conference (to appear).

[8] H. Fujimoto, M.Etoh, A.Kinno, and Y. Akinaga. Topic Analysis of Web User Behavior using LDA Model on Proxy Logs. In Proc. of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (to appear).

[9] J. Weng, E.P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In Proc. of the 3rd ACM international conference on Web search and data mining, 2010.

[10] K. Bessho. Text Segmentation Using Word Conceptual Vectors. In the Transactions of Information Processing Society of Japan vol.42(11), pp.2650-2662, 2001-11-15.

[11] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. In the Journal of Machine Learning Research archive, vol1. (September 2001), pp.211-244, 2001.

[12] S. Deerwester, S.T. Dumais, G.W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. In the Journal of the American Society for Information Science vol. 41, issue 6, pp. 391–407, September, 1990.

[13] T. Hofmann. Probabilistic Latent Semantic Analysis. In Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, California, 1999.

[14] X. Jin, Y. Zhou, and B. Mobasher. Web usage mining based on probabilistic latent semantic analysis. In Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, 2004.

[15] X. Wu, J. Yan, N. Liu, S. Yan, Y. Chen, et.al. Probabilistic latent semantic user segmentation for behavioral targeted advertising. In the Proc of the 3rd International Workshop on Data Mining and Audience Intelligence for Advertising, Paris, 2009.

Yahoo! Directory. http://dir.yahoo.com/, http://dir.yahoo.co.jp/

Appendix: Visualizing 24 topics and belonging students

The LDA output indicated 24 interesting topics. All the topics (named by authors) and their major words (or their description) are shown in Table 4. Each topic has distinctive words and they imply interests or tasks of belonging users.

Another interesting finding is that some topics are quite biased by the students' attributes such as grades or majors. To visualize them, we define a pair of attribute values "science degree (x_m)" and "higher grades degree (y_m)" that are implicit attributes of each user derived from the latent topics. Then we modeled associations between the latent topics and the implicit attributes for each user as a regression formula as follows:

Input:$\{\vartheta(d_m, z_k), a_k\}_{k=1}^{24}$, *Output:* a_m, *where* a_k *is a pair of attribute values* (x_k, y_k) *of each topic, and* a_m *is the pair of implicit attribute values* (x_m, y_m) *of the user* d_m.

The attribute value of each topic a_k derived as follows. We can know which topics each student belong to by choosing the topic with maximum probability on matrix Θ.We also prepare two **real** attribute values, i.e. "major" and "grade" for each user. The major is set from the students' major (science major set 1 and non-science major set -1), while the grade is set from their grade (1st grade set 1,..., 4th grade set 4). Then we can get a_k as follows where x_k is an average "major" and y_k is an average "grade" among the students belonging to the topic.

We chose a set of students for a learning set of the formula. In the learning phase, the output a_m was set as $a_{\tilde{k}}$ where $\tilde{k}$ was the belonging topic of each user. We learned the formula by Relevance Vector Regression using RVM [11], and we got a pair of implicit attributes a_m for all the students. The results are shown in Figure 5. The implicit attribute values of all the 7537 users are plotted where x-axis represents the "science degree" and the y-axis represents the "higher grades degree". Each point is color-coded by the user's belonging topic. The figure also represents distribution of the number of students belonging to each topic at the lower left of the figure where each topic number (1~24) correspond to the number in the Table 4.

The figure shows that points in the same topic tend to gather in a similar location. This indicates the fact that there is a strong relationship between belonging topics and attributes of students. Especially the points spread radially by highly attribute-biased topics. Examples of attribute-biased topics are "Full-Time Job Hunting" (#3), "Major in Bioscience" (#8), "Wikipedia User"(#19) or "Writing Report" (#21). We investigated their Web accesses on the proxy log and summarized as shown in the Figure. On the other hand, "SNS Addict"(#4) or "Twitterer"(#23) are not biased, i.e. students use these community sites regardless of their attributes.

Table 1. 24 topics and their major words.

Topic	Major Words	Topic	Major Words	Topic	Major Words
#1 MSN User	Hotmail, SkyDrive	#9 Search Books	Library of Osaka Univ.	#17 Net Shopping	Yahoo! Auctions, Amazon
#2 Video Freak	Youtube, MEGAVIDEO	#10 Internet Equity	Yahoo! Finance	#18 Major in Engineering	Site of Engineering Osaka Univ.
#3 Full-Time Job Hunting	Recruit Portals, Job search diaries	#11 Light User	Osaka Univ. Portal	#19 Wikipedia User	Wikipedia
#4 SNS Addict	SNS sites	#12 Anonymous-Forum Addict	Anonymous Forums	#20 Part-time Job Hunting	Part-time Job Portals
#5 Making Plans to go out	Weather forecasts, Google maps	#13 Geek	Sites for Geek	#21 Writing Report	Latex learning sites
#6 Newspaper Reader	Newspaper sites	#14 News Sensitive	Yahoo! News	#22 Information Search	Question Boards
#7 Sports Fan	Yahoo! Sports	#15 Blog Watcher	Yahoo! Blog	#23 Twitterer	Twitter Flickr
#8 Major in Bioscience	Sites about heredity or protein	#16 Geek Video Freak	Video sites for Japanese Geek	#24 Technology-Oriented	C-language learning sites

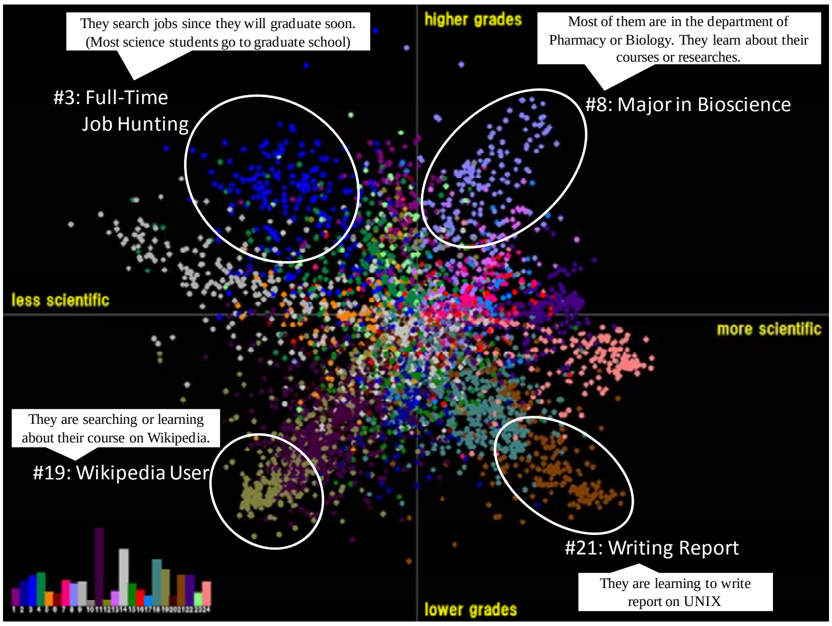

Figure 1. Plot of implicit attribute values for each user and their latent topics.