A survey on retrieving Contextual User Profiles from Search Engine Repository

[1]G.Ravi, [2]JELLA SANTHOSH
[1]Asst.Prof. CSE Dept.
[2] Mtech. (CSE Dept.)
Malla Reddy College of Engineering and Technology, Hyderabad.

Abstract
Personalized search is an important research area that aims to resolve the ambiguity of query terms. Therefore a Personalized Search engine return the most appropriate search results related to users interest. For example, a query "apple" a farmer would be interested in the apple fruits plants, farm etc. a technician would be interested in apple OS, Mac, Macintosh where as a gadget freak would be interested in latest apple products like ipad, iphone, iPod etc.

Here we focus on search engine personalization and develop several concept-based user profiling methods that are based on both positive and negative preferences. We evaluate the proposed methods against our proposed personalized query clustering method.

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.

Keywords-Negative Preferences, Personalization, Positive Preferences, User Profiling.

1.Introduction
A web search engine is designed to search for information on the World Wide Web and FTP servers. The search results are generally presented in a list of results and are often called hits. The information may consist of web pages, images, information and other types of files.A Query is the content the text that the user submits to the search engine. 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.

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).Is personalisation for every organisation? Probably not. If your website does not have enough content to personalise then there is little point in trying to fragment it into profiles or tracked experiences - but if your site is large and you are struggling to ensure users get presented with appropriate content - then it would be one very powerful way to improve the user experience.

Recommended systems technologies have been introduced to help people deal with these vast amounts of information, and they have been widely used in research as well as e-commerce applications, such as the ones used by Amazon and Netflix. The most common formulation of the recommendation problem relies on the notion of ratings, i.e., recommender systems estimate ratings of items (or products) that are yet to be consumed by users, based on the ratings of items already consumed. Recommender systems typically try to predict the ratings of unknown items for each user, often using other users' ratings, and recommend top *N* items with the highest predicted ratings. Accordingly, there have been many studies on developing new algorithms that can improve the predictive accuracy of recommendations. However, the quality of recommendations can be evaluated along a number of dimensions, and relying on the accuracy of recommendations alone may not be enough to find the most relevant items for each user. In particular, the importance of *diverse* recommendations has been previously emphasized in several studies. These studies argue that one of the goals of recommender systems is to provide a user with highly idiosyncratic or personalized items, and more diverse recommendations result in more opportunities for users to get recommended such items. With this motivation, some studies proposed new recommendation methods that can increase the diversity of recommendation sets for a given *individual* user, often measured by an average dissimilarity between all pairs of recommended items, while maintaining an acceptable level of accuracy. These studies measure recommendation diversity from an individual user's perspective (i.e., *individual diversity*).

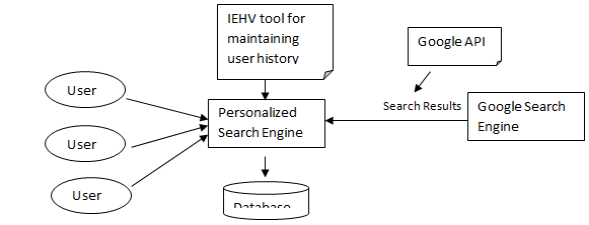
As a model for knowledge description and formalization, ontology's are widely used to represent user profiles in personalized web information gathering. However, user profiles, many models have utilized only knowledge from either a global knowledge base or user local information.

By this survey we conclude that

- Most of the present search engines are not personalized.
- Users queries that is what the user is interested in searching are not anticipated correctly.
- Users interested search results are not appropriately received.
- Users previous search record is not maintained. Whatever interested or his clicked are done by the user are not recorded.
- Few personalized search engine exists, but ask user to fill a detailed form of his personal information based on that information these search engines generate a near about personalized search results.
- Most of the users refrain from providing their detailed information to unknown website.

2.Proposed System

Architecture:



- User past search history is recorded.
- User is given a login account.
- User profile is created based on the previous searches done by the user.
- User positive and negative preferences are calculated. Positive preferences are the concepts or the topics in which a user is interested and negative preferences are the concepts in which the user is not interested at all.
- A track is maintained on what user clicked.
- Next time when a user login to his personalized search engine account he receives his most interested hits.
- Next time when he fires the same previous query he gets the results with his interested concept appended to the query.
- To extract users' positive preferences PClick algorithm is used and to extract users positives and negative preferences PmJoachims_C algorithm is used.
- Concept clustering is done using Personalized Agglomerative Clustering algorithm.

2. 1. CONSTRUCTING A USER PROFILE

Any personal documents such as browsing history and emails on a user's computer could be the data source for user profiles. This focus on frequent terms limits the dimensionality of the document set, which further provides a clear description of users' interest. This approach proposes to build a hierarchical user profile based on frequent terms. In the hierarchy, general terms with higher frequency are placed at higher levels, and specific terms with lower frequency are placed at lower levels.

