The Performance Study of Hyper Textual Medium Size Web Search Engine

Tarek S. Sobh and M. Elemam Shehab
Information System Department, Egyptian Armed Forces
tarekbox2000@gmail.com melemam@hotmail.com

Abstract

Despite the importance of med-scale search engines, 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. In this paper, we present MSSE (Medium Scale Search Engine), a prototype of a medium scale search engine which makes heavy use of the structure present in hypertext. MSSE is designed to crawl and index the Intranet and produce satisfying search results competing convenient systems. The prototype with a full text and hyperlink database was introduced of at least 500000 pages. We provided an in-depth description of MSSE. Moreover, MSSE is done in order to search over Intranet not on the Internet This work addresses this question of how to build a practical mid-scale system which can add additional information present in hypertext. Also we use OPNET modeler to simulate the environment of MSSE in order to verify the obtained results about the performance of the search engine.

Keywords: *World Wide Web, Search Engines, Information Retrieval, Page Rank and Scaling.*

1. Introduction

As the amount of information available on the websites increases, it becomes necessary to give the user the possibility to perform searches over this information. The using of search engine is one of the most activities used on Internet. Figure 1 show the top activities on Internet today [9].

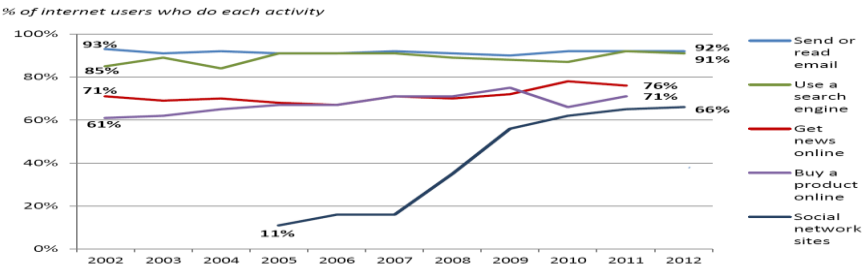

Fig.1: Percentage of Internet user's activities

When deciding to install a search engine in a website, there exists the possibility to use a commercial search engine or an open source one. For most of the websites, using a commercial search engine is not a feasible alternative because of the fees that are required and because they focus on large scale sites. On the other hand, open source search engines may give the same functionalities (some are capable of managing large amount of data) as a commercial one, with the benefits of the open source philosophy: no cost, software maintained actively, possibility to customize the code in order to satisfy personal needs, etc. Nowadays, there are many open source alternatives that can be used, and each of them have different characteristics that must be taken into consideration in order to determine which one to implement in your website. These search engines can be classified according to the programming language in which it is implemented, how it stores the index (inverted file, database, other file structure), its searching capabilities (boolean operators, fuzzy search, use of stemming, etc), way of ranking, type of files capable of indexing (HTML, PDF, plain text, etc), possibility of on-line indexing and/or making incremental indexes[10],[12].

This paper is structured as follows: Section 2 explains architecture of search engine. Section 3 illustrates the proposed MSSE search environment. Section 4 presents Evaluating MSSE Search Engine and metrics measures of the proposed MSSE. Section 5 introduces a performance study of MSSE and finally section 6 contains conclusion.

