Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Volume 22, No. 1, ISSN 1726-9679
ISBN 978-3-901509-83-4, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2011
Make Harmony between Technology and Nature, and Your Mind will Fly Free as a Bird
Annals & Proceedings of DAAAM International 2011

WEB MINING BASED ON USER PROFILE AND PREFERENCES

ACHSAN, H[arry] T. Y[ani]

Abstract: *Finding documents and information in the Internet becomes easier after the emergence of various search engines such as AltaVista, HotBot, Yahoo, and Google. But none of these search engines can give documents relevant to the user background knowledge and interests. Nevertheless, these search engines can be used in a meta search engine application that will predict user interests. This paper describes the idea of how to obtain user profiles or user interests implicitly from the query terms/keywords when searching in the Internet and also from documents/web pages opened by the user. Terms frequency and inverse documents frequency (tf-idf) are used to create the user profiles. Currently, this paper concentrates only on long term user profiles and neglects ad hoc interest. The software/application developed in this research aims to collect user profiles in two tables. First table records query terms and its frequencies, and second table records tf-idf extracted from documents opened by the user. The query output generated by this application is ranked based on three extended query. From the experiments, the application can provide information relevant to its user profile*

Key words: *information, retrieval, web, mining, preference, meta search engine, background knowledge, user profile*

1. INTRODUCTION

Before the advent of the Internet, it was difficult to find information because most of the sources are available not in digital format. At that time, most of information can only be obtained from the library or any other places which has a collection of data in hard copy format. Not many libraries can be visited due to limited time, distance or restricted membership.

The emergence of Internet has changed the difficulty in finding information. Today, most Internet users are experiencing the exceeded amount of information while trying to find data about certain topic. A single query creates huge amount of results. Then, users face another challenge to filter these results to meet their expectations.

Every person has their own preferences in finding information on a certain topic. For examples, searching with "oil refinery" keywords on a search engine will give equal results if it is done by three individual with different background knowledge. However, the information regarding "oil refinery" as required by a Refinery Engineer might be different as needed by a high school student or an entrepreneur. A Refinery Engineer might need more detail info about specific processes, raw materials, and the products of oil refinery plant. On the other hand, a high school student wants general and simple descriptions info related to oil refinery and its products. While an entrepreneur probably needs information related to business risks and profitability of oil refinery plant.

This paper discusses how to associate user profiles or user preferences in searching information using a special software. These user profiles can be used to expand keywords during query to meet user preferences. The software used to search information implements meta search engine, that is using commercial search engines to search information in the Internet.

2. RELATED WORKS

Several research papers relating to the User Profile are addressed in the information retrieval system based on the subjective / user side includes: (a) The acquiring of user background knowledge that taken implicitly from a large number of documents opened by users (Wu & Chen, 2009). The background knowledge of the user is used to retrieve documents (target document) relevant to the user's background. The information obtained in the form of association rules. (b) Developing iterative method to discovery user preferences using Vector Space Model and Fuzzy classification (Kiewra, 2005). The result is a vector of preferences that can be used as a measuring tool in finding information on the web. (c) Creating user profile in a client-server architecture for the personalized textual news for mobile users. User profile using two separate models, namely the long-term interest stored in the server and short-term interests stored in the handset. (d) Studying on the management of user profiles and personalization based on human factors in using information technology and telecommunications (Bartolomeo et al., 2008). (e) Analyzing proxy log to develop a web user profiling and clustering framework based on LDA-based topic modeling with an analogy to document analysis in which documents and words represent users and their actions (Fujimoto et al., 2011). (f) Personalizing web search according to user's geographic and temporal preferences can improve search results quality and satisfy user's different information needs (Yang et al., 2011). (g) Evaluating cross-site personalization across separately hosted open-source Web-based Content Management Systems (Koidl et al., 2011).

Those studies implement their own specific algorithm of search engines. This paper introduces a different approach. Retrieving information would be done by utilizing available related keywords from several available search engine, first. Then, using a software, those keywords will be used to assist searching process to get more accurate result of information.

3. USER PROFILE

User profile is a set of terms to represent background knowledge, preferences and interests of users. There are two techniques to obtain user profile: (a) submitted by user using a questionnaire, and (b) taken automatically from the query terms and from the results of user queries performed on information retrieval. This study emphasizes on second technique because many users feel uncomfortable to fill up form asking about their personal profile.

3.1 Term Frequency (Tf)

Terms or keywords are important words that describe the user's preferences and interests. Every keyword is captured whenever user performs a query. Thus, this captured data is stored in a table in database which has several columns, such as: login, password, key word, and term frequency. This table is indexed based on login name and password, so it is not possible for a user to save the same keywords. The higher the frequency of a word means it is more appropriate to be used as user profile.

$$t_i f = \sum_{k=0}^{n} t_{ik} \quad (1)$$

Tf calculation can be seen in equation (1) above, where f is frequency of term t_i. The value of t_{ik} is binary, where k is query number.

