International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

Design of compression and indexing techniques for documents of domain specific cluster

[1]Avi Kishore Soni [2] Ajay Kumar Yadav [3]Varun Kaushik

Mewar University, Chittorgarh(Raj.)

ABSTRACT

The World Wide Web is a huge source of hyperlinked information. Web is growing every moment in context of web documents. Search engines use web crawlers to collect web documents from web for storage and indexing. In the existing system, it has been observed that the repository resources are limited. So it has become an enormous challenge to manage the local repository (storage module of search engine) in a way to handle the web documents efficiently that leads to less access time of web documents and proper utilization of available resources.

This thesis proposes architecture of search engine with the clustered repository, organized in a better manner to make task easy for user to retrieve the web pages in reasonable amount of time. The thesis work focuses on cluster balancing issue by partitioning the repository into domain specific memory blocks called clusters. This technique forms the base for the division of repository into multiple blocks. The partitioning approach deals efficiently with the problem of limited size of storage resources and data searching complexity in repository. The main component of proposed architecture is coordinator module which not only indexes the documents but also decide the memory cluster for a web document but also uses compression technique to compress the size of document and increase the storage capacity of repository.

INTRODUCTION

The Internet [1,15] is an interconnection of millions of computers around the world. It is a global information system that is logically linked together by a globally unique address space based on the Internet Protocol (or its subsequent extensions). It is the biggest repository of online knowledge wherein the end-user employs the tool for searching and sharing information.

The Internet has revolutionized the computer and communications world like nothing before. The invention of the telegraph, telephone, radio, and computer set the stage for this unprecedented integration of capabilities. Within a span of few years, it has changed the way we do business and communicate. Today, it has become a world-wide broadcasting capability, a mechanism for information dissemination and a medium for collaboration and interaction between individuals and their computers without regard for geographic location.

INTERNET PROTOCOL

The Internet Protocol (IP) is the principal communication protocol [1,7] in the Internet protocol suite for relaying datagram across network boundaries. Its routing function enables internetworking and essentially establishes the Internet. IP, as the primary protocol in the Internet layer of the Internet protocol suite, has the task of delivering packets from the source host to the destination host solely based on the IP addresses in the packet headers. For this purpose, IP defines packet structures that encapsulate the data to be delivered. It also defines addressing methods that are used to label the datagram with source and destination information.

General Architecture of a Search Engine

A general web search engine [7,9] has three parts i.e. Crawler, Indexer and Query engine. The *web crawler* (also called robot, spider, worm, walker or wanderer) is a module that searches the web pages from the web world. These are small programs that peruse the web on the search engine's behalf, and follow links to reach different pages. Starting with a set of seed URLs, crawlers extract URLs appearing in the retrieved pages, and store pages in a repository database search engine

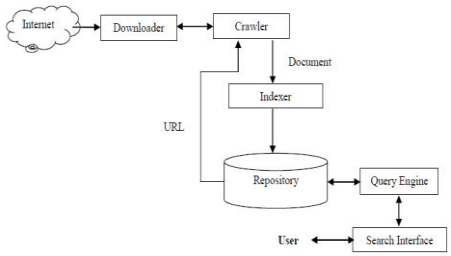

Figure 1. General architecture of search engine

DESIGN ISSUES OF SEARCH ENGINES

Designing a large-scale search engine is a non trivial task and entails huge challenges. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago.

A perfect automated search engine in the current scenario is one that crawls the web quickly and gathers all the documents to keep them up-to-date. Plenty of storage space is required to efficiently store indices or the documents themselves. The magnitude of data that has to be handled on the ever-growing internet includes billions of queries daily.

The indexing system of a search engine should be capable of processing huge amount of data by using the space most efficiently and handling thousands of queries per second. The best navigation experience should be provided to the users, in the form of finding almost anything on the web, excluding the junk results with the use of high precision tools, which is the main problem the users face.