2. Architecture of Search Engine

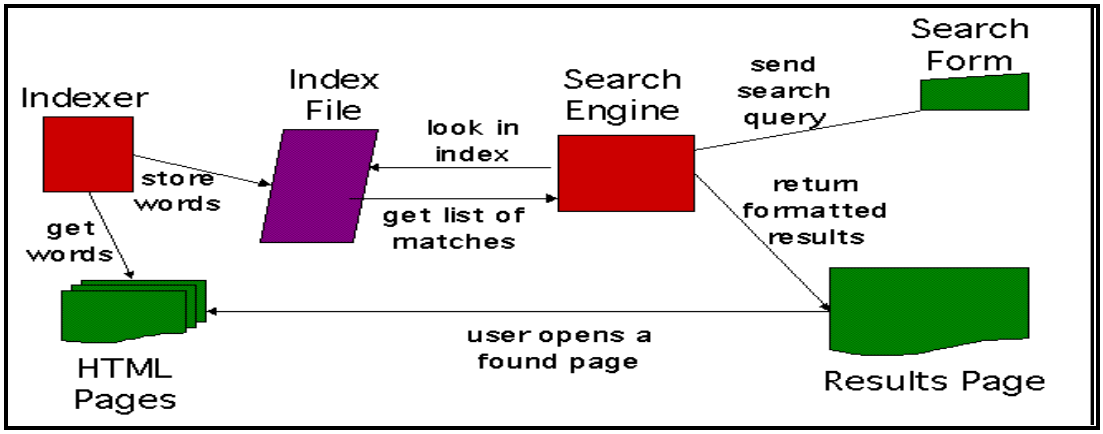

Fig.2 how search engine work

Our search engine architecture is used to present high level descriptions of the important components of the system and relationships between them. the architecture of MSSE is designed to ensure that it will satisfy the application requirements or goals. The two primary goals of MSSE are:

- **Effectiveness (Quality)**: We want to be able to retrieve the most relevant set of documents possible for a query.
- **Efficiency (Speed)**: We want to process queries from users as quickly as possible.

2.1 Basic Building Blocks

Search engine components support two major functions, which we call the indexing process and the query process. The indexing process builds the structures that enable searching and the query process is to respond the user's queries. Indexing process: have various steps:

Crawling: Finding and downloading web page automatically is called crawling and a program that downloads pages is called a web crawler.

Indexing: The index component takes the output of the text transformation component and creates the indexes or data structures that enable fast searching. Given the large number of documents in many search applications, index creation must be efficient, both in terms of time and space.Indexes must also be able to be efficiently updated when new document are acquired [6][1].

Ranking (Scoring): The scoring component also called query processing calculates the scores for documents using the ranking algorithm which is based on a retrieval model. The designer of some search engines explicitly state the retrieval model they use. For other search engine, only the ranking algorithm is discussed .Many different retrieval models and methods of deriving ranking algorithms have been proposed. The basic form of the document score calculated by many of these models is

$$\sum_{i} qi.\,d\,i$$

Where the summation is over all of the terms in the vocabulary of the collection, qi is the query term weight of i^{th} term and di is the document term weight.

3. MSSE Search Environment

The MSSE environment is established over Windows server 2008 R2 and hosted on WAMP Server (Windows Apache MySql PHP) the open source hosting server, MSSE consists of several components that are important for analyzing queries: the engine itself and the query logs, which store information about what queries are made to the engine.

3.1 MSSE Engine anatomy

MSSE is based on weighted Boolean search. There are two major search modes: simple querying and advanced querying. A simple query consists of a collection of words, which are ORed together. An advanced query is more explicitly boolean. In advanced query mode, and, or, and not are interpreted as boolean operators rather than as search terms. Advanced queries may also include tolerant search which find more alternatives for the keywords you are searching for and also containing phrase search which accept long string of text. After a query is entered and the various other restrictions processed, MSSE returns a screen consisting of 10 URLs and there is a pull down menu which enable you to choose how many links to appear in each page, with information about each URL such as the title.

These URLs are ranked in order of (relevance or top level domain or number of hits) to the query, as determined by the engine internal relevance function, the level of domain and number of hits respectively). The user may click on any URL to explore the associated web page [4].

3.2 MSSE Query Log

Our search engine query log has many components, only some of which concern us here. The query is a text file consisting of a series of requests. A request may consist of a new query or a new result screen for a previously submitted query. Each request includes the following fields:

- A timestamp indicating when the query was submitted. The timestamp is measured in milliseconds [10].
- A cookie, which can be used to say whether two queries come from the same user (this field is blank if the user has disabled cookies);
- The query terms, exactly as submitted;
- The result screen, that is the requested range of search results;
- Other user-specified modifiers, such as a restriction on the result (black listed or white listed)
- Submission information, such as whether the query is a simple or advanced query; and Submitter information, such as the browser the submitter is using and the IP address of the submitting host.

4. Evaluating MSSE Search Engine

Evaluation is the key to making progress in building better search engines. It is also essential to understanding whether a search engine is being used effectively in a specific application .One of the primary distinction made in the evaluation of search engine is between effectiveness and efficiency. Effectiveness, loosely speaking measures the ability of search engine to find the right information and efficiency measures how quickly this is done, we can more define effectiveness as a measure of how well the ranking produced the search engine corresponds to a ranking based on user relevance judgements.Efficiency is defined in terms of the time and space requirements for the algorithm that produce the ranking.Effectivness and Efficiency will be affected by many factors such as the interface used to display search results and query refinement techniques such as query suggestion. In this work we concerned with effectiveness metrics that judge the search engine more clearly than efficiency.

