IMPROVING PERSONALIZATION SOLUTION THROUGH CONCEPT BASED SEARCH ENGINE

Dr. SIVAKUMAR K

Abstract — User profiling is a fundamental component of any personalization applications. Most existing user profiling strategies are based on objects that users are interested in (i.e. positive preferences), but not the objects that users dislike (i.e. negative preferences). In this paper, we focus on search engine personalization and develop several concept-based user profiling methods that are based on both positive and negative preferences. Experimental results show that profiles which capture and utilize both of the user's positive and negative preferences perform the best. An important result from the experiments is that profiles with negative preferences can increase the separation between similar and dissimilar queries. The separation provides a clear threshold for an agglomerative clustering algorithm to terminate and improve the overall quality of the resulting query clusters.

I. INTRODUCTION

Most commercial search engines return roughly the same results for the same query, regardless of the user's real interest. Since queries submitted to search engines tend to be short and ambiguous, they are not likely to be able to express the user's precise needs. For example, a farmer may use the query "apple" to find information about growing delicious apples, while graphic designers may use the same query to find information about Apple Computer.

Personalized search is an important research area that aims to resolve the ambiguity of query terms. To increase the relevance of search results, personalized search engines create user profiles to capture the users' personal preferences and as such identify the actual goal of the input query. Since users are usually reluctant to explicitly provide their preferences due to the extra manual effort involved, recent research has focused on the automatic learning of user preferences from users' search histories or browsed documents and the development of personalized systems based on the learned user preferences. A good user profiling strategy is an essential and fundamental component in search engine personalization.

Most personalization methods focused on the creation of one single profile for a user and applied the same profile to all of the user's queries. We believe that different queries from a user should be handled differently because a user's preferences may vary across queries. For example, a user who prefers information about fruit on the query "orange", may prefer the information about Apple Computer for the query "apple". Personalization strategies such as [1], [2], [8], [10] employed a single large user profile for each user in the personalization process.

Existing clickthrough-based user profiling strategies can be categorized into *document-based* and *concept-based* approaches. They both assume that user clicks can be used to infer users' interests, although their inference methods and the outcomes of the inference are different. Document-based profiling methods try to estimate users' document preferences [1], [2], [8], [10]. On the other hand, concept-based profiling methods aim to derive topics or concepts that users are highly interested in. While there are document-based methods that consider both users' positive and negative preferences, to the best of our knowledge, there are no concept-based methods that considered both positive and negative preferences in deriving user's topical interests.

Most existing user profiling strategies only consider documents that users are interested in (i.e. users' positive preferences) but ignore documents that users dislike (i.e. users' negative preferences). In reality, positive preferences are not enough to capture the fine-grain interests of a user. For example, if a user is interested in "apple" as a fruit, he/she may be interested specifically in apple recipes, but less interested in information about growing apples, while absolutely not interested in information about the company Apple Computer. In this case, a good user profile should favour information about apple recipes, slightly favour information about growing apple, while downgrade information about Apple Computer.

Profiles built on both positive and negative user preferences can represent user interests at finer details. Personalization strategies include negative preferences in the personalization process, but they all are document-

Dr. Sivakumar K , Ph.D., MIEEE., Assistant Professor , College of Computer Science , King Khalid University - Kingdom of Saudi Arabia .(Email: rksivakumar@gmail.com)

based and thus cannot reflect users' general topical interests.

II. RELATED WORK

User profiling strategies can be broadly classified into two main approaches: *document-based* and *concept-based* approaches. Document-based user profiling methods aim at capturing users' clicking and browsing behaviors. Users' document preferences are first extracted from the clickthrough data and then used to learn the user behaviour model which is usually represented as a set of weighted features. On the other hand, concept-based user profiling methods aim at capturing users' conceptual needs. Users' browsed documents and search histories are automatically mapped into a set of topical categories. User profiles are created based on the users' preferences on the extracted topical categories.

1) Document-Based Methods

Most document-based methods focus on analyzing users' clicking and browsing behaviours recorded in the users' clickthrough data. On web search engines, clickthrough data is an important implicit feedback mechanism from users. Several personalized systems that employ clickthrough data to capture users' interest have been proposed [1], [2], [10]. Joachims [10] proposed a method which employs preference mining and machine learning to model users' clicking and browsing behavior.).

Ng et al. proposed an algorithm which combines a spying technique together with a novel voting procedure to determine users' document preferences from the clickthrough data. They also employed the RSVM algorithm to learn the user behavior model as a set of weight features. More recently, Agichtein et al. [1] suggested that explicit feedback (i.e. individual user behavior, clickthrough data, etc) from search engine users is noisy. One major observation is the bias of user click distribution toward top ranked results. To resolve the bias, Agichtein suggested to clean up the clickthrough data with the aggregated "background" distribution. RankNet [6], a scalable implementation of neural networks, is then employed to learn the user behavior model from the cleaned clickthrough data.

