ISSN No. 0976-5697

Volume 3, No. 6, Nov. 2012 (Special Issue)

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

Use of Grid Computing in Search Engine – A Survey

Dr. Laxmi Ahuja
Amity Institute of Information Technology,
Amity University, Uttar Pradesh, India
laxmiahuja1908@gmail.com

Abstract: The increasing requirement for computing power needs increased network bandwidth, more powerful computers and storage systems, and sophisticated software applications. But these new abilities demands dealing with the challenge of growing workload. Organizations face many challenges as they strive to remain in the competition. Grid computing is applying the resources of many computers in a network to a single problem at the same time-usually to a scientific or technical problem that requires a great number of computer processing cycles or access to large amounts of data. In web engineering also, grid computing can be applied to get more relevant information in an efficient and effective way. Present paper explores the existing literature in which grid computing is used in various search engine based develoment.

Keywords*:* Web Engineering, Search Engine Optimization, Grid computing.

I. INTRODUCTION

Reduced computing costs, greater throughput, faster time-to-market, and improved quality and innovation are all important. Investments in hardware need to be carefully justified, and organizations must find ways to accomplish more with available resources. Flexibility is the key, as enterprises need to handle dynamically changing workloads and quickly provide computing power where it is needed most. Even though the demand for computing resources is great, many existing systems are underutilized. While a few individual serves may be working at capacity, the vast majority of systems are not. As a result, many computing cycles are left unused.

II. GRID COMPUTING

A Grid is a collection of computing resources connected through network that perform tasks. It appears to users as a large system, providing a single point of access to powerful distributed resources. Grid middleware support a common set aggregates these resources and provides transparent, remote, and secure access to computing power wherever and whenever it is needed i.e., grid computing aggregates resources and delivers computing power to every user in the network. Users treat the Grid as a single computational resource. Users can submit thousands of jobs at a time without being concerned about where they run. Grid computing is currently used in technical computing environments, to provide more resources for compute-intensive tasks.

Grid computing enables organizations to use their distributed computing resources more efficiently and flexibly, providing more power out of existing systems and helping organizations gain a competitive business advantages. Grids enable the sharing selection, and aggregation of a wide variety of resources including supercomputer, storage systems, data sources, and specialized devices that are geographically distributed and owned by different organizations for solving large-scale computational and data intensive problems in sciences, engineering, and commerce. Thus enterprises or organizations come together to share resources and skills in order to better respond to business opportunities or large-scale application processing requirements, and whose cooperation is supported by computer networks. Grid computing is concerned with coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations. The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem-solving and resource brokering strategic emerging in industry, science, and engineering, associated configurations security groups and rules

III. GRID AND WEB

The grid can be assumed as next-generation internet. Though, it is not an alternative to the internet however it is rather a set of additional protocols and services that build on internet protocols and services to support the creation and use of computation or it is a layer of software and services that sits on top of operating systems and links different systems together, allowing them to share resources - and data-enriched environments. Any resource that is on the grid is also, by definition, on the Net. Search engines act as important services, providing the community with the information hidden in the Web and, due to their frequent use, stand as an integral part of our lives. The last decade has witnessed design and implementation of several state-of-the-art search engines. A traditional search engine is typically composed of three pipelined components: a crawler, an indexer, and a query processor [5].

Crawler is one of the main components in the search engines which use URLs to fetch web pages to build a repository of web pages starting with entering URL. Each web page is parsed to extract the URLs included in it and store the extracted URLs in the URLs Queue to fetch by the crawlers in sequential. The process of crawling takes long time to collect more web pages, and it has become necessary to utilize the unused computing resources and cost/time savings in organizations. The crawler component is responsible for locating, fetching, and storing the content residing within the Web. The downloaded content is concurrently parsed by an indexer and transformed into an inverted index, which represents the downloaded collection in a compact and efficiently query form.

MINERVA [6] is a peer-to-peer Web search engine, in which each peer independently executes a Web crawler. This system does not have any central coordinator, and hence there is no control over the coverage of each peer. As a result, the same pages may be crawled multiple times by different peers, which results in an overlap of pages. This overlap is a crucial problem in peer-to-peer Web search. MINERVA offers techniques that aim to solve this overlap problem and tries to aggregate the results of independent crawls to generate a global result.