4.1 Effectiveness Metrics

Recall and precision of the proposed MSSE are introduced. The two most common effectiveness measures ,recall and precision, Recall measure how well the search engine is doing at finding all the relevant documents for a query ,and precision measures how well it doing at rejecting non relevant documents.

$$Recall = \frac{|A \cap B|}{|A|} \qquad Precision = \frac{|A \cap B|}{|B|}$$

4.2 Relative Recall of MSSE

Recall is the ability of a retrieval system to obtain all or most of the relevant documents in the collection (Shafi & Rather, 2005). The relative recall can be calculated using following the formula:

$$Relative\ Recall = \frac{\text{no of sites retrieved by a search engine}}{\text{total number of sites retrieved by google and your search engine}}$$

In this work we make 2 experiments one on the internet and second on an intranet medium size scale in order to clarify the performance of MSSE.

4.2.1 Relative Recall of MSSE on Internet

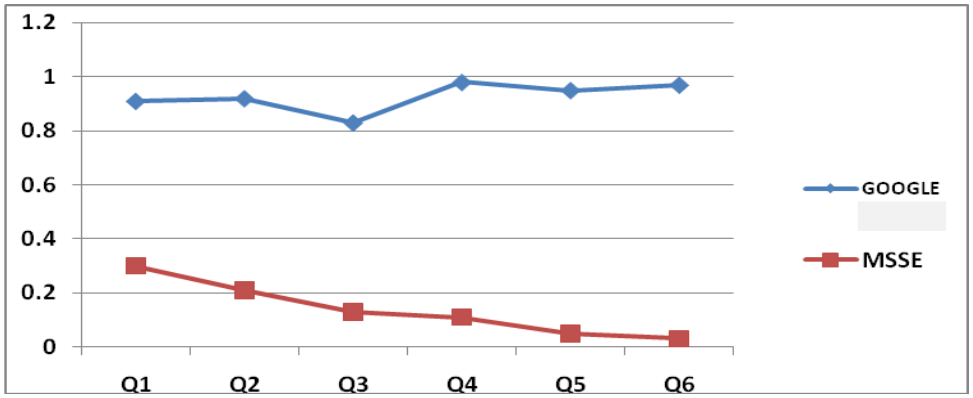

Fig.3 comparison of MSSE and Google Recall on Internet

Figure 3 shows relative recall of MSSE compared to Google search engine for six query done on both search engine ,the results shows that MSSE is not suitable for searching on large scale networks and there is a large different in performance between MSSE and Google . The number of queries (axis x) was represented against percentage of Recall (axis y).

4.2.2 Relative Recall of MSSE on Intranet

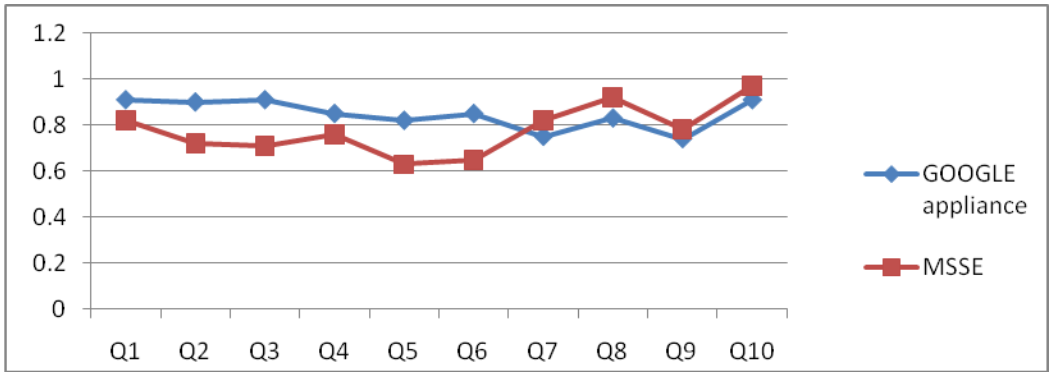

Fig.4 comparison of MSSE and Google Recall on Intranet

Figure 4 shows relative recall of MSSE compared to Google search engine appliance GSA7series for six query done on both search engine, the results shows that MSSE is very suitable and comparable to Google on medium size network (Intranet) and figure show that recall for both search engine is very close and leading to same level[11].

