CHAPTER I

INTRODUCTION

There are many important inventions that have changed the world for human being such as fire, but no other technology has made such a revolution like the computer in modern times (Naughton, 2000). After the invention of the computer, life started to change and this different life shaped many times with the improvements of computerized technology (Sanderson & Forcht, 1996). Today, factories are producing by computerized technology and they can provide more qualified products in a shorter time period. Many business environments are able to work faster and more secure. Many business environments are able to work faster and more secure because computer is the machine which comes to life and brings you another life (Naughton, 2000). With the electronic commerce (e-commerce) and electronic business (e-business), business world exceed the boundaries and became ubiquitous (Laudon & Traver, 2004).

With the creation of internet in the 1960s, the world started to change and with the improvements in the internet world, our world shaped many times because internet revolutionized computer and communications world like nothing before (Leiner, 1997). We can define the internet as a giant computer network protocol which connects the computer with a universal network. Internet started by J.C.R. Licklider as a defense aimed project Defense Advanced Research Project Agency (DARPA) in 1962. In that time the internet was known as the Advanced Research Projects Agency Network (ARPANET) and firstly, four main computers in USA connected as online at 1969. Those computers were located in Los Angeles and Santa Barbara Campuses of California University, University of Utah and Stanford Research Institute. On ARPANET, internet does not have any data and was working as a computer network. Data is stored on the computers which are connected to internet and internet is just a connection between computers (Bryant, 2000).

Until the beginning of 1990, it was forbidden to use internet for commercial purposes. Internet was serving only for education, research and governmental use. In the middle of 1995, the internet started to be used for commercial purposes as well. The internet service started with Delphi, and then it continue with American On-line (AOL), Prodigy and CompuServe. Then internet access opened for universities and even for nursery and primary schools. At the beginning, internet was used in companies just for file sharing but today

internet is the largest network of the world which connects more than 500 million computers. The position that internet has come to from 1962 to 2009 is unbelievable. Today, by online education systems, a student can graduate from a university which is in another country without going there. Moreover, a doctor can carry out surgery from miles away by using medical machines which are connected to the internet (Internet Society, 2000).

According to logic and history, we can classify electronic mail (e-mail), file transfer protocol (FTP) and remote login as three main services of internet. Moreover, we can describe e-mail as a starting of information society. E-mail brought a new model for communication, interaction and working together of people. FTP is one of the most used internet service and it takes its power from remote login. These two applications are the beginning of remote search (Palme, 1995).

The first example of indexing internet contents is Archie, short for archives (Frank,1996). The first search engine (SE) Archie is created in 1990 by Alan Emtage who was a student of Mc Gill University and improved by Alan Emtage, Bill Heelan and Peter Deutsch. Archie name comes from the achieve word in English. This search engine was finding the files which are provided by anonymous FTPs. Users was able to find which computer includes the needed file and download it to their own computer by FTP protocol (Tennant, Ober & Lipow, 1996). To make search on Archie, users was connecting to the Archie server by telnet or sending e-mail to Archie servers (Deutsch, 1992). If the user knows the filename that he or she is looking for, the Archie could be useful for them but sometimes filenames was not including enough information about the file contents. If the filename is something like readme.txt which can be placed in many computers, the search process could be a really and long process for a user. In many UNIX web sites, it is possible to call this software by writing “Archie”. The database is still presents on some web sites which called as “Archie Server” (Bitirim, Tonta & Sever, 2002).

After a while, the computer center of Minnesota University developed Gopher in 1991, according to campus-wide information system and it was a menu based system (Lindner, 1994). Gopher is formed by related graphic and text typed information resources and it has a menu and its own protocols (Alberti et al., 1992). Because of Gopher database was expanding fast, many indexing problems appeared and this problem eased by development of Very Easy Rodent-Oriented Net-wide Index to Computerized Archives (Veronica) (McCahill & Erickson, 1994; Anklesaria et al., 1993). Veronica was created by University of Nevada and it

is a kind of database which includes the keywords of thousand of Gopher databases. Users can search the keywords of Gopher menu's by entering a query to Veronica databases. The main objective is Veronica is finding which keyword exists on which Gopher menu (Tennant, Ober & Lipow, 1996). After a while, Jonzy's Universal Gopher Hierarchy Excavation And Display (Jughead) is created and it was working according to FTP. At June 1993, Matthew Gray from Massachusetts Technology Institute created the first web bot which called as "Wandex". Wandex was creating an index (Leiner et al., 1997). After that at November 1993 Aliweb SE is created without a web bot and it was the first search engine which was including the web site's data's. Aliweb is a kind of framework for automatic collection and processing of internet resources indices in the web. At November 1994, JumpStation is created. It was possible to use JumpStation as a web form and it was created as an interface to query program. The aim of JumpStation was finding web sites and creating an index of it (Koster, 1994).

The first full text browser was the WebCrawler which was created at 1994. WebCrawler is a computer program that browses the Web in a systematic and automated way. Other names of WebCrawlers are ants, indexer, bots, worms or Web Spider (Kobayashi & Takeda, 2000). Instead of previous SEs, it was enabling the user to search the every single word of the web site (Hu et al., 2001). SEs is not adequate to index more than 16% of the Web (Lawrence & Giles, 1999). All popular Web SEs uses powerfull WebCrawlers that traverse the Web continuously, trying to discover and retrieve as many Web pages as possible (Dikaiakos, Stassopoulou & Papageorgiou, 2005).

1.1 The Problem

After the creation of internet, World Wide Web (WWW or Web) invented by Tim Berners – Lee and it has rapidly gained popularity and became the second most widely used application of internet family after e-mail application which is the most used application of internet (Chu & Rosenthal, 1996; Byrant, 2000). The improvement of the Web is unequaled phenomenon. In 1990, this is after four years of Web's birth, millions of people was using Mosaic which is the first well known Web browser (Abbate, 1999). The growth of the Web was a result of highly increase of Web servers because with the improvement of Web servers, value and number of Web pages which are accessible by these servers are increased too (Can & Nuray, 2006). In 1999 the number of Web servers was approximately 3 million and estimated number

of Web pages was around 800 million (Lawrence & Giles, 1999). Just after three years, in 2002, the search engine AlltheWeb (www.alltheweb.com) announced that the number of Web pages on internet increased to approximately 2.1 billion. It means that number of Web pages increased 1.3 billion in 3 years. According to this ration we could be able to calculate and say that, today number of Web pages on internet is approximately 12 billion with 13% growth ratio but the current situation of Web and internet users, there is nearly impossible to estimate a number for the Web pages on the internet. Today, a primary school student knows at least how to open a blog on the internet. It is possible that internet users are creating millions of Web pages in one day. This situation makes estimation of total Web pages nearly impossible.

Friendly and easy interface and hypermedia features of Web have attracting all internet users and information providers to upload more and more data in every single day on internet. Today, internet became into a huge information reservoir and finding the needful data on internet is extremely difficult. The number of printed documents increases to double in every 14 years but the information on internet increases to triple every year. One of the biggest information stores of the world, American Congress Library has nearly 170 million documents. On internet, there is couple of billions of document which is open for public use (Bitirim, Tonta & Sever, 2002). In here, the importance of web search engines can be seen easily because there is nearly impossible to access to the needful document of information on the internet without any search engine (Broadbent, 1998). This situation can be explained with a very famous Turkish idiom. We can say that finding needful information on the Web without SE likes looking for a needle in the haymow.

The information which is given above is a kind of proof of the importance of SEs to access information on the Web. Because it is the most important part of finding information on the Web, researchers always trying to develop stronger SEs (Jansen, 1996; Adalı, Bufi & Temtanapat, 1997). If we check the statistics about search engines for last four years, we can see that Google and Yahoo! are leading the top search engines list since 2006. Since 2006, Google is the top and most used SE. Yahoo! follows Google at the second place. Between 2006 and 2008, Msn/Live was the third most used SE but in 2009, Msn/Live gave the place to their new and successful search engine Bing. These ranks determined according to the preferences of users. At the end of 2009, Google is most used SE, Yahoo! is the second one and Bing is the third one. Bing followed by Ask and AOLSearch is the fifth with (http://www.seoconsultants.com/search-engines; Hitwise Press Releases).

The case in here is not which SE has more users. A SE may retrieve 250 results for a query but there are only 20 of these results are relative. On the other hand, another SE may retrieve 100 results with 75 relevant. In this situation, which SE is better? The one which retrieve more results or the one which retrieve more relevant results? The case for SEs is not retrieving too many results. A successful SE should eliminate irrelevant results and dead links to provide relevant result list to the user. The problem in here is except which SE is most preferred one, which SE can provide more relevant results to the user (Hu et al., 2001).

1.2 The Purpose of the Study

Purpose of this study is evaluating performances of five popular SEs according to user's view. All these SEs may have a perfect architecture but most important thing in here is what they provide to the users. Architecture and technology of each SE is different and they all provide different service and different results to user. The main purpose of this thesis is to find out which SE gives the best performance to user, more specifically the SE usages of Near East University (NEU) students. The study attempts to find answers to the following questions:

1. What are the SE usage frequencies of students?
2. Which SE is the most preferred one in NEU?
3. What are the differences between SEs?
4. What are the students' criterions for SE prefers?
5. What are the students' opinions about SEs?
6. What is other prefers of students' request for SE?
7. When users leave search?
8. Which SE has the highest precision ratio for favorite search queries?
9. Which SE has the lowest currency ratio for favorite search queries?
10. Which SE has the highest precision ratio for IT and IS queries?
11. Which SE has the lowest currency ratio for IT and IS queries?
12. Which SE has the highest precision ratio for general performance test?
13. Which SE has the lowest currency ratio for general performance test?

1.3 Significance of the Study

We are living in the information world and people use internet as one of the main information reservoir but the most important issue is finding needful information in a shortest and most reliable way in this huge reservoir. In this point, SEs came into existence. Today, ultra-developed SE technology manages internet usage of people. When users want to get any data from internet, they are using SEs directly without using any other resource. Many users uses SE web pages as their homepage and even they know the address of any other web site that they want to enter, instead of writing the address to the address bar, they are entering name of the web sites to SE as keyword and searching from there (Kehoe & Pitkow, 1996; Sullivan, 2003). This kind of behaviors of users inspired SE engineers to develop more and more effective SEs and on the other hand, researches kept evaluating SEs. Since invention of SE, researchers highly interested with this technology and popular SE evaluated in different years. Those evaluations gave different results according to evaluation years, evaluation criteria and evaluated SE. Difference of this study from previous ones is it includes Bing as SE which started to the service at June 2009.

1.4 Limitations of the Study

- This thesis covered 10 months between September 2009 and June 2010.
- The study evaluates performance of 5 SE which are Google, Yahoo, Bing, Ask and AOLSearch.
- SE performance test performed between 28th January and 3rd February 2010.
- Study is limited with IT / IS Terms and favorite terms of 2009.
- Evaluated SEs and favorite terms of 2009 are according to information from www.hitwise.com.
- Results of performance test evaluated according to precision and currency.
- Questionnaires applied during January 2010.
- Research area is limited with NEU.
- Questionnaire applied on 300 students.
- Questionnaire results evaluated by SPSS.

1.5 Structure of the Study

First chapter of the thesis covers introduction part of the research, short information given about history and development of internet, Web and search engines beside short history of search engines. Also limitations and significance of study, the problem and purpose of the study explained. Second chapter of the thesis includes literature review. In this part, other research aims and results explained with details. Third chapter covers conceptual overview and Information Retrieval (IR) systems explained with details including components because SEs is sub-subject of IR systems. Also chapter includes architecture, indexing and result display specifications of search engines. Main search engine evaluation measurements explained in this part as well. Chapter four is about research methodology and application of research is explained including materials. Chapter five is results and discussions which explains research results. Also discussions about results took place in this part. The last chapter is includes conclusions and recommendations of this study.

1.6 Summary

During last years, the Web search engines turn into a highly commercial business area. Today many people earns important amount of money because of Search Engine Optimization (SEO). Beside SEO works, today advertisement which takes place in search engines can bring billion dollars in a year easily. Because of these commercial advantages, search engine business started to improve itself in every single day and became into a kind of trend. This research will focus on search engine evaluation to conclude and present the best search engine even it is hard to decide to the best one because of fast and extremely high changes in the web world.

CHAPTER II

REVIEW OF LITERATURE

Web search engines did not come into existence until 1994 (Chu & Rosenthal, 1996). Even the literature about search engines has a short time span; the number of researches about search engine evaluation and information retrieval systems are high. The researches started to evaluate Web search engines in order to describe them. At the beginning, no researcher paid attention to Web search engine technology as much as today but in the last years, search engines have turned into a kind of sector which earns a vast amount of money.

Stevenage and Babb (1976) studies about modern IR systems which explain the architecture and processes of their invention with details. Blair (1990) categorized IR rules in 12 different categories according to their types and processes. Taylor (1992) made another study about same subject but this time, a well developed IR system which works according to the central computer explained with details. In another study, Tonta (1995) examined IR systems in details by taking Blair's publication as reference.

Notess (1995) examined Lycos, WebCrawler, World-Wide Web Worm, Harvest Broker, CUI, and CUSI and InfoSeek. Notess recommended that "for single keyword searches of a large database; use Lycos". Also he defined that multiword searches with an AND, try WebCrawler and for a time-consuming comprehensive search, use CUSI. Notess also compared InfoSeek with Lycos and WebCrawler according to coverage, precision and currency.

In another research, Courtois, Baer and Stark (1995) evaluated 10 different search engines including CUI, Harvest, Lycos, Open Text, World-Wide Web Form and Yahoo. According to their research, Open Text was the best search engine "with its flexible, powerful search interface and quick response". Also they pointed that WebCrawler was offering the easiest interface. Chu and Rosenthal (1995) made a study about comparing and evaluation methodology. They evaluated Alta Vista, Excite and Lycos search capabilities according to precision and response times. As a result of this study, they discovered that all those SEs needs different methodologies to be evaluated according to their methodology.

In one of the researches, Scoville (1996) evaluated a wide range of search engines and concluded that Excite, InfoSeek, and Lycos should take place in the best search engines list because they can retrieve "accurate results from easy-to-use interfaces". Kimmel (1996)

evaluated World-Wide Web Form, Lycos, WebCrawler, Open Text, Jumpstation II, AliWeb, and Harvest according to documentation provided by the search engines with a single word testes (e.g., elections, Hilary). The author summarized that “of the robot-generated databases presented here, Lycos appears to be the strongest system overall”.

In a study, Gordon and Pathak (1999) evaluated the performance of 10 search engines using 33 information-needs. For measuring performance it calculates recall and precision at various document cut-off values (DCVs) and uses them for statistical comparisons. According to the result of study, “absolute retrieval effectiveness is low and there are statistical differences in the retrieval effectiveness of search engines”. Also study recommended seven features to maximize the accuracy and informative content of similar studies.

Brin and Page (2000) examined the anatomy of a large-scale hyper textual web SE. in the study, they presented Google, a prototypr of a large-sclae SE which makes heavy use of the structure presents in hypertext and they scale up 1994 – 2000. At the end of study, they remarked Google as an important research tool because it provides high quality research. In 2001, Aldred invented a more effective IR system and published it with United States (US) patent.

Hawking (2002) evaluated 20 search engines using Text Retrieval Conference (TREC) inspired methods and those 20 search engines tested with 54 queries which were taken from real Web search logs. The performance measures used was including precision at various DCVs and recall has not been used. This study proposes some more features in addition to the seven items which were specified in Gordon and Pathak (1999) study.

