Abstract

The voluminous amount of web documents has weakened the performance and reliability of web search engines. The subsistence of near-duplicate data is an issue that accompanies the growing need to incorporate heterogeneous data. Web content mining face huge problems due to the existence of duplicate and near-duplicate web pages. These pages either increase the
index storage space or increase the serving costs thereby irritating the users. Near-duplicate detection has been recognized as an important one in the field of plagiarism detection, spam detection and in focused web crawling scenarios. Here we propose a novel idea for finding near-duplicates of an input web-page, from a huge repository. We proposes a TDW matrix based algorithm with three phases, rendering, filtering and verification, which receives an input web-page and a threshold in its first phase, prefix filtering and positional filtering to reduce the size of records in the second phase and returns an optimal set of near-duplicate web pages in the verification phase after calculating its similarity. The experimental results show that our algorithm outperforms in terms of two benchmark measures, precision and recall, and a reduction in the size of competing record set.

Reference

- Dennis Fetterly, Mark Manasse and Marc Najork, Detecting phrase-level duplication on the world wide web, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp.170 - 177, Salvador, Brazil, 2005.
- D. Lowd and C. Meek, Good word attacks on statistical spam filters, Second Conference on Email and Anti-Spam, July 2005.
Index Terms

Computer Science  Information Retrieval

Key words

Near-Duplicate Detection  Term-Document-Weight Matrix
Prefix filtering  Positional
Singular Value Decomposition