MatchDetectReveal: Finding Overlapping and Similar Digital Documents

Krisztián Monostori, Arkady Zaslavsky, Heinz Schmidt
School of Computer Science & Software Engineering
Monash University
A.Zaslavsky@monash.edu.au
MDR: MatchDetectReveal

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

- Global resource
  - Matching engine
  - Format converter
  - Search engine
  - Visualizer

- Local repository
- Matching rule DB
- Indexes

- MDR users
- Internet
- Generator
- Document Set
- Base
- Local cluster
- NimrodM
- MDR customizer

IEEE DL
ACM DL
Match-Engine

- Suffix Tree
- Ukkonen’s Linear Time Algorithm
- Modified Suffix Tree
- Matching Statistics Algorithm by Chang
- Suffix Tree for Candidate Documents - STCD
- Suffix Tree for the Original Document - STOD
Matching Statistics with the Modified Suffix Tree

+First+they+went+home+and+then+they+went+out+$

+They+went+home+before+midnight+$
Running Time of the Sequential Algorithm

- **Configuration:** Intel Pentium II 433MHz, 128M RAM, Windows NT Workstation
- **Total size:** 9.84MB
- **Tree size:** app. 23 times of the file size
Speed-up vs Number of Processors

Running Time with Different Number of Nodes

Total size of files to be compared: app. 70MB in 35 batches
Document Generator

random text

Chunk 1
Size: n1

Chunk 2
Size: n2

... ...

Chunk n
Size: nn

Pos1

Pos2 + n1

Posn + Σ nx

Chunk 1 Chunk 2 . . . . . Chunk n