2) Concept-Based Methods

Most concept-based methods automatically derive users' topical interests by exploring the contents of the users' browsed documents and search histories. Liu et al. proposed a user profiling method based on users' search history and the Open Directory Project (ODP). The user profile is represented as a set of categories, and for each category, a set of keywords with weights. The categories stored in the user profiles serve as a context to disambiguate user queries. If a profile shows that a user is interested in certain categories, the search can be narrowed down by providing suggested results according to the user's preferred categories.

Gauch et al. [9] proposed a method to create user profiles from user browsed documents. The method assumes that terms exist frequently in user's browsed documents represent topics that the user is interested in. Frequent terms are extracted from users' browsed documents to build hierarchical user profiles representing users' topical interests. Liu et al. and Gauch et al. both use a reference ontology (e.g. ODP) to develop the hierarchical user profiles, while Xu et al. automatically extracts possible topics from users' browsed documents and organizes the topics into hierarchical structures. The major advantage of dynamically building a topic hierarchy is that new topics can be easily recognized and extracted from documents and added to the topic hierarchy, whereas a reference ontology such as ODP is not always upto-date. Thus, all of our proposed user profiling strategies rely on a concept extraction method, which extracts concepts from web-snippets2 to create accurate and up-to-date user profiles.

III. PERSONALIZED CONCEPT-BASED QUERY CLUSTERING

Our personalized concept-based clustering method consists of three steps. First, we employ a concept extraction algorithm, to extract concepts and their relations from the web-snippets returned by the search engine. Second seven different concept-based user profiling strategies , are employed to create concept-based user profiles. Finally, the concept-based user profiles are compared with each other and against our previously proposed personalized concept-based clustering algorithm.

1) Concept Extraction

Extracting Concepts from Web-snippets

After a query is submitted to a search engine, a list of web-snippets are returned to the user. We assume that if a keyword/phrase exists frequently in the web-snippets of a particular query, it represents an important concept related to the query because it co-exists in close proximity with the query in the top documents.

2) Query Clustering Algorithm

We now review our personalized concept-based clustering algorithm with which ambiguous queries can be classified into different query clusters. Concept-based user profiles are employed in the clustering process to achieve personalization effect. First, a query-concept bipartite graph *G* is constructed by the clustering algorithm with one set of nodes corresponds to the set of users' queries, and the other corresponds to the sets of extracted concepts. Each individual query submitted by each user is treated as an individual node in the bipartite graph by labeling each query with a user identifier. Concepts with interestingness weights (defined in Equation 1) greater than zero in the user profile are linked to the query with the corresponding interestingness weight in *G*. Second, a two-step personalized clustering algorithm is applied to the bipartite graph *G*, to obtain clusters of similar queries and similar concepts.

IV. USER PROFILING STRATEGIES

1) Click-Based Method (*PClick*)

The concepts extracted for a query *q* using the concept extraction method describe the possible concept space arising from the query *q*. The concept space may cover more than what the user actually wants. For example, when the user searches for the query "apple", the concept space derived from our concept extraction method contains the concepts "macintosh", "ipod" and "fruit". If the user is indeed interested in "apple" as a fruit and clicks on pages containing the concept "fruit", the user profile represented as a weighted concept vector should record the user interest on the concept "apple" and its neighborhood (i.e., concepts which having similar meaning as "fruit") , while downgrading unrelated concepts such as "macintosh", "ipod" and their neighbourhood.

2) Joachims-C Method (*PJoachims−C*)

Given a list of search results for an input query *q*, if a user clicks on the document *dj* at rank *j*, all the concepts *C*(*di*) in the *unclicked* documents *di* above rank *j* are considered as less relevant than the concepts *C*(*dj*) in the document *dj* , i.e., (*C*(*dj*) <*r_C*(*di*), where *r_* is the user's preference order of the concepts extracted from the search results of the query *q*).

3) mJoachims-CMethod

Given a set of search results for a query, if documents *di* at rank *i* is clicked, *dj* is the next clicked document right after *di* (no other clicked links between *di* and *dj*), and document *dk* at rank *k* between *di* and *dj* (*i* < *k* < *j*) is not clicked, then concepts *C*(*dk*) in document *dk* is considered less relevant than the concepts *C*(*dj*) in document *dj* (*C*(*dj*) <*r_* *C*(*dk*)) where *r_* is the user's preference order of the concepts extracted from the search results of the query *q*).

4) SpyNB-C Method (*PSpyNB−C*)

Both Joachims and mJoachims are based on a rather strong assumption that pages scanned but not clicked by the user are considered uninteresting to the user and hence irrelevant to the user's query. But instead assumes that unclicked pages could be either relevant or irrelevant to the user. Therefore, SpyNB treats clicked pages as positive samples and unclicked pages as unlabeled samples in the training process. The problem of finding user preferences becomes one of identifying from the unlabeled set reliable negative documents that are considered irrelevant to the user.