4.3 Precision of Search Engines

After a search, the user is sometimes able to retrieve relevant information and sometimes able to retrieve irrelevant information. The quality of searching the right information accurately would be the precision value of the search engine (Shafi & Rather, 2005). In the present study, the search results which were retrieved by the Google was classified as 'more relevant', 'less relevant', 'irrelevant', 'links' and 'sites can't be accessed' on the basis of the following criteria (Chu & Rosenthal, 1996; Leighton, 1996; Ding & Marchionini, 1996; Clarke & Willett, 1997):

Table 1: Rules used to measure the precision

Rule	score
If the web page is closely matched to the subject matter of the search query then it was categorized as 'more relevant'	2
If the web page is not closely related to the subject matter but consists of some relevant concepts to the subject matter of the search query then it was categorized as 'less relevant'	1
If the web page is not related to the subject matter of the search query then it was categorized as 'irrelevant'	0
If a web page consists of a whole series of links, rather than the information required, then it was categorized as 'links	.5
If a message appears "site can't be accessed" for a particular URL the page was checked again later. If the message occurs repeatedly the page was categorized as 'site can't be accessed'	0

Table.2 showed that 28.7% of the sites retrieved by MSSE were less relevant followed by links (18.2%) and irrelevant sites (29.6%). It was also observed that 17% sites were more relevant and only a small percentage of the sites (6.5%) "can't be accessed". The precision of the MSSE was calculated using the above formula. The overall precision of the MSSE was 71.8.

From the results obtained from our search engine and Google search engine obtained from Sampath & Prakash (2009) MSSE has a promising results compared that obtained from Google for one word query and for more tested sites we conclude that MSSE is more suitable to Intranet search not Internet.

$$precision = \frac{\text{Sum of the scores of sites retrieved by a search engine}}{\text{Total number of sites selected for evaluation}}$$

Table.2 precision of MSSE of a series of single word query

search QUERY	no of sites evaluated	more relevant	less relevant	irrelevant	links	cant accessed	precision
Q1	100	18	21	14	42	3	78
Q2	100	18	34	31	12	5	76
Q3	100	18	20	31	18	13	65
Q4	100	18	24	33	17	8	68.5
Q5	100	14	33	30	17	6	69.5
Q6	100	18	35	26	18	3	80
Q7	100	16	35	27	19	5	76.5
Q8	100	18	31	31	14	4	74
Q9	100	18	24	44	11	5	65.5
Q10	100	14	30	29	14	13	65
total	**1000**	**0.17**	**0.287**	**0.296**	**0.182**	**0.065**	**71.8**

5. Simulation Measurement

5.1 Simulation Environment

The simulation experiment is carried out using OPNET simulator under Windows XP as a platform, the OPNET instructions can be used to define the topology structure of the network.

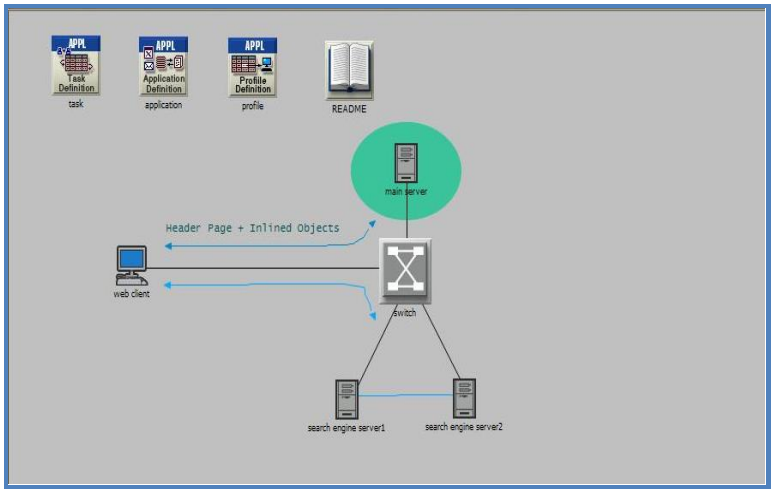

Fig.5 layout of the search engine simulation model.