M.E.ElAraby and others [9] used the crawler of search engine using grid computing. They p[resented the grid computing that has been implemented by Alchemi. Alchemi is an open source project developed at the University of Melbourne, provides middleware for creating an enterprise grid computing environment. The crawling processes are passed to Alchemi manager which distribute the processes over a number of computers as executors. The search engine crawler with the grid computing is implemented, tested and the results are analyzed. It was found that there is an increase in performance and less time over the single computer. Search Engine for South-East Europe (SE4SEE) is a socio-cultural search engine running on the grid infrastructure. It offers a personalized, on-demand, country-specific, category-based Web search facility. The main goal of SE4SEE is to attack the page freshness problem by performing the search on the original pages residing on the Web, rather than on the previously fetched copies as done in the traditional search engines. SE4SEE also aims to obtain high download rates in Web crawling by making use of the geographically distributed nature of the grid. Cambazoglu et. al [7] presented the architectural design issues and implementation details of this search engine. We conduct various experiments to illustrate performance results obtained on a grid infrastructure and justify the use of the search strategy employed in SE4SEE.

Haya et. al [8] proposed a tool named Grid Search And Categorization Engine (GRACE) that allows users to search through heterogeneous resources stored in geographically distributed digital collections. The GRACE toolkit will also provide a categorization engine which will dynamically integrate and categorize results from the various data sources. The categorization engine will be based on Automatic Idiomatic Representation (AIR) technology which is designed by Virtual Self, one of the GRACE partners. Results can be automatically categorized regardless of how they are formatted or whether they contain metadata. Moreover, the AIR technology is language independent, so with the aid of language lexicons it will work on sources in various languages. To begin with, GRACE will be capable of automatically identifying and then categorizing results in the following languages: English, German, Swedish and Italian. Additional languages may be added at a later stage.

The Web is not (yet) a grid: its open, general-purpose protocols support access to distributed resources but not the coordinated use of those resources to deliver interesting qualities of service. Grid computing has been proclaimed as the successor to the Web. Current internet technologies address communication and information exchange among computers but do not provide integrated approaches to the coordinated use of resource at multiple sites for computation. By adding the ability to extensively share computing power, applications and storage to the Web's current ability to share text and multimedia files, problems that require a lot of computing resources can be resolved, devices can work past their own limits and collaboration can become more intense. Let's investigate each of these more closely. The Web has a good test bed for grid computing, both through its successes and its Shortcomings. For grid computing to prosper, it will need to solve problems of standards, property rights, access and authorization and modularization and dispatching.

IV. CONCLUSION

The implementation of the crawler in search engine requires powerful computers to achieve high performance. With the ubiquity of the Internet and Web, search engines have been sprouting like mushrooms after a rainfall. However, innovative search engines and guided search capabilities have started appearing only in recent years. Internet computing and Grid-technologies combined promise to change the way we tackle complex problems. They will enable large-scale aggregation and sharing of computational resources, data and other resources across institutional and geographical boundaries.

V. REFERENCES

[1]. Foster, C. Kesselman, editors. The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, San Francisco, Calif. (1999).

[2]. I, Kesselman, C. and Tuecke, S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications

[3]. Ian Foster. The Grid: A New Infrastructure for 21st Century Science, Physics today

[4]. Rajkumar Buyya, Mark Baker. Grids and Grid technologies for wide-area distributed computing, Software Practices and Experiences (Wiley).

[5]. Arasu, A., Cho, J., Garcia-Molina, H., & Raghavan, S. (2001). Searching the Web. ACM Transactions on Internet Technologies, 1(1), 2–43.

[6]. Bender, M., Michel, S., Triantafillou, P., Weikum, G., & Zimmer, C. (2005). Improving collection selection with overlap awareness in P2P search engines. In Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (pp. 67–74). Salvador, Brazil.

[7]. B. Barla Cambazoglu, Evren Karaca, Tayfun Kucukyilmaz, Ata Turk, Cevdet Aykanat, Architecture of a grid-enabled Web search engine, Information Processing and Management (Elsevier), Vol. 43, pp: 609–623, 2007.

[8]. Glenn Haya, Frank Scholze and Jens Vigen "Developing a Grid-Based Search and Categorization Tool", High Energy Physics Libraries Webzine, issue 8, October 2003.

[9]. M.E.ElAraby, M.M.Sakre, M.Z.Rashad, O.Nomir, Crawler Architecture using Grid Computing, International Journal of Computer Science & Information Technology (IJCSIT), Vol. 4, No 3, June 2012