Abstract

Recent years have witnessed the drastic development of World Wide Web (WWW). Information is being accessible at the finger tip anytime anywhere through the massive web repository. The performance and reliability of web engines thus face huge problems due to the presence of enormous amount of web data. The voluminous amount of web documents has resulted in problems for search engines leading to the fact that the search results are of less relevance to the user. In addition to this, the presence of duplicate and near-duplicate web documents has created an additional overhead for the search engines critically affecting their performance. The demand for integrating data from heterogeneous sources leads to the problem of near-duplicate web pages. The detection of near duplicate documents within a collection has recently become an area of great interest. In this research, we have presented an efficient approach for the detection of near duplicate web pages in web crawling which uses keywords and the distance measure. Besides that, G. S. Manku et al.'s fingerprint based approach proposed in 2007 was considered as one of the "state-of-the-art" algorithms for finding near-duplicate web pages. Then we have implemented both the approaches and conducted an extensive comparative study between our similarity score based approach and G. S. Manku et
Performance and Comparative Analysis of the Two Contrary Approaches for Detecting Near Duplicate Web Documents in Web Crawling

We have analyzed our results in terms of time complexity, space complexity, Memory usage and the confusion matrix parameters. After taking into account the above mentioned performance factors for the two approaches, the comparison study clearly portrays our approach the better (less complex) of the two based on the factors considered.

References

large-scale study in the orkut social network”, in proceedings of International Conference on Knowledge Discovery and Data Mining, pp. 678 – 684, Chicago, Illinois, USA, 2005.


**Index Terms**

Computer Science  
Information Sciences
Performance and Comparative Analysis of the Two Contrary Approaches for Detecting Near Duplicate Web Documents

Keywords
- Near Duplicate Documents
- Similarity Score Measure
- Confusion Matrix
- Storage
- Space Complexity

Memory Usage Analysis

Computation Time Analysis