www.ijarcsse.com

Volume 3, Issue 6, June 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Personalized Web Search

Charanjeet Dadiyala* , Prof. Pragati Patil, Prof. Girish Agrawal

Department of M.Tech(CSE),RTMNU

India

Abstract— ***Today Internet and web search engines have become an important part in ones day to day life. Where Web search engines provides simple and yet user friendly interfaces, users give their queries in terms of simple words or keywords. Depending upon the keyword, the web search engine extract a list of web pages which rely primarily on the matching of the keywords. This approach has some limitations, like, the ambiguity of user needs, same keyword with different meaning in different context. Present search engines generally handle search queries or keywords without considering user preferences or contexts in which users submit their queries. At times, user also fail to use proper keywords that represent their information need accurately. Ambiguous keywords, different needs of users at different times, and the limited ability of user to precisely express what they need have been widely recognized as one of the challenging obstacle in improving search results. In this paper, we propose an approach to improve the retrieval quality of web engine and refining the search results depending on the users need.***

Keywords— ***Search engine, web search, web results, search queries.***

I. INTRODUCTION

Over recent years, the World Wide Web has become a new communication medium with Web information access. This incorporates with informational, cultural, social and evidential values to be specific. With the existence of various Search Engines e.g. Google, Yahoo and many more, the users are tend to use them for retrieving their desired Web pages and their information. Although today's search engines can meet a general request, they cannot distinguish different users' specific needs well. For example, a computer geek may use the search term Leopard to search for information on Apple OS X Leopard, but a biologist or any general user as well may use the same term Leopard to find information on the animal Leopard; however, a public search engine treat the two queries the same way. Alternatively personalized web search results provide customized results depending on each user's interests.

II. PROBLEM DEFINITION

A much more popular way for users to get easily their needed information over the web is Keyword based query rather than using a SQL queries in commercial search engines like Google and Yahoo! These search engines also provide a very user friendly, yet simple user interfaces to pose search queries simply in terms of keywords. Though the simple and user friendliness feature of a search engine fails at times to satisfy an individual's information goal. It has observed that the difficulty in finding only those which satisfy an individual's information goal increases with the keyword which has different meaning at different context. As the working of any search engine primarily based on the matching of the keywords to the desired documents to determine which Web pages will be returned given a search query. So, there are main two limitations of Keyword-based search queries. First, there are some keywords, which have different meanings in different context and hence the ambiguity of user need to be resolved as to get proper and relevant information over the Internet. The search engines currently available and used by the users generally handle search queries without considering user preferences or contexts in which users submit their queries. Another limitation is to choose proper and relevant search terms which expresses the user's need the best in the given context. Ambiguous keywords used in Web queries, the diverse needs of users, and the limited ability of users to precisely express what they want to search in a few keywords have been widely recognized as a challenging obstacle in improving search quality. One of the popular approach in recent data engineering field is encoding human search experiences and personalizing the search results using ranking optimization. This approach enhance the quality of information retrieval i.e. the quality of the search results of Web Search. The search results provided by the present search engines are primarily based on the matching of the keywords and hence another approach as result re-ranking can be seen for the refinement and quality improvement of the same. A general process of search result re-ranking can be used to re-order search results with the help of the personalized ranking criteria. Such criteria are typically observed, studied, derived and then can be implemented from the user's search history log or simply from the modelling of user's search behaviour and interests. There are basically two types of user's search interest depending upon time as a parameter as short-term and long-term interests. Short-term interest shows the behaviour of the user and desired search queries over shorter time period, whereas long-term interest shows the same over a considerably longer time period. For example, interest shown in the context of buying any commodity say, mobile phone comes under short-term interest, and interest shown in the context of let's say, planning a tour, comes under log-term interest. Even though short-term interests based personalized search uses the most recent search histories which gives efficient results at some times only, it is generally unstable and fails to capture the changing behaviour of the users. Again, most of existing long-term interests based personalization using the entire recent and previous search histories fails to differentiate the relevant search history from the irrelevant search history, making it difficult to be an effective parameter for search alone as well. On the other hand, commercial present search engines give related keyword suggestion based on the queries inputted by the users and hence it helps the users in rephrasing their query formulation for improving search quality. Related search terms in Google is based on the assumption that sometimes the best search terms for what a user is looking for are related to the ones the user actually entered. In the search box of Yahoo!, Search Assist compares an input query to searches all other Yahoo! users have composed and offers suggestions in real time. These related keyword suggestion based methods actually helps the users in order to clarify their information needs or to rephrase their query formulation for retrieving more related and relevant search results along with specifying some alternative related queries.