In a different kind of work, Mowshowitz and Kawaguchi (2002) measured performance of search engines using the overlap of URLs of the matching pages. Researchers used the similarity between the response vector of the collection of search engines and the response vector of particular search engine, which defined as bias, to evaluate the performance of that search engine. “The study defines the response vector of a particular search engine as the vector of URLs making up the result sets returned by that search engine and the response vector of the collection of search engines as the union of the response vectors for each of search engines. In order to calculate bias, norm vectors for each response vector for are determined by using the number of occurrences of each URL”. The study concluded that, search engines retrieves only URLs and number of occurrences in each URL, but do not

consider the content of these URLs. But according to many researchers, the content of the URLs is very important for the performance evaluation of search engines.

Chowdhury and Soboroff (2002) presented a method for search engine performance evaluation which is automatically based on how they rank the known item search result. In this performance evaluation method, initial query-document results are constructed randomly and for each search engine, reciprocal rank is computed over all query-document results. If results are reasonable and unbiased, then the method can be useful but these query-document results need a given directory and it cannot be possible every time. At the end of 2002, Schwartz invented a more efficient IR system which works according to the probabilistic approach.

Griesbaum (2003) evaluated three German SEs which are altavista.de, google.de and lycos.de according to their top 20 results. The test panelist were based on a collection of fifty randomly selected queries. According to the findings, Google reached to the best result values. Lycos also attained better values than Altavista.

In another research, Can, Nuray and Sevdik (2003) presented an automatic method for the search engine performance evaluation. They measured performance of search engines after examining various numbers of top pages returned by the search engines and check the consistency between human and automatic evaluations using these observations. In the experiments the researchers used 25 queries and look at their performance in eight different search engines based on binary relevance judgments of users. In the research, they concluded that their experiments shows a high level of statically significant consistency between the automatic and human-based assessments both in terms of effectiveness and also in terms of selecting the best and worst performing search engines. Sever and Tonta (2003) examined SEs with details including IR systems and main components of IR systems and SEs.

Mowshowitz and Kawaguchi (2004), examined a real-time measures of bias in web SEs. Differences between bias and classical retrieval measures are highlighted by examining the possibilities for bias in four extreme cases of recall and precisions. As a conclusion, they recorded that SEs need to develop their bias profiles.

Jansen and Spink (2005) made a research about how we are searching the Web and reported results from research that examines characteristics and changes in Web searching from nine studies of five Web search engines based in the US and Europe. They compared interactions

occurring between users and Web search engines from the perspectives of session length, query length, query complexity, and content viewed among the search engines. As a result, they concluded that users are viewing fewer result pages, searchers on US-based Web search engines use more query operators than searchers on Europe-based search engines. Also they pointed that there are statistically significant differences in the use of Boolean operators and result pages viewed and one cannot necessary apply results from studies of one particular Web search engine to another.

In 2005, Tonta made another study about IR systems and explained the subject including components. Pederson and Fain (2006) made a study about brief history of sponsored search. They categorized their subject descriptors according to information storage and retrieval and history of computing.

Jansen and Molina (2005) underlined the effectiveness of web SEs for retrieving relevant e-commerce links. The study examined the effectiveness of five different SEs in response to e-commerce queries by comparing the engines quality of e-commerce link using topical relevancy ratings. The findings showed that links retrieved used an e-commerce SE are significantly better than those obtained from most other engines types but do not significantly differ from link obtained from a web directory service.

Carterette and Jones (2007) proposed a model that leverages the million of clicks received by web search engines to predict document relevance. This model allows the comparison of ranking functions when clicks are available but complete relevance judgments are not. In the publication, Carterette and Jones have shown how to compare ranking functions using expected discounted cumulative gain. With just a few relevance judgments, they significantly increase their success at predicting whether a difference exists.

Sheperd (2007) described key features of next-generation information SE that will enable more powerful and rewarding searches to be made than is possible with current search technology. Proposal of the author for a new kind of SE which provides deep search prototypes a new idea combining logical linking, semantic analysis and clustering to overcome these problems and make possible a more powerful information search capabilities.

Hochstotter and Lewandowski (2009) investigated the composition of SE result pages. They defined what elements the most popular we SE use on their result pages and to which degree they used for popular versus rare queries. Findings include that SEs use quite different

approaches to result pages composition and therefore, the user gets to see quite different results sets depending on the SE and search query used. Also they found that all SE show Wikipedia results quite often, while other hosts shown depend on the SE used. Both Google and Yahoo prefer results from their own offerings such as YouTube or Yahoo Answers.

Croft, Metzler and Strohman (2010) declared a book which focus on biased towards the search rather than the engine as, in most places, discussion on effectiveness dominate those on efficiency by great margin. Ganzha, Paprzycki and Stadnik (2010) combined information from multiple SEs as preliminary comparison by using game theory, auction and consensus approaches. According to their results, auction method highly dependent on each individual result set and does not represent well the combined view of all SEs. Consensus method returned the result which represents the common view of participating SEs and game theory method seems to act in a way that positions it in between the two other approaches. It favors winners or if URL is at top places of more than one result set, it is incorporated into the final result set.

2.1 Summary

Since the invention of IR systems, researches interested with that subject and after outcome of web SEs, focal point of researchers moved to this area. With improved SE technology, even a detail about SE's like result page or crawler can be a topic alone. This research is another example for this kind of research and I do believe that since development of internet continue like this, SEs will be more and more important in every single day and researchers will not stop to work on this subject.

CHAPTER III

CONCEPTUAL FRAMEWORK

3.1 Information Retrieval Systems

IR is the science of searching documents and it is concerned with the representation, storage, and accessing of information items. Also IR can be known as data retrieval, document retrieval, and text retrieval but each of them has their own body of literature. IR is an interdisciplinary and it is based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and physics. Automated IR systems are used to reduce the information overload. Web search engines are most used and visible IR applications (Salton & McGill, 1983). The main purpose of the IR systems is accessing all relevant documents on databases and WWW while severing irrelevant documents (Tonta, Bitirim & Sever, 2002). An IR system performs information retrieval by using probabilities. When performing information retrieval, the IR system uses both the prior probability that a document is relevant independent of the query as well as the probability that the query was generated by a particular document given that the particular document is relevant (Salton, 1983).

An IR system needs two conditions to access to the needed information from the database. First of all, terms should be suitable with the indexed documents or objects. Second one is; the keywords which are entered by the user to the IR system must be match with the indexed objects and documents (Lawrence & Giles, 1999). The search and IR operations can be performed only with the matching of query and indexed objects and documents. During this operation, IR system follows a rule which is called as Retrieval Rule. We can define the Retrieval Rule like this; for every query, retrieve information from the indexed objects/documents and their sub-indexes. By this manner, we can define the main components of IR systems (Townler, 1976). These components are;

1- Indexed documents or their surrogates.
2- An interface for the users.
3- A Retrieval Rule to compare the queries and indexed documents or objects for the IR.

Another important point of IR system is a user group is needed to perform searches on the IR system (Maron, 1984).

Figure 3.1: Traditional IR system (Bitirim, Tonta & Sever, 2002)

Figure 3.1 shows the architecture of a traditional IR system and as we can see on the Figure 3.1, we can define retrieval process with three front-end and three back-end concepts which are formed and create the IR system. In this figure, concepts represented by rectangles and processes are showed by dashed ovals. Front-end part of the figure shows the external world part of the IR system. Back-end part of the system is transparent to the user and it is used for the communication between retrieval processes. Information need, text objects and retrieved objects are the front-end parts of the system. Back-end parts are queries, indexed objects/documents and terms (Bitirim, Tonta & Sever, 2002).

Information need can be state as a plain text or in can state with terms by using “and”, “or”, “not”, “if”, etc. text objects forms an entrance to automatic indexing process and results are shown as subjective in inverted file arrangement. In here, presentation of objects with terms shows diversity. A document can be shown in different ways and truly. It doesn’t matter if indexing is done as automatically or manually. At the end of search process, retrieved objects are listed according to relevance of information need. In another words, retrieved objects are arranged documents list which forms the retrieve function (Bitirim, Tonta & Sever, 2002).

Back-end concepts include documents, terms and queries. IR systems have more than one model but the only important thing is matching of terms and queries to perform IR. In Figure 1, clustering process is excessively loaded. Clustering process operates documents, queries and terms one by one and recursive. Clustering processes named same but their operations can be different. The aim of clustering documents is increasing the speed of the IR. Clustering terms creates flexible queries and it saves domain space. Term clustering follows Latent Semantic Analysis technique (Deerwester et al., 1990; Foltz, 1996). In the time perspective, evaluation/feedback is an expensive process. Clustering queries decreases the need of evaluation/feedback process (Deogun, 1998). Steepest descent algorithm method successfully applied on information filtering (Mettrop & Nieuwenhuysen, 2001). Another aim of the query clustering is increasing the performance of IR systems (Lee, 1995; Belkin, Stein & Thiel, 1995). In search engine technology, clustering is an important part of user interface. Search engines present relevant results to the user as a group, not one by one (Leuski, 2001).

As seen on Figure 3.1, IR systems have evaluation/feedback option. The user may need to give feedback to retrieve a better IR result. Recall and precision options defines the quality of IR systems. If you increase the value of these two options, the quality of IR system will increase as well. The aim of evaluation/feedback option is decreasing the error level to the minimum to satisfy the IR system users (Srinivassan, 1992).

3.1.1 Database

Database is a shared collection of logically related data (and a description of this data); designed to meet the information needs of an organization (Connoly & Begg, 1998). A database is a main condition for an IR system. Documents takes place in a database and IR system needs a database to retrieve documents. In here “document” word is used to represent various things as books, videos, 3D materials, electronic files, pictures and etc. In databases, terms of documents or full texts can be stored (Tonta, 1995).

3.1.2 Terms

Terms are used to represent a document or information need. Terms are also called as keyword, metadata or index term. The process of choosing term for representation of document is indexing. Terms represents the important part of documents. The important point

is; while deciding the terms we have to choose the word which are close to users (Srinivassan, 1992).

Because all documents are full with information, it is important to decide how terms will represent the document because this document should be retrieved when it is needed. Vocabulary difference creates an important paradox about information retrieval. Vocabulary of the document author and user is different. So sometimes users know what they need to find but they don't know how to define it in the search query. In this situation, the user can not retrieve the document even it is exist in the database (Blair & Maron, 1985). Also sometimes searches may go wrong according to keyword or concept. When user enters groom as a keyword, the IR system will retrieve horses or weddings. Concept searches tries to work out of meaning of the text rather than just using specific words like heart in medical and heart in love (Cooper, 1995).

Hans Peter Luhn is knows as a modern inventor of indexing with keywords. At the end of the 1950s, Luhn indexed the words of an article as in entry. This system is called as Key-Word-In-Context (KWIC) and this system is still in use for preparing bibliographic index (Svenonius, 2000).

3.1.3 Documents

In a typical IR system, documents are represented by terms. A traditional document indexing is forming as follows (Guinchat & Menou, 1983);

1- Non-letter characters are replacing with spaces
2- Single-lettered words are cancelling
3- All capital letters are changing into small caps
4- Keywords in first stop list are deleting [1]
5- Do stemming
6- Single-character stems are deleting

[1] Keywords in stop list has no importance for IR. This kind of words can be created as indepent from collection (or databases) or they can selected from terms of index which have high fequency (Bitirim, Tonta, Sever, 2002).

Term is automatically created stemmed words. We can use terms to represent queries and documents and we call them as document terms and query terms. If a term is exist in document weight of the term is 1 (relevant), otherwise it is 0 (irrelevant). We called this approach as Boolean. Another popular approach is *tf*idf* values which is used in vector based approaches. In here *tf* is term frequency which means the repetition of the term in relevant document and *df* is document frequency which is the repetition of document as relevant. In vector modeling, we consider documents and queries as vectors. If we consider that there are *"t"* unit terms in a document and "*i"* is document (Tonta, 1995). According to equation 1;

$$D_i = (ai1,ai2,\ldots ait) \quad (1)$$

And, if *j* is query,

$$Q_j=(qj1,qj2,\ldots,qjt) \quad (2)$$

Where:

D=Document

Q=Query

If any term's term frequency is high but frequency in the other documents is low, it relative weight of that term should be low. To provide this condition, we use inverse document frequency (*idf*). A typical *idf* parameter is equal to $log(N/df_j)$. N is total document number in the index and df_j. *j* is the frequency of the terms in document. If we want to calculate the frequency of t_j term for D_i document as w_{ij}, the equation is;

$$w_{ij}=tf_{ij}*log(N/df_j) \quad (3)$$

Where:

N= Total document number of index

tf= Term frequency in document

df= Document frequency

w= Frequency of t_j term for D_i document

In *tf*idf* method, relative weights are very important. There are many researches which discusses term weights with *tf*idf* and other methods (Salton & Buckley, 1988).

Like terms, documents can be divided into clusters. The aim of document clustering is fixing recall value to diminish the document searching space. Clustering document starts at the lowest level by comparing documents and clustering matching ones. This operation continues to the top level and at the end of the operation there is only one cluster. Query starts from the top level and goes down till find the best matched cluster. In literature, this operation is called as hierarchical clustering (Van Rijsbergen, 1979).

3.1.4 Queries

Query is the explicating of the information need of user in a formal way. User may express the information need in a various ways. Search terms or keywords are connected with Boolean operation. Boolean operators are “and”, “or”, and “and not”. If “and” operator is used, it means that retrieved objects will include all documents which are requested. “Or” operator means at least one of the requested documents must be in the requested objects. “Not and” operator means, retrieved object shouldn’t include that document (Salton, 1989; Van Rijsbergen, 1979).

The users may clarify their information need with a natural expression. In the naturally expressed queries, there is no condition which says retrieved objects should include all of the word in the query. In here, relevance of the retrieved objects is related with the correction of the entered query. So, the document which has the all keywords which is expressed in the query is the best document of the retrieved objects. Documents which exceed the threshold which is given by the user may take place in the retrieved documents. In another words, user may want to see other documents which are similar with their information need 80% or more (Bitirim, Tonta & Sever, 2002).

Probabilistic model weights search terms according to probability of their existence in document by using feedback and document terms has duo weights as 1 and 0 (Robertson & Jones, 1976; Crestani et al., 1998). In this model, at the beginning user enters the search words in the natural expression. If retrieved objects do not satisfy the user, the user may start the evaluation/feedback process to receive better results (Salton & Buckley, 1990).

In concept based models, user defines their information needs as rules (Alsaffar et al., 2000, 1999; McCune et al., 1985). Main and sub-concepts may connect to each other with “and”

and “or” operators. For example; the user may enter the query as ((<concept1> and <concept2>) or <concept3>). A sub-concept may define the main concept according to rules (Alsaffar et al., 2000). According to this approach, the weight of the search terms may defined by the user. The bridge between concept, vector and Boolean models can be formed by P-Norm words (Alsaffar et al., 2000; Salton, Fox & Wu, 1983).

3.1.5 User Interface

Every IR systems must have an interface to turn entered queries into a suitable form for IR system. In another words, the communication between system and user is performed by the interface. We can list the main functions of user interface as follows (Tonta, 1995);

1- Provide a possibility to users to enter queries by using natural language or query language.
2- Evaluate the query which is entered by the user.
3- Changing the query which is entered by the user into a suitable language for IR system and transferring the query to the system.
4- Showing the retrieved objects.
5- Receiving evaluation/feedback from the user about relevance of the retrieved objects.
6- Providing information about IR system, usage of the system and database.

