ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 11(20), DOI: 10.17485/ijst/2018/v11i20/123345, May 2018*

Analysing Performance in Retrieving Heterogeneous Information in Big Data Environment

N. Sulaiman[1]*, Osamah Ibrahim Khalaf[2], Ghaida Muttashar Abdulsahib[3] and R. Adel[1]

[1]College of Computer Science and Engineering, Taibah University, Medina 42353, Saudi Arabia; nsulaiman@taibahu.edu.sa, myexcel12@yahoo.com
[2]College of Information Engineering, Al-Nahrain University, Baghdad, Iraq; usama81818@yahoo.com
[3]Department of Computer Engineering, University of Technology, Baghdad, Iraq; gh961@yahoo.com

Abstract

Background/Objectives: The development of various communication media has generated few problems in retrieving information. The objective of the study is to analyze the performance of retrieving heterogeneous data. **Methods/Statistical Analysis**: A model was run to simulate the process of retrieving heterogeneous data from several servers. The information was distributed with different load and the servers were randomly selected. The performance had been analysed based on response time and CPU utilisation. A few types of load balancing techniques were applied to distribute the loads among the servers. The impacts on the overall system performance were discussed. **Findings**: Retrieving data requires high speed, where the response time must be very fast. The performance of retrieving heterogeneous data is a challenge, when servers have high load. When the load balancing techniques were not applied, some of the servers handle the entire load and the other servers have not been fully utilised. The results showed the response time decrease drastically when high load of data were applied to the server. When the load balancing was applied, the results were compared and presented. The results showed an improvement in the overall performance. **Improvements/Applications:** The load balancing techniques were applied based on several approaches. It allows an improvement in distributing the server load, which results in improvement in the performance.

Keywords: Analysing Performance, Big Data Environment, Heterogeneous Information

1. Introduction

Since the past few years, the Internet plays an important role in our lives. Many of our activities are searching for something via Internet, watching videos, writing in social media even make video or voice call. Variation of data is one of the features of big data. Retrieving enormous volumes of information from heterogeneous data resources still remains challenging.

Information can be retrieved frequently as a different type of documents. Restrictions of the distributed data sources for each user are based on the use of multiple data formats, bandwidth limitations as well as time constraints. When considering heterogeneous information retrieval system, the potential users with their information needs are important to be discussed. If the types of results to be used by the system can be identified, the responses from the retrieval process would be more relevant to the user needs.

The performance in retrieving information for heterogeneous data in big data environments has become a significant issue. Information retrieval involves searching

*Author for correspondence

information within documents, which includes searching unstructured and structured information. A study in information retrieval domain has been gradually increasing since the year 2000. This reflects the growing needs for studies of retrieving information in heterogeneousdata.

2. Web Search

The web can be modeled as a simple client-server model. The user web browser sends a request for web pages and the server responses. The web pages contain dynamic content from diverse sources. In order to display all the information, the web pages may be retrieved from many different servers. The web server receives a request from the user and collects the relevant information and provides the contents to the browser. The typical system architecture of a web application is as shown in Figure 1.

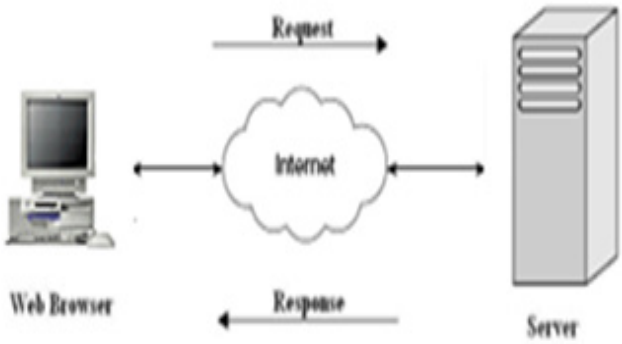

Figure 1. System architecture for web applications.

A few studies have been carried out on the retrieving information which includes heterogeneity of data and differences in access mechanisms. Some studies focus on the conceptual retrieval systems for heterogeneous information and a few studies focus on the availability of each data source.

Search engines allow users to find phrases, quotes, unique keywords and other information that are stored in the web pages. There are a few types of search engines available such as human powered directories, crawler based search engines, meta-search engines and hybrid search engines[1]. For crawler based search engines, the search engines index is created automatically. All pages retrieved can be rank using a computer algorithm. The search engines often retrieve a lot of information. Searching within the results of the previous search is allowed for complex searches. For human powered directories, the repositories are created by human, where the directories are organized into classification of pages in which subject are categorised.