The suggestion services supplied by those popular search engines highlight the importance of related keyword suggestion. But these techniques implemented by these commercial search engines are usually confidential and not revealed to anyone, whereas many academic researchers have showed large amount of interest in the study of the process of query suggestion.

III. GOAL/ PROPOSED WORK

A study of the log of a popular search engine reported that most search queries are about two terms per query. Therefore, the difficulty is that since Web users typically submit very short queries to search engines, the very small term overlap between queries cannot accurately estimate their relatedness. Given this problem, the technique to find semantically related queries (though probably dissimilar in their terms) is becoming an increasingly important research topic that attracts considerable attention. After the survey and research, it has been found that the need of having a search engine procedure or any searching technique which gives more refined and accurate search results in any of the user defined context. As the various search engines currently present in the market may or may not give the relevant or related search results. So to fill the gap between the output of a search engine from related search results to more related and relevant search results, a technique is required.

With the previous work and researches, the goal is to propose a technique or a procedure of learning the behaviour of a user surfing the net over a period of time and to refine the search results using the same click-through data.

IV. PERSONALIZED SEARCH

Kraft et al. [12] state that the context, in its general form, refers to any additional information associated with the query in the web search field, and also present three different algorithms to implement the contextual search instead of modelling user profiles. Generally speaking, if the context information is provided by an individual user in any form, whether automatically or manually, explicitly or implicitly, search engines can use the context to custom-tailor search results. The process is named as a personalized search. In this way, such a personalized search could be either server-based or client-based.

In the context of personalized search, one of the main component is learning user’s interest and their preferences. Many schemes for building and learning user profiles includes several schemes to figure user preferences from text documents. But the observation says that modelling user profiles or learning from text documents shows some amount of error which generally doesn’t consider the term correlations. Hence, a kind of a simple scheme is a taxonomic hierarchy, particularly generated as a tree structure, which also overcomes the drawbacks of learning from text documents, also called as the bag of words.

A. *Finding related keyword:*

The techniques to find semantically related queries is becoming an increasingly important research topic that attracts considerable attention. Existing techniques differ from one another in terms of how to improve the naive query term based suggestion which simple thinks that two Web queries are related if they share common terms. On the Web, recent studies are interested in using Web logs as an additional source to enrich short Web queries. There are two kinds of feature spaces commonly used in the literature, i.e., content-sensitive and content-ignorant features. Beeferman et al. [5] used single-linkage clustering to cluster related queries based on the common clicked URLs two queries share, a content-ignorant feature space. Wen et al. [13] further proposed three kinds of features to compute query to query relatedness: 1) based on terms of the query, 2) based on common clicked URLs, and 3) based on the distance of the clicked Web pages in a predefined hierarchy. The terms in a short Web query would not give reliable information, while the limitation of URL feature space is that two Web pages with different URLs may be semantically related in contents. The third features in [13] needs a concept taxonomy and requires Web pages to be classified into the taxonomy as well. Such taxonomy is not generally available. Baeza-Yates et al. [1, 2] find related queries based on the content of clicked Web pages using click frequency as a weighting scheme. Their experiments show that using the content information of a Web page (e.g., nouns) is a more accurate query enrichment way to measure query similarity than using the URL of a Web page.

B. *Query-URL context*

The content-based feature space, e.g., terms of a Web page, however, is not applicable, at least in principle, in settings including: non-text pages like multimedia (image) files, Usenet archives, sites with registration requirement, and dynamic pages returned in response to a submitted query and so forth. It is crucial to improve the quality of the URL (content-ignorant) feature space since it is generally available in all types of Web pages. The query-URL relationship can be represented by a bipartite graph. Finding biclique is a natural way of collecting the most related queries and URLs. A well-known problem related to biclique is the maximum clique, which is one of the most widely studied NP-complete problems in the literature [10].

Graph partitioning is an alternative for grouping which is done by cutting the set of vertices into disjoint sets. Beeferman et al. [1, 5] viewed the click-through data as a bipartite graph, and utilized an iterative, agglomerative clustering algorithm to the vertices of the graph for clustering queries and URLs, respectively. Their method just extracted connected components from the entire graph and used frequency to measure the similarity between queries.

