CHAPTER 1

INTRODUCTION

1.1 INFORMATION RETRIEVAL ON THE WEB

The World Wide Web (WWW) is emerging as an appropriate environment for business transactions and user-organization interactions, because it is convenient, fast, and cheap to use. Now more than ever, users rely on the Internet for information and news. With well over 433,193,199 websites (Internet Domain Survey 2007) and billions of individual Web pages, finding high-quality information is increasingly challenging. With such a large collection of semi-structured and structured information sources, Web users often suffer from information overload. Efficient tools are required to manage, retrieve and filter this information effectively. Within large Intranets, these tools are also used to extract or infer new information to support decision-making within the establishment.

Early definitions for Information Retrieval (IR), dating from the 1960’s, emphasize the very general nature of the task.

‘Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information’

- Salton’s classic textbook on ‘Automatic Information Organization and Retrieval’
(Salton 1968)
‘Information retrieval is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational stand-alone databases or hypertextually-networked databases such as WWW’.

- Wikipedia (Information_retrieval 2009)

A typical IR system will have the following basic components: Content, Metadata, Search Engine and Queries as shown in Figure 1.1.

Figure 1.1 A Typical IR System

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information need, for example, search strings in web search engines. In information retrieval a query does not uniquely identify a single document in the collection. Instead, several documents may match the query, perhaps with different degrees of relevancy. Depending on the application the document may be text documents, images or videos. Often the documents themselves are not stored directly in the IR system, but are instead represented by document surrogates.
Most IR systems compute a numeric score on how well every document in the collection matches the query, and rank the documents according to this value. The top ranked documents are then shown to the user.

In the context of web search, a critical goal of successful IR is to identify which web pages are of high quality and relevance to a user’s query. Current IR systems place most of the burden on the users, relying on them to identify sources likely to contain relevant information, compose an appropriate query, and sift through retrieved pages to extract relevant information. Clearly, as web pages continue to grow, it will become impractical for users to perform these tasks for all but the simplest requests.

1.2 WEB SEARCH

IR encompasses many types of information access. Web search is an important part of this spectrum of information systems. Web search now represents a significant portion of Web activity. There are three forms of searching the Web (Lawrence and Giles 2000). They are:

1. Using search engines that index a portion of the web documents as a full text database.
2. Using Web directories which classify web documents by subject.
3. Searching the web by exploiting its hyperlink structure (shared context).

General web search is performed predominantly through text queries to search engines. Because of the enormous size of web, text alone is usually not selective enough to limit the number of query results to a manageable size. Search engines for the web are one of the most publicly
visible realizations of information retrieval technology. Search engines are critically important to help users find relevant information on the World Wide Web. For some search tasks (e.g., home page finding), systems such as Google (2009) provide highly accurate results. However, the Web contains more than just home pages and web users are interested in more than just finding single pages. Therefore, significant challenges remain for improving general Web search.

In order to best serve the needs of users, a search engine must find and filter the most relevant information matching a user’s query, and then present that information in a manner that makes it readily palatable to the user. Moreover, the task of IR and presentation must be done in a scalable fashion to serve the hundreds of millions of user queries that are issued every day to a popular web search engines such as Google Danny Sullivan of Search Engine Watch (Searches Per Day 2009) estimates that eight major search services serve up over 625 million search requests per day. The following section briefs the issues of web search.

1.3 PROBLEMS ENCOUNTERED IN WEB SEARCH

The difficulties encountered during web search falls into four categories (Lawrence and Giles 2000). They are: 1) Problems with the data itself, 2) Problems faced by the users trying to retrieve the data they want, 3) Problems in understanding the context of search requests and 4) Problems with identifying the changes in user’s information need.

1.3.1 The Data

Data is distributed on many computers and on different platforms. The data is volatile, i.e. computers and data are easily added or removed.
When domain names or file names change a problem exists with hanging links (30 percent of the Web changes every month). Web growth is exponential which causes scaling problems that are difficult to cope with. Data is generally unstructured and redundancy exists due to duplication of pages and sites. The quality of data must be questioned. The lack of editorial process on the Web results in false and invalid data, poorly written text, and documents full of errors. And finally, the Web is heterogeneous due to the different languages and alphabets that exist.