Designing a search engine is a tricky task and there are differences in the ways they work. There are different ways to improve performance of search engines but three main characteristics are improving algorithms to search the web, using filters towards the user's results; and improving the user interface for query input**.**

General Architecture of a Web Crawler

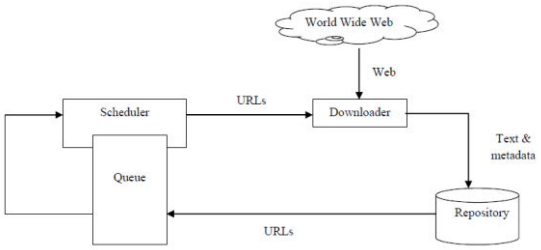

Figure 2: General architecture of web crawler

The browser parses the document and makes it available to the user. Normally, a crawler starts by placing an initial set of seed URLs [10] in a queue, where all URLs to be retrieved are kept and prioritized. From this queue the crawler extracts a URL, downloads the page, extracts URLs from the downloaded

page, and places the new URLs in the queue. This process is repeated and the collected pages are used by a search engine. The browser parses the document and makes it available to the user

Parallel Crawlers

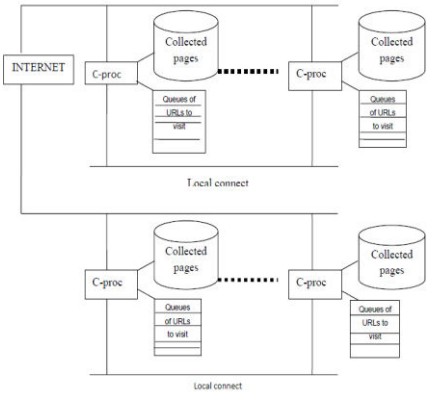

Figure 3: General architecture of parallel crawler

A parallel crawler [37] is very useful as web size is growing at a fast rate. In a parallel crawler (fig. 3) multiple crawling processes (C-Proc's) run in parallel to perform the crawling task, which maximizes the download rate. Each C-proc performs the basic tasks that a single process's crawler conducts. Here, all C-Proc's run on the same local network and communicate through a high speed interconnection, and use the same local network to download web pages. The network load from C-Proc's is centralized at a single location where they operate. To improve the quality of downloaded pages and to prevent overlap, the crawling processes need to communicate with each other.

WEB REPOSITORY

There is still considerable confusion across education, training, research and information management communities regarding what actually constitutes a repository, given the wide utility of the concept. Many HE practitioners implicitly equate digital repositories with learning object repositories, which may contain both content and metadata or may only contain metadata which references external resources. The term "referatory" [33] has been coined in recent years to identify repositories that contain metadata alone. But a web repository is a storage module that [29,31,39] stores and manages a large collection of 'data objects' in this case web page. So a web repository is basically collection of web pages or web documents. The repository receives web pages from a *crawler*, which is the component responsible for mechanically finding new or modified pages on

the web. Repository contains the full HTML of every web page. A web repository need to provide the following functionalities [38]:

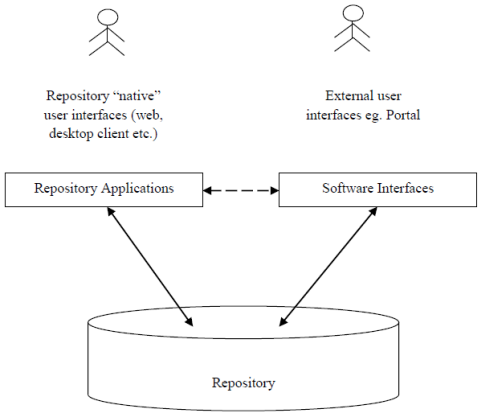

Figure 4: General communication to Web Repository

ARCHITECTURE

The proposed work is an architecture as shown in figure of crawler with enhanced repository technique which deals efficiently with the problem of limited size of storage resources and data searching complexity in repository. This technique forms the base for the division of repository into multiple blocks. The memory is partitioned into numbers of fixed size blocks and these blocks are called as clusters as represented in figure

and the partition of memory into blocks is done on the foundation of different domains like .com, .edu, .org etc. that belongs to web. As the memory block creation is based on specific domain so these clusters are also called as domain specific clusters.

Each domain has its own cluster and the particular cluster contains the documents that only belong to its specific domain type. Partitioning the repository into number of clusters and balancing these clusters, two different modules i.e. coordinator module and ranking module are used

