Overview

- Copy-detection
- MatchDetectReveal (MDR)
  - Matching Engine
  - Performance Analysis
  - Visualiser
- Document Generator
- Conclusion
Copy-detection

- Copy-prevention
  - Physical isolation
  - Hardware for authorisation
  - Active documents
- Copy-detection
- Existing systems: SCAM, Glatt, Plagiarism.org

MatchDetectReveal (MDR)

- Internet
- MDR users
- Base Document Set
- Generator
- MDR customizer

Global resources

- Matching engine
- Format converter
- Search engine
- Visualiser

Local cluster

- Similarity & overlap rule interpreter

IEEE DL

ACM DL

Base Document Set

Matching engine

Format converter

Search engine

Visualiser

- Local repository
- Matching rule DB
- Indexes
Converter

- Converting from popular formats (PDF, PS, Word, RTF, HTML) into pure text
- Pure text is converted further
  - converting to lowercase
  - eliminating multiple whitespaces
  - prepare it to be processed by matching engine
- Chunks identical before the conversion remain identical after the conversion

Converter Example

Mr. X. plagiarised a lot of documents according to (Garcia Molina et al., 1996b)
**Matching Engine**

- Suffix tree
- Modified suffix tree
- Matching statistics

**Algorithms**

- **Suffix Tree for Candidate Documents (STCD)**
  - Build a suffix tree for each candidate document

- **Suffix Tree for Original Document (STOD)**
  - Build only one suffix tree
  - Applying the matching statistics algorithm in a reverse fashion
Performance Analysis (STOD)

Configuration: Intel Pentium II 433MHz, 128M RAM, Windows NT w/s
Total size: 9.84MB

Parallel and Distributed Algorithms

- Monash Parallel Parametric Modelling Engine
  - Clustor
  - Data-intensive jobs
  - MPI

- Distributed Approaches
  - Mobile agents
  - Globus project
    - Resources close to documents
The Clustor Tool

SCAM - hashing with different chunk sizes:
- word
  - detects only whole document match
- sentence
  - sentence boundaries are easily shifted
- hashed breakpoint chunking
  - different chunk sizes

“Shingling” approach
min_file_size - defines the minimum size of the document to be generated in kilobytes
max_file_size - defines the maximum size of the document to be generated in kilobytes
min_overlap - defines the minimum overlap content in percentage of the total size of the document
max_overlap - defines the maximum overlap content in percentage of the total size of the document
num_files - defines the number of files to be used for generation. It must be less than or equal to the number of base documents
min_char - the minimum number of characters in a single chunk
max_char - the maximum number of characters in a single chunk
num_words – the number of words to be substituted in each chunk
Conclusion and Future Work

- MatchDetectReveal
  - Matching Engine
  - Visualiser
  - Search-Engine
- Document Generator
- More Efficient Parallel Algorithm
- Performance Analysis of Distributed Approaches
- Submission System

Objectives

- Repository of assignments
- Indexing
- On-line submission
- Plagiarism-detection
  - comparing documents to each other
  - comparing documents to the Internet
Repository of Assignments

- Assignments of Current Semester
  - indexing
  - browsing
  - plagiarism-detection

- Assignments of Previous Semesters
  - indexing
  - browsing
  - current documents are also compared to these documents

On-Line Submission

- Students Log On to the System and Submit Documents
- Students Are Able to See Information on Subjects They are Enrolled to
- Lecturers Are Able to See Information on Subjects They Teach
- Lecturers Are Able To Initiate Plagiarism-Detection on Submitted Documents
- Web-based
Existing systems - The Platypus Project

- Previously WebFace
- Developed at GSCIT
- Implemented in Perl
- Not Finished Yet - Founding Problems
- About to Go Open Source

Existing Systems - CSSE System

- E-mail-Based
- Scripts Process and Forward Mails
- Sending Individual Files
- Sending Contents of a Given Directory
Existing Systems - Problems

- No Web Interface
- Limited Repository
- No Indexing
- No Plagiarism Detection

Plagiarism-Detection

- Comparing Local Documents
- Comparing Internet Documents
- Suffix Trees
- Local Cluster
- Internet Resources - Globus
Scenario

Browse Submissions
Initiate Detection
Report

Submission Plagiarism Detection MDR Repository Browse

Document Overlap Detection System for Distributed Digital Libraries

K. Monostori, A. Zaslavsky, H. Schmidt
School of Computer Science & Software Engineering
Monash University
A.Zaslavsky@monash.edu.au

DL'2000, San Antonio, 2-7 June, 2000