Search engines, individuals, commercial corporations, and others use web search agents to retrieve information from the Web on their behalf. A Web search agent is a program that automatically traverses the Internet using Web’s hypertext structure. The search agent can either retrieve a particular page or use some specified search algorithm to recursively retrieve all web pages that are referenced from a web page (Koster 1998). Most Web search engines, such as Alta Vista (2009) and Google (2009), employ crawlers as web search agents (Sullivan 2002).

The hypertext structure, distinguishes the Web from a traditional document collection. This affects the evaluation methodology and retrieval models. In other words, the independent relevance assumption is clearly violated due to the links; i.e. a page pointing to a relevant document can be regarded as partially relevant, even if it is not relevant by itself. The amount of data that the links carry is constantly being modified or removed. Retrieving the most recent versions of a page becomes an important task of every information retrieval system.

It is especially challenging to maintain freshness and coverage in a centralized search engine. The current approach is to have different re-visit frequencies for different types of pages/websites. There is something
inherently not very much appealing with waiting for a crawler to come around and pick up the new content before it can be ‘found’ by people and as the web grows the issues of freshness will get worse. Cho and Garcia-Molina (2003) and Brewington and Cybenko (2000) performed a number of experiments to discover how Web documents are updated. Their models and experiments indicate that web page updates can be modeled as independent poisson processes. Their experiments on the web indicate that on an average, a web page is changed once in every 10 days and 50% of all web pages are changed after 50 days (Cho and Garcia-Molina 2000).

Furthermore, the links between documents are constantly being established and removed. The dynamics of the link structure is important for search engines as long as link structure is an important component in ranking of search results. It is likely that using link structure in ranking creates a slow working positive feedback loop in the entire web to make popular sites even more popular at the expense of less known or new sites. A number of studies of link structure have been made in the literature (Broder et al 2000). However, these studies hardly cover the dynamics of link structure. The Hyperlink Induced Topic Selection - HITS algorithm (Kleinberg 1999) is a well known approach in information retrieval. According to HITS algorithm a node that has many outgoing links is a good hub, whereas a node that has many incoming links is a good authority. The number of links pointing to and pointing out of a node determines the node’s authoritiveness and hubness, respectively. However, the HITS does not account for the freshness of a web page.

1.3.2 The User

The problems of users during web search fall into two areas: 1) query specification, and 2) interpretation of search results. It is hard to
specify a query unless it is very simple. Many natural language questions, e.g., ‘What is a hard disk’, are submitted to search engines on the web every day, and an increasing number of search services on the web specifically target natural language questions. For example, AskJeeves (2009) uses databases of pre-compiled information, metasearching, and other proprietary methods, while services such as AskMe (2009) and Google Answers (2009) facilitate interaction with human experts.

Web search engines such as AltaVista (2009) and Google (2009) typically treat natural language questions as lists of terms and retrieve web pages similar to the original query. However, web pages with the closest match may contain few of the terms from the original query and be ranked low by the search engine. These queries could be answered more precisely if a search engine recognized them as questions. Consider the question ‘What is a hard disk?’, the best results for this query are probably not company websites of disk storage manufacturers, which may be returned by a general-purpose search engine, but rather hardware tutorials or glossary pages with definitions or descriptions of hard disks. A good response might contain an answer such as: ‘Hard Disk: One or more rigid magnetic disks rotating about a central axle with associated read/write heads and electronics, used to store data...’.

This shows that there is no proper existence of a conceptual relation(s) identified between the search queries and web pages by today’s search engines.

Generally, there are a large number of results returned from a query. The question of ranking the web pages for relevance and to efficiently browse large documents is another dimension to the same problem. Current web search engines are similar in operation to traditional information retrieval systems – they create an index of words within web pages, and return a ranked list of pages in response to user queries. Traditional database and IR
systems as well as two recent lines of database research inspired from IR, namely keyword searches (Florescu et al 2000; Agrawal et al 2002; Bhalotia et al 2002; Hristidis and Papakonstantinou 2002) and best-match query answering (Chaudhuri and Gravano 1999; Agrawal and Wimmers 2000; Hristidis et al 2001; Chomicki 2002) fall under this category.

