Digital Documents in Educational Environment: Misuse, Appropriation, and Detection Issues

Krisztián Monostori, Arkady Zaslavsky, Heinz Schmidt
(Monash University)


Contact: A/Prof Arkady Zaslavsky; email: A.Zaslavsky@Monash.edu.au; phone: (IDD + 61 3/Aus 03) 9903 2479
School of Computer Science & Software Engineering, Monash University, Melbourne, Australia

MDR: MatchDetectReveal

- Identifying Overlapping Text
- Plagiarism Detection
- Match-Engine: String-Matching Algorithms and Suffix Trees
- Parallel Comparison of Documents
Overlap Detection

- Copy Detection or Copy Prevention
- Copy Prevention
  - Physical Isolation
  - Hardware for Authorisation
  - Active Documents
- Plagiarism Detection
  - Glatt, Plagiarism.org, SCAM, Eve

Partial Copy Detection

- Index on Chunks
- Chunking Primitives
  - word
  - sentence
  - k-word or hashed breakpoint
  - shingles

The shingling approach uses Rabin’s fingerprinting method.
**MDR Architecture**

- **Nimrod**
- **Local cluster**
- **Internet**
- **Global resources**
- **MDR users**
- **mdr architecture**
- **matching engine**
- **format converter**
- **search engine**
- **visualizer**
- **local repository**
- **matching rule DB**
- **indexes**
- **Similarity & overlap rule interpreter**

**Comparison**

- **Submission Index Report**
- **Plagiarism Detection**
- **MDR**
- **Index**
- **Submission**

**IEEE DL**

**ACM DL**
Distributed Repository

1. Comparing Local Documents
2. Transferring Documents from Other Nodes

Submission System

Assignments Submitted

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Submitted</th>
<th>Plagiarism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Database Design</td>
<td>4/4/2000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Database Implementation</td>
<td>3/11/2000</td>
<td></td>
</tr>
</tbody>
</table>

Assignments to Submit

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Due Date</th>
<th>Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Distributed Databases</td>
<td>6/1/2000</td>
<td>Submit</td>
</tr>
</tbody>
</table>
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 & database archiving
- Case studies