University of Macau

Abstract

LOG MINING TO SUPPORT WEB QUERY EXPANSIONS

by Ngok Man Chan

Thesis Supervisor: Associate Professor Gong Zhiguo
E-Commerce Technology Master Degree Program

In the modern world, information is abundantly available on the Internet. Despite the availability, finding useful information is not an easy task, especially when information of such volume is poorly organized. The growing user needs, together with the accumulation of information at an exponential rate, have made the development of better search algorithms increasingly pressing.

Although most popular search engines have been very efficient in responding to a user search query by generating hundreds or even thousands of documents as search results, the relevance of these documents are largely unsatisfactory.

Traditionally, the query suggestions that are provided simultaneously by the search engines in response to a submitted query are usually syntactically similar (e.g., a submitted query is "encryption", then the query suggestion is "encryption algorithm"), but not semantically related.

There are two reasons for such situation, e.g., 1) the query is too short; 2) the query meaning may not be expressed well enough for users' needs.

This project therefore aims to propose a new mechanism to analyze the search queries in the hope of providing more precise suggestions and improving the search quality.

In order to overcome the first problem, query expansion will be used for the inputted query. As a result, the returned results quantity can be reduced to help users to review more effectively.

Furthermore, to solve the second problem, query expansion will be done in three directions in order to have better expression for users' needs. They are: (1) association method, (2) query information in a query log, and (3) a thesaurus method. Specifically,

1. A log file that has recorded the queries history in a search engine is processed and grouped using association technique. These are done so that the relationships (association rules) between the query terms can be effectively established, which are subsequently used in the query expansion. As a result, every new query can be reformulated, and expanded by the established association rules.

2. Other information (e.g., similar queries comparable to the new input query in the log file) in the query log is used as query expansion.

3. A thesaurus, i.e., WordNet, is used to expand the submitted query by making use of its synonym, hypernym, and/or hyponym.

The above three approaches, e.g., query expanded by association and WordNet, can collectively help to solve the second foregoing stated problem that there may be keywords which are not syntactically similar as the inputted query, but are related to users' needs.