1.3.3 Understanding the Context of Search Requests

Despite some recent attention to the above mentioned problem, little progress has been made due to the difficulty of capturing and representing knowledge about users, context, and tasks in a general web search environment. For web search to be more effective it must be context-oriented and incorporate query features to infer characteristics of the information need such as query type, answer type, and answer level, and use these characteristics in retrieval models to rank potential answers such as sentences, passages, documents, or combinations of documents.

In general, search engines decide the web page relevance using several factors. Google employs a number of techniques to improve search quality including page rank, anchor text, and proximity information (Lawrence et al 1998). Modified page ranks namely Topic Sensitive Page rank (Haveliwala 2002) and Probabilistic Page rank with content and link information (Richardson and Domingos 2002) were used to compute the page rank. These factors are ‘general’ in the sense that they do not consider personal preferences. In other words, given input keywords, the search output is identical for every user.

On the other hand if we collect information of each user, for example, his expertise, experience and web usages we can provide better search results. Web search is meaningful only in the context of a particular task and/or class of users. Greater focus on the user will enable major
advances in information retrieval technologies (Mobasher et al 1999). The focus may involve better modeling of the user’s background and type, personalization, the context in which information access occurs, the interactive information seeking process, explanation of what happened, results presentation, or combinations of all of those. Systems which employ web usage mining techniques to build a set of rules that represent the user’s interests (Ngu and Wu 1997; Mobasher et al 1999). Having discovered these rules the system can recommend new or updated web pages to the users according to their interests.

Thus with a context-aware web search system, if a user enters a query such as ‘Taj Mahal’, and if the user is at a desktop computer and has spent time earlier planning a conference trip to India (reading emails on the trip; examining travel web pages; placing entries in a diary), then the system will be aware of this context and will be more inclined to retrieve pictures and videos of the Indian mausoleum, while music from the jazz band with the same name would be less likely to be retrieved.

1.3.4 Identifying the Changes in User’s Information Need

Users of the web search often come up with different queries, which represent different information needs at different intervals of time. For example a user who has so far visited web pages speaking about cooking with apples, at one point of time might search for details regarding upgrading his/her personal lap top say, Apple make. Then the user might use queries containing the term ‘apple’. Now the user needs information regarding apple machines and not recipes involving the fruit apple. Identifying this change in the user’s information need will help to filter out irrelevant search results that are returned by any search engine. But keeping track of users’ information need is a very difficult task in a dynamic environment like the WWW where
the content as well as the users keeps changing. Tracking the users’ information and representing the tracked information as knowledge will become more difficult when it comes to multiple users.

One common solution to all the above-mentioned problems is personalization (Shahabi and Chen 2003), which customizes the web environment for users and helps them search their information need easily. Web personalization is currently known as the key to success in business today and in the future (Allen et al 2001).

1.4 NEED FOR PERSONALIZATION

Personalization is about tailoring information about anything, like products, web pages, services, to better fit the user. Personalization can be achieved by modeling the information either to suit large user groups, smaller interest groups, or the individual user. There exist several techniques to realize personalization like utilizing content and/or collaborative approaches, query refinement and personalized page ranking. In all these techniques the emphasis lies in various factors like focusing on the user needs, preferences, interests, expertise, workload, tasks etc. The roots of personalization of information systems can be traced back to the early adaptive user-interfaces, personal assistants/agents, and adaptive information retrieval. Relevant readings in these areas can be found in (Edmonds 1981; Schneider et al 1993; Myrhaug and Thomasen 1997). Most of the approaches started with users’ needs, preferences and expertise. Other approaches involve detecting patterns in user behavior when searching for information.

Personalized Information Retrieval cover a range of spectrum. At one end of the spectrum, there are filtering systems, which filter input from an information resource; from the input, information of possible interest is marked. Another end of the spectrum encompasses web search systems that
enhance user’s web experience by modeling user behavior to predict users’ interests better.

According to a Pew Internet and American Life Project research (Pew Internet 2009), web search is now the number-two activity for web users, Email remains the top internet activity for web users, and news is the third most popular web activity. In order to best serve the needs of users, a search engine must find and filter the most relevant information matching a user’s query, and then present it to the user in a palatable manner.

