Latent Topic Model LDA User Profiling

Collaboration Projects

User Profiling based on Latent Topic Modeling

We have developed a method for modeling Web-access behavior using a topic model with the aim of performing user profiling using Web-access logs. This method achieves highly accurate user-profile modeling by extracting from previous user Web accesses only those URLs that best reflect user intention. This research was conducted jointly with the Collaborative Research Division at the Osaka University Cybermedia Center.

Service & Solution Development Department *Hiroshi Fujimoto*
Minoru Etoh
Yoshikazu Akinaga
Strategic Marketing Department *Akira Kinno*

1. Introduction

User profiling by analyzing Web-access logs is an effective means of enhancing and personalizing Web applications such as targeted advertising and content recommendation.

Our aim in this study is to establish a Web-user profiling method using a latent topic model[*1] that has found widespread use in the field of document categorization. Specifically, we aim to generate user profiles by analyzing a proxy log[*2] that records a wide range of user Web-access behavior and modeling that behavior.

In short, we perform user profiling by applying a latent topic model used in document analysis to the analysis of a proxy log. Here, to obtain superior profiling results, it is essential that the set of URLs input into the latent topic model reflect well the intentions of users. Accordingly, the theme of this study is to determine how to extract a group of URLs that best reflects user intention from a proxy log that records accesses to a large number of Web pages.

In the field of document categorization, a key issue is how to extract from a large number of documents sets of words that best reflect what individual documents mean. It has been pointed out that configuring a dictionary to perform abstraction by extracting word attributes is an effective means to this end [1]. It has been considered that this approach may also be effective in the analysis of proxy logs, but no studies in this regard have been performed to date. In this study, we propose Cross-Hierarchical Directory Matching (CHDM) as a technique for generating from a proxy log a word set that best reflects user intention for each URL session. In CHDM, it is assumed that the most semantically abstract URL accessed during the same URL session is the best reflection of user intention. For this reason, CHDM uses a hierarchical dictionary to extract only highly abstract URLs from a Web-access log. This dictionary should cover URLs in a broad Web space and assign a semantically hierarchical relationship to all URLs that it registers. A dictionary of this type can be generated by a directory-type search engine such as Yahoo! JAPAN Directory[*3].

[*1] **latent topic model**: A model widely used in document categorization based on the concept that a document is generated by latent topics each represented by a distribution of words.

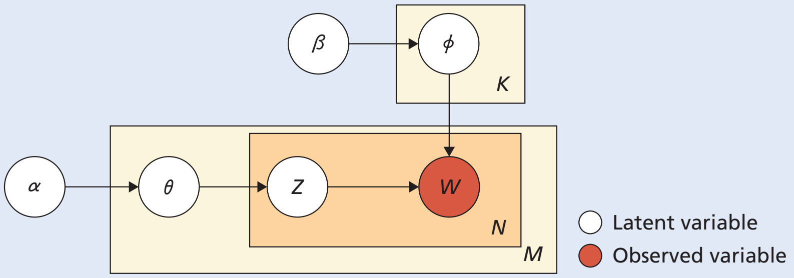
The development of the proposed method required the preparation of a large Web-access log based on a variety of user intentions, interests and preferences, and to collect this information, we conducted joint research with Osaka University using their computing environment.

2. Formulation by LDA

In this study, we used Latent Dirichlet Allocation (LDA)[*4] [2] as a latent topic model for modeling user Web-access behavior from a user proxy log.

When using LDA to analyze a proxy log, the assumption is made that latent topics[*5] exist in user Web-access behavior. For example, a user might access an educational site having the topic “C language” under the latent topic “programming.” This assumption enables LDA to be applied by substituting “user” for “document” and “URLs” for “words.” Document analysis and proxy-log analysis are compared in **Figure 1**. Here, by inputting user access frequency per URL (N) to the LDA model, the user in question can be expressed as a probability distribution of latent topics (θ) and each latent topic by a probability distribution of URLs (ϕ).