There are various user interface models like menu or command based model, graphic based model and blank filling model to help users to enter query and retrieve information (Shneiderman, 1986). Also there are some IR systems which accept voice as query entrance but the most important thing is the users have to know how to use the IR system. In here; the model of IR system is not important if the user does not know how to use it (Tonta, 1995).
User interface is a tool for users to access to the reach information store of the IR systems. The main purpose of the user interface is, helping to the user to retrieve information from IR system without dealing with the complex architecture of the systems. Mooers rule is valid for all of the IR systems. If the retrieving information is harder and troubled than not having that information, users will stop to use IR systems (Mooers, 1960).

3.1.6 Retrieval Rules

The matching between document indexing and queries can be defined by only retrieval rules. Blair (1990) examined 12 different retrieval rules detailed. Those rules can be classified in 3 main groups.

1- Vector space retrieval rule which terms processes as vectors in n-dimensioned space.
2- Boolean rule which requires exact match between query and index terms.
3- Probabilistic rule which depends on weighting queries and index terms according to probability theory.

Table 3.1: Summary of retrieval rules (Blair, 1990)

Model	Search Need	Documents	Retrieval Rule
1	Single Query Term	Documents has one ore more than one index terms	If query term is matching with document term, you can retrieve the document.
2	Multiple Query Term	Index terms set	If all query terms are exists in document index record, you can retrieve the document.
3	Query Terms and Threshold Value	One or more than one index term set	If term which is over threshold matches with query term, you can retrieve the document.
4	Same as Model 3	Same as Model 3	Documents which matches more than term number, presented according to matching number.
5 **Weighted Queries**	Query terms set with positive values	Same as Model 3	Documents are listed according to their total weights on queries and indexing terms.
6 **Weighted Indexing**	Query terms set	Index terms set with positive values	Same as Model 5
7 **Weighted Queries and Indexing**	Same as Model 5	Same as Model 6	Documents are listed according to product of term's query weight and indexing weight.
8 **Cosines Rule**	Same as Model 5	Same as Model 6	Considers term weights on index and query as vectors. Value of retrieved document is cosine of angle between two vectors.
9 **Query sentences according to Boolean Approach**	Query words are formed by Boolean operators	One or more than one index term set	AND: Documents which matches with all terms of query must be retrieved. OR: Documents which matches at least one term of query must be retrieved. NOT: Documents which does not match with terms of query must be retrieved.
10 **Full Text Retrieval**	Same as Model 9	Search on full texts of documents is possible (excluding irrelevant words)	Same as Model 9. Also it is possible to use proximity operators.
11 **Simple Conceptual Index**	Single Terms	One or more than one index term set	An online index is checking and adding synonymous terms with terms in query.
12 **Weighted Conceptual Index**	Single Terms	One or more than one index term set	Terms which are over threshold is adding from an online index as disjunctively. A user may define the threshold value.

As explained before, IR systems have 3 main clusters. These are terms, queries and documents. Terms can be used to represent both document and queries because of this it is possible to see them as a point in vector space (Tonta, 1995).

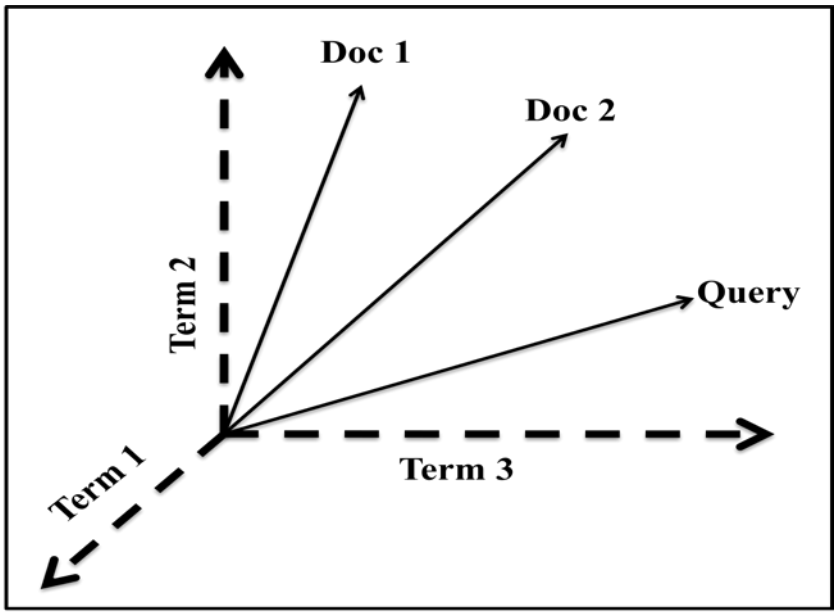

Figure 3.2: Vector space IR system mode (Tonta, 1995)

Figure 3.2 explains the vector space approach. In this approach, there are at least two distinct vectors; document vector and query vector. The vector product of these two vectors gives the degree of similarity of query and document. This coefficient is also called as Cosine coefficient because it is equal to cosine of angle between two points. Scalar product and inner product is other two names of this calculation. These coefficients are given below (Ingwersen, 1992):

$$\text{Inner product } (D_r,Q_s) = \Sigma^t d_{ri} * q_{si} \quad (4)$$

$$\text{Vector Product } (D_r,Q_s) = (\Sigma^t d_r i * q_{si}) / (\Sigma^t (d_{ri})^2 * \Sigma^t (q_{si})^2)^{1/2} \quad (5)$$

Where:

D= Document vector

Q= Query vector

d= Weight of i component on document

q= Weight of i component on query

In the formulas, D_r is document vector, Q_s is query vector and d_{ri} and q_{si} represents the weight of *i* component on document and query vectors.

In Boolean model, we can think a document or query as sub-cluster of term clusters. In this condition, matching degree between two clusters (document-query) forms the value of retrieve function. Jaccard coefficient gives intersection ratio between two clusters. On the other hand, Dice coefficient is related with average sizes of intersection ratios of D_r and Q_s clusters. The official definition of these coefficients is given below (Bitirim, Tonta, Sever, 2002);

$$\text{Jaccard Coefficient } (D_r, Q_s) = \left| (D_r \times Q_s) \right| / \left| (D_r + Q_s) \right| \quad (6)$$

$$\text{Dice Coefficient } (D_r, Q_s) = 2 * \left| (D_r \times Q_s) \right| / (\left| D_r \right| + \left| Q_s \right|) \quad (7)$$

Where:

D= Document cluster

Q= Query cluster

As explained before; probabilistic model weights search terms according to probability of their existence in document by using feedback and weight of documents is duo. Assume that distribution of terms in relevant and irrelevant documents is independent from each other 1[2]. Moreover, let's consider prior conditional for t_i document term variables (Bitirim, Tonta & Sever, 1995);

$$p_{ri} = (a_{ri} = 1: \text{ relevant } (Q_s)) \text{ and} \quad (8)$$

$$q_{ri} = (a_{ri} = 0: \text{irrelevant}(Q_s)) \quad (9)$$

Where:

Q=Query

t= Term

[2] In dual independent retrieval model (Robertson & Jones, 1976), independence of terms in relevant and irrelevant document hypothesis is always criticizes with a reason which supports that this hypothesis does not represent the truth. However, Cooper (1995) suggested that there is no need to this assumption in dual independent retrieval model and older version of this assumption; linked dependence is a better assumption for this kind of situations. We can define linked dependence like this; possibility ratio of existence of a document in relevant and irrelevant classes is equal to product with possibility ratio of existence of a query in relevant and irrelevant classes.

In here, *relevant* (Q_s) and *irrelevant* (Q_s) are functions which retrieves relevant and irrelevant documents for Q_s query. In this time, p_i gives the probability of t_i is equal to 1 if document is relevant and q_i gives the probability of t_i is equal to 0 if document is irrelevant. When the probability retrieve function which is given below is using, it is proved error probability of system is decreased to minimum (Robertson & Jones, 1976; Crestani et al., 1998).

Probability retrieval equation:

$$(D_r:Q_s): \Sigma\, t_i \log((p_i*(1-q_i))/(q_i*(1-p_i))) \quad (10)$$

Where:

D= Document

Q=Query

t= Term

The p_i and q_i values which are given above are estimated according to user evaluation for Q_s query but it is not practical to estimate prior probability values according to feedback (Yu & Lee, 1986).

3.2 Search Engines Technology

SE is IR system based web site that helps users to retrieve any information from huge internet database and it is a kind of tool that crawls in the web according to user direction and it will record everywhere it has been and everything user look for (Capra & Quinones, 2005). The SE software is a kind of IR program and it has two major task; Searching through the billions of terms recorded in the index to find matches to a search and ranking retrieved records in order to decide most relevant (Chowdhury, 1999) Usually, internet users prefer SEs to access required information from the internet because SEs are open for public use with billions of web sites and during last years, there are many important researches about this area. Bases of search engines are IR systems which are improving for 50 years but according to architecture and process specifications, search engines shows some differences form IR systems (Lavrence & Giles, 1999).

3.2.1 Architecture

One of the main components of SE is a robot which is called as Web Crawler (or Spider) and it works as a network surfer and it downloads a searched web site to local disk. Web crawler is a kind of computer program that browses the Web in a methodical, automated way. This process is called as Web Crawling or spidering. Search engines use spidering to provide up-to-date information. The most important aim of web crawler is copying all visited web pages for later searches to make next searches faster. Web crawlers can also used for automating maintenance task on a web site like checking links or validating code. Also web crawlers are used to collect specific information from Web pages (Batzios et al., 2007).

Web crawler starts with a list of Uniform Resource Locator (URL) to visit which is called as seeds. While visiting URLs, it identifies all the hyperlinks in the page and adds them to the list of visited URLs which is called as crawler frontier. URLs which are placed in the frontier are visited again according to some policies (Dikaiakos, Stassopoulou & Papageorgiou, 2005).

Even web crawlers are very easy programs, they finds million of documents and helps to IR systems to retrieve correct information in easy way. Also sometimes, crawler can find the information which is hiden by website owner or webmaster. Because of this, many web crawlers has to work according to robots exclusion protocol. Some search engines use more than one web crawler for different purposes but not all web crawlers are works to find information. Web crawlers also may work as link checker, page change monitor, validator, File Transfer Protocol (FTP) client or web browser (Dolowitz, Buckler & Sweeney, 2008).

Figure 3.3: High-level architecture of a standard web crawler
(Dolowitz, Buckler & Sweeney, 2008)

Figure 3.3 explains the architecture of a standard web crawler. The basic principle of web crawling method arranged on hypertext which are represented as URL which includes information about unique location referenced web resource. Downloader starts to work from root node and gets URL of processed web document from processing queue then downloads the document and parse document's content to extract set of URL links to other resources and update processing queue. At the end of the process, crawler stores web documents for future processing.

There are two types of SEs: first type is the search index which is a vast catalog made up of every word taken from all the web pages searched by crawler. Google is an example for this kind of SEs. Other type is the web directory is compiled by real people who organize web pages into categories and subcategories and they lets user to search very effectively. Yahoo is a kind of web directory and a good example for this kind of SEs. Most popular SEs is combination of these two principles (Cooper, Milner & Worsley, 2000).

3.2.2 Indexing

Indexing is the process of examining information items according to an algorithm to build a data structure that can be searched in a fast way (Hu et al.; 2001). In traditional IR systems, indexed documents are static and a document can be indexed only one time but internet resources changes very fast. Full life of a web link is approximately 44 days (Brake, 1997;

Kahle, 1997). Volume of the web increase in every single day and half-life of web pages on search engines is represented by days. This situation makes architecture more complex and search engines started to index fewer web pages day by day. Different search engines indexes different web resources so now it is hard to guess the matching ratios of documents. So every single day makes job of search engines harder (Lawrence & Giles, 1998; Bergman, 2001; Kobayashi & Takeda, 2000).

In traditional approach, the quality of documents is very high but in search engines, documents may include many mistakes. Wrong indexing is also possible. Another problem of search engine indexing is; sometimes a document can be indexed is search engine more than one time. According to researches, 30% of web pages in search engine indexes are repeated documents (Kabayashi & Takeda, 2000).

3.2.3 Representation of Documents

After the SE finds relevant documents in the search, it represents those documents in the result page according to some rules (Laursen, 1998). Search engines does not display all document on the result page like traditional IR systems (Kobayashi & Takeda, 2000; Laursen, 1998). Generally one or two sentences which include search query are listing with metadata or header of web page. Conception of search engine result pages should be address to the user's eye but also it has to include all possible document retrievals. This situation affects the efficiency and precision of search engines negatively (Olgun & Sever, 2000). First step about this subject made progress with HTML 3.2. The metadata area which is placed at the beginning of HTML code and limited as <head>…..</head> can not be displayed by search engines and directly related by web crawler (Küçük, Olgun & Sever, 2000).

One of the most important problems of using metadata is spam. At the beginning, metadata was a solution for indexing web pages but later, webmasters started to use metadata in a bad and they developed some spam techniques to make website to take place on upper rows in the result pages (Henshaw, 2001). Webmasters started to write most searched terms or keywords as metadata even it is not related with the concept of the web page. By this method, the web page will be displayed more frequently on upper rows in the search engine result pages. In this situation, the efficiency of search engines will decrease. On the other hand, search engine services develop some algorithms to stop spam (Notess, 2001). Even it is nearly impossible to stop spam 100%, with the hard works of search engine services, spam users can not be fully

successful. At the beginning, some search engines like Excite and Lycos, started to not using metadata. In today's technology, there is no need to metadata because in previous technology, search engines was retrieving information according to metadata but today, web crawler can read the whole document or web page without checking the metadata. But it does not mean that this technology can stop the spam because webmasters started to place the spam into the body of document or web page (Menczer, 2002).

3.2.4. Efficiency

There are various methods which are used to evaluate IR system's efficiency and efficiency of a SE measures the success of that SE. Precision, recall and wrong alarm are three of those methods and these to methods are the most prevalent ones (Tonta, 1995). Sometimes, online IR systems can not retrieve relevant documents. On the other hand, it is possible for online IR systems to retrieve irrelevant documents. We can summarize IR process as follows (Blair, 1990).

IR is a trial and error process. As user can access to relevant document, it is also possible to retrieve an irrelevant document as well. This situation causes a kind of indefiniteness and there is no problem which comes from this. On the other hand, it is possible that user will not be able to retrieve other relevant documents after research. As Blair stated in 1990, we can divide the documents in database into four different groups:

1- Retrieved and relevant.
2- Retrieved and irrelevant.
3- Un-retrieved and relevant.
4- Un-retrieved and irrelevant.

Table 4.1: Presentation of search results (Blair, 1990)

	Relevant (P)	**Irrelevant (¬ P)**	
Retrieved (R)	a	b	a+b
Un-retrieved (¬ R)	c	d	c+d
	a+c	b+d	a+b+c+d

In Table 4.1; *a* is retrieved relevant documents and *b* is retrieved irrelevant documents (with another name; false drops). c is unretreived relevant documents and d represent unretreived irrelevant documents. a+b+c+d is total document number in the index. So; *a+b* is represents the total of retrieved relevant and irrelevant documents. According to this, recall is the ratio of retrieved relevant documents (*a*) to total retrieved and un-retrieved relevant documents (*a+c*) (Van Rijsbergen, 1979). Precision is the ratio of retrieved relevant documents (*a*) to retrieved relevant and retrieved irrelevant documents (*a+b*) (Van Rijsbergen, 1979). Recall and precision values change between 0 and 1. If these values are high, it means that efficiency of the IR system is high as well (Salton, 1989). Wrong alarm is the ratio of retrieved irrelevant documents (b) to total retrieved and un-retrieved documents (b+d). This ratio measures how the IR system declines irrelevant documents (Blair, 1990). About precision and recall, there are four cases to examine.