5) Click+Joachims-C Method (*PClick+Joachims−C*)

we observed that *PClick* is good in capturing user's positive preferences. In this paper, we integrate the click-based method, which captures only positive preferences, with the Joachims-C method, with which negative preferences can be obtained. We found that Joachims-C is good in predicting users' negative preferences.

6) Click+mJoachims-C Method (*PClick+mJoachims−C*)

Similar to Click+Joachims-C method, a hybrid method which combines *PClick* and *PmJoachims−C* is proposed.

7) Click+SpyNB-C Method (*PClick+SpyNB−C*)

Similar to Click+Joachims-C and Click+mJoachims-C methods, create a hybrid profile *PClick+SpyNB−C* that combines *PClick* and *PSpyNB−C*:

V. EXPERIMENTAL RESULTS

In this section, we evaluate and analyze the seven concept based user profiling strategies (i.e., *PClick*, *PJoachims−C*, *PmJoachims−C*, *PSpyNB−C*, *PClick+Joachims−C*, *PClick+mJoachims−C* and *PClick+SpyNB−C*).

1) Experimental Setup

To evaluate the performance of our user profiling strategies, we developed a middleware for Google3 to collect clickthrough data. We used 500 test queries, which are intentionally designed to have ambiguous meanings (e.g. the query "kodak" can refer to a digital camera or a camera film). We ask human judges to determine a standard cluster for each query. The clusters obtained from the algorithms are compared against the standard clusters to check for their correctness. 100 users are invited to use our middleware to search for the answers of the 500 test queries (accessible at [3]). To avoid any bias, the test queries are randomly selected from 10 different categories.

2) Comparing Concept Preference Pairs Obtained using Joachims-C, mJoachims-C and SpyNB-C Methods

In this Section, we evaluate the pairwise agreement between the concept preferences extracted using Joachims-C, mJoachims-C and SpyNB-C methods. The three methods are employed to learn the concept preference pairs from the collected clickthrough data as described in Section 5.1. The learned concept preference pairs from different methods are manually evaluated by human evaluators to derive the fraction of correct preference pairs. We discard all the ties in the resulted concept preference pairs to avoid ambiguity in the evaluation.

3) Comparing

PClick, *PJoachims−C*, *PmJoachims−C*, *PSpyNB−C*, *PClick+Joachims−C*, *PClick+mJoachims−C* **and** *PClick+SpyNB−C*

4) Termination Points for Individual Clustering to Community Merging

As initial clustering is run, a tree of clusters will be built along the clustering process. The termination point for initial clustering can be determined by finding the point at which the cluster quality has reached its highest. The same can be done for determining the termination point for community merging. The change in cluster quality can be measured by *Similarity*, which is the change in the similarity value of the two most similar clusters in two consecutive steps. For efficiency reason, we adopt the single-link approach to measure cluster similarity. As such, the similarity of two cluster is the same as the similarity between the two most similar queries across the two clusters.

VI. CONCLUSIONS

An accurate user profile can greatly improve a search engine's performance by identifying the information needs for individual users. The techniques make use of click through data to extract from web-snippets to build concept-based user profiles automatically. We applied preference mining rules to infer not only users' positive preferences but their negative preferences, and utilized both kinds of preferences in deriving users profiles. Our experimental results show that profiles capturing both of the user's positive and negative preferences perform the best among the user profiling strategies studied. We plan to take on the following two directions for future work. First, relationships between users can be mined from the concept-based user profiles to perform collaborative filtering. This allows users with the same interests to share their profiles. Second, the existing user profiles can be used to predict the intent of unseen queries, such that when a user submits a new query, personalization can benefit the unseen query. Finally, the concept-based user profiles can be integrated into the ranking algorithms of a search engine so that search results can be ranked according to individual users' interests.

REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proc. of ACM SIGIR Conference*, 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning user interaction models for predicting web search result preferences," in *Proc. of ACM SIGIR Conference*, 2006.

[3] Appendix: 500 test queries. [Online]. Available:http://www.cse.ust.hk/˜dlee/tkde09/Appendix.pdf

[4] R. Baeza-yates, C. Hurtado, and M. Mendoza, "Query recommendation using query logs in search engines," vol. 3268, pp. 588–596, 2004.

[5] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," in *Proc. of ACM SIGKDD Conference*, 2000.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. of the International Conference on Machine learning (ICML)*, 2005.

[7] K. W. Church, W. Gale, P. Hanks, and D. Hindle, "Using statistics in lexical analysis," *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 1991.

[8] Z. Dou, R. Song, and J.-R. Wen, "A largescale evaluation and analysis of personalized search strategies," in *Proc. of WWW Conference*, 2007.

[9] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-based personalized search and browsing," *ACM WIAS*, vol. 1, no. 3-4, pp. 219–234, 2003.

[10] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. of ACM SIGKDD Conference*, 2002.