Parallel and Distributed Overlap Detection on the Web

Krisztián Monostori, Arkady Zaslavsky, Heinz Schmidt
Monash University, Melbourne

Overview

- Overlap Detection
- Underlying Algorithm
- Using a Local Cluster
- The Globus Project
- Using the GRID
- Conclusion and Future Work
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.
Suffix Trees

- All Suffixes of a Text
- Searching
- Matching Statistics

Performance of the Sequential Algorithm

- STCD - Suffix Tree for Candidate Documents
- STOD - Suffix Tree for Original Document
Comparison

Local Cluster

- Monash Parallel Parametric Modelling Engine
- Clustor - Batching Jobs
Distributed Repository

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

The Globus Project

www.globus.org
Globus Architecture

Applications

High-level Services and Tools
GlobusView
DUROC
Gloperf
Nexus
MPI
MPI-IO
Globus Directory Service
Nimrod/G
Globus Security Interface
GASS
Testbed Status
CC++
globusrun

Core Services
Heartbeat Monitor
GRAM
Global Resource Allocation Manager
Monitoring

Local Services
Condor
LSF
Easy
NQE
TCP
UDP
AIX
Irix
Solaris

Globus Components - GRAM

• Global Resource Allocation Manager
• Authentication
• Processing Resource Requests
• Allocation of Required Resources
• DUROC - Dynamically Updated Request Online Co-allocator
**Globus Components - GIS**

- Grid Information Services
- Publishes Information via LDAP
- “White Pages” Directory
- “Yellow Pages” Directory

**Globus Components - GASS**

- Globus Access to Secondary Storage
- Copies Files between Globus Nodes

- MPICH-G - the Globus Version of MPI
Nimrod/G

- Resource Management
- Scheduling System
- Parametric Modelling
- GUSTO Testbed
- Cost Model
- Deadline

Utilising Globus

- Distributed Index
- Distributed Comparison
- Resources in the Close Proximity of Candidate Documents
Conclusion and Future Work

- Overlap Detection Algorithms
- Local Clusters
- Globus

- Performance Analysis
- Agent Technologies