1. High recall and low precision.
2. High precision and low recall.
3. Low recall and low precision.
4. High recall and high precision.

First case eventuates when most of the relevant items in the index have been retrieved but irrelevant ones included as well. Second case occurs when few of the relevant items are retrieved from the database, but even fewer irrelevant ones are retrieved in response to the given query. Third case occurs when both precision and recall is low; which means few relevant items have been retrieved from the index and many of retrieved documents are irrelevant. Fourth case eventuates, if nearly all relevant items in the database are retrieved and very few irrelevant ones are included. (Mowshowitz & Kawaguchi, 2004).

3.2.5 Ranking and Retrieval Function

Retrieval rules and functions which are explained in details in this chapter are also valid for search engines. When user enters information need with natural language, query engine creates a query from this information need or take the sentence as a query. Then, system matches the query and documents or web sites on the web and display results in a descending order according to frequency. Query engines may use more than one retrieval function to perform this operation. Traditional IR systems retrieve static documents but search engines are dealing with hyper-dynamic web resources. Also search engine gathers data about links

between websites and they can store the algorithm or architecture of web sites (Bitirim & Sever, 2003).

Because it is a commercial secret and it may cause spam, search engines do not prefer to explain their indexing techniques or retrieval functions but many search engines created in academic places so it is not that much hard to guess some search engine services' retrieval functions. For example; Alta Vista uses weighted Boolean search (Silverstain et al., 1999). Google considers hub and authoritative connections of web sites beside document statistics (Kleinberg, 1998; Kobayashi & Takeda, 2000).

Ranking is another important issue for web sites which takes places in search engines result pages. Usually search engines returns thousand or sometimes millions of results but users are not willing to view more than a few. Because of this, first five result page is very important (Jansen & Resnick, 2006; Jansen & Spink, 2006, Lorigo et al., 2005). If the click frequency of a document or web site is high on the web, this situation increases the rank of that document or web site on search engine result pages. Page rank is an indicator of a web site or document for their value on the web. We can point Google page rank technique as a good example. Google uses a link to make connection from Page A to Page B and Page B to Page A. Also it makes some content analysis to protect the rights of the web site or document. If the web site or document contains some important criteria that make them important, they will have a higher page rank than others for Google. So Google will always remember this high ranked websites in every related search (Cicone & Serra-Capizzano, 2010). Formal page rank equation of Google is explained as follows:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn)) \quad (11)$$

Where:

PR(A)= Page rank value of a website

d= Damped down factor (0.85)

In this equation, *PR(A)* is the page rank value of web site *A*. At the beginning, this value is equal to 1 for all web sites. *d* is a special coefficient which is called as damped down factor and it is always equal to 0.85. This is a fixed coefficient like *Π* number. *PR(Tn) = A* is the

page rank value of a web site which gives link to web site *A* and *C(Tn) = A* is equal to number of links which goes from the web site that gave link to web site *A*, to another web sites (Lin, Shi & Wei, 2008). Page rank value can be updated anytime and this apdates effects search results. In despite of this situation, official page rank are announced approximately in every 3 months and new values are announced (Cicoine & Serra-Capizzano, 2010).

3.3 Evaluated Search Engines

3.3.1 Google

At the beginning, Standfor University PhD students Larry Page and Sergey Brin were working on a project as a BackRub search engine a thesis. Their aim was separate the internet into the parts because internet was a huge data pile and it was very hard to find what you are looking for. They developed a new system for this at google.stanford.edu address and their new system was searching internet the web sites in a different style in comparison with classic search engines.

At the same time, they were looking for an investor to apply their project to real life and they arranged a meeting with the founder of Yahoo! David Filo. Filo advised them to improve their systems and start to look for investor later. After this, Page and Brin decided that they are not good enough to take the attention of big companies and on September 1998, they founded Google Company at one of their friend's car garage. At the same year PC Magazine showed www.google.com in the best 100 web sites and announced it as the best search engine.

Larry Page and Sergey Brin started to Google at 1998 in a car garage but today they have over 10,000 workers and some of best and most experienced technology experts of the world are prefers to work with Google. In 2001, Eric Schmidt joined to Google as manager and CEO. While they were creating the Google, they developed the search results according to copyrighted PageRank™ technique because of this, copyright of PageRank™ belongs to Stanford University, not to Google (Cicoine & Serra-Capizzano, 2010; www.google.com).

3.3.2 Yahoo!

Yahoo! is a main portal which is founded by Stanford University students Jerry Yang and David Filo in 1995. At the beginning, Yahoo! was giving service just as a search engine but then it started to be more popular with different services like e-mail and instant messaging.

www.yahoo.com is the most visited website of the world with 7 billion clicks. Yahoo! Messenger service is very popular especially in USA. Yahoo e-mail service is the first in the world with its unlimited storage. Today, Yahoo! Music and Yahoo! Movie are the biggest achieves of the world. According to the researches, Yahoo! Music is strong enough to stand against to the rest of music sector of the world.

Also people can search for job by using Yahoo! Hotjobs, learn what is happening in all over the world by Yahoo! News, provide much valuable information about stock exchange or bones from Yahoo! Finance, play online games with other users by Yahoo! Games. On the other hand, with Yahoo! LAUNCHcast, users can create their own radio channel by voting their favorite songs and they are able to listen it as online with a very high sound quality.

On 4th of February 2008, Microsoft offered 44,6 billion dollars for Yahoo! Inc but Yahoo! Refused this offer. (www.yahoo.com)

3.3.3 Bing

Bing is today's popular SE which is powered by Microsoft company. The SE published at the first half of 2009 as "Kumo" and it started with "Changing Habits" claim. Just after 3 months of its publication, it took the place of Windows Live search and today it is the most dangerous rival of Google and Yahoo. At 1st of June 2009, the SE published as Bing Beta and Bing started to give service in 58 different languages including Turkish. Another feature of Bing is its daily changing background picture. Today, MSN is one of the most used portals of the web (www.hitwise.com) and many people uses Bing directly from MSN web site without opening Bing web page. But the difference of MSN/Live and Bing is in comparison with MSN/Live, Bing has a very high SE technology (www.bing.com).

3.3.4 Ask

Ask or Ask Jeeves is powered by Garrett Gruener and David Warthen in 1996 at California but the original software belongs to Gary Chevsky. At the beginning, company had hundred editors. Those people were gathering websites from internet according to users demand to provide best information to Ask users because that was the ideology of the SE. User will ask something and they will provide information as user is asking to the guru of that subject. But that process was very hard and expensive. In these days, Ask has around 10 editors and they

are still using publishing editor advices but today, their main information resource is Tahoma as their web crawler (www.ask.com).

3.3.5 AOLSearch

AOL or American Online is American global Internet service and Media Company and founded in 1983 as Quantum Computer Services. AOL has franchised its services to companies in several nations around the world or set up international versions of its services. AolSearch is the SE service of AOL Inc. and founded around 1990s to give data search service for American Online users buy today this SE is used by people from all around the world. According to the big deal with Google Company in 2009, comprehensive web results of AolSearch are enhanced by Google. In 1990s, many SEs has been developed but they could not continue in the race with new generation SE (Preston, 2002). Instead of other 90s SEs, AOLSearch could achieve to improve its technology in a perfect way and even it is an old SE, it is still in use and it could take it's place in top 5 for many years including 2009 (www.hitwise.com; www.aolsearch.com).

3.4 Summary

IR systems invented to retrieve searched file from database to user in an easy and fast way. Later, improved computer technology needed more detailed and complex systems and then those systems turn into SEs. IR systems are still in use for databases. As an example; your computer's hard disk is a kind of database of you that you store your information in it. Search option in your start menu helps you to find your files or folders by using keywords. Another example; universities has a huge databases to keep records of their students. In here an IR system is a must to retrieve information about students. According to me, IR systems technology will continue to improve with developed computerized technology. At the beginning, SEs was just a service for internet users but today, it became into a very important and profitable sector. Especially SEO is a kind of business area that people are getting education to work in this subject. SEs has a very complex architecture and famous engines like Google or Bing hides their page ranks or architecture as a secret because competition between SEs is really big and important. Like IR systems, SEs has to develop their architecture without stopping because high file upload of users to internet made SEs a must to retrieve information from internet.

CHAPTER IV

METHODOLOGY

In this chapter, detailed information is given about methodology of this thesis. Aim and method of performance evaluation test, selection of SEs and test queries, aim and application of precision and currency test and methods are explained in this chapter. Also applied research on NEU, aim and data collection of research is clarified.

4.1 Research Model

The study investigated which SEs are the most widely used ones in among students, what are the criterions that direct students to use those SEs, do students use any other SE except engines which are selected for performance evaluation test, which SE gives the best performance to users, and which SE has the highest currency. The research was conducted in the frame of general survey model, questionnaires and performance evaluation test. A deep literature review has been done in order to create background of this study. To gather data from students, a questionnaire has been prepared and applied on random volunteers. On the other step of the thesis, queries have been prepared in two different groups. First group includes queries which are gathered from top search queries of 2009 and other group includes queries about information technology and information systems.

4.2 Students

This research was conducted at NEU in Turkish Republic of Northern Cyprus (TRNC) during 2009-2010 fall semester and 300 students from 15 different faculties of NEU took part in the study. Twenty different students were selected randomly from each faculty with 20-22 age average. Faculties that took part in the research were Faculty of Atatürk Education, Faculty of Maritime Studies, Faculty of Dentistry, Faculty of Pharmacy, Faculty of Arts and Sciences, Faculty of Fine Arts and Design, Faculty of Law, Faculty of Economics and Administrative Sciences, Faculty of Communication, Faculty of Architecture, Faculty of Engineering, Faculty of Health Sciences, Faculty of Medicine, Faculty of Performing Arts and Faculty of Tourism. Departments of students are Medicine, Computer Education and educational Teaching, Guidance and Psychological Counseling, Elementary Teaching, History Teaching, Deck, Maritime Business Administration and Governance, Dentistry, Pharmacy, Turkish Language and Literature, Psychology, Graphic Design, Law, Business Administration, Economics, International Relations, Computer Information Systems,

European Union Relations, Radio-Television-Cinema, Journalism, Public Relations and Advertising, Architecture, Interior Design, Computer Engineering, Mechanical Engineering, Nursing, Nutrition and Dietetics, Theater, Author and Tourism and there are students from undergraduate, masters and PhD degrees. Faculty and department details of the research are shown in Table 4.1.

Table 4.1: Faculty and department details of students

Faculty of Communication (20 Students)	# of Students
Department of Radio, Television & Cinema	14
Department of Public Relations & Advertising	5
Department of Journalism	1
Faculty of Economics and Administrative Sciences (20 Students)	
Department of Business Administration	3
Department of Economics	9
Department of Computer Information Systems	5
Department of International Relations	2
Department of European Union Relations	1
Faculty of Dentistry (20 Students)	
Department of Dentistry	20
Faculty of Medicine (20 Students)	
Department of Medicine	20
Faculty of Architecture (20 Students)	
Department of Architecture	5
Department of Interior Decoration	15
Faculty of Pharmacy (20 Students)	
Department of Pharmacognosy / Pharmaceutical Botany	20
Faculty of Law (20 Students)	
Department of Law	20
Faculty of Fine Arts and Design (20 Students)	
Department of Graphic Design	20

Faculty of Performing Arts (20 Students)	
Department of Theatre	19
Department of Author	1
Faculty of Health Sciences (20 Students)	
Department of Nursing	7
Department of Nutrition and Dietetics	13
Faculty of Tourism and Hospitality Management (20 Students)	
Department of Tourism and Hospitality Management	20
Faculty of Engineering (20 Students)	
Department of Computer Engineering	15
Department of Machine Engineering	5
Faculty of Arts and Sciences (20 Students)	
Department of Psychology	2
Department of Turkish Language & Literature	18
Faculty of Maritime Studies (20 Students)	
Maritime Business Administration and Governance	16
Department of Deck	4
Faculty of Atatürk Education (20 Students)	
Department of History Teaching	4
Department of Psychological Counselling and Guidance	9
Department of Computer Education and Educational Technology	5
Department of Pre-School Teaching	2

4.3 Data Collection

The Opinions of University Students about SEs named questionnaire was prepared by the author to learn which SEs is preferred by students in NEU and what their criterions to prefer those SEs are. The questionnaire has reliability Cronbach's alpha of 0.89. The questionnaire consisted of 3 parts. Part 1 aimed to collect personal information from the respondents with 6 questions. The part aimed to gather general information about some subjects like faculty, department and etc. Part 2 of the questionnaire consisted of 6 questions again and it is focused

on gathering information about computer skills of respondent students. This part brought information about why and how students are using computer and internet. Part 3 of the questionnaire was about SE usage and it consisted of 9 questions. This last part reveals which SE is the most used one and why it is preferred. Also this part gathers information about complaints of students about SEs.

SE performance test was used to calculate the dead-link and precision ratios of SEs in the thesis. SE performance test is finding the SE which has the lowest dead-link ratio and the highest precision ratio in order to classify selected SEs from the best to the worst. If a web site does not exist anymore on internet, a SE should not find this link and shows it in the result page because that links is dead. A SE should update itself in order to find new links and not to find dead-links. A low dead-link ratio shows the success of the SE about renewing itself and this is a really important issue for SEs because facing with dead-links will not attract the user to use that SE. Precision is the ratio of total retrieved relevant links to total retrieved links. High precision ratio shows the success of that SE. Otherwise it means that the SE is retrieving irrelevant links which means that the SE cannot perform its job (Egghe, 2006).

4.4 Data Analysis

- The performance evaluation test is applied to reveal which SE gives best performance according to student's aspect.
- Currency of evaluated SEs is calculated according to dead-link number and it shows the success of SEs about eliminating dead-links while representing search results.
- Precision ratio of evaluated SEs is calculated according to retrieved relevant links and it shows the success of SEs about bringing correct results to user.
- One sample *t*-test applied on questionnaire results in order to designate means of SE which are preferred by NEU students.
- Other features and criterions of NEU research is represented by graphics according to percentages.
- Results of performance test calculated by Microsoft Excel 2007 (Formulas such as average, general total).
- Questionnaire results evaluated by SPSS 12.0 (Frequency, percentage, one sample *t*-test, mean, standard deviation (SD)).

4.5 Duration and Resources

The study started in June 2009 and was completed in June 2010. The work was carried out during this period and weekly duration is given in Table 4.2. First 3 months of thesis is appointed only for literature research and during later processes, literature search is done continuously. Books, internet, and academic journals are main information resources of this thesis.