Figure 5 shows the reference structure of our network model built .The shown network topology consists of three different sites, 'site1' to 'site3', every pair of which is interconnected by an switch. The main server with a co-located location server acts as the functional core. The first site act as web client, the second site act as a main server which is the middle layer between the web client and search engine servers and the third site the search engine servers.

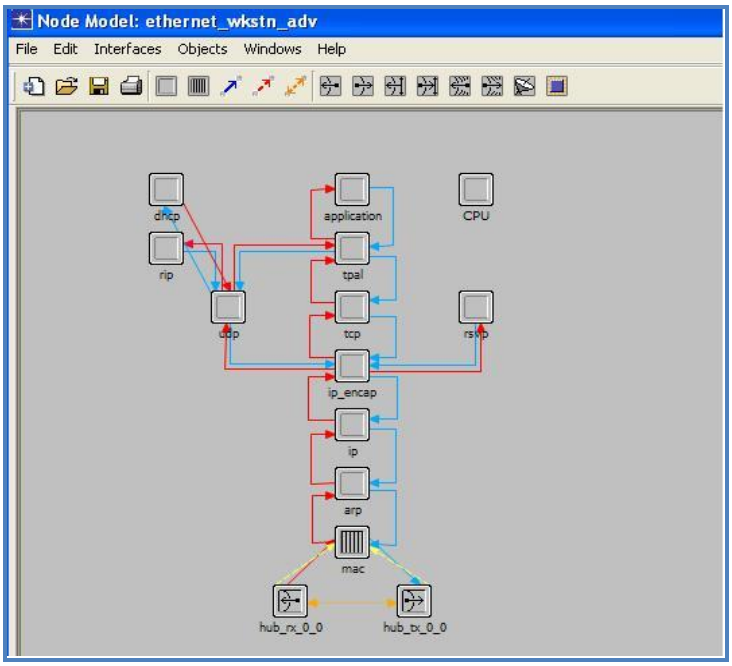

Fig.6 Node model represented in all devices in the network topology

Node model which represent the layers stack for every device used in the simulation model

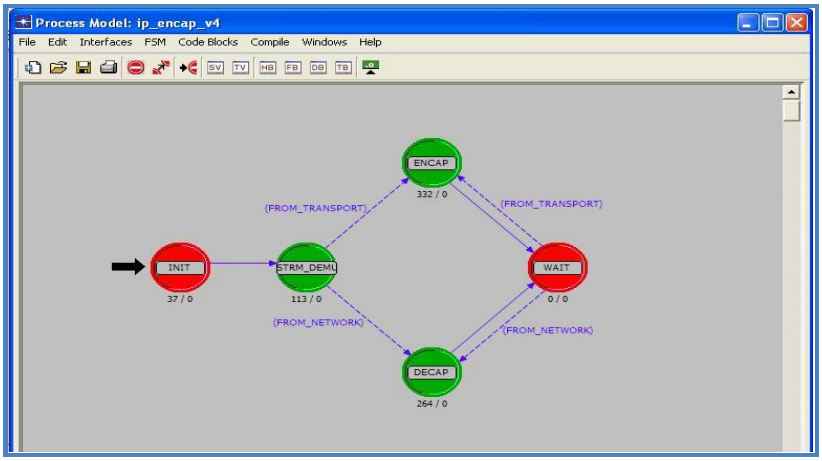

Fig.7 Process model represented in all devices in the network topology

Process model simulate all the process done in the model and its status like ENCAP (encapsulated) and DECAP (decapsulated).

5.2 Simulation Results and Analysis

This section reports the results obtained to examine a prototype of a search engine regarding all specification of real machines used to deploy MSSE on the OPNET simulator. The measuring criteria's used to evaluate the mentioned protocol are HTTP response time, Application load, Application response time and packet delay.