For meta-search engines, the results of multiple primary search engines are accumulated. Searches can be sent to several search engines all at the same time. The results will be combined together as one page result. On the other hand, hybrid search engine is different from traditional text oriented and directory based search engines. This type of search engine is good for one type of listing over the other. In general, a combination of a crawler based search engine and a directory service is used by many search engines.

It was claimed that normal text searching is economically feasible to be used over the past few years[2]. This is due to the fact that in dissemination and retrospective search request programs, normal text searching has an ability to gain meaningful information retrieval in searching.

A few methods have been proposed which allow information retrieval systems to learn and predict what information is being retrieved. There is a growing interest in improving the search process towards the user needs. The system ensures the interaction by integration of the informational needs of the user and the process of annotation, which makes annotating retrieved information and the contexts of use of retrieved information possible for user[3].

A collection of documents need to be searched and the retrieved document will be either relevant or non-relevant to a particular query. Algorithms to evaluate the retrieval system can be classified into the ranked or unranked retrieval results[4]. The users' needs and behaviours in information retrieval were discussed[5]. It was found that in developing a good information system, the user behavior and understanding their needs are important

The search engines can be categorized into several categories. In primary search, the entire sections of the web are scanned and the results are produced from the database which is created by computers. Search engines for business and service seem to be the important directories for search engines. For Employment and Job search engines, information about potential employers and prospective employees with information on job availability is provided.

Searching for specific information about companies is facilitated by finance-oriented search engines. Similarly, the image search engines help in searching all kind of images. For searching newspaper and news in website archives, news search engine is required. The people search engines helps in searching information about people details such as names, addresses, and e-mail addresses. Subject guides will be used in selecting and organizing resources, which provide more focus and guidance in selection. For searching specialized databases, users are allowed to enter search terms in an easy way based on the user needs[1].

Challenges with big data are described by author in[6], which include typical analytical problems in big data analysis with the ways to solve the problems. Retrieving medical data from heterogeneous information were discussed[7,8]. Ontologies are used for managing semantic knowledge in which distributed heterogeneous data sources can be queried by hiding details like information structure, access pattern and the data semantic structure[9]. A structural approach was proposed where heterogeneous information networks can be constructed almost in any domain[10]. Mobile agents for distributed and heterogeneous information retrieval were suggested for large, heterogeneous, and distributed data sources[11].

The basic principle of data integration was proposed in which to integrate chosen information sources from a specific domain to generate a new data source[12]. However, it is not cost effective to create new data source based on information which is integrated from heterogeneous data sources. Achievements and new challenges in big data were described, which focus on the social networks. The challenges include data processing, data storages, and data tracking[13].

The challenges and opportunities in big data of various types were discussed[14]. Lack of structure, heterogeneity, error handling, privacy, timelines and visualization are some of the challenges that had been identified. It involves all stages from data acquisition process, which includes the result interpretation. Video transmission over heterogeneous networks was described[15]. The effect of using various types of error correction algorithms over heterogeneous networks was analysed. It was shown that the method affect the received video quality.

3. System Models

A system model was simulated to analyse the use of HTTP protocol in retrieving information from the Internet. The model was simulated with all HTTP requests are sent to one server. A web search typically contains the title and URL of the link. A source will send a request to the server. A server will respond with an HTTP response message, after receiving and interpreting a request message.

Search engines provide an interface that enables users to specify the query about terms to be retrieved and the engine find the matching information. For text search engines, a set of words will be required to identify the desired information contains in the documents. To get the information from the web, fast crawling technology is required. Speed is the main priority in searching and response time must be very fast as suggested by Lavania[16].

3.1 Response Time

Response time for web applications is influent by many factors including server capabilities, protocol, and network characteristics. The response time is affected by the situation where information may be placed on a single server. The performance and availability are some of the factor affecting the response time. By using multiple servers at a different location, information can be distributed and this will improve the performance and availability. The

server passes additional information about the response using the response-header fields. Information about the server and further access to the resource identified by the requests can be identified from on these header fields.