The aim of this study is to derive θ and ϕ that optimally models user Web-access behavior. To obtain highly accurate modeling here, it is essential to generate a set of URLs (W) from the proxy log that are most suitable for input to LDA.



	Standard document analysis	LDA in proxy-log analysis
α	Parameter for generating prior distribution θ	Same
β	Parameter for generating prior distribution ϕ	Same
θ	Latent-topic probability distribution	Same
ϕ	Word probability distribution	URL probability distribution
M	Number of documents	Number of users
N	Number of appearances per word in one document	User access frequency per URL
K	Number of topics	Same
Z	Word latent topic	URL latent topic
W	Word set	URL set

* Other latent variables can be inferred by using observed W.

Figure 1 Comparison of document analysis and proxy-log analysis in the LDA model

3. Proposed Method

3.1 Word Set Generation

Our study assumes a proxy log as shown in **Figure 2**. Each record of this proxy log includes at least a user ID, content access time, and URL of the accessed content, with log records appearing in order of access time. The set of records for each user, moreover, are divided into URL sessions (hereinafter referred to as “sessions”) according to the session ID indicated for each record. Sessions are defined by a specific timeout period. In the proxy log shown in the figure, user Web access is divided into three sessions, that is, sessions 1 – 3.

As shown by past research in the field of document categorization, abstraction by extracting attributes from words is an effective means of generating words [1]. In this study, we consider that this method can also be effective in generating URLs from sessions. Thus, for any particular session, we extract those URLs that are the most conceptual in nature and treat them as the word set generated from that session. We will later describe the specific technique that we use for extracting a word set from each session for input to the LDA model

Fig. 2 also shows the relationship between sessions and the word sets extracted from those sessions. The

[*2] **proxy log**: The access log on a proxy server recording Web-page accesses made by users from that server.

[*3] **Yahoo! JAPAN Directory**: A directory-type Web-page search engine that classifies Web pages into categories having up to 18 hierarchical levels and that allows Web-page searching by category. Yahoo! is a trademark or registered trademark of Yahoo! Inc.

[*4] **LDA**: A latent topic model in which topics are probabilistically determined for each word and for each document.

[*5] **latent topics**: In the latent topic model, latent variables that are assumed to exist and that are each expressed by a frequency distributions of words. In this model, a document is expressed as a probability distribution of latent topics.

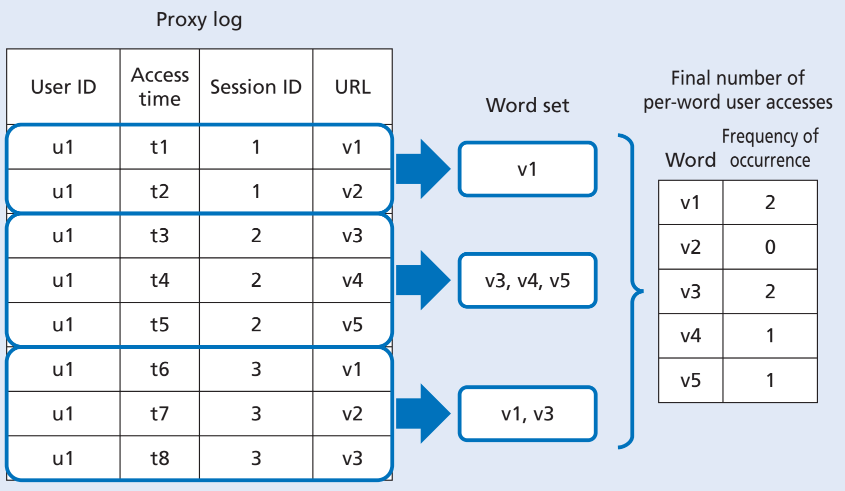
proxy log shown consists of three sessions labeled sessions 1, 2 and 3 for user u1. In session 1, for example, the proxy log records URLs v1 and v2, but only v1 is extracted. In session 2, moreover, URLs v3, v4 and v5 are extracted, and in session 3, v1 and v3 are extracted. The following section describes how an optimal word set is actually extracted from each session. The three results of this extraction process can be combined as shown in the figure to derive the final number of times that the user has accessed each word (URL).