Traditional search engines obtain the results by keyword based techniques and finally retrieve thousands of result pages of which only a handful are relevant. Today, most search engines are not personalized, therefore, irrespective of the users and their information need, the results returned and their ranks are the same for every user. In addition, WWW data and the web users are so dynamic and it is difficult to capture the user preferences and interests without interfering with the normal activity of the user. With the cost of running a large scale search engine already very high, it is likely that such a full-scale personalization is currently too expensive for the major web search engines. Hence a layer of personalization above any existing search engine is a cost effective way of effective personalized web search.

The discussion so far, presented a complete scenario of what is further required to improve a user’s search through the World Wide Web from personalization viewpoint and this requirement in fact makes personalized web search a compelling area for research.

1.5 MOTIVATIONS OF RESEARCH

Personalization is effective only when it captures the conceptual relation between search query and its relevant pages. The relevancy of a page
must be evaluated from the user’s search context and user interests. The context and interest of user may be predicted from the content browsed and it may be further confirmed from the user’s behavior. In the process of building such an effective personalized search system, the various factors which motivated this research are as follows:

- Collection of user data without explicit user intervention.
- Integration of various factors that affect personalization.
- Representation and modeling of user data mathematically to aid effective page recommendations.
- Representation of web page content by utilizing conceptual representatives to aid context oriented search.
- Assigning weights to representative concepts, based on the factors of personalization.
- Re-rank the search results utilizing concept weights.
- Conceptually index search queries and its relevant web pages based on content features and factors of personalization.
- Tracking user actions and page view time, and thereby deriving user search behavior, from mathematical modeling of user search process.
- Identifying relevant and irrelevant pages from user search behavior.
- Filtering irrelevant search paths and irrelevant web pages from the results.
- Identifying and ranking semantic search paths which have lead to relevant search results.
- Providing semantic link between visited and unvisited semantic search paths and to aid recommendation of unvisited relevant pages.
• Tracking the changes in individual user’s interests and constructing interest based User Profiles.
• Prediction of user’s long and short-term interest.
• Utilizing individual user profiles in collaborative personalized page recommendations (i.e. user group personalization).
• Utilizing group user profiles in naïve user personalization.
• Evolve evaluation methodologies for personalized web search.

1.6 CONTRIBUTIONS

This thesis has contributed towards context-oriented search by exploiting the conceptual relation between search queries and contents viewed by the user. Such conceptual relations are quantified as User Conceptual Index (UCI) which provides a searchable index for the user data. This work provides an architecture that binds the various factors of personalization like hit counts, content viewed, page-view time, user actions, semantic link between pages, conceptual relation between search queries and contents viewed and semantic search paths to provide an effective and efficient web search. Moreover this work proposes a novel filtering methodology for filtering irrelevant pages by utilizing graph based techniques aided by implicit factors. Navigation based analysis is also performed for recommending shortest search path that lead to relevant pages. Such navigational analysis help to identify relevant pages that are missed by search engines. Additionally, the UCI-based search exploits user’s long and short-term interests by taxonomical information. Thereby this thesis attempts to eliminate the problem of recommendation to a new user through interest based user groups which is evolved by exploiting common interest among the users.
1.7 ORGANIZATION OF THESIS

The thesis is organized as follows: Chapter 2 gives an overview of classifications of web search and the related works and their implications. Chapter 3 gives a detailed view of the proposed integrated personalized search architecture and narrates about the various layers and their functionalities. This chapter on a whole provides an overall view of our work.

Chapter 4 explains the new search index called the User Conceptual Index (UCI) and its computation. This chapter also discusses about the utilization of implicit factors for filtering irrelevant pages based on graph based techniques. Chapter 5 discusses about the various semantic analyses which are used for search path analysis and ranking of semantic search paths. This chapter also explains the establishment of semantic link between the visited and unvisited pages in the form of conceptual graphs and how such pages are utilized for personalized page recommendations.

Chapter 6 illustrates about the user interest analysis by grouping the long and short-term user interests and usage of such interests to evolve interest based user groups for new user recommendations. Chapter 7 speaks about the UCI based page recommendation algorithms, sample search results and result evaluation. This chapter introduces novel interest oriented measures for evaluation and a comparative study of normal and UCI based search. Chapter 8 concludes the various contributions of our work and the possible enhancements that may be carried out in the near future.

The following chapter explains the classification of web search from a personalized and non-personalized viewpoint and then goes on to explain the construction of user profiles, the importance of query refinement and personalized page recommendations in re-ranking the normal search results to aid personalized web search.