Assume that the time $t_{i,\ 1,\ \ldots,}t_{i,xi}$ to be chosen to minimize the average time delay which is estimated to be a_i $(t_{i,\ 1,\ \ldots,}t_{i,xi})$, given that there are$n_i$ crawls of page i in the system[17].

Hence,

$$A_i\,(x_i) = a_i\,(t_{i,\ 1,\ \ldots,}\ t_{i,xi}) \qquad (1)$$

The optimal values of the x_i variables will be chosen, with the aim to minimize the function

$$\sum_{i=1}^{N} wiAi(xi) \qquad (2)$$

depends on the constraints

$$\sum_{i=1}^{N} xi = R \qquad (3)$$

Where

$xi \in \{0, \ldots, R\}$ and the weight

will be used to determine the importance of each web page i.

3.2 Server Load

Several servers are used where the information can be distributed. This server will handle the loads that have been allocated. However, if this server fails, then all requests will be denied until it recovers.

Let's consider:
s is the number of servers
t is the time to process a request
N is the number of jobs
p is the number of jobs processed by each server

$$p_i = \frac{\frac{1}{t_i}}{\sum_{i=1}^{s}\frac{1}{t_i}} \times N$$

Assume the following an algorithm that give allocation of jobs to the server:

For i = 1 to s
while (p_1 + ... + p_p< N)
find k in {1,...,s} such that
$t_k(p_k + 1) = \min\{t_i(p_i + 1)\}$
$p_k = p_k + 1$

3.3 Load Balancing

To improve the performance, server capabilities was simulated where the response time will depend on the

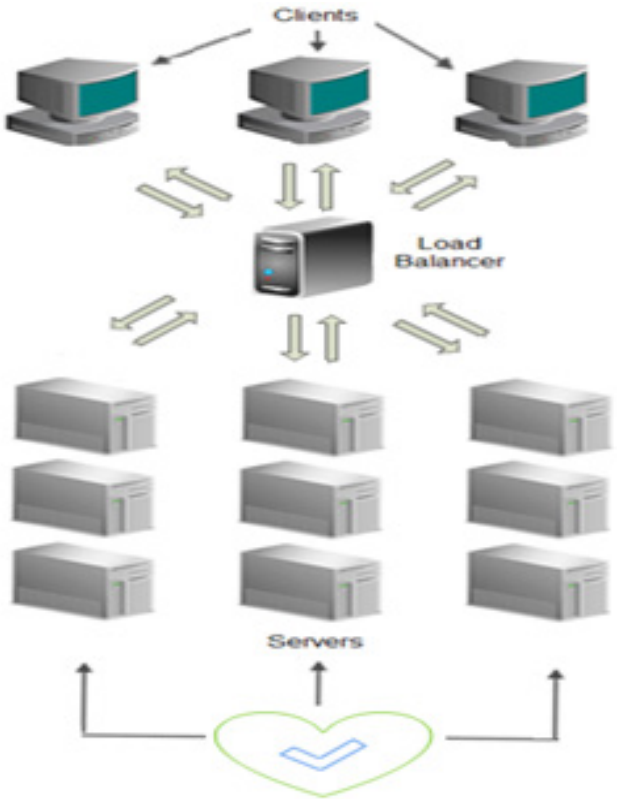

Figure 2. Health check load balancing.

processing load. The model was simulated with the load balancing techniques were applied. Load balancing is a way to improve the system performance by distributing loads among the servers. The load balancer will choose one of the candidates as shown in Figure 2.

An appropriate server will be chosen based on the lowest server load and client's request will be forwarded to that server. The load balancer forwards traffic to the healthy pool of servers. The servers will be checked regularly and attempt to connect to the server are tested to ensure that servers are listening. If a server fails, it will not be able to serve the request. Hence, it will be removed from the pool and traffic will not be forwarded to it until it recovers.

These scenarios are also simulated with a different configuration in which the server will be selected based on weight and the least number of connections. Based on the types of balancing applies. The model was run based on a few scenarios and the results are analysed.

4. Results and Discussion

A simulation model was run based on different scenario. The model was run with the lowest server load was applied. The model was simulated with different types of load. Then, the load balancing approach was used to study the impact on the performance. The response time to retrieve all the information from the server was recorded and compared. The CPU utilizations have also been monitored.

4.1 Effect on Response Time

Figure 3 shows the response time for web users. The results show that the server can handle the traffic. However, as the load increases very high, the response time increases drastically.

