

POSTER

MatchDetectReveal: Finding Overlapping and Similar Digital Documents

Krisztián Monostori, Arkady Zaslavsky, Heinz Schmidt

School of Computer Science and Software Engineering
Monash University,
Melbourne, Australia

Email: {krisztian.monostori, arkady.zaslavsky, heinz.schmidt}@infotech.monash.edu.au

Abstract

The Internet provides easy access to large collections of semi-structured digital documents. Student assignment papers as well as conference papers can easily be prepared by "cut & paste" techniques since it is easy to find relevant documents with the help of WWW browsers and search engines. The proposed MatchDetectReveal (MDR) system is capable of finding overlap between documents on the Internet. The proposed architecture of the MatchDetectReveal system is presented in this poster.

The core component of the system has already been developed and it uses string-matching algorithms based on suffix trees to compare suspicious documents to candidate documents. Ukkonen's algorithm was modified to build a more space-efficient representation of a document and Chang's matching statistics algorithm was also tailored to utilize this new representation.

With the large number of digital documents available in digital libraries finding candidate documents is a hard task. Document indexing techniques will be analysed to efficiently index documents and retrieve a limited number of candidate documents, which can be handled by the system. The proposed search-engine component of the system is responsible for finding and retrieving candidate documents.

Powerful processors, high-speed networks and standard tools for distributed computing make clusters of PCs or workstations an appealing platform for parallel computing. We have already experimented parallel approaches with the Clustor tool used by the Monash Parallel Parametric Modelling Engine (PPME) at the School of Computer Science and Software Engineering, Monash University. Performance results are presented in the poster as well as other possible approaches e.g. using the MPI library.

The Document Generator is a supplementary component of the MDR system and it is capable of generating arbitrary documents out of a given set of documents (base document set). The main purpose of this component is to generate sufficient number of documents to test our copy-detection algorithm. Varying the parameters you can generate documents of different size, different amount of plagiarism, you can use different number of files and different sizes of chunks.

Other components of the system are still under development including Visualizer, which will present the user with the result, Similarity & Rule Interpreter, which will manage inexact matches ie. changing the names of localities or substituting synonyms, and the search-engine. Implementation issues of these components will also be presented in the poster.

Keywords

Document Management Systems, Algorithms, Information Retrieval