Measuring User Preference Changes in Digital Libraries

Yang Sun[1], Huajing Li[2], Isaac G. Councill[1],
Wang-Chien Lee[2] and C. Lee Giles[1,2]
[1]College of Information Sciences and Technology
[2]Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA, USA
{ysun, icouncill, giles}@ist.psu.edu, {huali, wlee}@cse.psu.edu

ABSTRACT

Much research has been conducted using web access logs to study implicit user feedback and infer user preferences from clickstreams. However, little research measures the changes of user preferences of ranking documents over time. We present a study that measures the changes of user preferences based on an analysis of access logs of a large scale digital library over one year. A metric based on the accuracy of predicting future user actions is proposed. The results show that although user preferences change over time, the majority of user actions should be predictable from previous browsing behavior in the digital library.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*User issues*

General Terms

Algorithms, Measurement, Experimentation, Performance

Keywords

personalization, web usage mining, user preference, stability

1. INTRODUCTION

Extracting implicit user preference feedback becomes an attractive method to obtain personalized services since the method typically do not require additional user actions and generate much more data compared to explicit feedback methods[3, 6]. Research shows that clickthrough data obtained from information retrieval systems has a strong correlation with user explicit feedback [4]. However, most user preference research to date has not dealt with measuring the changes of user preferences. These changes will typically result in the need to weight document features differently over time. To study the temporal properties of user preferences, we extract implicit feedback for more than 4,000 discrete users based on one year's worth of web access logs from a large scale academic digital library. We propose a metric to measure the longitudinal changes of user preferences. The changes of user preferences is analyzed on one year's worth of web access log of a large scale academic digital library.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

2. RELATED WORK

The user interests have been modeled as profiles of categories to study the impact of changes of user interests [1, 7, 5]. All these models represent user interests by categories and keywords and the results are evaluated based on user feedback instead of the actual usage in a system. Our research investigates the actual usage of a web based academic digital library system and models user preference with document features.

3. USER PREFERENCE MODEL

We model user preferences as vectors in the document feature space where features include all the metadata presented to users in a system. One assumption in this feature space vector model is that users can only rely on the document features (document metadata) to determine the relevance of the document. The notions of the user preference model are introduced as the following:

- A *feature* is defined to be an item from a complete set of features F that are presented to users to determine relevance of documents.
- A document D is defined as a N-vector $D = \{d_1, d_2, ..., d_N\}$ in the feature space where d_i is the score of feature i for the document.
- For N features, a user is defined as a N-vector $U = \{u_1, u_2, ..., u_N\}$ such that $u_i > 0$ if the user prefers higher score of the feature i and $u_i < 0$ otherwise. u_i is the weight that user U evaluate feature i of documents.
- For a given document D and a preference vector U of a user, a score is defined as the inner product of the two vectors. $Score(D, U) =< D \cdot U >= \sum_{i=1}^{N} d_i \times u_i$.

Any two documents D_1, D_2 in a website requested by or presented to a user U for the same purpose form a document pair p. A user prefers document D_1 over D_2 if he/she requests additional information (such as download documents in digital libraries) of D_1 but not D_2 ($p = \{D_1 > D_2\}$). Thus, the preference vector of the user U should rank D_1 higher than D_2 ($Score(p, U) = Score(D_{i_1}, U) -$

$Score(D_{i_2}, U) > 0$). The preference vectors trained based on the implicit feedback extracted from user clickstreams can be studied as a function of time by segmenting the web access logs. For example, an one-year-long log can be split into 12 sections for each month. Therefore, the preference vectors trained on each section of logs represent the user preference in each month, and the series of preference vectors for each month capture the changes of each user information needs.