Figure 4 shows the response time for web users when the server with the least number of connections will be selected. By specifying a maximum number of connections for each server, it will avoid the server from exceeding the capacity limit. The results show that the response time is reduced to about 50%.

Figure 3. Response time for web client.

Figure 4. Response time for web client when load balancing is applied.

Figure 5 shows the response time for web users when weight balancing is applied. The results show that the response time for all the clients is further reduced. The server with less weight will be selected and it improves the network performance with affects the response time.

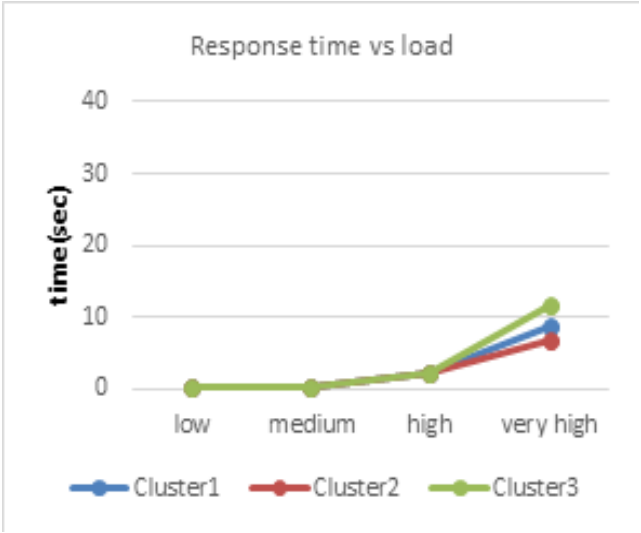

Figure 5. Response time for web client when load balancing is applied.

4.2 Effect on CPU Utilisation

Figure 6 shows the effect of load on CPU utilisation for each server. The results show that when low and medium usages are applied, server 2 and server 3 have 0% of utilisation. In addition, when the high load and very high usage are applied, the percentage of utilisation has increased.

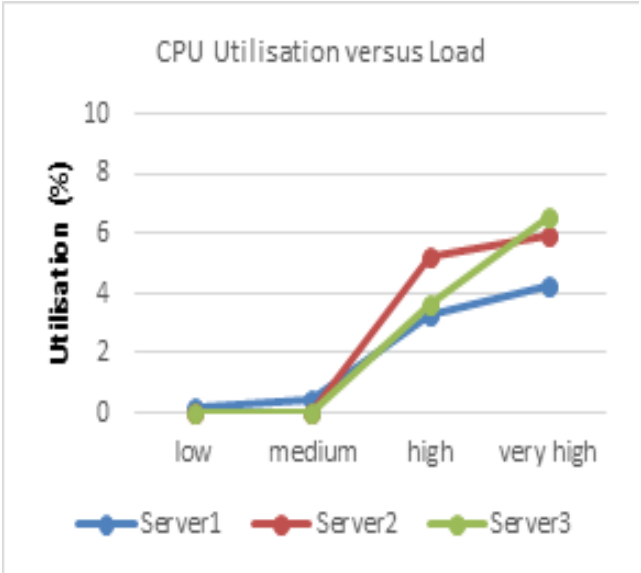

Figure 6. CPU utilisation.

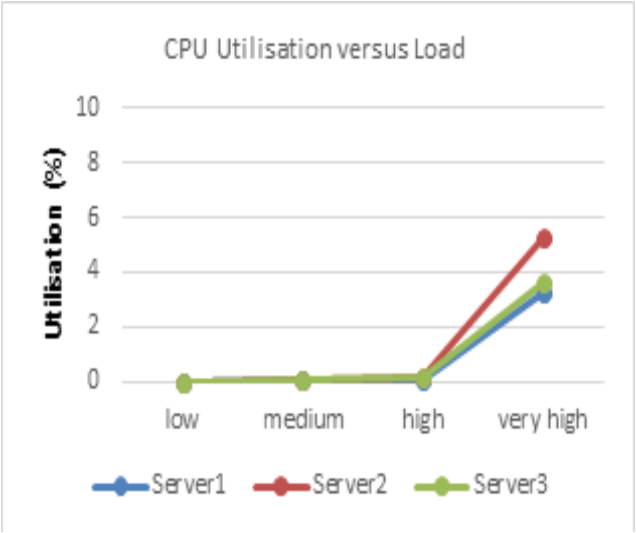

Figure 7. CPU utilisation when load balancing is applied.

