Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.555 – 558

RESEARCH ARTICLE

Modified Approach of User Profiling from Search Engine Logs

C. DEEPA[1], S. PRATHIBA[2]

[1]Department of Information Technology, Bharath University, India
[2]Department of Information Technology, Bharath University, India

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 references), 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. We evaluate the proposed methods against our previously 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.***

Key Terms: - Search Engine; ranking; concept based search engines; Logs

I. Introduction

A key factor for the popularity of today's Web search engines is the friendly user interfaces they provide. Indeed, search engines allow users to specify queries simply as lists of keywords, following the approach of traditional information retrieval systems [1,2,3]. Keywords may refer to broad topics, to technical terminology, or even to proper nouns that can be used to guide the search process to the relevant collection of documents.

Despite that this simple interaction mechanism has proved to be successful for searching the Web, a list of keywords is not always a good descriptor of the information needs of users. It is not always easy for users to formulate effective queries to search engines. One reason for this is the ambiguity that arises in many terms of a language.

Queries having ambiguous terms may retrieve documents which are not what users are searching for. On the other hand, users typically submit very short queries to the search engine, and short queries are more likely to be ambiguous. From a study of the log of a popular search engine, Jansen *et al* [4], conclude that most queries are short (around 2 terms per query) and imprecise.

Users searching for the same information may phrase their queries differently. Often, users try different queries until they are satisfied with the results. In order to formulate effective queries, users may need to be familiar with specific terminology in a knowledge domain.

The specific contributions of this paper include: Analysis of alternatives for incorporating user behavior into web search ranking. An application of a robust implicit feedback model derived from mining millions of user interactions with a major web search engine. A large scale evaluation over real user queries and search results, showing significant improvements derived from incorporating user feedback. We summarize our findings and discuss extensions to the current work, which concludes the paper.

II. LITERATURE SURVEY

Baeza-Yates [1] presents a survey on the use of Web logs to improve different aspects of search engines. Wen et al. [5] propose to cluster similar queries to recommend URLs to frequently asked queries of a search engine. They use four notions of query distance: (1) based on keywords or phrases of the query; (2) based on string matching of keywords; (3) based on common clicked URLs; and (4) based on the distance of the clicked documents in some pre-defined hierarchy. Befferman and Berger [6] also propose a query clustering technique based on distance notion (3). Notions (1)-(3) are difficult to deal with in practice, because distance matrices between queries generated by them from real query logs are very sparse, and many queries with semantic connections appear as orthogonal objects in such matrices. Ad-hoc clustering algorithms are needed to deal with this problem. Notion (4) needs concept taxonomy and requires the clicked documents to be classified into the taxonomy as well. Fonseca *et al* [3] present a method to discover related queries based on association rules. Here queries represent items in traditional association rules. The query log is viewed as a set of transactions, where each transaction represents a *session* in which a single user submits a sequence of related queries in a time interval. Their notion of query session is different than the notion we use in this paper. The method shows good results, however two problems arise. First, it is difficult to determine sessions of successive queries that belong to the same search process; on the other hand, the most interesting related queries, those submitted by different users, cannot be discovered.

This is because the support of a rule increases only if its queries appear in the same query session, and thus they must be submitted by the same user. Zaiane and Strilets [7] present a method to recommend queries based on seven different notions of query similarity. Three of them are mild variations of notion (1) and (3). The remainder notions consider the content and title of the URL's in the result of a query. Their approach is intended for a meta-search engine and thus none of their similarity measures consider user preferences in form of clicks stored in query logs.

Another approach adopted by search engines to suggest related queries is *query expansion* [2, 8]. The idea here is to reformulate the query such that it gets closer to the term-weight vector space of the documents the user is looking for. Our approach is different since we study the problem of suggesting related queries issued by other users and query expansion methods construct artificial queries. In addition, our method may recommend queries that are related to the input query but may search for different issues, thus redirecting the search process to related information of interest to previous users.

III. PROPOSED SYSTEM ARCHITECTURE

A. EXISTING SYSTEM

Existing click through-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 (i.e., users are interested in some documents more than others) On the other hand, concept based profiling methods aim to derive topics or concepts that users are highly interested. 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.

DRAWBACK IN EXISTING SYSTEM

Most existing user profiling strategies only consider documents that users are interested in (i.e., users' positive preferences) but ignore documents that users dislike

Personalization strategies include negative preferences in the personalization process, but they all are document-based, and thus, cannot reflect users' general topical interests.

B. PROPOSED SYSTEM

In this paper, we address the above problems by proposing and studying seven concept-based user profiling strategies that are capable of deriving both of the user's positive and negative preferences. All of the user profiling strategies is query-oriented, meaning that a profile is created for each of the user's queries. The user profiling strategies are evaluated and compared with our previously proposed personalized query clustering method. Experimental results show that user profiles which capture both the user's positive and negative preferences perform the best among all of the profiling strategies studied. Moreover, we find that negative preferences improve the separation of similar and dissimilar queries, which facilitates an agglomerative clustering algorithm to decide if the optimal clusters have been obtained. We show by experiments that the termination point and the resulting precision and recalls are very close to the optimal results. The proposed system architecture is as shown in Figure 1

ADVANTAGES IN PROPOSED SYSTEM

We extend the query-oriented, concept-based user profiling method *proposed* to consider both users' positive and negative preferences in building users profiles.

