Near-duplicates detection for Vietnamese documents in large database

Cong T.T., The D.B., Bao S.P.
Vietnam National University, Hanoi

Abstract: Near-duplicate documents exacerbate the problem of information overload. Research in detecting near-duplicates has attracted a lot of attention from both industry and academia. In this paper, we focus on addressing this problem for Vietnamese documents which, to the best of our knowledge, has not been done before. Most of the current algorithms have been designed for English which are not directly applicable to Vietnamese - a monosyllabic language. We propose to combine Charikar's algorithm [2] with a "weighting scheme" and Vietnamese specific features to address the language intricacy. Experimental results indicate that our scheme is effective for detecting near-duplicates in a corpus of Vietnamese documents. © 2008 IEEE.

Author Keywords: Charikar; Hash scheme; LSH; Near-duplicate Vietnamese detection; Weighting scheme

Index Keywords: Information technology; Technology; Charikar; Hash scheme; Information overloading; International conferences; Language processing; Large databases; LSH; Near-duplicate Vietnamese detection; Web information; Weighting scheme; Weighting schemes; Linguistics

Year: 2008
Source title: Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology
Art. No.: 4584344
Page : 70-75
Link: Scopus Link
Correspondence Address: Cong, T. T.; Vietnam National University, Hanoi; email: Thanhtruongcong@gmail.com
Conference name: ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology
Conference date: 23 July 2008 through 25 July 2008
Conference location: Liaoning
Conference code: 73559
DOI: 10.1109/ALPIT.2008.76
Language of Original Document: English
Abbreviated Source Title: Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology
Document Type: Conference Paper
Source: Scopus
Authors with affiliations:
  • Cong, T.T., Vietnam National University, Hanoi
References:

- C. Lyon, R Barrett, J Malcolm. Plagiarism is easy, but also easy to detect. Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and FalsificationBroder, A., On the resemblance and containment of documents SEQS: Sequences, 91