3.2 CHDM

In this study, we use Yahoo! JAPAN Directory to generate a hierarchical URL dictionary (hereinafter referred to as "dictionary") for use in generating from a session a set of URLs abstracted to superordinate concepts. This dictionary consists of categories arranged in a hierarchical format so that categories in upper levels correspond to concepts with higher levels of abstraction. Multiple URLs are registered to each category. For example, the category of World Cup lies under the category of sports news, and URLs are registered under each of these categories. This arrangement makes it possible to determine the hierarchical semantic relationship between registered URLs. In this study, we call the process of session abstraction using a dictionary of the type described above CHDM.

The basic operation of CHDM is divided into two steps. The first step extracts a word set registered in the dictionary from URLs included in the session in question (matching step). The second step extracts a word set corresponding to superordinate concepts for that session (abstraction step). This is accomplished by using the dictionary to discover sets of words having a hierarchical semantic relationship and to extract from each set URLs corresponding to superordinate concepts. The operation of CHDM and definition of the dictionary are explained in detail in [3].

We here describe an example of CHDM operation referring to **Figure 3**. In this figure, a certain user session as recorded in the proxy log is shown on the left and a diagram of the dictionary is shown on the right. This session includes six URL accesses and the dictionary includes five categories from c1 to c5 registering URLs v1, v3, v4 and v5.

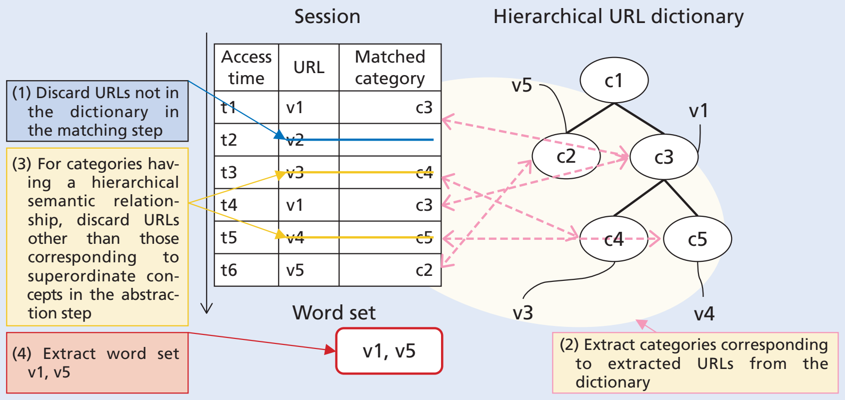


User ID	Access time	Session ID	URL
u1	t1	1	v1
u1	t2	1	v2
u1	t3	2	v3
u1	t4	2	v4
u1	t5	2	v5
u1	t6	3	v1
u1	t7	3	v2
u1	t8	3	v3

Word set: v1 / v3, v4, v5 / v1, v3

Word	Frequency of occurrence
v1	2
v2	0
v3	2
v4	1
v5	1

Figure 2 Relationship between sessions and words



Access time	URL	Matched category
t1	v1	c3
t2	v2	
t3	v3	c4
t4	v1	c3
t5	v4	c5
t6	v5	c2

(1) Discard URLs not in the dictionary in the matching step

(2) Extract categories corresponding to extracted URLs from the dictionary

(3) For categories having a hierarchical semantic relationship, discard URLs other than those corresponding to superordinate concepts in the abstraction step

(4) Extract word set v1, v5

Figure 3 CHDM operation

First, in the matching step, CHDM extracts from accessed URLs those that match URLs in the dictionary. In this example, this means URLs accessed at times t1, t3, t4, t5 and t6 (step (1) in Fig. 3).

Next, in the abstraction step, CHDM extracts a word set corresponding to superordinate concepts from the above extracted URLs. This step begins by using the dictionary to extract the category corresponding to each extracted URL (step (2) in Fig. 3). Among the categories obtained in this way, categories c3 and c4 have a hierarchical relationship as do categories c3 and c5. Thus, for these two sets of categories, CHDM extracts only the URL corresponding to c3 since that category corresponds to a superordinate concept (step (3) in Fig. 3). As a result, the final word set obtained for this session consists of the URL set (v1, v5) corresponding to categories c2 and c3 (step (4) in Fig. 3).