2.2. PRIVACY VS SEARCH QUALITY

In this experiment, users are required to try different privacy thresholds to explore the relationship between privacy preservation and search quality. For each query, all parameters are fixed. A group of search results is presented to show how search quality is affected by the amount of private information that is exposed. These include general interest terms like "research", "search", "sports" and websites frequently visited such as "Google" and "NYTimes". Experiments showed that these general terms are especially helpful in identifying ambiguous queries like "conference" and "IT news". At the opposite extreme, over 100 terms are exposed. Most of these terms indicate specific events that happened recently, such as "websites that are occasionally visited (such as friends' blogs) which are too detailed to help refine the search.

2.3. QUERY CLUSTERING

User's queries can be classified into different query clusters. Concept-based user profiles are employed in the clustering process to achieve personalization effect. The personalized clustering algorithm iteratively merges the most similar pair of query nodes, and then, the most similar pair of concept nodes, and then, merge the most similar pair of query nodes, and so on. Each individual query submitted by each user is treated as an individual node and each query with a user identifier.

2.4. CLICK THROUGHS

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"),

Concept Preference Pairs Obtained Using Joachims-C Methods

Concept Preference Pairs for d_1	Concept Preference Pairs for d_5	Concept Preference Pairs for d_8
Empty Set	apple store $<_{r'}$ product macintosh $<_{r'}$ product	macintosh $<_{r'}$ product catalog $<_{r'}$ product
	apple store $<_{r'}$ mac os macintosh $<_{r'}$ mac os	macintosh $<_{r'}$ mac os catalog $<_{r'}$ mac os
	macintosh $<_{r'}$ apple store apple store $<_{r'}$ iPod macintosh $<_{r'}$ iPod	macintosh $<_{r'}$ apple store catalog $<_{r'}$ apple store macintosh $<_{r'}$ iPod catalog $<_{r'}$ iPod
		macintosh $<_{r'}$ fruit catalog $<_{r'}$ fruit macintosh $<_{r'}$ apple hill catalog $<_{r'}$ apple hill
		macintosh $<_{r'}$ fruit catalog $<_{r'}$ fruit

while downgrading unrelated concepts such as"macintosh," "ipod," and their neighborhood. The click-based profile PClick in which the user is interested in information about "macintosh." Hence, the concept "macintosh" receives the highest weight among all of the concepts extracted for the query "apple." The weights of the concepts "mac os," "software," "apple store," "iPod," "iPhone," and "hardware" are increased based on method we used, because they are related to the concept "macintosh." The weights for concepts "fruit," "apple farm," "juice," and "apple grower" remain zero, showing that the user is not interested in information about "apple fruit."

3.Conclusion

An accurate user profile can greatly improve a search engine's performance by identifying the information needs for individual users. In this paper, we proposed and evaluated several user profiling strategies. 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 also their negative preferences, and Utilized both kinds of preferences in deriving users profiles.

The user profiling strategies were evaluated and compared with the personalized query clustering method that we proposed previously. 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. Apart from improving the quality of the resulting clusters, the negative preferences in the proposed user profiles also help to separate similar and dissimilar queries into distant clusters, which helps to determine near optimal terminating points for our clustering algorithm. We plan to take on the following two directions forfuture work. First, relationships between users can be mined from the concept-based user profiles to perform collaborative filtering. This allows users with th 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," Proc. ACM SIGIR, 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning User Interaction Models for Predicting Web Search Result Preferences," Proc. ACM SIGIR, 2006.

[3] Appendix: 500 Test Queries, http://www.cse.ust.hk/~dlee/ tkde09/Appendix.pdf, 2009.

[4] R. Baeza-yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Workshop Current Trends in Database Technology, pp. 588-596, 2004.

[5] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. ACM SIGKDD, 2000.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to Rank Using Gradient Descent," Proc. Int'l Conf. 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, Lawrence Erlbaum, 1991.

[8] Z. Dou, R. Song, and J.-R. Wen, "A Largescale Evaluation and Analysis of Personalized Search Strategies," Proc. World Wide Web (WWW) Conf., 2007.

[9] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing," ACM Web Intelligence and Agent System, vol. 1, nos. 3/4, pp. 219-234, 2003.

[10] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD, 2002.

[11] K.W.-T. Leung, W. Ng, and D.L. Lee, "Personalized Concept- Based Clustering of Search Engine Queries," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp. 1505-1518, Nov. 2008.

[12] B. Liu, W.S. Lee, P.S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," Proc. Int'l Conf. Machine Learning (ICML), 2002.

Jella Santhosh

Pursuing Mtech. (CSE Dept.) Malla Reddy College of Engineering and Technology, Hyderabad.

Mr. G.Ravi

Asst.Prof. CSE Dept. Malla Reddy College of Engineering and Technology, Hyderabad.