Table 4.2: Time schedule

WORK DONE	DURATION
• Literature Search	June 2009 – June 2010
• Preparation of the Research Proposal	10 Weeks
• Preparation of Questionnaires	2 Weeks
• Application of Questionnaires	4 Weeks
• Applying Questionnaire Results to SPSS	2 Weeks
• Performance Evaluation Test	1 Week
• Data Analysis	2 Weeks
• Writing the Thesis	20 Weeks
• Reading, discussion, and correction of the thesis based on the feedback by the thesis supervisor.	4 Weeks

4.6 Application

At the beginning of research, a problem defined during the literature research. A pilot study applied at NEU students in order to reveal their popular SEs and criterions to prefer a SE. Based on the literature, a questionnaire was prepared by the author with help of thesis advisor to gather information and asked to randomly students to answers questionnaires. During application of questionnaires students directed to choose the most appropriate answers for themselves and results evaluated in SPSS by determined appropriate statistic methods. In the next process, performance evaluation test applied on 5 popular SEs which are Google, Yahoo, Bing, Ask and AOLSearch by considering students prefers and results evaluated according to user aspect by calculating precision and currency ratios. The performance of five SE has been evaluated by asking same queries to the SEs. There are 20 queries totally. First 10 queries covers top searched terms of 2009 and Hitwise Releases which is an international statistics and data company (www.hitwise.com). Other 10 query is about Information Technology (IT) and Information Systems (IS). First 5 pages of each SE which is equal to 50 links have been

checked for each query. Results have been categorized as relevant, irrelevant and dead link. Average results retrieve and precision of SEs has been calculated. Test queries are as below:

Queries about top search terms of 2009

1- 2009 worst actress
2- How to sell clothes on e-bay
3- Facebook login
4- Rachael Ray videos on Youtube
5- New laws for 2010
6- Van Morrison father
7- Mel Gibson Mexican Jail
8- Mapquest driving directions
9- Rosie O'Donell girlfriend
10- Myspace login account

Results for top search terms are shown in Table 5.

Queries about IT and IS

1- Search engine performance evaluation
2- Googled: The end of the world as we know it (Best Seller in Amazon – Computer Books)
3- What is information technology
4- How to connect Oracle Database with ASP
5- Java, mobile technologies
6- How can we create relationship between SQL and programming
7- Near East University, CIS Department
8- Visual Studio 2008 download
9- Top ten search engines 2009
10- C++ free e-book

According to the general results, Google, Yahoo, Bing, Ask and AOLSearch are the leaders of SE sector for 2009 and also these SEs are the most used ones during 2009 (www.seoconsultants.com, www.hitwise.com). Performance test has been applied to SEs in order to determine which SE gives best performance to user according to precision and dead-

link ratios. At the end of the research, result reports have been taken by comparing performance evaluation test results and students prefers.

4.7 Summary

Questionnaires and test queries are used as data collection tools in this study and questionnaires prepared by the author and thesis supervisor. Only willing applicants joined to the questionnaires and collected by researcher after applicants fill them completely. Test queries applied all 5 SEs and totally 5000 hyperlink has been checked. Precision and dead-link ratios have been calculated according to the results and SEs has been ranked in order to clarify which SE has the best performance for user.

CHAPTER V

RESULTS

In this chapter, the results of the study and the performance evaluation test are presented in detail through the use of tables and figures.

5.1 SE Usage of Students

According to the findings students use internet for various aims, such as research, homework, games, instant messaging services, social networking sites, e-mail services, video channels, and downloading or watching movies. But whatever the aim of use of internet, 100% of students use web SE. In the research, whether students use any other SE except these is also examined and according to the results, 96.33% of students are not using any other SE. The remaining 3.67% is using SE for hackers or Bear Share SE which is just an advertisement page and directing to the Google result page. During the research, students evaluated MSN/Live and Bing in different statues because when the study started, Bing was on the test drive and MSN/Live Search was still in use. Then MSN/Live replaced its place with Bing and MSN portal started to direct its searches to Bing. Frequency of results are checked in order to define SE prefers of students. In Table 5.1 and Figure 5.1, preferred SEs frequency ratios have been given.

Table 5.1: SE usage frequency of students

	Always Use	**Mostly Use**	**With Another**	**Rarely Use**	**Never Heard**	**Never Use**	**Don't Like**
Google	74.30%	15.70%	9.30%	0.30%	0%	0.30%	0%
Yahoo	2.30%	5%	9.30%	20.70%	52.70%	1.30%	8.70%
Bing	0.30%	0.70%	3%	5.70%	51%	34.70%	4.70%
MSN/Live	4%	5%	13.30%	9.30%	49.70%	6%	12.70%
Ask	0.30%	0.30%	1.30%	3.70%	52.70%	36.70%	5%
AOLSearch	0.30%	0.30%	0.70%	1.70%	56%	36%	5.30%

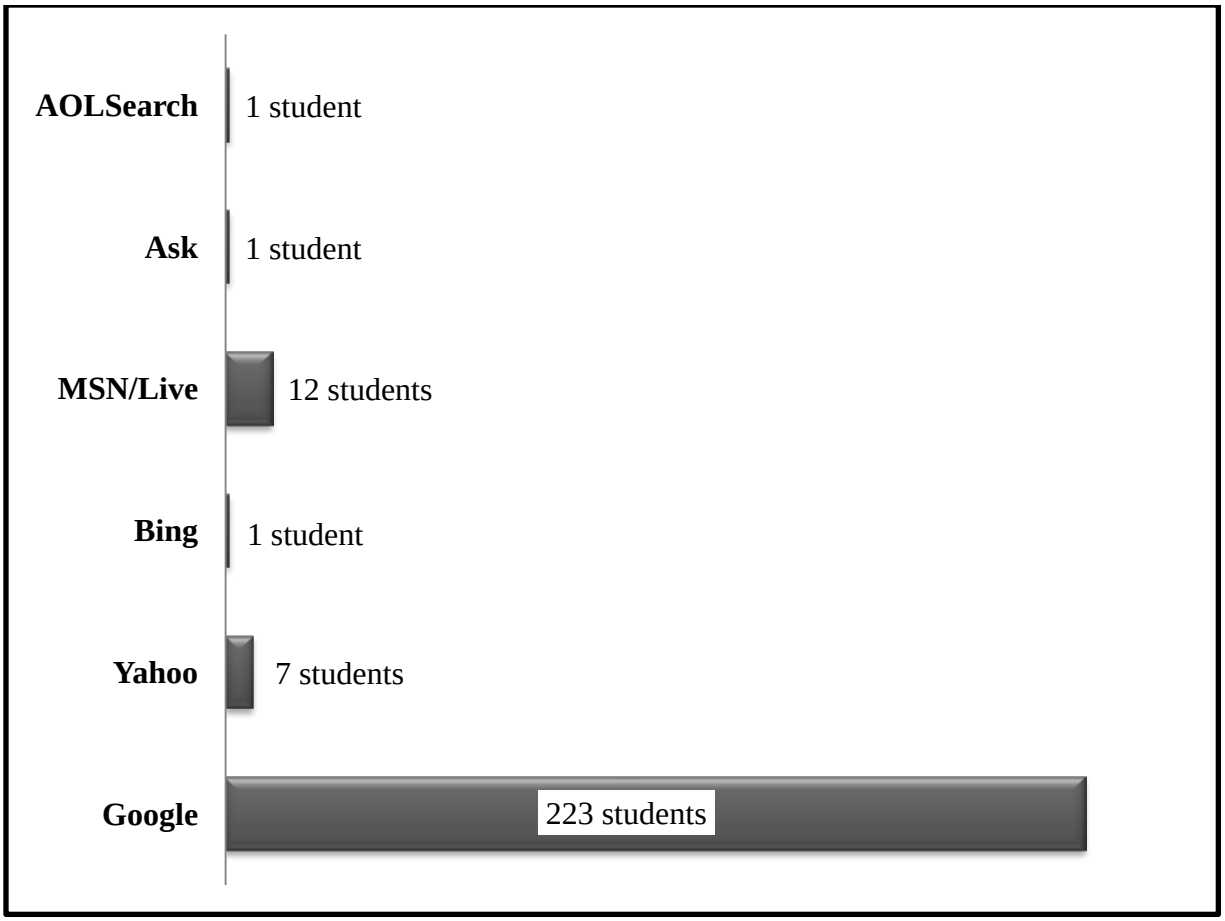

Figure 5.1: Favorite SEs of students in numbers

According to Table 5.1 and Figure 5.1, 223 students (74.30%) always use Google as SE. 48 students (15.70%) of mostly use Google. The Google is enough alone for students and only 27 students (9.30%) of use Google with another search engine if they can not find what they are looking for. Google is working as primary information resource of internet users (Ganzha, Paprzycki & Stadnik, 2010). This situation makes Google the most famous SE for students in this research. With respect to Yahoo, 159 students (52.70%) of said that they never heard this SE and 4 students (1.3%) never used Yahoo. Only 7 students (2.3%) indicate Yahoo as their favorite and 15 students (5%) indicated Yahoo as mostly used one. 62 students (20.70%) use Yahoo rarely and 26 students (8.70%) tried but didn't like Yahoo. 27 students (9.30%) use Yahoo with another SE. On the other hand, results indicate that Bing has never been used by 104 students (34.70%) and 153 students (51%) indicating that they have never heard of such a SE. 14 students (4.70%) tried but did not like the Bing. 9 students (3%) use Bing with another SE and interesting result that, only 1 student (0.30%) indicated Bing as favorite and 17 students (5.7%) use this SE rarely. Furthermore, MSN/Live has not been used by 18 students (6%). 49.70% never heard about it and 12.70% of students tried and didn't like it. 9.30% of

students rarely use MSN/Live and 13.30% of students use it with another SE. Only 12 students (4%) said they always use this SE and 5% prefers to use MSN/Live mostly. The other SE Ask have not been used by 36.70% of students and 52.70% of students never heard about this SE. 5% of students tried but didn't like Ask. 3.70% of students rarely use this SE. Ask is favorite SE of 1 student (0.30%). On the other hand, results indicate that 36% of students have never used AOLSearch as a SE and 56% have never heard of it. 5.30% of students tried and did not like AOLSearch. And 1 student (0.30%) chooses AOLSearch as favorite SE.

5.2 Differences between SEs

Applied One-sample *t*-test results for SE usages of students are shown in Table 5.2 and Table 5.3.

Table 5.2: One-sample *t*-test for SE usage frequencies among students

						95% Confidence Interval of the Difference	
SEs	N	Mean	SD	Sig. (2-tailed)	*T*	Lower	Upper
Google	300	1.37	0.69	0.00	34.15	1.29	1.45
Yahoo	300	4.55	1.25	0.00	62.98	4.41	4.69
Bing	300	5.29	0.86	0.00	106.39	5.19	5.39
Ask	300	5.38	0.77	0.00	120.54	5.29	5.47
AOLSearch	300	5.43	0.68	0.00	137.66	5.35	5.50
MSN/Live	300	4.64	1.47	0.00	54.68	4.48	4.81

*Significant at the 0.05 level of confidence

According to the result of One Sample *t*-test for SE usage of students, there is significant difference between selections of SEs. Google is the one which students always use with 1.37 mean differences. Beside Google, other SEs' means stacked between being rarely used SE or never used SE. Yahoo and MSN/Live are rarely used SEs with 4.55 and 4.64 mean difference. On the other hand; Bing, Ask and AOLSearch has never used SE mean with 5.29, 5.38 and 5.43 mean differences.

Google is the most used search engine in the world. After Google, the order continues as Yahoo, Bing, Ask and AOLSearch (www.hitwise.com). According to the research results,

Google has a crushing superiority on other SE in Near East University as the rest of the world, but the research results indicate the order after Google as MSN/Live, Yahoo, Bing, Ask and AOLSearch. In the world listing, MSN/Live is not in the list anymore because Bing is powered by Microsoft Company like MSN/Live and today the active SE in the MSN page is Bing but Bing and MSN/Live asked to students separately and according to the results, even some students uses MSN/Live, they don't know that they are using Bing actually. Also, according to the SPSS mean test, all the factors that influences a user's choice of a SE is more meaningful for Google users because we can easily say that everybody is a Google user at the university.

5.3 Students Criterions for SE Prefers

There are some factors which influence students in their preference of a SE. These factors include homepage style, result page style, number of retrieved results, number of retrieved relevant results, popularity of search engines, and easy user interface (Bitirim, Tonta, Sever, 2002). Figure 5.2 represents students' most important preference criterions for this study.

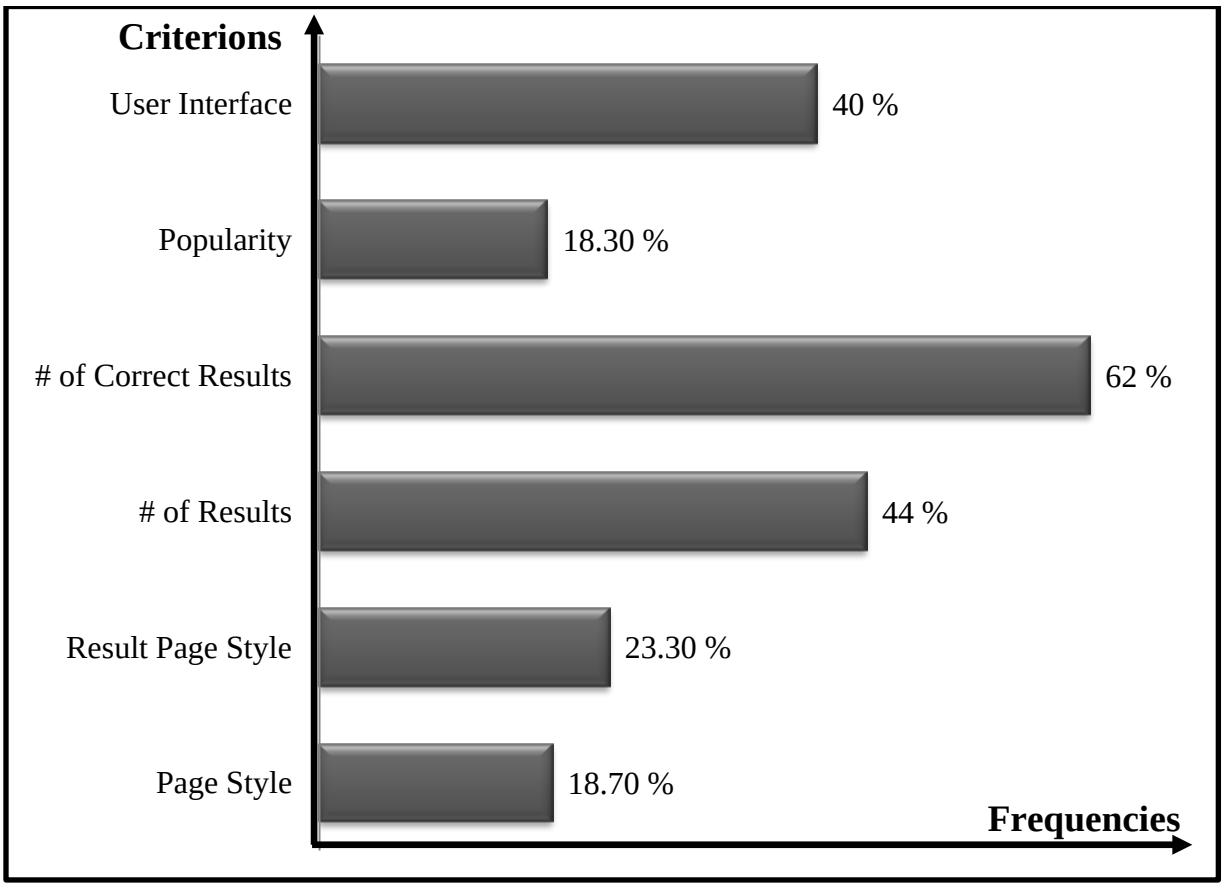

Figure 5.2: Important criterions for SE preferring

As Figure 5.2, the most important criterion for students is number of correct results with 62% frequency. Number of results is another important criterion and has 44% frequency. User interface has 40% frequency and it is another important criterion for students that influence them for SE preferring. While result page style is important for 23.30% of students, 18.70% gives importance of page style. Popularity of SE has the lowest importance with 18.30% frequency for students.

