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: Scorpus 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
ISBN: 9.78E+12
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

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

References:

- Charikar, Similarity Estimation Techniques from Rounding Algorithms (2002) Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, , ACM Press
- Andoni, A., Indyk, P., Near-Optimal Hashing Algorithms for Near Neighbor Problem in High Dimensions (2006) Proceedings of the Symposium on Foundations of Computer Science (FOCS'06)
- S. Brin, J. Davis, H. Garcia-Molina. Copy detection mechanisms for digital documents. In Proceedings of the ACM SIGMOD Annual Conference. San Francisco, CA, May 1995Shivakumar, N., Garcia-Molina, H., SCAM: A copy detection mechanism for digital International Conference in Theory and Practice of documents (1995) Proceedings of 2nd Digital Libraries, , Austin, Texas, June
- Lyon, C., Barrett, R., Malcolm, J., A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector. Plagiarism: Prevention (2004) Practice and Policies Conference, , June
- 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
- Kolcz, A., A. Chowdhury, et al. (2004). Improved stability of I-Match signatures via lexicon randomization, AOL. 1998Manku, J., Sarma: Detecting Near-Duplicates for Web Crawling (2007) Proceedings of the 16th international conference on World Wide Web, , ACM Press
- Henzinger, Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms (2006) Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, , ACM Press
- Detecting Near-Duplicates in Large-Scale Short Text Databases by Cai chun Gong, Yulan Huang, Xueqi Cheng, Shuo BaiU. Manber. Finding similar files in a large file system. In Proc. 1994 USENIX Conference, pages 110, Jan.1994Broder, A., On the resemblance and containment of documents (1998) SEQS: Sequences, 91
- Broder, A., Glassman, S., Syntactic clustering of the Web (1997) Proceedings of the 6th International Web Conference
- Heintze, N., Scalable document fingerprinting (1996) Proc. USENIX Work-shop on Electronic Commerce
- Yao, A.C., Yao, F.F., Dictionary look-up with one error (1997) J of Algorithms, 25 (1), p. 194202
- A. Broder. Some applications of Rabins fingerprinting method. In Renato Capocelli, Alfredo De Santis, and Ugo Vaccaro, editors, Sequences II: Methods in Communications, Security, and Computer Science, 993:143152Dinh Dien, Hoang Kiem. POS-Tagger for English - Vietnamese Bilingual Corpus. Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, CA. 2003Thi Minh Huyen Nguyen, Hong Phuong Le, Xuan Luong Vu. A case study of probability tagger QTAG for tagging Vietnamese text. Proceeding of ICT.rda'03, Hanoi, Vietnam. 2003Phuong, L.H., Vinh, H.T., A Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts (2007) IEEE International Conference on Research, Innovation and Vision for the Future RIVF, , Vietnam
- Thi, M., Nguyen, H., Hong, M.P.L., Thang Dinh, Q., Luong, X., Vu and Cam Tu Nguyen. Word segmentation of Vietnamese texts: A comparison of approaches (2008) Proceedings of the 6th Language Resources and Evaluation Conference LREC, , Marrakech Morocco