CHAPTER 3
STUDY OF SEARCH ENGINE SEARCH PROCESS

3.1 INTRODUCTION

Search results of today's search engines are the outcome of users hunt for keywords to suit their query and filtering the relevant document from huge number of search results. Relevance of search engine results is still an active area of research for scientists even after remarkable contributions in the area for a decade. The growing size of the web, lack of regulated procedures and the presence of operational secrets are some of the reasons that compel the requirement of further research in the area. This chapter deals with the search process, its stages, issues at each stage that affect the relevance of search results. It also attempts to discuss in detail the ranking algorithms and analyses the pitfalls of them. Finally it concludes with importance of context in the search process of search in providing relevant search results.

3.2 SEARCH PROCESS

Offline data acquisition components and online data retrieval components of search process are responsible for producing relevant search results. Web data acquisition components gather data from the web pages collected by the web crawlers and build the index. Data retrieval components are involved in
the process of retrieving, ranking and displaying the search results from the index generated by the data acquisition components.

The requirements for web document searches discussed by Hu and Chen (2001) include the systems effectiveness in locating and ranking web documents, the efficiency of the system's web document search and ranking algorithms, the systems unbiased ranking to pages, the systems access to up-to-date web information, web coverage and the systems adaptability to the user queries. The notoriously low precision of web search engine coupled with the ranked list presentation makes it hard for users to find information they are searching for.

### 3.2.1 Phases of search process

Search process operates in phases to transform the user query to search results. The phases (Fig. 3.1) include

- Query formulation
- Query processing
- Repository indexing
- Searching
- Ranking

All phases of search process are interdependent to produce the relevant results. Query formulation phase involves users in framing the query with or without boolean operators. The accuracy of the search results
depends on the form of the query. A specific query produces more accurate search results than a general query. The process of transforming the query to the format required by the search algorithm is done by the query-processing phase. Search engines index the web documents in the form of identifiers and the query is converted to wordID and searched in the index. Repository indexing is an offline operation, which gathers information from the web pages harvested and generates the index. This phase uses tools such as parser to parse the documents, word extractors to extract the words to build an inverted index.

The other online processes then use the index. The inverted index is searched by the search algorithm to locate the documents in which the query word occurs. The set of documents which are retrieved in the search phase are ranked based on the relevance of the document to the search query. Different search engine follow different ranking algorithms to rank the search results producing different search results for the same query word in spite of using the same the same search providers.

3.2.2 Role of phases on relevance of search results

Each phase of the search process contribute to the relevance of the search results. The following sections analyses the role of phases, which contribute to search process in generating search results with more precision.
3.2.2.1 Query formulation

Keyword which can be any word on a web page is the most common form of text used to search the web. Most search engines receive the keyword and perform the text query and retrieval for the user. The keywords are intended to convey something about the subject and content of the page required by the user.

The user is responsible for choosing precise word as the keyword. The relevance of the search results in current search engines exactly depend on the keyword. If the keyword has another word, which spells the same but mean different, or words that means the same but are not actually entered in
the query, then the chances of relevant results produced by the search process is reduced.

Moreover search process provides an user interface with a text input box to receive input keyword. The user interfaces in most of today's search engines still rely on a grammar of Boolean operators and keywords, and for good results, the user is expected to be able to fill the box with the right keywords and in right combination (GOOSE). Research has reported that about 10% of web searchers utilize advanced query operators, with the other 90% using extremely simple queries.

Google's (2003) advice on advanced search states that the accuracy of searches can be increases by adding operators that fine-tune the user keywords. America Online search's (AOL) advice page (AOL, 2003) states, “there are times when you might want the precise results that Boolean query provides”. Most search engines now offer ways to narrow the searches to produce fewer and more focused results. The operators provided by the search engines are tabulated in Table 3.1.

3.2.2.2 Repository indexing

Once the spiders have completed the task of finding information on web pages, the search engine must store the information in a way that makes it useful. There are two key components involved in making the gathered data
Table 3.1 Features supported by various search engines (Sullivan, 2003)

<table>
<thead>
<tr>
<th>ENGINE NAME</th>
<th>BOOLEAN</th>
<th>PHRASE</th>
<th>WORD STEM</th>
<th>LOCATION IN DOCUMENT</th>
<th>DATE</th>
<th>SEARCH WITHIN RESULTS</th>
<th>DOMAIN TYPE</th>
<th>LANGUAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altavista</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Alltheweb</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Google</td>
<td>Yes except OR</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Hotbot</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes, enable in advance search</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Lycos</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

accessible to users are

- The information stored with the data
- The method by which the information is indexed

Search engines store more than just the word and URL in order to produce relevant results. They store the number of times that the word appears on a page. The engine assigns a weight to each entry, with increasing values assigned to words as they appear near the top of the document, in subheading, in links, in the meta tags or in the title of the page.