For a set of n document pairs $P = \{p_1, p_2, ..., p_n\}$ representing the implicit feedback of a user, we use the preference vector that maximizes the correctly ranked document pairs of a user as the user's preference vector (see Eq1).

$$U = \arg\max \sum_{i=1}^{n} |Score(p_i, U) > 0| \quad (1)$$

4. MEASURING CHANGES

To measure the changes of user preferences, we define a metric S of user preferences as Equation 2.

$$S = \frac{\overline{A}}{\sigma(A) + 1}. \quad (2)$$

Let $a(t)$ be the accuracy of using preference vector trained on access logs prior to time t to predict the user actions in time t. $A = \{a(t = 1), a(t = 2), ..., a(t = T)\}$ is the series of accuracy for the time section from 1 to T. $\overline{A}$ is the average prediction accuracy and $\sigma(A)$ is the standard deviation of the series. Since the prediction accuracy is always less than 1, the metric S will be a real number between 0 and 1. According to the definition, user preferences are more stable if their average prediction accuracy is higher or the deviation is lower. As an example, given the preference vectors for three continuous time sections, if the first vector predicts 80% of the second section, and the second vector predicts 60% of the third section, the measure S is 0.636.

5. EXPERIMENTS AND RESULTS

We analyze user preferences based on one year's worth of access logs from CiteSeer [2]. CiteSeer is a large scale academic digital library and search engine hosting 767,558 academic documents primarily in computer science field. The users of CiteSeer are typically computer scientist including faculty and students in academia as well as researchers in related research institutions. CiteSeer receives more than two million visits per day including both users and web crawlers. In our research, web crawler generated log records are filtered out by their identification and access behavior. The users are identified by unique IP addresses. By applying the user preference model and computing the measure S, the changes of each user preference is given as a real number that higher value represents stable user preference. Thus, the metric S can also be used as an indicator of user preference changes to investigate user related problems such as ranking and recommendations. The distribution of S for about 4,000 CiteSeer users is shown in Figure 1. The distribution peak is at 0.94 which indicates the majority user preferences are stable.

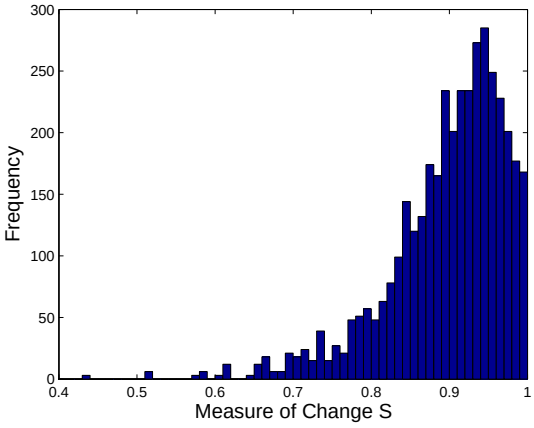

Figure 1: Histogram of the measure of changes S for 4,000 CiteSeer users.

6. CONCLUSIONS

The temporal properties of user preferences are seldom studied. We analyze one year's worth of web access logs from a widely used academic digital library and propose a metric to measure the changes of user preferences. In our study, user preferences are represented by feature space vectors. The metric S is defined based on the vector model and the accuracy of predicting future user actions using preference vectors trained on prior access logs.

Our results show that the majority of the user actions may be predicted by the user preference vectors trained from prior access logs. Different user groups may present differences in stability properties and it is possible that the user stability for other information retrieval systems will present different patterns. Our future work will extend the current method to a broader domain to study the differences in all types of information systems.

7. REFERENCES

[1] A. DÍłaz and P. GervÍćs. Adaptive user modeling for personalization of web contents. *Adaptive Hypermedia and Adaptive Web-Based Systems*, 3137:65–74, 2004.

[2] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: an automatic citation indexing system. In *DL '98: Proceedings of the third ACM conference on Digital libraries*, pages 89–98, New York, NY, USA, 1998. ACM Press.

[3] S. Holland, M. Ester, and W. Kiebling. Preference mining: A novel approach on mining user preferences for personalized applications. In *Knowledge Discovery in Databases: PKDD*, pages 204–216. Springer Berlin, 2003.

[4] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. of the 28th ACM SIGIR conference*, pages 154 – 161, 2005.

[5] W. Lam and J. Mostafa. Modeling user interest shift using a bayesian approach. *JASIST*, 52(5):416–429, 2001.

[6] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proc. of CIKM'05*, pages 824–831, 2005.

[7] D. H. Widyantoro, T. R. Ioerger, and J. Yen. Learning user interest dynamics with a three-descriptor representation. *JASIST*, 52(3):212–225, 2001.