論文の内容の要旨

論文題目

ENHANCING WEB SEARCH BY PERSONALIZED RE-RANKING AND RELATED KEYWORD SUGGESTION

（検索順位の個人化及び関連語提示を用いた Web 検索の高度化に関する研究）

氏名　李　琳

（本文）

After a few decades since its inception, the World Wide Web has become a new communication medium with informational, cultural, social and evidential values. Search engines are widely used on the Web and they are making more information easily accessible than ever before. Although the general Web search today is still performed and delivered predominantly through search algorithms, e.g., Google's PageRank based query independent ranking algorithms, the interests in helping Web users effectively get their desired Web pages in ranking have been growing over the recent years.

In commercial search engines like Google and Yahoo!, keyword based query is a much more popular way to let Web users easily specify their information needs than SQL queries for Web information access. The simple and yet friendly Web user interfaces provided those commercial search engines allow users to pose queries simply in terms of keywords. However, the difficulty in finding only those which satisfy an individual's information goal increases. This is because search engines primarily rely on the matching of the query terms to the document terms in the desired documents to determine which Web pages will be returned given a keyword-based query. The main limitation with keyword-based search is two folds. First, due to the ambiguity of user needs, some keywords have different meanings in different context, such as mouse trap, Jaguar, Java and so on. Present search engines generally handle search queries without considering user preferences or contexts in which users submit their queries. Furthermore, users

often fail to choose proper terms that best express their information needs. 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.

Currently, encoding human search experiences and personalizing the search result delivery through ranking optimization is a popular approach in recent data engineering field to enhancing the result quality of Web search and user experience with the Web today. A general process of search result re-ranking is to devise efficient mechanisms to re-order the search result ranking using the global importance by personalized ranking criteria. Such criteria are typically derived from the modeling of users' search behavior and interests. However, most of existing long-term interests based personalization using the entire recent and previous search histories fails to distinguish the relevant search history from irrelevant search history, making it harder to be an effective measure alone for search personalization.

On the other hand, commercial search engines give keyword suggestions on the queries input by users, thus assisting them in rephrasing their query formulation to improve 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. The purpose of these methods is to help users specify alternative related queries in their search process in order either to clarify their information needs or to rephrase their query formulation to retrieve more related search results. The services supplied by those popular commercial search engines highlight the importance of query recommendation. Although the techniques used in these proprietary commercial search engines are usually confidential, researchers have showed growing interests in query recommendation. A study of the log of a popular search engine reported that most 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.

In this thesis, we study how to enhance the retrieval quality of Web search. We are mainly interested in personalized search and related keyword suggestion. Following, we will give a detailed outline of the thesis.

First, the thesis starts by introducing our research background and motivation. We

review some related work, including personalized search, learning users' interests, building ontology-based user profiles, Web query clustering, Web query expansion, Web query classification, rank aggregation, and so on.

Second, in personalized re-ranking for Web search enhancement, we present four re-ranking strategies by augmenting user search history. Our works focus on studying learning user profiles and utilizing the learned user profiles to re-rank search results. Because user preferences change over time, it becomes important to keep the user profile up-to-date, and for a search engine to adapt accordingly. In addition, a user profile covers both short-term and long-term user preferences. Using one model to represent two differently featured parts of the user profile will be far from perfect. To address this problem, one of our works is that we design use profiles which contain a taxonomic hierarchy for the long-term model and a recently visited page-history buffer for the short-term model. We also propose updating strategy to capture the changes of user preferences. For re-ranking search results, we study score and rank aggregation methods by using ontology-based user preferences, i.e., the hierarchy of the ODP structure. The other work is that we discuss how metadata like ODP (Open Directory Project) can be further exploited to achieve high quality personalized Web search. The metadata expressing topical categorizations of web pages in ODP is manually entered and edited, and commonly considered as useful means for automated learning of user interests and user search behaviors. We propose a query context window (QCW) based framework for query-centric Selective uTilization of search history in personalized leArning and re-Ranking (STAR). Our STAR framework consists of three design principles and a suite of algorithms for learning and encoding users' short-term and long-term search interests and re-ranking of search results through a careful combination of recent and previous search histories.

Third, in finding related keywords as suggestion for Web search enhancement, we are interested in finding semantically related queries. Since Web queries are usually a couple of words, we utilize query enrichment as an effective method to enrich the representation of a Web query by alternative feature spaces, instead of using terms in the query. We mainly study how to get suitable feature spaces in this thesis. One of our works is that we study two kinds of feature spaces from search results of a query, i.e., content-sensitive (e.g., nouns) and content-ignorant (e.g., URLs). Our experiment results show that the URL feature space produces lower precision scores than the noun feature space which, 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. We are inspired to propose a novel content-ignorant feature space, i.e., Web community which is a collection of Web pages with different URLs, but sharing common interest on a specific topic. Experiment results show that the novel Web community feature space generates much better results than the traditional URL feature space. The other work is that we improve the performance of the URL feature space from another viewpoint. We generate query affinity graph based the similarity measure between pairs of queries using query-URL vector model. We propose to utilize the monotonicity of the merging distances of a hierarchical agglomerative clustering (HAC) which can capture the propagation of similarity (distance) implicitly. By using HAC based algorithms we can globally capture the diffusive transition of similarity on each connected component extracted from query affinity graph to rank queries. Furthermore, instead of fixing the degree of similarity of queries based on their query-URL vector similarity, our approach adaptively controls the output of different lists of related queries in terms of the level of users' satisfaction.

Finally, the thesis gives a summary of main results and a discussion of the future work and remaining open problems.