Each commercial search engine has a different formula for assigning weight to the words in its index. This is one of the reasons that a search for the same word on different search engines will produce different lists, with the pages presented in different orders.
Search engines encode the data to save storage, regardless of precise combination of additional pieces of information. The purpose of an index is to find information as quickly as possible. Hash table is used to build the index effectively. The key to the effectiveness of a hash table is that the numerical distribution is different from the distributions of words.

The Web page author can have some influence over which keywords are used to index the document, and even in the description of the document that appears when it comes up as a search engine hit, using the meta tags in the HTML document. This is obviously very important when the author is trying to draw people to the website based on how the site ranks in search
engines hit lists. There is a lot of conflicting information out there on meta tagging. Different search engines look at meta tags in different ways. Some search engine rely heavily on meta tags, others don't use them at all. The general opinion seems to be that meta tags are less useful than they were a few years ago, largely because of the high rate of spamdexing i.e., web authors using false and misleading keywords in the meta tags. It is also called as spamming.

Repository index should contain data aimed at providing information to improve the relevance of the search query. It is also important to ensure that the data contained in the index must not be prone to misguide the search process. The data present in the indices of the current search engines does not contain information to produce relevant search results rather they are aimed to produce search results containing the words. This thesis proposes to extend the search engine index to provide information useful in increasing the precision of search results.

3.2.2.3 Ranking of search results

Once a search engine has used the search terms to gather hits from its database, it ranks them according to its formula for determining their relevance. It is often difficult to understand exactly the ranking of the search results because the factors of ranking search results are guarded secrets. Ranking factors can be classified as one within a page and one external to a page. The classification of ranking factors is discussed below.
Ranking factors within a page

- Word frequency factor determines how many search terms appear in the document and how often they occur.
- Location of the search terms in the document determines search terms found in title, heading, the first 200 words of text, etc.
- Relational clustering determine how many pages from the site contain the search terms.
- HTML design determine the characteristics such as frames, broken links, loading speed, ALT and meta tags.

Ranking factors external to a page

- Link popularity of a web page is determined by sites with more links pointing to them rank higher.
- Click popularity is based on the sites visited more often rank higher.
- "Sector" popularity is based on how popular a site is with a given demographic or social group.
- Business alliances among services -- Results from a partner search service will be ranked higher.
- Pay-for-placement -- Site owners pay for high rankings.

Ranking algorithms

Many of existing search engines use a two-step process to retrieve pages related to a user's query. In the first step, traditional text processing is done to find all documents using query terms, or related to query terms by semantic meaning. This is done by a lookup into an inverted file, with a vector space...
method, or with a query expander that uses thesaurus. With the massive size of the web, this first step can result in thousands of retrieved pages related to the query. To make this list manageable for a user, many search engines sort this list by some ranking criterion. Pagerank and HITS are considered to be the most successful ranking algorithms, which found their way from research labs to commercial search engines. Pagerank is trademark of the top search engine Google and HITS algorithm is used by Teoma search engine. HITS algorithm belongs to different classification of ranking algorithm. Ranking algorithms can be classified into link popularity and click popularity algorithms.

Link popularity ranking algorithms

A simple heuristic that can be viewed as the predecessor of all link analysis ranking algorithms is to rank the pages according to their popularity. The number of pages that link to this page measures the popularity of a page. This algorithm is referred as the INDEGREE algorithm or simply link popularity, since it ranks pages according to their in-degree in the graph. Kleinberg (1999) makes a convincing argument that the INDEGREE algorithm is not sophisticated enough to capture the authoritativeness of a node, even when restricted to a query dependent subset of the web. The concept of link popularity often avoids good rankings for pages which are only created to deceive search engines and which don't have any significance within the web, but numerous webmasters elude it by creating masses of inbound links for doorway pages from just as insignificant of other web pages.
Pagerank Algorithm

Pagerank is not simply based upon the total number of inbound links. The basic approach of pagerank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally. First of all, a document ranks high in terms of pagerank, if other high ranking documents link to it. So, within the pagerank concept (Fig. 3.3), the rank of a document is given by the rank of those documents which link to it. Their rank again is given by the rank of documents which link to them.

