MinerSoft: A Software Search Engine for the Grid

Asterios Katsifodimos, George Pallis, Marios D. Dikaiakos - University of Cyprus
{asteriosk, gpallis, mdd}@cs.ucy.ac.cy

Abstract

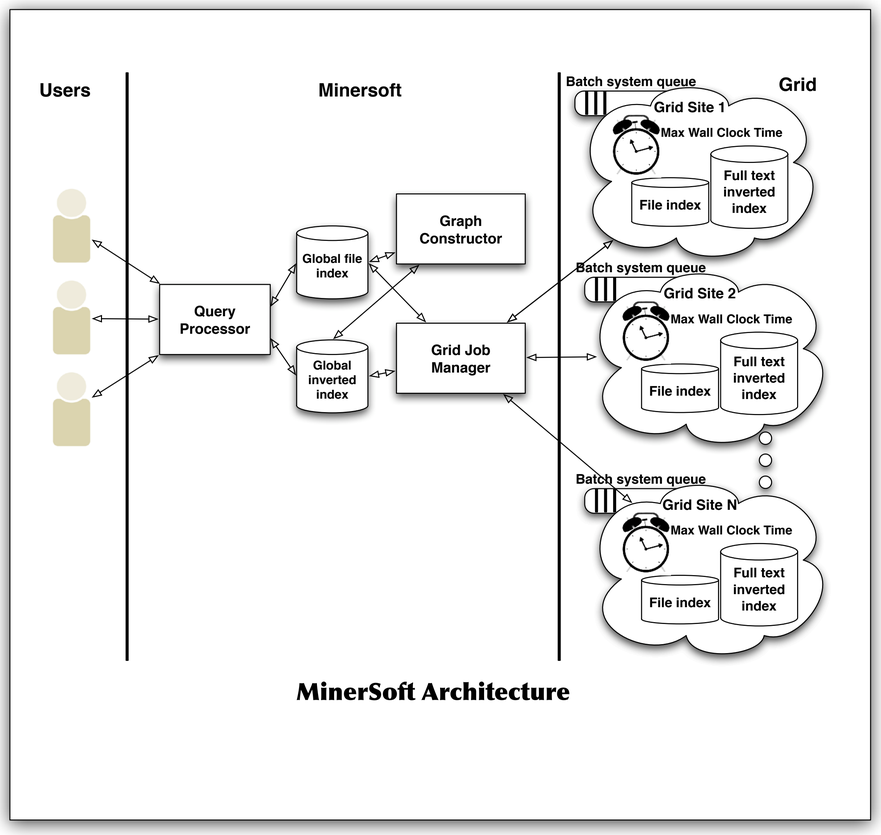
We present the design, architecture and implementation of an open-source keyword-based paradigm for the search of software resources in Grid infrastructures, called **Minersoft**. A key goal of Minersoft is to annotate automatically all the software resources with keyword-rich metadata. Experiments were conducted in EGEE. Results showed that Minersoft successfully crawled 12.3 million valid files (620 GB size) and sustained, in most sites, high crawling rates.

Workflow

Step 1: Crawling & Files classification
Every site is visited(through Grid Jobs) and all filesystem's files are examined for their file type. (binaries, libraries, man pages, scripts etc).

Step 2: Association Mining: The Software Graph
Binaries require libraries to be dynamically linked during runtime, README files document or describe other files, man pages describe executable files. Those relationships make the Software Graph.

Step 3: Full text Indexing
In order to search all Software files, we need to have a full text index of those we are interested in. A full text index is being created for each site and the results are uploaded to each site's storage element. The full text index will be used to answer users queries.



MinerSoft Architecture

Step 4: Content enrichment
Some certain type of software files do not contain many words that can be used during queries. In order to enrich those files' content so that they will be content-rich, we user the software graph and the files' relationships.

Step 5: Results merging
After all the above processes are finished, indexes and Software Graphs from all the Sites are merged to a global index that can be searched through keywords.

Step 6: Answer user queries
After all of the indexes and Software Graphs are created, MinerSoft is ready to give responces to users' queries.

Experiments

Indexing Performance:
MinerSoft's indexers thoughput ranges between 14 and 92 files per second.

Grid Sites	File Rate (files/sec) Files/Run time	Size Rate (MB/sec) MB/Run time
HG-03-AUTH	38,65	1,75
RO-08-UVT	24,03	0,82
MK-01-UKIM-II	14,94	0,31
AEGIS01-PHY-SCL	59,04	1,98
HG-05-FORTH	10,75	0,40
BG01-IPP	92,25	1,93
CY-03-INTERCOLLEGE	37,09	1,82
HG-02-IASA	65,07	2,92
CY-01-KIMON	66,50	3,36

Indexing Performance

Crawling Performance:
MinerSoft's crawlers thoughput ranges between 27 and 339 files per second.