5.2.1 HTTP response time

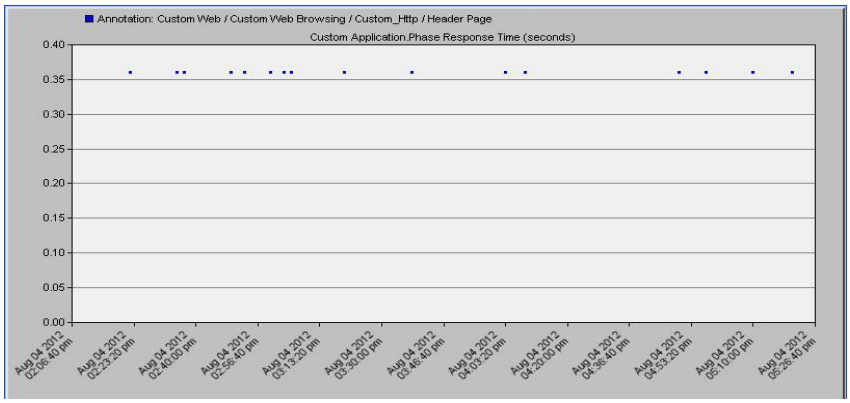

Fig.8 Application response time in seconds

The first metrics is to measure how much time the search engine will to response to a HTTP requests and from figure 8 we conclude that the application response time is varying from 0.35 and 0.40 seconds and we see that it is a reasonable response time.

5.2.2 Application load

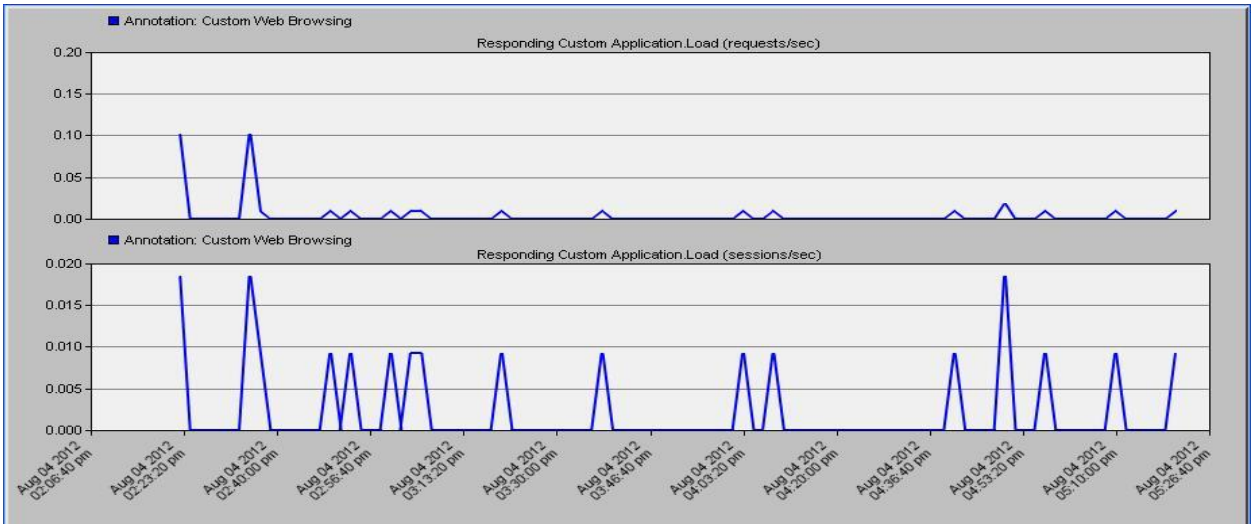

Fig.9 Application load (request/sec) and (sessions/sec)

Figure 9 show the application load in both (request/sec) and (session/sec) and the both provide that the load for our application is in normal ranges and very promising results.