The pagerank of a document is always determined recursively by the pagerank of other documents which is given by

$$PR(A) = (1-d) + d \left( \frac{PR(T1)}{C(T1)} + ... + \frac{PR(Tn)}{C(Tn)} \right)$$

Figure 3.3 Schematics of pagerank algorithm (Haveliwala, 2003)
where

\( PR(A) \) is the pagerank of page \( A \),

\( PR(T_i) \) is the pagerank of pages \( T_i \) which link to page \( A \),

\( C(T_i) \) is the number of outbound links on page \( T_i \) and

\( d \) is a damping factor which can be set between 0 and 1.

The pagerank of pages \( T_i \), which link to page \( A \), does not influence the pagerank of page \( A \) uniformly. Within the pagerank algorithm, the pagerank of a page \( T \) is always weighted by the number of outbound links \( C(T) \) on page \( T \). This means that the more outbound links a page \( T \) has, the less will page \( A \) benefit from a link to it on page \( T \).

The weighted pagerank of pages \( T_i \) is then added up. The outcome of this is that an additional inbound link for page \( A \) will always increase page \( A \)'s pagerank. Finally, the sum of the weighted pagerank of all pages \( T_i \) is multiplied with a damping factor \( d \), which can be set between 0 and 1. Thereby, extend of pagerank benefit for a page by another page linking to it is reduced.

**The HITS Algorithm**

Kleinberg (1999) argued that it is not necessary that good authorities point to other good authorities. Instead, there are special nodes that act as hubs that contain collections of links to good authorities. He proposed a two-level weight propagation scheme where endorsement is conferred on authorities through hubs, rather than directly between authorities. In his framework, every page can be thought of as having two identities. The *hub* identity captures the
quality of the page as a pointer to useful resources, and the authority identity captures the quality of the page as a resource itself. If two copies of each page were made graph $G$ is visualized as a bipartite graph, where hubs point to authorities. There is a mutual reinforcing relationship between the two. A good hub is a page that points to good authorities, while a good authority is a page pointed to by good hubs. In order to quantify the quality of a page as a hub and an authority, Kleinberg (1999) associated with every page a hub and an authority weight. Following the mutual reinforcing relationship between hubs and authorities, Kleinberg (1999) defined the hub weight to be the sum of the authority weights of the nodes that are pointed to by the hub, and the authority weight to be the sum of the hub weights that point to this authority.

The SALSA Algorithm

An alternative SALSA algorithm proposed by Lempel and Moran (2001), combines ideas from both HITS and pagerank algorithms. In HITS the graph $G$ is visualized as a bipartite graph where hubs point to authorities. The SALSA algorithm performs a random walk on the bipartite hubs and authorities graph, alternating between the hub and authority sides. The random walk starts from some authority node selected uniformly at random. The random walk then proceeds by alternating between backward and forward steps. When at a node on the authority side of the bipartite graph, the algorithm selects one of the incoming links uniformly at random and moves to a hub node on the hub side. When at node on the hub side the algorithm selects one of the outgoing links uniformly at random and moves to an
authority. The authority weights are defined to be the stationary distribution of this random walk.

**Implementation of pagerank in Google**

The pagerank is integrated into the general ranking of web pages by the Google search engine. The ranking of web pages by the Google search engine was determined by three factors:

- Page specific factors
- Anchor text of inbound links
- pagerank

Page specific factors are, besides the body text, for instance the content of the title tag or the URL of the document. It is more than likely that since the publications of Page and Brin, more factors have joined the ranking methods of the Google search engine. But this shall not be of interest here.

In order to provide search results, Google computes an IR score out of page specific factors and the anchor text of inbound links of a page, which is weighted by position and accentuation of the search term within the document. This way the relevance of a document for a query is determined. The IR-score is then combined with pagerank as an indicator for the general importance of the page. To combine the IR score with Pagerank the two values are multiplicated. It is obvious that they cannot be added, since otherwise pages with a very high Pagerank would rank high in search results even if the page is not related to the search query.
Especially for queries consisting of two or more search terms, there is a far bigger influence of the content related ranking criteria, whereas the impact of pagerank is mainly visible for unspecific single word queries. If webmasters target search phrases of two or more words it is possible for them to achieve better rankings than pages with high pagerank by means of classical search engine optimisation.

If pages are optimised for highly competitive search terms, it is essential for good rankings to have a high Pagerank, even if a page is well optimised in terms of classical search engine optimisation. The reason therefore is that the increase of IR scores diminishes the more often the keyword occurs within the document or the anchor texts of inbound links to avoid spam by extensive keyword repetition. Thereby, the potentialities of classical search engine optimisation are limited and Pagerank becomes the decisive factor in highly competitive areas.