Grid Sites	File Rate (files/sec) Files/Run time	Size Rate (MB/sec) MB/Run time
HG-03-AUTH	117,82	6,80
RO-08-UVT	231,82	19,99
MK-01-UKIM-II	275,24	5,52
AEGIS01-PHY-SCL	339,66	11,62
HG-05-FORTH	172,65	9,43
BG01-IPP	79,15	2,06
CY-03-INTERCOLLEGE	225,43	6,50
HG-02-IASA	27,74	1,60
CY-01-KIMON	34,93	2,02

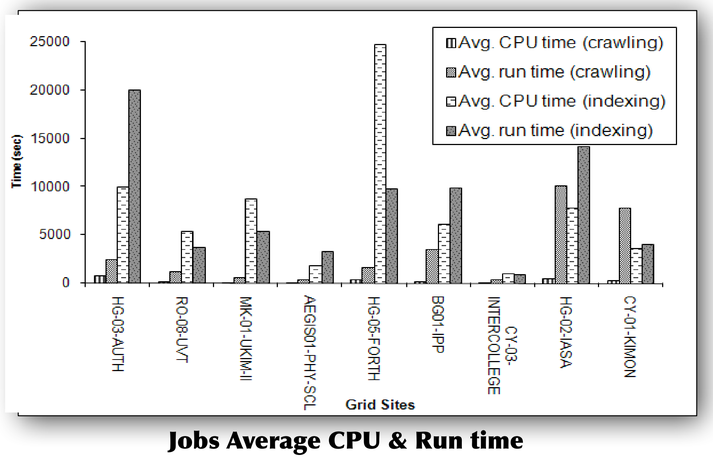
Crawling Performance

Categories of software files per site:
Different categories of files exist in Grid Sites like: binaries, libraries, scripts, souces etc.

Grid Site	Binaries	Sources	Libraries	Docs
HG-03-AUTH	32276	1354421	96698	2312900
RO-08-UVT	8134	66093	4199	136400
MK-01-UKIM-II	15083	71245	8010	135431
AEGIS01-PHY-SCL	6064	41257	7669	122615
HG-05-FORTH	26175	510430	49067	975504
BG01-IPP	28684	1013642	83332	1912159
CY-03-INTERCOLLEGE	26971	13636	3644	40090
HG-02-IASA	50096	1820096	121979	3038387
CY-01-KIMON	28690	232131	22571	478029
Total	222173	5122951	374598	9151515

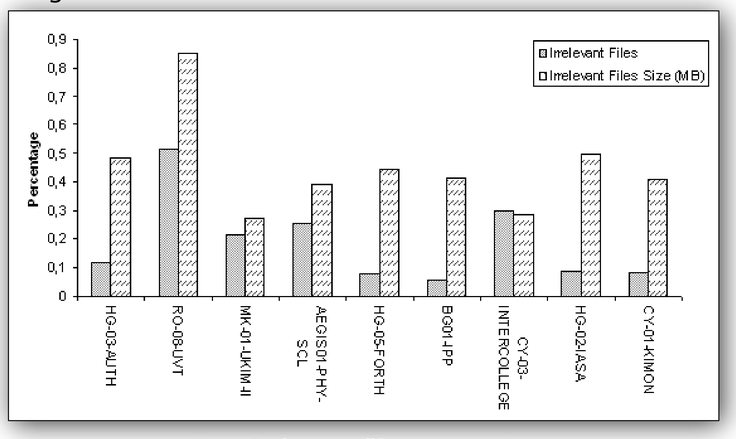
Files per software category

Grid Jobs run time:
Each site had to be crawled and indexed. Below, you can see the average run time/CPU time used by our crawlers and indexers.



Jobs Average CPU & Run time

Irrelevant files percentage:
The fact that more than 50% of the Sites' files, are Software files, shows the need for a Software Search Engine for Grid Infrastructures.



Non Software files percentage

Conclusions

-- **MinerSoft successfully crawled 12.3 million files(620 GB)** sustaining high crawling and indexing performance.

-- **There is a large percentage of duplication of files among Grid Sites**(~33%).

-- It is very important to establish advanced software discovery services for the Grid since **more than 50% of sites' filesystems are Software Files.**

eGee Enabling Grids for E-sciencE
SEARCHiN

High Performance Computing Systems Lab
Department of Computer Science – University of Cyprus