

ADVANCED SEARCH ENGINE USING GRID

Tool : Globus Toolkit

Platform : Linux

With the growing number of web search engines coming up, it has become difficult to choose the best amongst them. People have their own favorites, but even they may not come up with user relevant results for selected search strings. It becomes more confusing when each search engine ranks the links (search results) differently for the same search string.

To avoid this problem we have tried to integrate the search results from different search engines chosen by the user and display the results based on our own Ranking Algorithm. This algorithm will rank the links based on a number of parameters that will help the user to arrive at accurate results. The advantage of this method is - it will eliminate the dependency on a single search engine and will pool data from different search engines and rank them. For this we make use of the Grid.

Grid computing is an emerging computing model that distributes processing across a parallel infrastructure. Throughput is increased by networking heterogeneous resources (personnel computers, servers, Workstations, PDA's, Laptops, clusters) across administrative boundaries to model virtual computer architecture. What is meant by heterogeneous resources is that we can have computers with different CPU speeds, different RAM's and different operating systems forming the Grid. The use of Grid for this purpose will take care of large amount of computation required while searching the web on multiple search engines for multiple clients and also to rank them with the user preferences.

For its implementation we used **Globus Toolkit** on Linux platform (Fedora Core 4/6). The working of this search collator proceeds by taking search string from the user at a client machine. This search string will be passed on to the Scheduler program. The Scheduler program will centrally administer the nodes in the grid. The Scheduler

program will send this search string to each node (computers) in the Grid by first checking if the node is up. Each node will search on different search engine for the same search string and the results will be collected at the central database. By applying our ranking algorithm on these results we rank the results with at most accuracy as compared to using a single search engine. The ranking of the search results will also be done in a distributed manner.

Parameters for Ranking:

1. Individual Ranking of search result in the respective search engines
2. Presence of typed search string in URL
3. Presence of typed search string in excerpt and if present its frequency
4. Depth of URL
5. Number of search results returned by each search engine.

The user has the options to change any of the ranking preferences according to own needs. For example, a user can set priorities (higher/lower) to any of the above parameters to reach at results that are most relevant to him/her. The user can give more preference to any of the parameters and arrive at results accordingly.

The user will no longer have to rely on a single search engine but can take advantage of results obtained from multiple search engines which tend to be more reliable and are based on the user requirements. Our search engine will provide users with one more layer of abstraction and help them arrive at accurate and user relevant results at the top in the order.

The advantage of grid can further be described with the ease with which addition and deletion of nodes is possible. Clusters on the other hand use homogenous resources (same kind of hardware and operating system on each node) and addition and deletion of nodes require reconfiguration of the entire network. This is one of the primary reasons for choosing Grid over cluster.

The use of Grid is not confined only to a dedicated LAN but to have a high powered Internet. It allows users to share their computing resources on the internet and use resources of other end users for high computation. The basic requirement of Grid is to use the idle resources to its fullest. Companies around the world have started moving from clusters to Grids because of its numerous advantages.

We acknowledge with thanks the expert academic advice and assistance rendered by our internal guide Prof. Ms. V Y Kulkarni who was a constant source of encouragement to us during the course of this project. We would like to thank our external guide Mr. Gaur Sunder, Head of Medical Informatics, CDAC Pune, for his valuable guidance without which this project would not have been possible. We would like extend our vote of thanks to our Vice Principal and Head of Computer Engineering Department, MIT Pune, Prof. Mr. A K Pathak for providing us with all the facilities and resources required for completion of our project. Finally, we would like to extend our gratitude to all the teachers who helped us during the course of this project.

PROJECT TEAM: -

Abijit Bej (abej@usc.edu, getabijit@yahoo.co.uk)
Ashwinkumar Ganesan (gashwin_k86@yahoo.co.in)
Kedar Kadlaskar (kedarkadlaskar@gmail.com)
Pritesh Baviskar (priteshh22@yahoo.com)