5.4 Students' Opinions about SEs

Opinions of students about other features of SEs are indicated in the Figure 5.3.

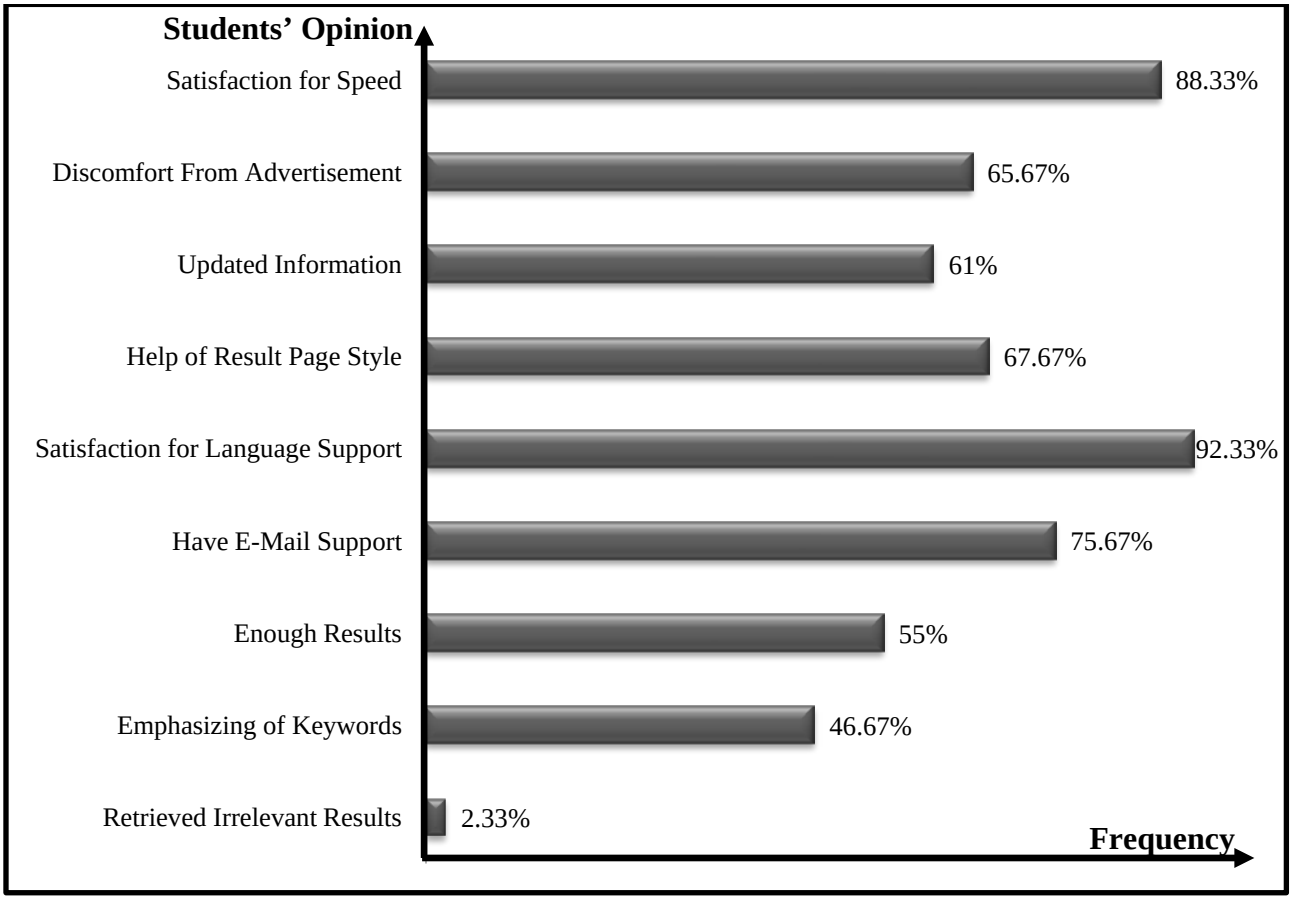

Figure 5.3: General opinions for SEs

As seen in Figure 5.3; students satisfied with the speed of their favorite SE is a total of 88.33%. Nearly all SE publish some advertisements, especially on the result page. Students evaluated the published advertisement on SEs and 65.67% of them underlined that

advertisements bother them. 61% of students feel that their favorite SE is updated daily. As clarified before, style of result page is very important for user. Also 92.33% of students are satisfied from language support of their favorite SE. 75.67% of students use e-mail support of their favorite SE. 55% of students needs to see enough results at the end of their searches. Emphasizing of keywords helps 46.67% of students and 2.33% of students complains about retrieved irrelevant results during search process.

5.5 Other Prefers of Students' Request for SEs

Students' request for homepage of SEs is indicated in Figure 5.5.

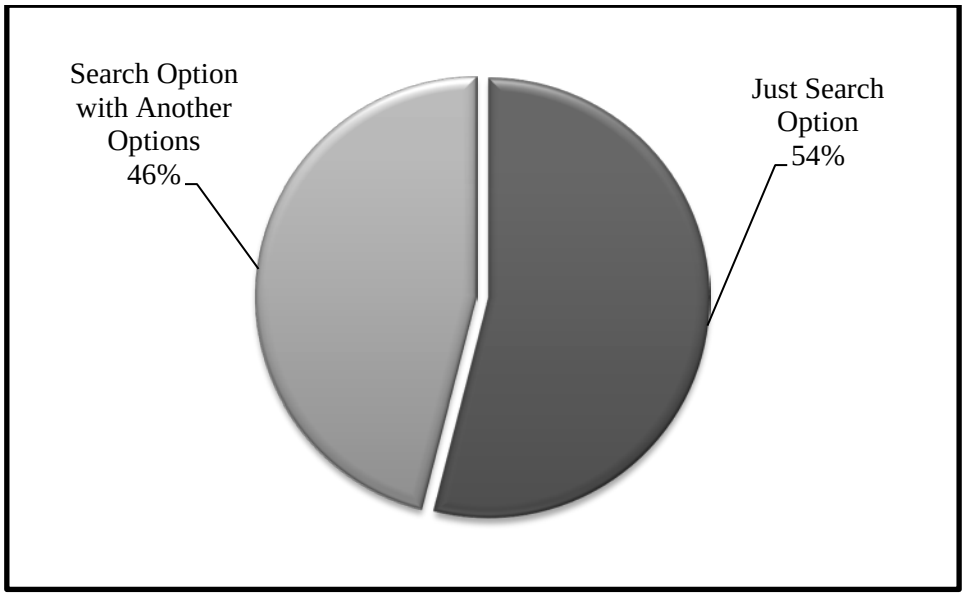

Figure 5.4: User's request for homepage

As can be seen in Figure 5.4, 54% of students prefer plain homepages and they prefer to see only search option in the page. However, the remaining 46% prefer to see other options such as news, galleries, games and etc.

On the result page, some SE display general information about retrieved hyperlinks and some other display quotations which includes user's keyword. Users are available to choose most relevant hyperlink or hyperlinks with their search criteria by reading those information or quotations. Figure 5.6 displays students' preferences about this criterion as follows.

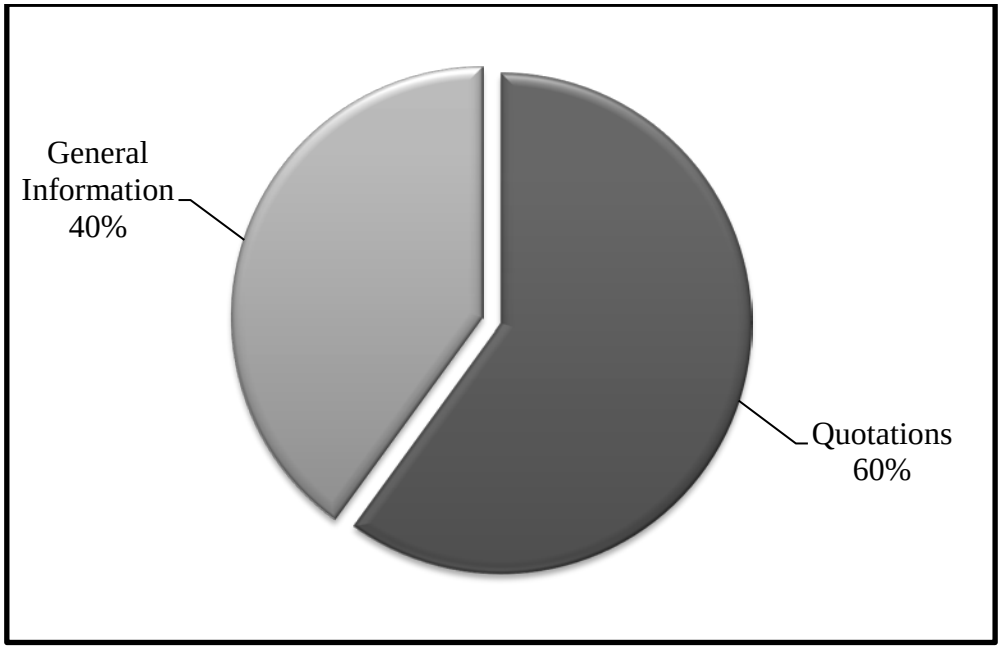

Figure 5.5: General information or quotation prefers of users

According to Figure 5.5, 60% of students want to see general information about retrieved documents with the 40% preferring to see quotations.

Users have different behaviors during search process. One of these behaviors is leaving the search process. Figure 5.7 displays a subject from NEU students' behaviors and states when students leaves search process as follows.

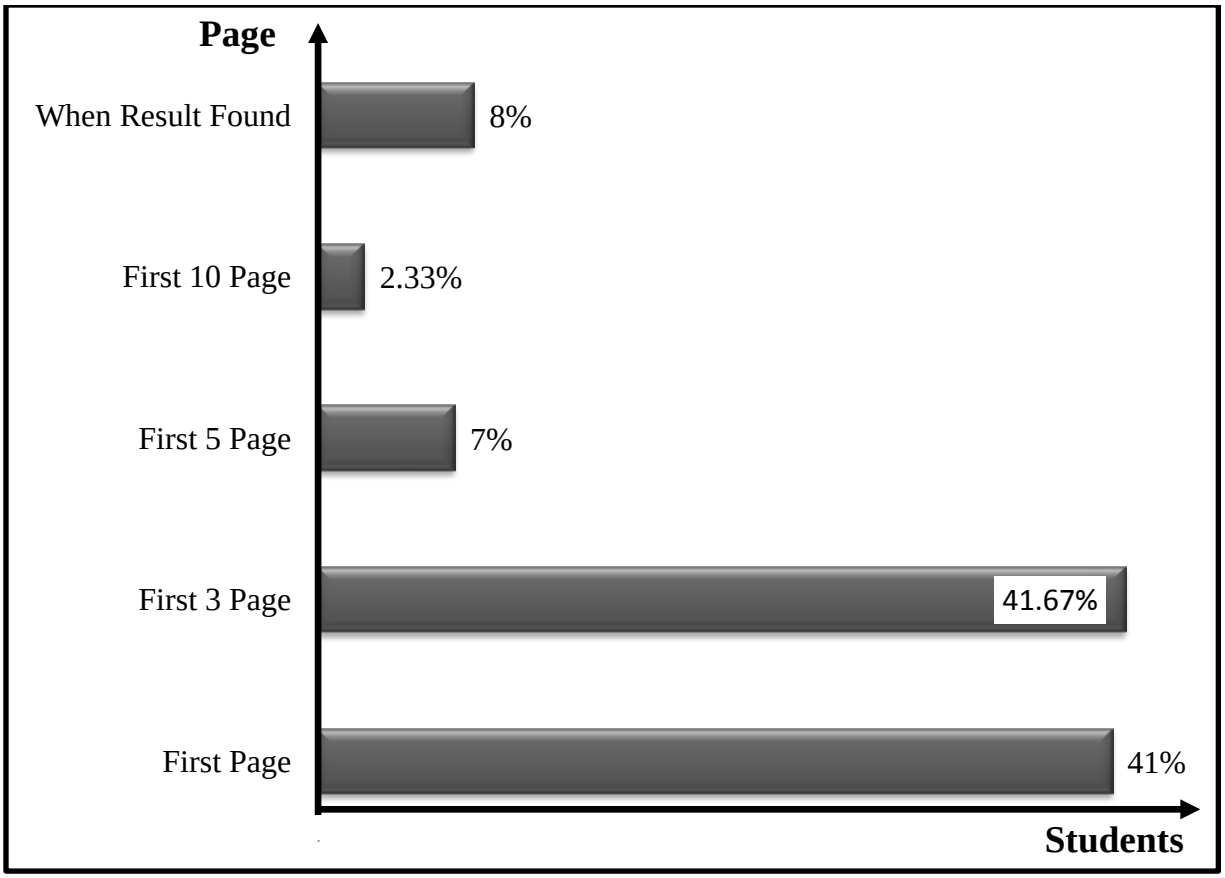

Figure 5.76 Percentage for users when leave search

According to the results in Figure 5.6, 41% of students leave search process at the first page, 41.67% in the first three pages, 7% in the first 5 pages, 2.33% in the first 10 page, and 8% continue to search until they find what they are looking for.

5.6 Findings and Interpretations of Search Engine Performance Evaluation Test

This part of the study explains the results of SE performance evaluation. SE performance evaluation test was getting number of relevant, irrelevant and dead links by entering same queries to 5 different SEs in the study. A successful SE can be measured by checking precision. Precision is the ratio of retrieved relevant documents to total retrieved documents. A precision ratio is changing between 0 and 1. 0 means weak precision and 1 refers to the perfect precision ratio (Bitirim, Tonta & Sever, 2002). In this research, first 5 pages tested from each SE which covers 50 from each. So in this study, number of retrieved relevant documents calculated ratio to 50. Average of precision ratios calculated by dividing sum of the precision ratios into 10. Currency of SEs can be measured by checking dead link ratios. If any web site or document does not exist on the internet anymore, the SE should not retrieve it.

Otherwise the link will be dead link. If a search engine does not have a good currency, it will continue to display retrieve dead links. The well updated and designed web crawler may clean dead links but we should remember that internet resources are changing in every single minute (Jansen & Molina, 2006). In the study, currency ratios represented in percentage and average of currency calculated by dividing sum of the currency ratios into 10. In general results precision ratios calculates by dividing sum of the precision ratios into 20 and currency ratios calculated by dividing sum of the currency ratios into 20.

5.6.1 Precision and currency ratios for top search queries of 2009

Table 5.3 explains the results of precision calculation of performance evaluation test for top search queries for 2009.

