A User Profile-based Approach for Personal Information Access: Shaping Your Information Portfolio

Lo Ka Kan
Dept. of Sys. Eng. and Eng. Mana.
The Chinese Univ. of HK
Shatin, Hong Kong
kklo@se.cuhk.edu.hk

Xiang Peng
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Shatin, Hong Kong
xpeng@cse.cuhk.edu.hk

Irwin King
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Shatin, Hong Kong
king@cse.cuhk.edu.hk

ABSTRACT
In the spread of internet, internet-based information service business has started to become profitable. One of the key technologies is personalization. Successful internet information services must realize personalized information delivery, by which the users can automatically receive highly tuned information according to their personal needs and preferences. In order to realize such personalized information services, we have developed an automatic user preference capture and an automatic information clipping function based on a Personalized Information Access technique. In this paper, those techniques will be demonstrated by showing a deployed personalized webpage service application.

Categories and Subject Descriptors
D.3.3 [**Information System**]: Information Search and Retrieval – *Web Engineering, Web Design Patterns and Pattern Mining*

General Terms
Design, System, Algorithm, Experimentation

Keywords: Personal Information Access, Information Retrieval, User Profile, Internet Behavior

1. INTRODUCTION
With the explosive growth of World Wide Web, the public is gaining access to massive amount of information. The scale of the web has grown to such an extent that we, as human beings, can hardly pick a single piece of the mostly wanted information from the zettabyte size of the future web. However, recent successes in search engine technologies have provided some new research directions and insights in solving some of our information need problems.

From the dawn of the commercialization of the Internet business, search engines have been there, absorbing information as the web grew from several thousand pages to billion of pages. From their invention, search terms have been the major medium between human beings and the search engines and even today, we are still relying on these terms to solve our information needs. We must admit that this approach has been working for some time and it still works for a short future. As our information needs are increasingly complex and diverse, is it still possible for the current search model to sustain in the future?

Copyright is held by the author/owner(s).
WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

Why are our information needs increasingly complex and diverse? Every web users are subjected to different past experience and even given the same set of terms to the search engines, the expected information returned will vary from person to person. As the web grows huge, the web users will expect increasingly fine-grained answers to their information needs. The current "One-keyword-fits-all" approach of search technology will definitely fail to sustain.

Personalization is the solution. Beforehand, we must differentiate what are explicit information (I_e) and implicit information (I_i) of the web users. Explicit information is the information the user explicitly inputs to the web to search for the required information: search terms, relevancy feedback and so on. Implicit information is the information in which the web users interact with the web and the users are usually not aware of: browsing duration, viewed page sequence and so on. The portfolio containing both the explicit and implicit information acts as the information portfolio or the user profile, which contains highly personalized information requirement of the users.

By utilizing the new strategy of information portfolio, we built up a new architecture for effectively fulfilling information needs of users that provides a new insight to a new information retrieval strategy.

2. THE MODEL
In this section, we explain how we can capitalize the information portfolio of an user to enhance the information retrieval process, precedes by the general system architecture.

2.1 The whole system architecture
Every time the user interacts with the web in some ways, e.g. clicking the links on the web pages, entering web site addresses and so on, their interactions with the web will be encoded in their own information portfolio that contains both explicit and implicit information.

Their information portfolio will be periodically sent back to a server, where the web pages have been indexed. The portfolio is then analyzed and merged with the old portfolio to discover what kind of topic will be most interested to the user at that moment or what kind of situation the user is in by investigating the portfolio.

This portfolio will then act as a query to the indexing system to retrieve a ranked list of pages that the user would be interested at that particular moment. By observing how the user interacts with the returned result, the portfolio will be appended with information to fine-tune the parameters of the portfolio.

2.2 Information Portfolio

The following are some of the implicit information captured for effective retrieval strategy.

2.2.1 Time and Duration Factor

These factors model how temporally important the pages are to the user. The more time the users spent on a page, the page will be potentially more interesting to the user and fits more precisely to the user information needs.

$$T_f = \alpha / (| t_{current} - t_{url} |)$$

$$D_f = \beta * (d_{url} / \sum d)$$

2.2.2 MAX-K-factor

For some locally most interesting pages, there exists some k pages that are also arousing interest to the user but no as interest as the locally most interesting page.

$$K_{\max} = \gamma * (n / k) \quad \text{for } n = 1,2,\cdots k$$

In this formula, α, β, γ are training parameters for further machine learning purpose.