After extracting word sets from all sessions for all users, the final URL set (*W*) to be given to LDA is obtained as a union of sets. For each user, moreover, the access frequency of each URL, that is, the value of N to be input to LDA, can be derived as a sum total of the number of sessions accessing that URL.

4. Performance Evaluation

4.1 Data Sets

To evaluate the accuracy of the model obtained by CHDM, we used a proxy log recording the Web accesses of 7,537 students of Osaka University for a four-month period running from April to July 2010. This was a 40 GB log consisting of approximately 130,000,000 records. The timeout period used to establish sessions was 1,800 sec resulting in a total of 175,831 sessions. We created a dictionary consisting of 570,000 URLs by crawling through the Yahoo! JAPAN Directory in July 2010, and from among these URLs, we extracted those that had been accessed by five or more users according to the proxy log resulting in 4,500 URLs. On applying CHDM between this dictionary and the 175,831 sessions described above, we extracted word sets for more than 80% of all sessions.

4.2 Evaluation Results

We compared the accuracy of the model obtained by CHDM with the accuracy of the models obtained by two other methods: a non-abstraction method that generates a model without applying CHDM at all, and a directory matching method that generates a model by applying only the matching step of CHDM. These two methods are used to evaluate the performance of the abstraction step in CHDM. In this evaluation, we used perplexity[*6] as an evaluation index and evaluated model accuracy for each method by comparing the model generated using the first three months of the log with the recorded activity in the last month of the log.

Results are shown in **Figure 4**. It can be seen that perplexity changes with number of latent topics for each of the evaluated methods and that CHDM exhibits the best performance among all methods. It can be seen, in particular, that the directory matching method in itself improves performance by about 10% and that the abstraction step has a significant effect.

At the same time, CHDM is based

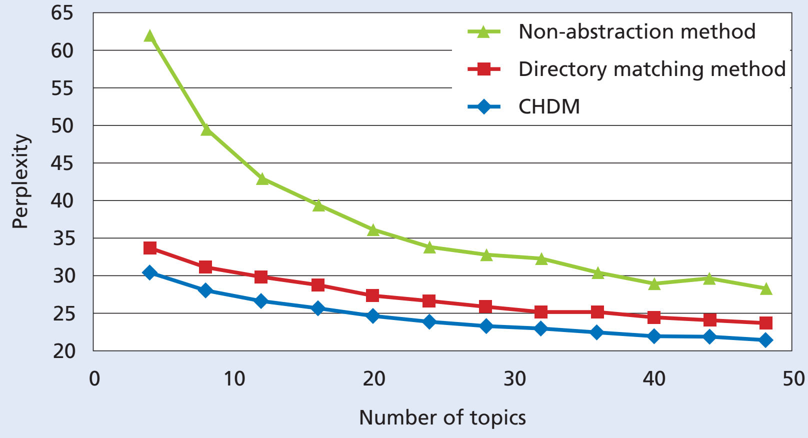
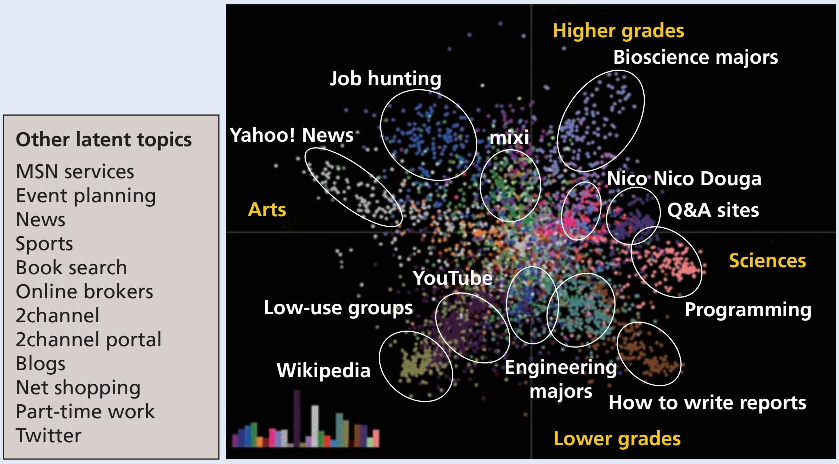