Table 5.3: Precision ratios for top search queries of 2009

	Google		Yahoo		Bing		Ask		AOLSearch	
Query	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio
1	34	0.68	28	0.56	41	0.82	23	0.46	42	0.84
2	35	0.70	38	0.76	47	0.94	30	0.60	48	0.96
3	31	0.62	29	0.58	49	0.98	25	0.50	46	0.92
4	30	0.60	34	0.68	50	1.00	35	0.70	44	0.88
5	41	0.82	35	0.70	50	1.00	31	0.62	47	0.94
6	36	0.72	34	0.68	50	1.00	25	0.50	47	0.94
7	50	1.00	29	0.58	50	1.00	37	0.74	50	1.00
8	39	0.78	26	0.52	47	0.94	24	0.48	46	0.92
9	37	0.74	30	0.60	47	0.94	32	0.64	47	0.94
10	40	0.80	27	0.54	50	1.00	30	0.60	45	0.90
Avg.		**0.75**		**0.62**		**0.96**		**0.58**		**0.92**

On favorite queries test, Google reaches the best precision only one time in query number 7 as AOLSearch. Bing reached to the best precision 5 times. Other evaluated SEs could not reach this level. Results indicated that Bing has the best precision ratio average which is 0.96 and AOLSearch follows Bing with 0.92. Google is 3rd with 0.75. Yahoo ranks 4th with 0.62 precision ratios, and last one is Ask with 0.58.

Other output of this performance evaluation test is currency in the study. The Table 5.4 shows dead link ratios and calculated currencies of evaluated search engines for top search queries.

In the table; first column represents the queries that used for performance test. Numbers in tables represents number of dead links for each query and SE. Last 2 rows represent total number of dead links for each SE and in last row dead link ratio is calculated in per cents.

Table 5.4: Dead link ratios for top search queries of 2009

Query	Google	Yahoo	Bing	Ask	AOLSearch
1	8	11	3	0	2
2	3	1	0	2	1
3	5	3	0	1	1
4	9	5	0	4	2
5	4	3	0	4	0
6	0	6	0	8	0
7	0	9	0	1	0
8	0	7	0	1	0
9	4	11	1	3	1
10	2	4	0	5	1
Total #	35	60	4	29	8
Ratios	7%	12%	0.8%	5.8%	1.6%

According to the test results in Table 5.4, for top search queries of 2009, Bing has the best currency with 0.08% dead link ratio. AOLSearch follows Bing with 1.6% dead link ratios and Ask follows AOLSearch with 5.8%. World's most popular SE Google is on the 4th rank about currency with 7% and worse currency belongs to Yahoo with 12%.

5.6.2 Precision and Currency Ratios of IT and IS Queries

Table 5.5 explains SE performance test results for IT and IS queries which were used in the research by the author.

Table 5.5: Precision ratios for IT and IS queries

	Google		Yahoo		Bing		Ask		AOLSearch	
Query	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio
1	35	0.70	23	0.46	38	0.76	16	0.32	33	0.66
2	42	0.84	32	0.64	47	0.94	32	0.64	45	0.09
3	21	0.42	25	0.50	43	0.86	19	0.38	39	0.78
4	41	0.82	26	0.52	44	0.88	39	0.78	43	0.86
5	41	0.82	36	0.72	46	0.92	43	0.86	43	0.86
6	45	0.90	46	0.92	47	0.94	43	0.86	44	0.88
7	24	0.48	4	0.08	10	0.20	4	0.08	19	0.38
8	45	0.90	38	0.76	50	1.00	37	0.74	45	0.90
9	32	0.64	31	0.62	48	0.96	24	0.48	43	0.86
10	40	0.80	37	0.74	48	0.96	35	0.70	46	0.92
Avg.		**0.73**		**0.60**		**0.84**		**0.58**		**0.80**

For information system and technology queries, Bing has the best precision again with 0.84 and AOLSearch is 2nd with 0.80, Google ranks 3rd with 0.73. Yahoo's precision ratio is 0.60 and this ratio makes Yahoo the 4th one. Ask has the worst precision again with 0.58. In this test, only Bing could reach to the perfect precision with query number 8 and no other SE could reach to this level. On query number 7, Yahoo and Ask got the worst precision ratio with 0.08.

Table 5.6 shows dead link ratios of queries about information technology and information system queries in the research. In the table; first column represents the queries that used for performance test. Numbers in tables represents number of dead links for each query and SE. Last 2 rows represent total number of dead links for each SE and in last row dead link ratio is calculated in per cents.

Table 5.6: Dead link ratios for information technology and information system queries

Query	Google	Yahoo	Bing	Ask	AOLSearch
1	3	0	1	6	2
2	1	6	1	2	1
3	7	4	1	4	1
4	3	10	1	1	3
5	1	5	0	3	1
6	3	2	0	3	0
7	0	0	0	0	0
8	5	5	0	7	2
9	0	0	0	0	0
10	3	2	2	2	0
Total#	26	34	6	28	10
Ratios	5.2%	6.8%	1.2%	5.6%	2%

About information technology and system queries, Bing has 1st rank again with 1.2% and AOLSearch follows it closely with 2% currency. Google has 3rd rank with 5.2% and Ask is coming after Google with 5.6%. Yahoo has the worst currency with 6.8%.

5.6.3 Precision and Currency for General Test Results

Test results of SE performance test also revealed a general view for precision and currency ratios. Results are shown in Table 5.7 and 5.8 as follows. In the table; first 10 queries represents the queries of top search queries and other 10 queries represents IT and IS queries.

Table 5.7: General results for precision ratios

	Google		Yahoo		Bing		Ask		AOLSearch	
Query	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio	Retrieved	Ratio
1	34	0.68	28	0.56	41	0.82	23	0.46	42	0.84
2	35	0.70	38	0.76	47	0.94	30	0.60	48	0.96
3	31	0.62	29	0.58	49	0.98	25	0.50	46	0.92
4	30	0.60	34	0.68	50	1.00	35	0.70	44	0.88
5	41	0.82	35	0.70	50	1.00	31	0.62	47	0.94
6	36	0.72	34	0.68	50	1.00	25	0.50	47	0.94
7	50	1.00	29	0.58	50	1.00	37	0.74	50	1.00
8	39	0.78	26	0.52	47	0.94	24	0.48	46	0.92
9	37	0.74	30	0.60	47	0.94	32	0.64	47	0.94
10	40	0.80	27	0.54	50	1.00	30	0.60	45	0.90
1	35	0.70	23	0.46	38	0.76	16	0.32	33	0.66
2	42	0.84	32	0.64	47	0.94	32	0.64	45	0.09
3	21	0.42	25	0.50	43	0.86	19	0.38	39	0.78
4	41	0.82	26	0.52	44	0.88	39	0.78	43	0.86
5	41	0.82	36	0.72	46	0.92	43	0.86	43	0.86
6	45	0.90	46	0.92	47	0.94	43	0.86	44	0.88
7	24	0.48	4	0.08	10	0.20	4	0.08	19	0.38
8	45	0.90	38	0.76	50	1.00	37	0.74	45	0.90
9	32	0.64	31	0.62	48	0.96	24	0.48	43	0.86
10	40	0.80	37	0.74	48	0.96	35	0.70	46	0.92
Avg		**0.74**		**0.61**		**0.90**		**0.58**		**0.82**

According to Table 5.7, general evaluation scenario is not that much different. Bing is the leader again with 0.90 and AOLSearch follows Bing as 2nd with 0.82. 3rd one is Google with 0.74 precision ratios. After Google, Yahoo comes as 4th and Ask as 5th with 0.61 and 0.58 precision ratios. Other hand of performance test general currency ratios is shown in Table 5.8. In the table; first 10 queries represents the queries of top search queries and other 10 queries represents IT and IS queries.

Table 5.8: General results for currency ratios

Q.	Google	Yahoo	Bing	Ask	AOLSearch
1	3	0	1	6	2
2	1	6	1	2	1
3	7	4	1	4	1
4	3	10	1	1	3
5	1	5	0	3	1
6	3	2	0	3	0
7	0	0	0	0	0
8	5	5	0	7	2
9	0	0	0	0	0
10	3	2	2	2	0
1	8	11	3	0	2
2	3	1	0	2	1
3	5	3	0	1	1
4	9	5	0	4	2
5	4	3	0	4	0
6	0	6	0	8	0
7	0	9	0	1	0
8	0	7	0	1	0
9	4	11	1	3	1
10	2	4	0	5	1
Total#	61	94	10	57	18
Ratios	6.1%	9.4%	1%	5.7%	1.8%

In general evaluation, Bing has the best currency with 1% and AOLSearch follows Bing with 1.8%. Ask has the 3rd rank in general currency evaluation with 5.7%. Google's currency percentage is 6.1 and it has the 4th rank. Yahoo's currency is the worst one with 9.4% and it has 5th rank in general currency evaluation.

5.8 Summary

Two different researches have been done in this study in order to clarify popular SEs used by students and SE which gives best performance to user. Applied questionnaires clarified students' prefers and reasons about SE usage and performance evaluation test help to the author to evaluate performance of SE in user view. According to the research and evaluation results, Google is the most popular and well known SE between students in but in performance test revealed that Bing gives the best performance with lowest dead link ratio and highest precision.

CHAPTER VI

CONCLUSION, DISCUSSIONS AND RECOMMENDATIONS

This chapter covers the conclusions, discussions and recommendations of the study.

6.1 Conclusion and Discussions

Since SEs has come into existence, much research has been conducted into this subject, and they have taken place in the literature. Between 1976 and 1995 research into IR systems increased until SEs come into existence in 1994. SE performance evaluation and SE comparison studies started after 1995. Since SE usage became popular, researchers started to pay more attention to SE examinations but with developed SE technology, studies started to deepen. After 2002, most researches preferred to work in relation to technological details of SEs. In the search for the literature performance evaluation studies conducted with the user aspect were very few in number with existing studies having been conducted several years ago. There is no more research in the literature which covers performance evaluation of Google, Yahoo, Bing, Ask and AOLSearch until second half of 2010. This study focuses on SE performance evaluation with the user aspect and. Also it covers performance evaluation of the five most used SEs of 2009 with the study being the first one on this subject in Cyprus and in the literature.

In NEU, Google is the most used SE and Yahoo is the rarely used one but students generally have never heard of the rest of the SEs including Yahoo with 52.7%. These SEs are the ones that are leading the sector in the world but NEU students may not well enough informed about these SEs. Exactly as in our study Tezer and Bicen (2009) stated Google as mostly used SEs in NEU.

In this research, Bing provides perfect output for users. Also outputs of other SEs are successful as well. Vaughan (2003) stated that Google performed best. This situation can be explained in two ways; Google was founded in 1998 and was announced as the best SE in same year (PC Magazin, 1998) and since then Google has been able to meet the expectations of users because a large number of students do not need to look for another SE.

On the other hand, Bing has been in use since 2009 and even though it is a new SE it has a large amount of users, however students have never heard of Bing. Only 4% of students

frequently use MSN/Live as SE but they have not realized that MSN/Live has started to offer services as those offered by Bing. Also significant difference found between SE prefers of students. In this study, Bing reached to the best values and Bing followed by AOLSearch and Google but in previous researchers, Griesbaum (2003) evaluated three German SEs and Google reached the best result ratios. Jansen and Molina (2005), evaluated 5 SEs including Google for ecommerce links and authors found out that Froogle and Yahoo has better performance from the rest three SEs. In this study, Yahoo could take only rank four in performance evaluation test.

Most important criterion for users is number of retrieved correct results and also number of results in another important criterion. Because user interface is another important criterion, homepage style and result page style is important as well. Students prefer to see only search option in the homepage of SE and seeing quotations in result page is other prefers of students. As indicated in the results, students are satisfied from their favorite SE's language and e-mail support. Also students are pleased with the speed of their favorite SE. Big amount of users checking first 3 result page or just first page during search and if they can not retrieve any relevant link, they are changing the keyword or SE.

During the performance evaluation test of this research, Bing gave the best results. It has the highest precision ratio and lowest dead link ratio. Google could be the 3rd one in the classification and AOLSearch is the second one in performance test and its precision ratio is so close to the Bing. But results of AOLSearch are powered by Google since 2009. Google experiences major problems because of incredible upload numbers and spam pages. But results of AOLSearch are clean and successful. From here, we can make some assumptions. Google could not achieve to clean its index from spam pages so they preferred to power and develop AOLSearch. If they swap their addresses in one day, how a user can understand this? So Google may prefer to develop AOLSearch and instead of repairing their own system because if the problem is very big, repairing may cost more. Foundation of Bing has the same date with the arrangement between AOLSearch and Google. Even it is a really new system; Bing can be a big threat for Google with its improved high technology. Ask has the low precision and high dead link ratios. Before that Ask was giving service as a web directory but then they could not continue with this system and preferred to work as an indexing service. According to the results, Ask could not achieve to improve their new systems in a Good way. Yahoo comes after Google in performance evaluation test. Even it can not give a well SE service, its huge movie and music archive keeps Yahoo popular.

According to the results, if Google can not find a good solution to ease its spam and ranking problems, it will give its place to the Bing in future years because the creators of Bing has improved their systems in order to keep users away from Google's problems. Also we can see that Google changed design of its result page during the last weeks and Google's new result page shows many similarities with the result page of Bing. Google is a kind of habit for users and it is designed perfectly in order to give this habit for users. As you can understand from the slogan of Bing, this SE's first aim is to get the place of Google in the world sector by changing user's habits.

6.2 Recommendations

In addition, we can make the following proposals to other researches in this field:

- Some conferences or seminars may arranged by experts in order to give information to students about SEs.
- Spreading this research among universities in Cyprus.
- Applying the research on Turkish Republic universities.
- Applying the research on European universities.
- Increasing number of SEs.
- Increasing number of checked links.
- Increasing number of queries.

REFERENCES

Abbate, J. (1999). Inventing the Internet. 1st Edition, Wellington Graphics.

Adalı, S., Bufi, C. & Temtanapat, Y. (1997). *Integrated Search Engine. IEEE Knowledge and Data Engineering Exchange Workshop:Proceedings: November 4, 1997, Newport Beach, California,. Los Alamitos, CA: IEEE Computer Society Press.* 140-147

Alberti, B., Anklesaria, F., Lindner, P., McCahill, M. & Torrey, D. (1992). The Internet Gopher Protocol: A Distributed Document Search and Retrieval Protocol. University of Minnesota.

Alsaffar, A.H., Deogun, J.S., Raghavan, V.V. & Sever, H. (2000). Enhancing Concept-Based Retrieval Based on Minimal Term Sets. *Intelligent Information Systems Journal,* 14(2/3), 155-73.

Anklesaria, F., Lindner, P., McCahill, M., Torrey, D., Johnson, D. & Alberti, B. (1993). Gopher+: Upward Compatible Enhancements to the Gopher Protocol. University of Minnesota.

Batzios, A., Dimou, C., Symeonidis, A.L., & Mitkas, P.A. (2007). BioCrawler: An Intelligent Crawler for the Semantic Web. *Expert Systems with Applications*, 53(1/2), 524-530.

Begg, C.E. & Connolly, T.M. (1998). Database Systems: A Practical Approach to Design, Implementation and Management. Addison Wesley.

Belkin, N.J., Stein, C.C., & Thiel U. (1995). Cases, Scripts, and Information-seeking Strategies: On the Design of Interactive Information Retrieval Systems. *Published by Elsevier Science Ltd*, 9(3), 379-395.

Bergman, M.K. (2001). The Deep Web: Surfacing Hidden Value. (White Paper), *The Journal of Electronic Publishing*, 7(1).

Bitirim, Y., Tonta, Y. & Sever, H (2002). Türkçe Arama Motorlarında Performans Değerlendirme, Damla Matbaası.

Blair, D.C. & Maron, M.E. (1985). An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communication of the ACM*, 28(3)0, 289-299.

Blair, D.C. (1990). Language Representation in Information Retrieval. Elsevier.

Brake, D. (1997). Lost in Cyberspace. *New Scientist*, 154(2088), 12-13.