3.2 TF-IDF

TF-IDF is the classic way in determining the rating of terms. The more a term is used in the document and the few documents which use that term, then the term / keyword has a higher ranking. Tf-idf calculation is more precise when performed on the documents that have been selected by the user, i.e. a document which has been opened / read by user.

$$d_i f = \sum_{k=0}^{n} d_{ik} \quad (2)$$

Where $d_i f$ is the number or frequency of documents that have been opened / read by user and contain term i. Variable n is total documents that have been opened by user. Dividing equation (1) by equation (2) gives tf-idf.

$$t_i f.id_i f = \frac{\sum_{k=0}^{n} t_{ik}}{\sum_{k=0}^{n} d_{ik}} \quad (3)$$

3.3 Bookmarked Documents

Sometimes users are very interested in the document being read, so users bookmarked the document or include it in the list of favorite documents. The contents of these favorite documents are indexed. Then, this index can be added to the user profile table.

3.4 Downloaded Documents

Downloading documents means that those documents are important to a user. The contents of these downloaded documents are indexed. So, it can be added to the user profile table as well.

3.5 Dictionary

Online dictionary and thesaurus can be used to find the synonym of keywords and to expand the keywords before a certain word would be sent to search engines. There are various methods in determining the content of user profiles, we need a term weighting to the ranking obtained from each of the above methods. Weighting can be done after conducting tests on information retrieval system. In other words, the weighting is based on empirical data.

4. SOFTWARE

This system emphasizes security and user privacy, because the user profile is type of data that should be kept confidential. Authentication and authorization should be implemented in the system. The software using three search engines: Google, Yahoo and Bing. By typing keyword on the toolbar, user can get the search results using those search engines simultaneously. The keywords itself has been expanded using user profile before sent to the search engines to meet user

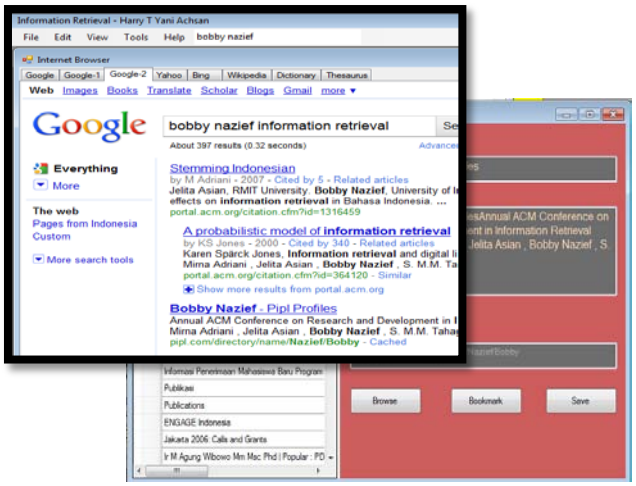

Fig. 1. Software interface

expectation. In Fig.1., the keywords was written as “Bobby Nazief”, but sent to search engines as “Bobby Nazief information retrieval”.

The application needs to be refined (fine tuned) as described at the end of section 3. In addition, if the application has been used for months or years, it will get long term preferences (user background knowledge) and ad-hoc preference (temporary profile). These applications have the ability to perform "self-learning" process from every word the user types and each document that he/she reads, inserted in the favorites or downloaded by users.

5. NEXT RESEARCH

A suggestion for subsequent research is develop an application to include an information from thesaurus, encyclopedia and dictionaries to improve the accuracy of search results. Improving application could be done by implementing client-server model, where users can search documents anywhere using either: a laptop, PDA or mobile phone. With this model, long-term preferences can be stored on the server while the ad-hoc preference stored on the client side.

6. REFERENCES

Bartolomeo, G., Petersen, F., Pluke, M. (2008). Personalization and User Profile Management

Fujimoto, H., Etoh, M., Kinno, A., Akinaga, Y. (2011). Web User Profiling on Proxy Logs and Its Evaluation in Personalization. Beijing, China, April 18-20, 2011

Kiewra, M. (2005). Iterative Discovering of User's Preferences Using Web Mining. *International Journal of Computer Science and Applications Vol II No. II, pp. 57-66*. Madrid-Technomathematics Research Foundation

Koidl, K., Conlan, O., Wade, V. (2011). Towards User-Centric Cross-Site Personalisation. *Proceedings of 11th International Conference of Web Engineering, ICWE 2011*. Paphos, Cyprus, June 20-24, 2011

Papadogiogaki, M., Papastathis, V. (2008). Two-Level Automatic Adaptation of a Distributed User Profile for Personalized News Content Delivery in *International Journal of Digital Multimedia Broadcasting*

Wu, Yi-fang Brook and Chen, Xin. (2009). Discovering Personalized Novel Knowledge. In: *Handbook of Research on Text and Web Mining Technologies*. New York: Information Science Reference

Yang, D., Nie, T., Shen, D., Yu, G., Kou, Y. (2011). Personalized Web Search with User Geographic and Temporal Preferences. *Proceedings of 13th Asia-Pacific Web Conference*. Beijing, China, April 18-20, 2011