User Profiling and Personalized Information Delivery on the Static and Mobile Web

Dik Lun Lee
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
dlee@cse.ust.hk

ABSTRACT

Imagine a system that can push highly selective information right to our hands when and only when we need it. This requires a mind-reading machine, but unfortunately we don't have one — yet. User profiling attempts to estimate what is most important to a user at a particular point in time and space. In this talk, I will start with simple raw data such as the users' queries and clicks on the web and places they have visited to estimate what they might be interested in. We further divide user interests into content-based and location-based. We discuss issues involving the transformation of raw activities to conceptual needs, identifying user groups for collaborative filtering and the roles of locations in personalized information delivery.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*information search and retrieval*

General Terms

Algorithms, Design

Keywords

user profiling, user clustering, query clustering, clickthrough

1. BEHAVIOR MINING

Data mining deals with the discovery of semantics from data, whereas behavior mining discovers semantics from the actions of the users on the data. In this talk, I will focus on actions that can be easily collected by search engines, which include queries submitted by the users and the results they clicked on (i.e., clickthroughs), although the ideas generally apply to web browsing.

There is no doubt that user queries are driven by user interests. However, since user queries tend to be short and ambiguous, it is hard for a system to figure out what the users actually want in terms of, say, the types and levels of details of the desired information. Clickthroughs, that is, the actions taken by the users on the information presented to them, can provide the answer. When a query is ambiguous, the search engine will likely return a mix of results covering several aspects of the query. As users go through the results, they click on the results that are believed to contain interesting information. From the clicks, the system can further understand the users' interests. Different dimensions can be considered: the type of information (images and text), the association between different interests (e.g., labtops made by Sony), locations (products made in China), etc. The challenge is how to derive high-level information from raw queries and clicks and how to dissect the interests into categories. The ultimate goal of understanding a user's interests is of course to find and deliver the most important information to the user.

A vast amount of research has been done on user profiling, personalization and information delivery. In this talk, I will highlight some research issues related to these problems, with particular focus on work performed in our research group. Figure 1 illustrates the basic components of the personalization process, which will be briefly described below.

2. THE CLICK SPARSITY PROBLEM

The general assumption that two persons clicking on the same page share similar interests suffers from the click sparsity problem, which means that the number of clicks are much much smaller than the number of pages on the web, making the chance for two persons browsing the web to click on the same page extremely small. To alleviate the click sparsity problem, we need to extract the concepts (themes or topics) embodied in the pages. In other words, the assumption that "two persons clicking on pages that cover the same topics share similar interests" has a much better chance to identify users with similar interests [2].

3. GROUP-BASED RECOMMENDATION

Personalization purely based on the user's past actions is not enough because a person's actions are quite limited in scope compared to the vast amount of information on the web. In particular, when a user has a new interest, the system will not be able to personalize information for the user as it has no information about the user on his new interest. The value of personalization is diminished if improvement can only be made on topics that the user has already encountered. Group-based personalization can solve this cold-start problem, but it could be done effectively only if coherent communities can be identified.

4. CHARACTERIZING THE USERS

A difficulty in making use of clickthroughs for behavior

WIDM'09, November 2, 2009, Hong Kong, China.
ACM 978-1-60558-808-7/09/11.

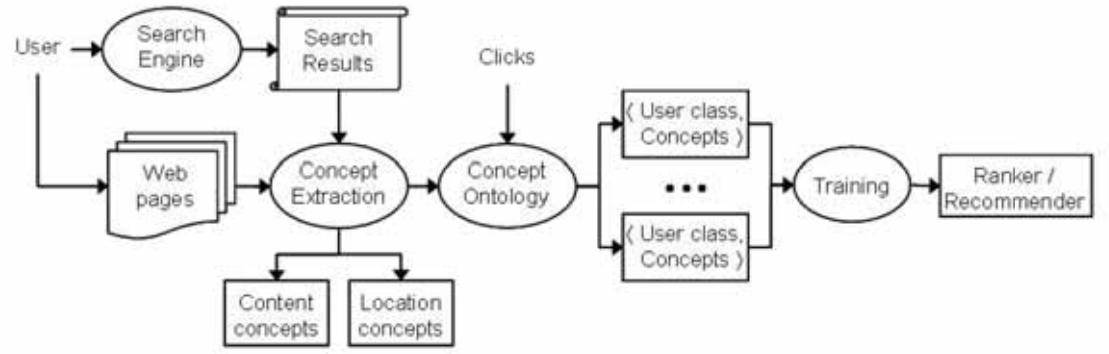

Figure 1: Architecture in deriving user classes.

