FROM JASIST, V. 54 (3)

Study and Results: Vast numbers of documents are created and published in electronic form, many of which are copies or alternative versions of one kind or another. Our research concerns the problem of identifying documents that originate from the same source, which we refer to as co-derivative. We compare new and existing techniques for identifying co-derivatives, and show that our novel identity measure is the most effective for finding plagiarized documents.

What's New? We investigated two approaches to determination of co-derivation. The first is a ranking method, based on information retrieval technology; we developed a new similarity measure designed specifically for identifying co-derivative documents. The second approach, fingerprinting, is based on work pioneered by Manber. We investigated many variations of the fingerprinting approach and evaluated their performance against the ranking method. While both methods were able to identify most of the co-derivative documents in our test collections, we found that the ranking method was the best at separating the correct from the incorrect matches.

Limitations: All of the techniques that we tested were used to measure whole-document similarity. We anticipate that further improvements may be made to the effectiveness of these plagiarism-detection methods by comparing short sections of the documents.


Study and Results: Rule-based information filtering systems maintain user profiles where the profile consists of a set of filtering rules expressing the user’s information filtering policy. This study compares the effectiveness of the two alternative rule-based filtering methods: stereotype-based rules vs. personal rules. The results show that the stereotype-based filtering outperforms the personal rule-based filtering.

What's New? Although, intuitively, personal filtering rules seem to be more effective because each user has his own tailored rules, this comparative study reveals that stereotype filtering rules yield more effective results. We believe that this is because users find it difficult to evaluate their filtering preferences accurately, while the stereotype generation process smooths the subjective evaluations of the users. The results imply that by using a stereotype it is possible not only to overcome the problem of user effort required to generate a manual rule-based profile, but also at the same time even to provide a better initial user profile.

Limitations: The experiment was performed with a small group of users, and since it addressed initial user profile generation scenario, it was static in nature. No adaptation was performed.

FROM JASIST, V. 54 (6)

Study and Results: There is a lot of interest in counting links between websites for the purpose of creating indicators or discovering patterns of collaboration or information use. Previous studies have found that simple link counting is problematic within the academic domain because individual websites can duplicate links thousands of times, swamping the total counts and making it difficult to extract meaning from the results. A previous paper by the first author [JASIST, 53(12), 995-1005] created models of Web documents on a larger scale than that of the page by binding pages together into directories, domains or university-wide sites. These models do not account for cases where knowledge of a general information URL is passed among members of a university and repeatedly used. This phenomenon has caused anomalies in previous counts. We have produced new modified counting schemes based upon the new models by disallowing repeated counts of the same target document.

What’s New? A method for counts of links between websites that attempts to exclude multiple links between sites due to information sharing in the source site.

Limitations: No simple document model can on its own eliminate all anomalies in Web-publishing behavior. The data set for the study only covers the UK academic Web.


Study and Results: In this paper we investigate utilizing user’s search behavior as additional evidence to relevance feedback algorithms. Our main hypothesis is that how a user interacts with an information retrieval system can provide useful information on what is of interest to the user. We carry out five interactive experiments to investigate how aspects of search behavior, such as precision, and aspects of relevance assessments, such as the use of partial relevance assessments, can be used to make information retrieval systems more responsive to individual searches.

What’s New? This article furthers research into relevance feedback by not only considering the content of relevant documents but also considering the context in which relevance assessments are made - the user’s search behavior. One of the most popular features of our approach is that user search behavior can be used to explain to users why the system makes individual query modification decisions. This explanatory role increases user satisfaction and also increases the use of relevance feedback.

Limitations: The experiments presented in this paper are preliminary pilot experiments, using only six subjects per experiment.

FROM JASIST, V. 54 (8)

Study and Results: Systems such as Web search engines and site-based search tools provide rapid access to vast quantities of text data. Efficient construction of inverted indexes is essential to provision of such search. In this paper, we review the principal approaches to inversion, analyze their theoretical cost and test them experimentally on collections of up to 20 gigabytes. Of the previous approaches, the most efficient is a sort-based single-pass method of Moffat and Bell (published in JASIS in 1995), which however has the severe drawback of requiring that the full vocabulary of the lexicon be held in memory. This drawback is also a problem for the two-pass approaches, which have the additional limitation that the data must be processed twice. As an alternative, we explore a straightforward one-pass approach and show that it is more efficient than the alternatives.

What’s New? Our one-pass approach involves building partial indexes in memory, dumping them to disk (including their vocabularies) and then merging. Efficiency is achieved through a range of refinements, leading to a fast approach that does not need the complete vocabulary of the indexed collection main memory and can operate within limited resources.

Limitations: Using this approach, indexes can be built for larger collections, faster and in less memory than with previous methods. While larger collections or other kinds of full-text databases, such as path-indexed XML, were not explored, such data is not expected to present difficulties.