Figure 7 shows the CPU utilisation for each server after the load balancing is applied. The percentage of utilisation shows that the load is spread uniformly among the available servers. The load on a particular server is much lower compared to when no balancing was applied.

5. Conclusion

The way people communicate and doing business have changed with the Internet technology. The evolution of technology enables us to access a huge amount of information anytime and anywhere. The World Wide Web is a hyperlinked repository of data in different server all over the world. It is no longer the lack of existence of information, but the problem is the difficulty is in accessing the information. Internet evolution has become unlimited and the heterogeneous information is difficult to manage. Retrieving heterogeneity of data involves major issues including the different access mechanism, response time and the server performance.

The models were simulated to compare the response time for different scenarios. The response time for retrieving information was simulated and the results were analysed. The results show that the loads on the servers are not fairly distributed. When one of the servers fails, the request will be rejected and the response time will be

longer. When the load balancing was applied to improve the performance of the system, the results show that the response time is reduced. The results indicate that the performance in retrieving data in heterogeneous big data environments is affected by the types of information retrieval as well as the server performance in managing the process. An appropriate way to reduce the response time is recommended.

6. References

1. Kumar D, Kumar A. Design issues for search engines and web crawlers: A review. IOSR Journal of Computer Engineering (IOSR-JCE). 2013; 15(6):34–7. Crossref.
2. Bama SS, Ahmed MSI, Saravanan A. A survey on performance evaluation measures for information retrieval system. International Research Journal of Engineering and Technology (IRJET). 2015May; 2(2):1015–20.
3. Guo K, et al. An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval. Journal of Systems and Software. 2015Apr; 102:207–16. Crossref.
4. Zuva K, Zuva T. Evaluation of information retrieval systems, International Journal of Computer Science and Information Technology (IJCSIT). 2012 Jun; 4(3):3–43. Crossref.
5. Lee JH, Cunningham SJ. Toward an understanding of the history and impact of user studies in music information retrieval. Journal of Intelligent Information Systems. 2013 Dec; 41(3):499–521. Crossref.
6. Alguliyev RM, Gasimova RT, Abbash RN. The obstacles in big data process. I.J. Modern Education and Computer Science. 2017; 3:28–35. Crossref.
7. Quellec G, et al. Case retrieval in medical databases by fusing heterogeneous information. IEEE Transactions on Medical Imaging. 2011 Jan; 30(1):108–18. Crossref.PMid:20693107
8. Huang DS, et al. Multi-modality medical case retrieval using heterogeneous information. Proceeding of International Conference on Intelligent Computing; China. 2014. p. 80–91.
9. Munir K, et al. Semantic Information Retrieval from Distributed Heterogeneous Data Sources, International Workshop on Frontiers of Information Technology; 2006.
10. Sun Y, Han J. Mining heterogeneous information networks: A structural analysis approach. ACM SIGKDD Explorations Newsletter. 2012 Dec; 14(2):20–28. Crossref.
11. Brewington B, et al. Mobile agents for distributed and heterogeneous information retrieval. Intelligent Information Agent, Springer Verlag. 1999. p. 355–95.
12. Zaman M, Quadri SMK, Butt MA. Information integration for heterogeneous data sources. IOSR Journal of Engineering. 2012 Apr; 2(4):640–3. Crossref.
13. Orgaz GB, Jung JJ, Camacho D. Social big data: Recent achievements and new challenges. Information Fusion. 2016; 28:45–9. Crossref.
14. Ammu N, Irfanuddin M. Big data challenge. International Journal of Advanced Trends in Computer Science and Engineering. 2013; 2(1):613–5.
15. Khalaf OI, et al. Improving video transmission over heterogeneous network by using ARQ and FEC error correction algorithm. Indian Journal of Science and Technology. 2015 Nov; 8(30). Crossref.
16. Lavania KK, et al. Google: A Case Study (Web Searching and Crawling). International Journal of Computer Theory and Engineering. 2013 Apr; 5(2):337–40. Crossref.
17. Wolf JL, et al. Optimal crawling strategies for web search. Proceeding of 11th International World-Wide Web Conference; USA. 2002. p. 136–47. Crossref.