**Click popularity ranking algorithms**

Click popularity algorithms track how many users click on a link and stickiness measurement calculates how long they stay at a web site. Properly used and combined, this data can make it possible for users, via a passive feedback, to help search engines organize and present relevant results.

Click popularity is calculated by measuring the number of clicks each website receives from a search engines results page. The theory is that the more often the search result is clicked, the more popular the website must be.
For many engines the click through calculation ends there. But for the search engines that have enabled toolbars, the possibilities are enormous.

Stickiness measurement is a really great idea in theory, the premise being that a user will click the first result, and either spend time reading a relevant webpage, or will click on the back button, and look at the next result. The longer a user spends on each page, the more relevant it must be. This measurement does go a long way to fixing the problem with spoofing click popularity results. A great example a search engine that uses this type of data in their algorithms is Alexa.

Alexa’s algorithm is different from the other search engines. Their click popularity algorithm collects traffic pattern data from their own site; partner sites and also from their own toolbar. Alexa combines three distinct concepts: Link popularity, click popularity and click depth. Its directory ranks related links based on popularity, so if the web site is popular it will be placed in Alexa. The Alexa toolbar does not just allow searches it also reports on peoples internet navigation patterns. It records where people who use the alexa toolbar go. For example, their technology is able to build a profile of which web sites are popular in the context of the search topic and display the results sorted according to overall popularity on the internet.

Ranking bias
According to Kohelar (2003) US sites were much better covered than the others in the study sites from China, Taiwan and Singapore. The author
analyses the possible technical causes of the differences and concluded that the language of a site does not affect its coverage by search engines. However the visibility of the site, measured by number of links to it, affects its chance to be covered by search engines. He also concluded that coverage bias does exist due not to deliberate choices of search engines but occurs as a natural result of cumulative advantage effects of US sites on the web. Nevertheless, the bias remains a cause for international concern.

Table 3.2 Countrywise coverage of web pages (Kohelar, 2003)

<table>
<thead>
<tr>
<th></th>
<th>U.S</th>
<th>CHINA</th>
<th>SINGAPORE</th>
<th>TAIWAN</th>
<th>AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>87%</td>
<td>70%</td>
<td>56%</td>
<td>75%</td>
<td>72%</td>
</tr>
<tr>
<td>Alltheweb</td>
<td>83%</td>
<td>61%</td>
<td>50%</td>
<td>75%</td>
<td>67%</td>
</tr>
<tr>
<td>Altavista</td>
<td>80%</td>
<td>52%</td>
<td>41%</td>
<td>4%</td>
<td>44%</td>
</tr>
<tr>
<td>Average</td>
<td>83%</td>
<td>61%</td>
<td>49%</td>
<td>51%</td>
<td>61%</td>
</tr>
</tbody>
</table>

Silvestein showed that for 85% of the queries only the first result screen is requested. Thus inclusion in the first result screen, which usually shows the top 10 results, can lead to an increase in traffic to a web site, while exclusion means that only a small fraction of the users will actually see a link to the web site. For commercial oriented sites whose income depends on their traffic, it is in their interest to be ranked with the top 10 results for a query relevant to the content of the web site. This concludes that variation in user behavior in turn affects the ranking of the results.
Search engines make an unprecedented amount of information quickly and easily accessible - their contribution to the web and society has been enormous. Lawerence (2000) concludes another aspect of ranking bias as the “one size fits all” model of web search, which limits diversity, competition and functionality. The author suggests that an increased use of context in web search may help improve the ranking bias. As web search becomes a more important function within the society, the need for even better search services is becoming increasingly important.

3.3 Problem of relevance

Current search engines are unable to effectively filter the content in a way that provides a high level precision combined with high recall. The previous section discussed about the role of the phases of the search process in improving the relevance of the search results. This section analyses various other perspectives of improving the precision of search results.

3.3.1 Context problems

Large-scale search engines do not have access to most information about the context of the searchers, and return generic results for everyone. This is not always appropriate. According to authors Stephen Antony Zaman and Mundeep Singh Rehill (2001), context has direct relevance for web searching, since when people enter a query term they are searching on the meaning or semantics of that term in relation to their own context. The context has several layers and dimensions to it. For an individual, context is determined
by his/her current task, geographical location and socioeconomic background. The context of an organization is determined by its type whether business, educational, government and so forth.

Moreover, generic search engines often return results that lack relevance to the users search query, especially where ambiguous or general query terms are used. The paper “Web search your way” states that web search engines generally treat search requests in isolation. The results for a given query are identical, independent of the user, or the context in which the user made the request.