Figure 4 Perplexity evaluation

[*6] **perplexity**: A measure of distance between two distributions, used here as in index to evaluate the degree to which a model fits measured values; a smaller value of perplexity indicates a better model.

on the heuristic assumption that a good model can be obtained through abstraction using a dictionary. This assumption was shown to be correct by a performance evaluation, but there is no guarantee that similar results will be obtained if using another dictionary.



Nico Nico Douga®: A registered trademark of Dwango Co., Ltd.
mixi™ : A registered trademark of mixi, Inc.
MSN® : A trademark or registered trademark of Microsoft Corp. in the United States and other countries.
Twitter : A registered trademark of Twitter, Inc. in the United States and other countries.
Yahoo!® : A registered trademark of Yahoo! Inc.
YouTube™ : A trademark or registered trademark of Google Inc., in the United States.

Figure 5 Visualization of user profiling

5. Visualization of User Profiling

We here present the results of user profiling using the model obtained by CHDM and subjectively evaluate the validity of the model. We used the same data as that described in the previous chapter and used the model obtained from LDA when setting 24 latent topics.

To assess the validity of the obtained model, we examined the relationship between the latent topics and student attributes (major, year of study). Specifically, we performed a non-linear projection over two axes—major (arts/sciences) and year of study (upper-grades/lower-grades)—in order to visualize the results of user profiling. The projection method is explained in detail in [3].

The results of mapping the projection onto a two-dimensional graph are shown in **Figure 5**. Each point signifies a user. Latent topics associated with a user situated on the positive side or negative side of the x-axis tend to be scientific or artistic, respectively, and latent topics associated with a user situated on the positive side or negative side of the y-axis tend to be those of a higher-grade or lower-grade student, respectively. For this graph, we prepared colors to represent each of the 24 topics and assigned each user one of these 24 colors according to the latent topic that dominated the Web-access behavior of that user. We also assigned names to each of these 24 colors and labeled those locations in the graph where users associated with a certain latent topic clustered. We listed all other non-clustering latent topics to the left of the graph. A bar graph representing the number of students associated with each latent topic is also shown below the main graph.

This visualization reveals that points of the same color tend to cluster at the same location, which indicates a strong correlation between latent topics and student attributes. This tendency is especially evident for latent topics like “job hunting,” “bioscience majors,” and “programming.” We can therefore conclude that the results of profiling obtained from the generated model reflect student attributes well and that this model is qualitatively valid.

6. Conclusion

We proposed a method for generating words from URL sessions for the purpose of modeling a wide range of user Web-access behavior. This method abstracts a group of URLs using a hierarchical URL dictionary and extracts a word set from which a highly accurate model can be obtained. The proposed method was applied to a proxy log for 7,537 users, and the results of evaluating its word-prediction accuracy showed it to be effective. The results of profiling obtained from the model were also visualized, which demonstrated the model to be subjectively valid.

The performance of the proposed model differs according to the hierarchical URL dictionary used. In future research, we plan to compare performance between dictionaries with different configurations to study techniques for creating more accurate dictionaries.

REFERENCES

[1] Z. Elberrichi, A. Rahmoun and M. A. Bentaalah: "Using WordNet for Text Categorization," The International Arab Journal of Information Technology, Vol. 5, No. 1, Jan. 2008.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan: "Latent dirichlet allocation," The Journal of Machine Learning Research archive Vol. 3, pp. 993-1022, 2003.

[3] H. Fujimoto, M. Etoh, A. Kinno and Y. Akinaga: "Topic Analysis of Web User Behavior Using LDA Model on Proxy Logs," ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, LNCS Vol. 6634/2011, pp. 525-536, 2011.