The limitation of this method is the weakness of selecting queries from the most frequently occurred connected component that contains the input query (keyword list). However, the selected queries may not be the best query suggestion, as the frequency is not always the best descriptor of relatedness because it does not discern the individual targeted queries.

In addition, an alternative representation for query-URL data can be given by a contingency matrix whose rows correspond to queries and columns to URLs. This matrix is sparse, since the majority of queries retrieve only a small number of URLs. The elements of the matrix can be set as binary or weighted according to a measure (e.g., each entry is the probability of choosing same query and same URL).

C. *Query clustering*

Query clustering also helps find related Web queries, which appears to be less explored than clustering. Web pages or documents [1, 2, 5, 13]. Wen et al. [1, 13] proposed to cluster similar queries to recommend URLs to frequently asked queries of a search engine. They combined similarities based on query contents and user clicks, and regarded user clicks as an implicit relevance feedback but not the top ranked Web pages. The distilled search-related navigation information from proxy logs to cluster queries. The data they relied on differed from those used in the above other studies. In addition, there are three URL-based similarity measures analytically and empirically to provide better understanding of the propagation of similarity from query to query by inducing an implicit topical relatedness between queries.

V. Conclusion and future work

The use of Internet in the recent years is growing rapidly which makes the need of a technique which can give accurate and relevant results to the user. Although there are several search engines currently present, it has been observed that they fails to capture user's preference and behaviour and hence the search results may or may not be related with the context of the user. In this paper, hence we proposed a possible technique which can give users an experience of personalized web search and ultimately users can get what they want in a crisp manner in shorter time and fewer clicks as well. In future, the concept of query keyword suggestion can be added and with the feature of query formulation and query expansion, which helps the user at those times when users are not sure about the search query terms, so that these feature will guide the user to get the desired information in a very specific context with comparably less effort.

Acknowledgment

We would like to thanks all the authors whom we referred for giving their suggestions and making the material available for us to refer.

References

[1] C.S.Dadiyala, Pragati Patil and Girish Agrwal. A review of Query Log and Query Clustering. In Proceedings of the International Journal of IJERT, vol.1-issue 10, December 2012.

[2] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In Proceedings of the 16th International Conference on World Wide Web (WWW'07), pages 581–590, Banff, Alberta, Canada, 2007.

[3] R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza. Improving search engines by query clustering. JASIST, 58(12):1793–1804, 2007.

[4] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. In Proceedings of the 14th International Conference on World Wide Web - Special interest tracks and posters (WWW'06), pages 801–810, Chiba, Japan, 2005.

[5] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Trans. Knowl. Data Eng., 15(4):784–796, 2003.

[6] L. Li, Z. Yang, B. Wang, and M. Kitsuregawa. Dynamic adaptation strategies for long-term and short-term user profile to personalize search. In Proceedings of A Joint Conference of the 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management, pages 228–240, Huang Shan, China, 2007.

[7] Y. Li, Z. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng., 15(4):871–882, 2003.

[8] F. Liu, C. T. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. IEEE Trans. Knowl. Data Eng., 16(1):28–40, 2004.

[9] Y. Lv, L. Sun, J. Zhang, J.-Y. Nie, W. Chen, and W. Zhang. An iterative implicit feedback approach to personalized search. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL'06), pages 585–592, Sydney, Australia, 2006.

[10] S. Otsuka and M. Kitsuregawa. Clustering of search engine keywords using access logs. In Proceedings of 17th International Conference on Database and Expert Systems Applications (DEXA'06), pages 842–852, Krak´ow, Poland, 2006.

[11] S. Otsuka, M. Toyoda, J. Hirai, and M. Kitsuregawa. Extracting user behaviour by web communities' technology on global web logs. In Proceedings of 15th International Conference on Database and Expert Systems Applications (DEXA'04), pages 957–968, Zaragoza, Spain, 2004.

[12] R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar. Searching with context. In Proceedings of the 15th International Conference on World Wide Web (WWW'06), pages 477–486, Edinburgh, Scotland, UK, 2006.

[13] J.-R. Wen, J.-Y. Nie, and H. Zhang. Query clustering using user logs. ACM Trans. Inf. Syst., 20(1):59–81, 2002.