Broadbent, B. (1998). Using the Internet Smarter and Faster, Crisp Publications.

Bryant, S. (2000). The Story of Internet. Penguin Readers.

Can, F., Nuray, R. (2006). Automatic Ranking of Information Retrieval Systems Using Data Fusion. *Information Processing & Management*, 42(3), 595-614.

Capra, R.G., III Perez-Quinones, M.A. (2005). Using Web Search Engines to Find and Refind Information. *IEEE Computer Society*, 38(10), 36-42.

Chowdhury G.G. (1999). Introduction to Modern Information Retrieval, 1st Edition, Facet Publishing.

Chu, H. & Rosenthal, M. (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. Global Complexity: Information, Chaos and Control. ASIS'96: *Proceedings of the 59th ASIS Annual Meeting*, Baltimore, Maryland, October .

Cicone, A. & Serra-Capizzano, S. (2010). Google PageRanking problem: The Model and the Analysis. *Journal of Computational and Applied Mathematics*, 234(11), 3140-3169.

Cooper, B., Milner, A., & Worsley, T. (2000). Essential Internet Guide. Dk Publications.

Cooper, W.S. (1995). Some Inconsistencies and Misidentified Modeling Assumptions in Probabilistic Information Retrieval. *ACM Transactions on Information Systems*, 13(1), 100-111.

Crestani, F., Lalmas, M., Van Rijsbergen, C.J. & Campbell, I. (1998). Is This Document Relevant? Probably: A Survey of Probabilistic Models in Information Retrieval. *ACM Computing Surveys*, 30(4), 528-552.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. ve Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Deogun,J.S., (1998). Structural Abstractions of Hypertext Documents for Web-based Retrieval. *A Min Tjoa and Roland R. Wagner (Eds.), Proceedings: Ninth International Workshop on Database and Expert Systems.*

Deutsch, P. (1992). Resource discovery in an Internet environment – The Archie approach. *Electronic Networking: Research, Applications and Policy*, 2(1), 45-51.

Dikaiakos, M.D., Stassopoulou, A. & Papageorgiou, L. (2005). An Investigation of Web Crawler Behavior: Characterization and Metrics. *Computer Communications*, 28(8), 880-897.

Dolowitz, D., Buckler, S., & Sweeney, F. (2008) Researching Online. Palgrave Study Skills. Palgrave Macmillan.

Egghe, L. (2006). The Measuring Precision, Recall, Fallout and Miss as a Function of the Retreived Documents and Their Mutual Interrelations. *Information Processing and Management*, 44, 856-876.

Egghe, L. (2007). Existence Theorem of the Quadruple (P, R, F, M): Precision, Recall, Fallout and Miss. *Information Processing and Management*. 43(1), 265-272.

Foltz, P.W. (1996). Latent Semantic Analysis for Text-Based Research. *Behavior Research Methods, Instruments and Computer*, 28(2), 197-20.

Frank, A. (1996). Internet services. *LAN Magazine/Network Magazine*, 94.

Ganzha, M., Paprzycki, M & Stadnik, J. (2010). Combining Information from Multiple Search Engines—Preliminary Comparison. *Information Science*, 180(10), 1908-1923.

Gromow, G.R. (1996). History of Internet and WWW: The Roads and Crossroads of Internet History.

Guinchat, C. and M. Menow. (1983). General Introduction to Science and Technology of Information and Documentation. Havana: UNESCO.

Henshaw, R. (2001). What Next for Internet Journals? Implications of the Trend Towards Paid Placement in Search Engines. *First Monday*, Vol.6, Iss.9.

Hu, W., Chen, Y., Schmalz, M. & Ritter, G.X. (2001). An Overview of World Wide We Search Technologies. Auburn University, University of Florida.

Ingwersen, P. (1992). Information Retrieval Interaction. Taylor Graham Publishing.

Internet Society. (2000), A Brief History of Internet, (Online). Electronic Address: http://www.isoc.org/internet/history/brief.shtml, Accessed 26 July 2009.

Jansen, B.J. & Molina, P.R. (2006). The Effectiveness of Web Search Engines for Retrieving Relevant Ecommerce Links. *Information Processing & Management,* 42(4), 1075-1098.

Jansen, B.J. & Resnick, M. (2006). An Examination of Searcher's Perceptions of Nonsponsored and Sponsored Links During E-Commerce Web Searching. *Journal of the American Society for Information Science and Technology,* 57(14), 1949–1961.

Jansen, B.J. & Spink, A. (2006). How are we Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs. *Information Processing and Management,* 42(1), 248-263.

Jansen, J. (1996). Using an Intelligent Agent to Enhance Search Engine Performance. *First Monday*, 2(3).

Kahle, B. (1997). Preserving the Internet. *Scientific American*. 276(3), 82-83.

Kleinberg, J.M. (1998). Authoritative Source in a Hyperlinked Environment. *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 668-677.

Kobayashi, M. & Takeda, K. (2000). Information Retrieval on the Web. *ACM Computing Surveys*, 32(2), 144-172.

Koster, M. (1994). ALIWEB – Archie-Like Indexing in the Web. *First World-Wide Web Conference*, 27(2), 175-182.

Küçük, M.E., Olgun, B. ve Sever, H. (2000). Application of Metadata Concepts to Discovery of Internet Resources. *Advances in Information Systems: First International Conference ADVIS*, 1909, 304-313.

Laudon, K.C. & Traver, C.G. (2004). E-Commerce. Business, Technology, Society. 2nd Edition, Pearson Education Inc.

Laursen, J.V. (1998, February/March).Somebody Wants to Get in Touch with You: Search Engine Persuasion. *Database*, 21(1), 43-46.

Lavrence, S., & Giles, C.L. (1999). Accessibility of Information on the Web. *Intelligence*, 32(6), 32-39.

Lawrence, S. & Giles, C.L. (1999). Accessibility of Information on the Web. *Nature 400*, 11(1), 107-109.

Lee, J.H. (1995). Combining Multiple Evidence from Different Properties of Weighting Schemes. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, Seattle, Washington, USA, July 9-13, 1995, 180-188.

Leiner, B.M., Cerf, V.G., Clark, D.D., Kahn, R.E., Kleinrock, L., Lynch, D.C., Postel, J., Roberts L.G. & Wolff, S. (1997). A Brief History of Internet. *The Internet Society*, May/June 1997.

Leuski, A. (2001). Evaluating Document Clustering for Interactive Information Retrieval. *Conference on Information and Knowledge Management Proceedings of the Tenth International Conference on Information and Knowledge Management*. 33-40.

Lin, Y., Shi, X. & Wei, Y. (2009). On Computing Page Rank via Lumping the Google Matrix. *Journal of Computational and Applied Mathematics*, 224, 702-708.

Lindner, P. (1994). Internet Gopher User's Guide. University of Minnesota.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L. & Gay, G. (2005). The Influence of Task and Gender on Search and Evaluation Behavior Using Google. *Information Processing & Management*, 42(4), 1123-1131.

Maron, M.E. (1984). Probabilistic Retrieval Models. *Progress in Communication Sciences*, 5, 145-176.

McCahill, M. & Erickson, T. (1994). A Preliminary Design for a 3-D Spatial User Interface for Internet Gopher. University of Minnesota.

McCune, B.P., Tong, R.M., Dean, J.S. & Shapiro, D.G. (1985). {RUBRIC}: A System for Rule Based Information Retrieval. *IEEE Transactions on Software Engineering*, 11(9), 939-944.

Menczer, F. (2002). Growing and Navigating the Small World Web by Local Content. *Science Sessions*, 99(22), 14-19.

Mettrop, W. & Nieuwenhuysen, P. (2001). Internet Search Engines -Fluctuations in Document Accessibility. *Journal of Documentation*, 57, 623-51.

Mooers, C.N. (1960). The Tape Typewriter Plan. ASLIB Proc.

Mowshowitz, A & Kawaguchi, A. (2004). Measuring Search Engine Bias. *Information Processing & Management*, 41(5), 1193-1205.

Naughton, J. (2000). A Brief History of the Future, The Origins of the Internet. Phoeniz & Paperback.

Notess, G.R. (2001). Joining the in-crowd. *E-Content*, 24(3), 60.

Olgun, B. ve Sever, H. (2000). Kaynak Keşif Yeteneğinin Artırılması için Internet Kaynaklarının İçeriklerinin Standart Biçimde Tanımlanması. *Bilgi Dünyası*, 1, 56-88.

Palme, J. (1995). Electronic Mail. Artech House Inc.

Pitkow, J.E. & Kehoe, C.M. (1996). Emerging Trends in the WWW User Population. *Communications of the ACM*, 39(6), 106-108.

Robertson, S.E. & Jones, K.S. (1976). Relevance Weighting of Search Items. *Journal of the American Society for Information Science*, 27(3), 129-146.

Salton, G. & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval*1. *Information Processing & Management*, 24(5), 513-523.

Salton, G. & Buckley, C. (1990). Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41, 288-97.

Salton, G. & McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw Hill.

Salton, G. (1989). Automatic Text Processing. Massachusetts: Addison-Wesley.

Salton, G., Fox, E.A. & Wu, H. (1983). Extended Boolean Information Retrieval. *Communications of the ACM*, 26(11), 22-36.

Sanderson, E. & Forcht, K. A. (1996). Information Security in Business Environments. *Information Management and Computer Security*, 4(1), 32-37.

Sever, H. & Bitirim, Y. (2003). FindStem: Analysis and Evaluation of a Turkish Stemming Algorithm. *Lecture Notes in Computer Science*, 2857(2003), 238-251.

Shneiderman, B. (1986). Empirical Studies of Programmers: The Territory, Paths, and Destinations, Keynote Address for Workshop. *Empirical Studies of Programmers*, 1-12.

Silverstain, C., Marais, H., Henzinger, M. & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. *ACM SIGIR Forum*, 33(1), 1-6.

Srinivassan, P. (1992). Thesaurus Construction. *Bill Frakes ve Ricardo Baeza Yates (Eds.), Information Retrieval Data Structures & Algorithms*, 161-218.

Svenonius, E. (2000). The Intellectual Foundation of Information Organization. MIT Press.

Tennant, R., Ober, J. ve Lipow, A.G. (1996). Internet Handbook. Ankara: Türk Kütüphaneciler Derneği.

Tonta, Y. (1995). Information Retrieval Systems. *Türk Kütüphaneciliği*, 9(3), 302-314.

Townler, H.M. (1976). System Analysis for Information Retrieval, A Grafton Book.

Van Rijsbergen, C.J. (1979). Information Retrieval. Butterworths.

Yu, C.T. & Lee, T.C. (1986). Non-Binary Independence Model. *ACM Conference on Research and Development in Information Retrieval, Palazzo dei Congressi, Via Matteotti, 1, Pisa, Italy, September 8-10*, 265-268.

http://www.aolsearch.com, 3.04.2010

http://www.ask.com, 3.04.2010

http://www.bing.com, 3.04.2010

http://www.seoconsultants.com, 5.01.2010

http://www.google.com, 3.04.2010

http://www.hitwise.com, 5.01.2010

http://www.yahoo.com, 3.04.2010

APPENDICES A

Performance Test Results for Favourite Queries of 2009

Query	Test Date	Google	Yahoo	Bing	Ask	AOLSearch
1	28/01/2010	34+/8-/8o	28+/11-/1o	41+/6-/3o	23+/27-/0o	42+/6-/2o
		6,240,000	18,900,000	22,000,000	1,420,000	1,090,000
2	29/01/2010	35+/12-/3o	38+/11-/1o	47+/3-/0o	30+/18-/2o	48+/1-/1o
		22,800,000	93,800,000	3,750,000	13,800,000	10,300,000
3	29/01/2010	31+/14-/5o	29+/18-/3o	49+/1-/0o	25+/24-/1o	46+/3-/1o
		496,000,000	1,800,000,000	1,750,000	313,000,000	196,000,000
4	29/01/2010	30+/11-/9o	34+/11-/5o	50+/0-/0o	35+/9-/4o	45+/3-/2o
		966,000	3,120,000	745,000	327,000	247,000
5	30/01/2010	41+/5-/4o	35+/12-/3o	50+/0-/0o	31+/15-/4o	47+/2-/1o
		163,000,000	268,000,000	25,100,000	21,400,000	15,900,000
6	30/01/2010	36+/14-/0o	34+/10-/6o	50+/0-/0o	25+/17-/8o	47+/3-/0o
		475,000	4,550,000	1,280,000	113,000	79,800
7	30/01/2010	50+/0-/0o	29+/12-/9o	50+/0-/0o	37+/12-/1o	50+/0-/0o
		55,800	909,000	146,000	22,500	19,100
8	30/01/2010	39+/11-/0o	26+/17-/7o	47+/3-/0o	24+/25-/1o	46+/4-/0o
		1,100,000	56,200,000	1,480,000	274,000	206,000
9	30/01/2010	37+/9-/4o	30+/9-/1o	47+/2-/1o	32+/15-/3o	47+/2-/1o
		256,000	2,640,000	552,000	43,700	33,000
10	30/01/2010	40+/8-/2o	27+/19-/4o	50+/0-/0o	30+/15-/5o	45+/4-/1o
		45,000,000	188,000,000	607,000	69,700	21,200,000

(+) relevant links (-) irrelevant links (o) dead links

APPENDICES B

Performance Test Results for IT and IS Queries

Query	Test Date	Google	Yahoo	Bing	Ask	AOLSearch
1	31/01/2010	35+/12-/3o	23+/27-/0o	38+/11-/1o	16+/28-/6o	33+/15-/2o
		27,100,000	14,800,000	3,170,000	1,216,000	557,000
2	01/02//2010	42+/7-/1o	32+/12-/6o	47+/2-/1o	32+/16-/2o	45+/4-/1o
		172,000,000	10,900,000	975,000	95,400,000	71,900,000
3	01/02//2010	21+/22-/7o	25+/21-/4o	43+/6-/1o	19+/27-/4o	39+/10-/1o
		212,000,000	2,400,000,000	348,000,000	99,500,000	75,000,000
4	01/02//2010	41+/6-/3o	26+/14-/0o	44+/5-/1o	39+/10-/1o	43+/4-/3o
		1,760,000	12,500,000	1,870,000	309,000	233,000
5	02/02//2010	41+/8-/1o	36+/9-/5o	46+/4-/0o	43+/3-/3o	43+/6-/1o
		26,100,000	85,300,000	24,900,000	15,200,000	11,500,000
6	02/02//2010	45+/2-/3o	46+/2-/2o	47+/3-/0o	43+/4-/3o	44+/6-/0o
		210,000	1,660,000	300,000	45,000	25,600
7	02/02//2010	24+/26-/0o	4+/46-/0o	10+/40-/0o	4+/46-/0o	19+/31-/0o
		469,000	904,000	239,000	141,500	66,800
8	03/02//2010	45+/0-/5o	38+/7-/5o	50+/0-/0o	38+/6-/7o	45+/3-/2o
		18,100,000	50,700,000	18,400,000	9,390,000	7,080,000
9	03/02//2010	32+/18-/0o	31+/19-/0o	48+/2-/0o	24+/26-/0o	43+/7-/0o
		91,600,000	10,600,000	12,100,000	47,900,000	36,100,000
10	03/02//2010	40+/7-/3o	37+/9-/2o	48+/0-/2o	35+/13-/2o	46+/4-/0o
		4,690,000	5,000,000	1,380,000	1,970,000	1,490,000

(+) relevant links (-) irrelevant links (o) dead links