Our proposed methods use an RSVM to learn from concept preferences weighted concept vectors representing concept-based user profiles

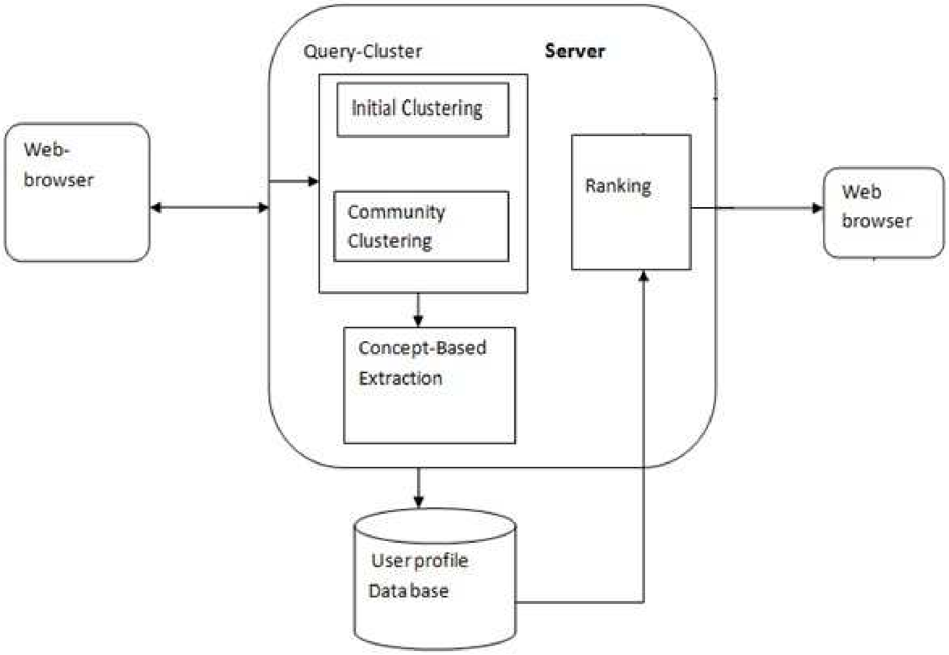

Fig 1 System Architecture

IV. Implementation

In this paper, our approach consists following Modules

- User Interface
- Collection of unrelated data
- Data search

- Concept-based user profiling method
- Result calculation

A. *User Interface:*

In this module we are going to design all the user interface for search by which the user can communicate with the server and make query for their search and can get the data from the server and also user login.

B. *Collection of data*

First we have to collect the specific data about an object and it is stored in related database in order to get the data for the users' queries. The data may be conflict but we have to retrieve the data according to the user query.

C. *Query Clustering:*

In this module we are going to review our personalized concept based clustering in order to achieve personalization effect by differentiating different queries in their cluster. In this module user retrieve the specific data about a query.

D. *Concept Extraction:*

These methods consider both users' positive and negative preferences in building users profiles. We proposed six user profiling methods that exploit a user's positive and negative preferences to produce a profile for the user using Ranking SVM (RSVM) .these methods are concentrated on extracting the concepts for the users query.

E. *Result calculation:*

In this result calculation module we have to retrieve the data which is the Expected output from the user for their query and also we have to compare the result of the proposed Technique with the existing.

V. CONCLUSIONS

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' utilized both kinds of preferences in deriving users' profiles. Positive preferences but also their negative preferences and utilized both kinds of preferences in deriving users profiles.

REFERENCES

[1] R. Baeza-Yates. Query usage mining in search engines. Web Mining: Applications and Techniques, Anthony Scime, editor. Idea Group, 2004.

[2] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval, chapter 3, pages 75–79. Addison-Wesley, 1999.

[3] B. M. Fonseca, P. B Golgher, E. S. De Moura, and N. Ziviani. Using association rules to discovery search engines related queries. In First Latin American Web Congress (LAWEB' 03), November, 2003. Santiago, Chile.

[4] M. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the web. ACM SIGIR Forum, 32(1):5-17, 1998.

[5] J. Wen, J. Nie, and H. Zhang. Clustering user queries of a search engine. In Proc. at 10th International World Wide Web Conference, pages 162–168. W3C, 2001.

[6] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In KDD, pages 407–416, Boston, MA USA, 2000.

[7] O. R. Zaiane and A. Strilets. Finding similar queries to satisfy searches based on query traces. In Proceedings of the International Workshop on Efficient Web-Based Information Systems (EWIS), Montpellier, France, September, 2002.

[8] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with the local context analysis. ACM Transaction of Information Systems, 1(18):79–112, 2000.