Sample of impact of context on precision

This is best illustrated with a generic query example. Consider the search engine query "information about cats" and consider how relevant search results will be in context of the following scenarios. (taken form Watson paper).

Scenario 1

Consider a veterinary student writing a term paper on animal cancer. In this case, the most appropriate resources would be on feline cancer, its diagnosis and treatment.

Scenario 2

In case of contractor working on a proposal for a new building the most likely reference may be to caterpillar corporation, a major manufacturer of
construction equipment, usually shortened to “Cat” by people in construction business.

**Scenario 3**

When a primary school student writing a paper about Egypt, the student would like to see information about cat mummies, laden with pictures and descriptions that are appropriate for a grade school student.

The search engines are not able to differentiate the need of different users since they are not provided with the context of the requirements. The search engines only provide ways of including boolean operators to refine a query to convey the requirements in an more accurate manner. This thesis focuses to provide context information in the search process to improve the precision of search results.

**Classification Of Context Problems**

*Relevance of active goals*

The goals of the user contribute significantly to the interpretation of the query and to the criteria for judging a resource relevant to the query, however are not included within the query itself.

*Word sense ambiguity*

The word sense of “cat” is different from the others in scenarios 2. The context of the request provided a clear choice of word sense.
**Audience appropriateness**

The audiences in each of the scenarios also constrain the choice of results. Sources appropriate for a veterinarian probably will not be appropriate for a student in a grade school.

Finkestein states that the search provides a better match to the users current needs than just relying on the user’s fixed personal profile. He also state that searches should be processed in the context of the information surrounding them, allowing more accurate search results that better reflect the users actual intentions.

**3.3.2 Problem of personalization**

Relevance of search results is typically person dependent, so personalization will become important in future search engines. Most current personalization approaches place the burden of constructing a user profile on the user. This approach is viable in some restricted domains, but it does not scale for internet search problems. In addition, users may be vocal in demanding what they want, but they often have difficulty defining what they actually need or how they behave (Henry Tirri, 2002).

According to previous research (e.g., Leake & Scherle, 2001, McGowan et al., 2002), one of the key reasons for low user satisfaction with the results of search engines is that the searching process is not personalized or context-sensitive. This causes two kinds of difficulty. First, different users have distinct goals and backgrounds and so will have different views of what
is relevant. Secondly, even a single user will have different needs at different times according to his or her currently active task. Yang et al. (2000) outlines three aspects to be improved in the third generation of search engines. First, search engines must have the ability to detect user conceptual intention; secondly, they must have the ability to allow the user to provide more information as required; and thirdly, they must be able to allow the user to have more sophisticated interactions with the system. All of these criteria suggest the importance of research towards personalized and context-sensitive search.

3.4 DISCUSSION

Pre-processing or post-processing of the context of keyword is carried out in the existing search process. However, context has not been used in actual search process (Jing Su and Mark Lee, 2003). The Personalised Search Service of Google allows the users to create a profile of the categories such as science, internet, mobile computing and astrology that the users are interested in. While searching, users are permitted to control the degree of personalization in search results using a slider. The Google personalization approach demands user interaction in creating the user profile. The degree of the personalization introduced is limited to the category of the user query and not the context of the search query, which has direct relevance on web search process.
The research work of McGowan et al (2002) on personalized information access aims to model users with multiple profiles, each corresponding to a distinct topic. The personal constructor component uses hierarchical clustering techniques over web page content. Once the persona is identified offline, it is utilized when the user is online. However, due to transient browsing behavior it is found that this poses several difficult user interface challenges.

This research work includes semantic information in the ranking process of the search results. The topic sensitive ranking proposed by Havilewala (2003) computes different pageranks for different topics. However, the topic-sensitive pagerank does not consist of millions of pageranks for different terms, but of a few pageranks for different topics. Haveliwala assigns a higher value to the pages of those Open Directory Project (ODP) categories for which the Pagerank is calculated. If, for example, Pagerank for the topic health is calculated, all the ODP pages in the health category receive a relatively higher value and the value is passed in the form of Pagerank on to the pages which are linked from there. This pagerank is passed on to other pages and, if it is assumed that health-related websites tend to link more often to other websites within that topic, pages on the topic health generally receive a higher pagerank. However, one crucial point in Haveliwala's work on topic-sensitive-pagerank is the identification of the user's preferences. Having a topic-specific pagerank is useless as long as the topic of the user's interest is not known. In the end, search results must be based on the pagerank that
matches the user's preferences best. The thesis proposes to incorporate semantics in all phases of search process.