Abstract

The development of the WWW challenges the effectiveness of classical methods and requires further improvement in Web Information Retrieval (IR) techniques. A major problem with current Web IR systems is that only keywords are used to express search requests. This causes these systems' failures in clearly understanding users' information needs. An Empirical method based on global scale user behavior analysis is proposed to solve the problem and improve the performance of Web IR systems. The major contributions of our work are:

1. A learning-based approach is proposed to esitmate Web page quality. This approach judges a page's usefulness by whether it can meet search users' information needs and Naïve Bayes learning is used to combine page features. Experiment on a corpus made up of 37 million Chinese Web pages show that the approach can reduce Web IR systems' indexing amount by 95% as well as retaining over 90% high quality pages and improving its retrieval performance.

2. A classification approach is proposed to indentify user's information need according to his search queries. Click-through behavior analysis is adopted to finish this task. This approach is proved effective in successfully classifying over 80% user queries and gains 20% improvement compared with traditional methods.

3. An automatic search engine performance evaluation method is proposed based on click-through data analysis. This method generates query topics and answers automatically based on search users' querying and clicking behavior. Experimental results based on a Chinese search engine's user logs show that the automatically method gets a similar evaluation result with traditional assessor-based ones.

4. A new Web IR system archititeture is proposed based on our research in global scalse user behavior analysis. A hierarchical index structure and an information need oriented querying model are included in this architecture. It is designed to solve the search request expression problem as well as improve search performance.

Key words: Web IR; Behavior analysis; Quality estimation; Performance evaluation