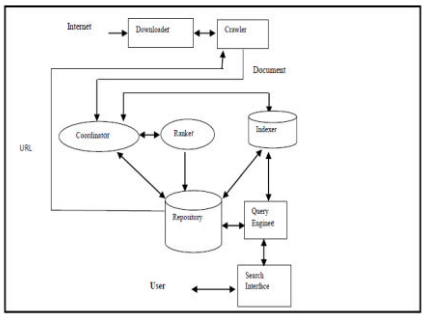

Figure5: General Architecture of clustered Webbase Search engine

Conclusion

The proposed work presented clustered web repository, and meta-data management. Clustered Web base uses repository and working modules to distribute data among domain specific clusters which compose a clustered web repository. Distribution of web pages to different domain specific blocks reduces searching complexity. The proposed work also representing an indexing mechanism to store the keyword present in the document in compressed form by mapping variable length Huffman code. It also focuses on the presence of keyword in different document because the keyword in maximum number of documents will be mapped to less length Huffman code. The mechanism reduces the size of document and after updates the index. The data structure of the indexer fastens the search for matched results from the Inverted Index with the cluster information. It also helps the user to process the user query with fast and more relevant results

References

[1] Gary C. Kessler, "An Overview of TCP/IP Protocols and Internet",2007.

[2] Grossan, B, "Search Engines : What they are, how they work, and practical sugg estions for getting the most out of them", 1997.

[3] A. Ntoulas, J. Cho, and C. Olston, "What's new on the web ? : the evolution of the web from a search engine perspective.", In Proceedings of the 13th conference on World Wide Web, pages 1-12, 2004.

[4] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, Sriram Raghavan, "Searching the Web.", ACM Transactions on Internet Technology, 2001.

[5] Brewington, B. E., Cybenko, G, "Keeping up with the changing web", IEEE Computer, 2000.

[6] Komal Kumar Bhatia, A. K. Sharma, "A Framework for Domain-Specific Interface Mapper (DSIM)", International Journal of Computer Science and Network Security (IJCSNS), Vol 8, No12, Dec 2008.

[7] Niraj Singhal, Ashutosh Dixit, "Retrieving Information from the Web and Search Engine Application", in proceeding of National conference on "Emerging trends in Software and Networking Technologies ETSNT'09", Amity University, Noida, India, pp. 289-292,2009.

[8] S. Chakrabarti, M. van den Berg, and B. Dom., "Focussed Crawling: a New Approach to Topic-specific Web Resource Discovery", In Proceedings of the Eight International World-Wide Web Conference, pages 545-562, Toronto, Canada, May 1999.

[9] Niraj Singhal, Ashutosh Dixit, "Web Crawling Techniques : A Review", in proceedings of National Seminar on "Information Security : Emerging Threats and Innovations in 21st Century", Ghaziabad, March 2009, pp. 34-43.

[10] A. Heydon, Najork M., "Mercator : A scalable, extensible Web crawler.", World Wide Web, vol. 2, no. 4, pp. 219-229, 1999.

[11] Berners-Lee and Daniel Connolly, "Hypertext Markup Language. Internetworking draft", http://www.w3.org/hypertext/WWW/MarkUp/ HTML.html, 13 Jul 1993

[12] B. Kahle,"Archiving the Internet", Scientific American, 1996.

[13] M. Gray, "Measuring the growth of the Web", http://www.mit.edu /people/mkgray/growth/, 1993.

[14] Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard, Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, Stephen Wolff, "A Brief History of the Internet", www.isoc.org/internet/history.

[15] Bernardo A. Huberman, Lada A. Adamic, "Growth dynamics of the World-Wide Web. Nature",1999.

[16] Fred Douglis, Anja Feldmann, Balachander Krishnamurthy, "Rate of change and other metrics: a live study of the world wide web", USENIX Symposium on Internetworking Technologies and Systems, 1999.

[17] Franklin, Curt, "How Internet Search Engines Work", 2002, www.howstuffworks.com.

[18] Gulli A., Signorini A.,"The Indexable Web is More than 11.5 billion pages", Proceedings of the Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan. pp. 902-90, 2005.

[19] Yates R. Baeza, B. Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley, 1999.

[20] B. Pinkerton, "WebCrawler : Finding What People Want," Ph.D. dissertation, University of Washington, 2000Brian E. Brewington, George Cybenko, "How dynamic is the web.", In Proceedings of the Ninth International World-Wide Web Conference, Amsterdam, Netherlands, 2000.

[21] Junghoo Cho, Hector Garcia-Molina, "The evolution of the web and implications for an incremental crawler", In Proceedings of the 26th International Conference on Very Large Databases, 2000.

[22] Junghoo Cho, R. Adams. "Page quality:Ins search of an unbiased Web ranking", Technical report, UCLA Computer Science, 2004.

Ajay Kumar Yadav (Associate professor)

College of Engineering and Rural Technology Meerut (U.P.)