mining (in fact, any kind of mining) is that a click could mean many things. Let's ignore for the time being cases in which users click on pages out of curiosity (i.e., pages that are interesting but not relevant to the user's information needs on hand) or simply due to mistake. In normal cases, users issuing the same query may look for different kinds of information about the topic. Thus, a click made by one user may not be a good recommendation for another user even though they issued the same query. For example, a high school student, a professional engineer and a financial analyst may all be looking for information on "clean energy" but the pages they click on could have very different natures.

One way to tackle the problem is by classifying users into different classes with the objective that users in the same class benefit each other more than users in other classes. How to classify users when only clickthrough data are available? One way is to look at how diversify a user's clicks are. We can assume that, with respect to the topic being investigated, users who click on pages that cover a large number of concepts under that topic are novice searchers whereas users who click on pages covering a few concepts are expert searchers. According to this assumption, users can be classified into different groups with different levels of focuses. We can envisage that recommendations (or collaborative filtering) are best made between users of the same group, and users with broad focus may benefit from users with narrow focus but probably not the other way around. For example, users who are looking for information to write a survey article on "clean energy" may benefit from users who had surveyed the topic before or from users who had studied in depth some aspects of the topic. After user classes are created, group-based collaborative filtering can be conducted with higher effectiveness than when all users are treated as homogeneous.

5. FROM STATIC TO MOBILE

As everybody has a mobile phone nowadays, the ability to deliver personalized, highly relevant information to a user has become extremely important. Location is an important notion for mobile users and can be divided into two kinds: a user's physical locations (current and past) and his location interests derived from online activities [1]. Notice that a user's physical locations can be considered results of his offline activities (e.g., commuting between home and office) while location interests are derived from his searching and browsing on the web.

Let's consider an example. A user located in USA searches for information about Hong Kong, such as hotels, tourist places and the CIKM 2009 conference. The user could have searched for "Hong Kong" and "CIKM 2009" and then clicked on the tourist information and the conference pages to prepare his trip to Hong Kong. A month later, the user arrives Hong Kong. Knowing that the user is in Hong Kong, the system should realize that he had acquired a lot of information about Hong Kong, including the CIKM 2009 conference, a month ago. The pages he had browsed a month ago now have much higher importance to the user due to his new location. The system will refresh the pages and make them available on his mobile phone. This example illustrates an important concept that a user's online and offline activities must be linked together to provide a full coverage of the user's behavior.

6. CONCLUSIONS

This talk primarily advocates the importance of transforming user clicks into conceptual interests and discusses the opportunity to characterize users by analyzing their clicking patterns. We suggest that users can be classified based on the focus of their clicks but clearly there are many other ways to characterize users with different tradeoffs in cost effectiveness. Finally, we discuss the roles of a user's physical locations and his location interests in delivering important information to the user.

7. ACKNOWLEDGMENTS

This work was supported by grants from the Research Grant Council, Hong Kong SAR, China (Grant Nos. 615707 and CA05/06.EG03) and is a collaborative work with K.W.T. Leung and W.C. Lee.

8. REFERENCES

[1] D. L. Lee. To find or to be found, that is the question in mobile information retrieval. In *Proceedings of the SIGIR 2008 Workshop on Mobile Information Retrieval (MobIR 2008)*, pages 7–10, July 2008.

[2] K. W.-T. Leung, W. Ng, and D. L. Lee. Personalized concept-based clustering of search engine queries. *IEEE Transactions on Knowledge and Data Engineering*, pages 1505–1518, Nov 2008.