2.2.3 Block factor B_f

Usually a web page is divided into a number of blocks containing particular piece of information. By investigating how the user interacts with different blocks of information on a page, we can append higher score to the block with more information in which the user is interested in.

2.2.4 Browse Tree

By observing the sequence of the pages viewed by the users and the factor score of the pages viewed, a hierarchical tree of the browsing experience is captured. The relation provides hint to the relationship between sets of the pages users are interested in. By examining these sets, we can deuce the situation the user may have and be more sensitive to particular information.

2.2.5 Formula for machine learning

The final formula - so called web psychological formula that provides a new retrieval strategy, is as follows:

Ordinary Page Factor:

$$P_{fo} = (T_f + D_f) * K_{max} + B_f$$

The page factor:

$$P_f = P_{fo} + \Sigma\, 1/2^i\, \Sigma \log_2 p_{fij}$$

where P_{fo} is the ordinary page factor, $\Sigma\, 1/2^i\, \Sigma \log_2 p_{fij}$ is the weighted sum of the page factors for the page viewed linked from this page and i represents the level of the browse tree.

3. EVALUATION AND CONCLUSION

We collected a small web corpus of 10,000 pages to perform the experiments where each pages are of least 10KB for sufficient information for indexing. We then clustered these pages into about 40 categories for testing. 40 sets of information portfolio between the user and the web are produced and fits into the system. In these 20 sets of information portfolio, the emulated pages viewed are from the 10,000 pages corpus and the factors as stated in Section 2 are generated randomly.

The returned pages, after the information portfolio is analyzed by the server, are compared with the information portfolio. A high recall and precision values are obtained, which is 50% better than in another setting when no information portfolio is present.

This result suggested that the concept of information portfolio can make the information request more accurate. Further work is under progress in modeling the factor more precisely to fit our web psychological behavior and expanding the work outside web searching but to the general information retrieval strategy.

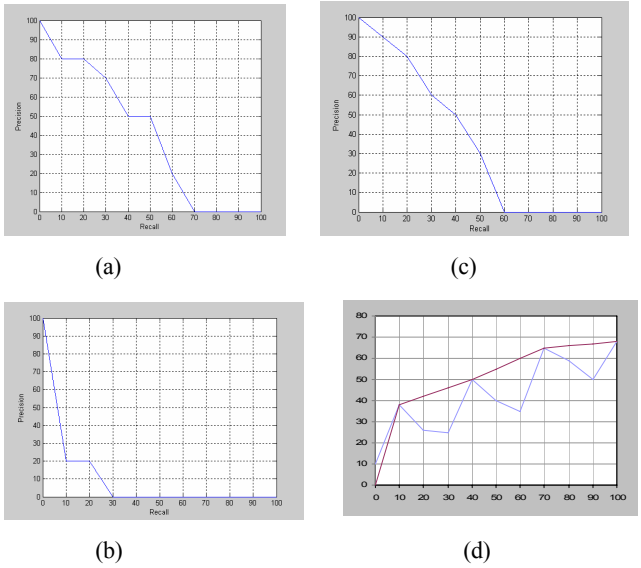

Figure 1.Recall vs Precision (a) shows the initial R vs P; (b) shows the R vs P before the profile is updated and the information request is changed; (c) shows the R vs P after the profile is updated; (d) shows the average R vs P

4. ACKNOWLEDGMENTS

We thank Prof. Lam Wai for his useful comments to the experiments. The work described in this paper is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4235/04E) and is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing & Interface Technologies.

5. REFERENCES

[1] K Andrew Edmonds, James Blustein, Don Turnbull. A Personal Information and Knowledge Infrastructure Integrator. Journal of Digital Information, vol. 5, 2004.

[2] Da Silveiva G., Borenstein D., Fogliatto F.S.. Mass customization: Literature review and research directions. International Journal of Production Economics, vol. 72, 2001.

[3] Carl Shapiro, Hal R. Varian. Information Rules. Boston: Harvard Business School Press, November 1998.