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

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)

- MDR users
- Global resources
- Internet
- Local cluster
- Similarity & overlap rule interpreter
- Matching engine
- Format converter
- Search engine
- Visualiser
- Generator
- Base Document Set
- MDR customizer

IEEE DL
ACM DL
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
Parallel Algorithm

Running Time with Different Number of Nodes

Search-engine

- **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
The files in scenario 1 and 2 are 11K and 14K respectively, while the file in scenario 3 is a 1.67M document. We compare these three documents to 19 documents, which are 3.5M altogether and of course contain those 3 documents, which are used in Scenario 2 and the one used in Scenario 3. We used the algorithm, which builds only one suffix tree, for analysing the documents. This algorithm correctly reported the files in Scenario 2 and 3 as plagiarised while the document in Scenario 1 as genuine.

Optimal substructure: This is sometimes called the principle of optimality. It states that for the global problem to be solved optimally, each subproblem should be solved optimally.

Polynomially many subproblems: An important aspect to
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

Submission

Plagiarism Detection

MDR

Repository

Browse

Report

Submit