5.2.3 Application response time and packet network delay

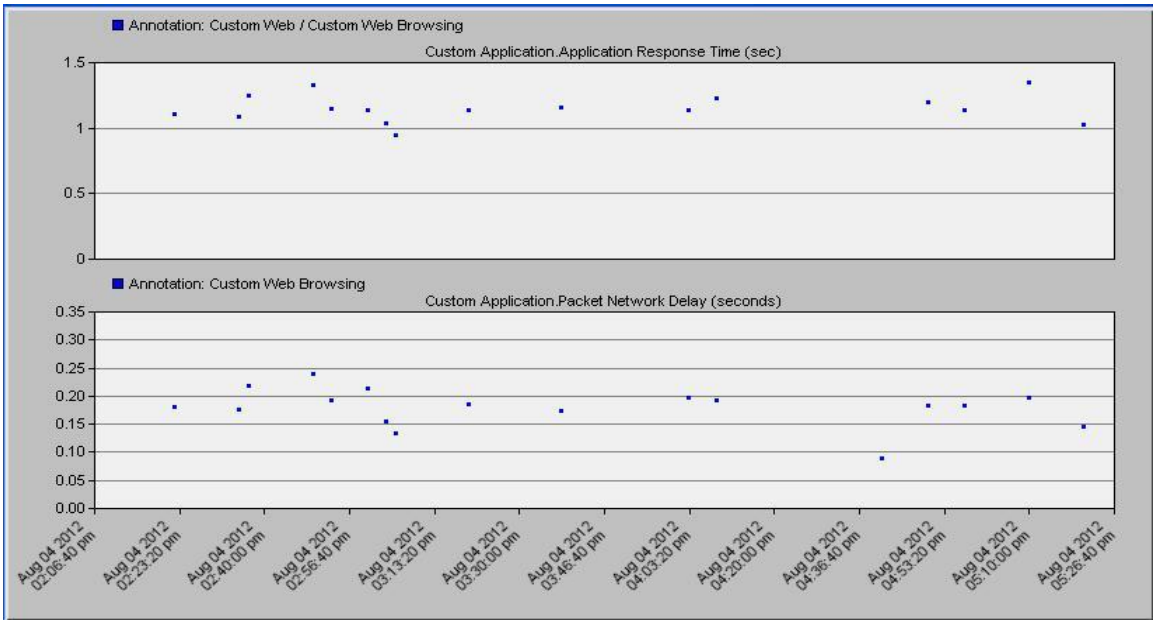

Fig.10 Application load (request/sec) and (sessions/sec)

Figure 10 show the application response time and it provide that the response time for our application is in normal ranges and very promising results.

6. Conclusion

Web search engine opens the door to explore a huge amount of information. There is a variety of search engines which offer diversified services to its users. This paper draws a clear picture of the designing and implementation of a medium size search engine (MSSE) and disproves the notion that all web search engines have same search capability, coverage, ranking and indexing techniques. Web search engines differ from each other in multiple aspects such as the searching strategy, coverage of the web, relevance of the search results with respect to the search query, ranking of the search results etc. and this paper also provide a performance study of MSSE and regarding many parameter in both real testing parameters and simulation results.

References

[1]A. Mo_at, W. Webber, J. Zobel, and R. A. Baeza-Yates. A pipelined architecture for distributed text query evaluation. Inf. Retr., 10(3):205{231, 2007.

[2]A. Tomasic and H. Garcia-Molina. Query processing and inverted indices in shared: nothing text document information retrieval systems. The VLDB Journal, 2(3):243{276, 1993.

[3]Amanda Spink, Bernard J. Jansen, Chris Blakely, and Sherry Koshman. A study of results overlap and uniqueness among major Web search engines. Information Processing and Management 42 (2006) 1379–1391

[4]Bar-Ilan, J. (2002). Methods for assessing search engine performance over time. Journal of the American Society for Information Science and Technology, 53(4), 308–319.

[5]Bernard J. Jansen and Paulo R. Molina. The effectiveness of Web search engines for retrieving relevant ecommerce links. Information Processing and Management 42 (2006) 1075–1098

[6]Jean Véronis. A comparative study of six search engines. Université de Provence, Version 1.0 (en) –22 février 2006:

http://www.up.univ-mrs.fr/veronis, *http://aixtal.blogspot.com*

[7]Jin Zhang and Wei Fei, Search engines' responses to several search feature selections. The International Information & Library Review (2010) 42, 212e225

[8]Liwen Vaughan, New measurements for search engine evaluation proposed and tested. Information Processing and Management 40 (2004) 677–691

[9]Search Engine Watch *http://www.searchenginewatch.com/*

[10]S. Spencer, The overlap between Google and Yahoo! Results is less than you might think, Natural Search Blog, 29 August 2004, Available from:

<http://www.naturalsearchblog.com/archives/2004/08/29>.

[11]Wu, G., & Li, J. (1999). Comparing Web search engine performance in searching consumer health information: Evaluation and recommendations. Bulletin of the Medical Library Association, 87 (4), 456-461.

[12]Sullivan, D. (2003, February 23). Nielsen/NetRatings search engine ratings [website]. SearchEngineWatch.com. Retrieved, 21 October,2002. Available from *http://www.searchenginewatch.com/reports/netratings.html*.