Chapter 1

INTRODUCTION

1.1 GENERAL

World Wide Web (WWW) [1, 3], a huge source of hyperlinked documents containing useful information for millions of users, entered the information retrieval [3] world in 1989. This event caused the evolution of a branch of information retrieval that is different from traditional IR in the sense that it searches the required information within new document collection.

WWW can be broadly classified into Closed Web and Open Web [3]. Closed Web comprises high quality controlled collections on which traditional IR techniques can be fully applied while Open Web contains wide collection of documents on which traditional IR techniques concept and methods can not be directly applied and different IR tools (such as IR search engines) [2] are used to retrieve information.

An Information Retrieval (IR) system retrieves information about a subject from a collection of data objects. This is in contrast to Data Retrieval [6], wherein a collection of documents containing the user specified keywords are delivered. Information Retrieval can be precisely defined as:

"... the IR system must somehow 'interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This 'interpretation' of document content involves extracting syntactic and semantic information from the document text..." [5]

For management of relatively small and coherent collections such as newspaper articles or book catalogs several IR techniques were reported [6, 10, 11] even before the advent of World Wide Web (WWW). However, owing to unique characteristics of WWW, these techniques were found unsuitable and inefficient i.e. WWW is massive in size, much less coherent, changes more rapidly and spread over geographically distributed computers [2].

IR search engines [2] available on the Web, allow a user to submit queries and retrieve (usually ordered) links of relevant Web pages. Virtually all IR engines work by downloading and indexing the Web document collection(s) to which retrieval is to be applied. With a relatively small collection, it may be possible to generate the index terms for a given document dynamically when the collection is being searched. On the other hand, for a huge set of collections such as the information sources available through the Web, the search engines generate indexes in advance and due to dynamic nature of the Web, the indexes are constantly updated.

1.2 WEB SEARCHING

The typical design of a search engine [5] mainly comprises of three activities: Web crawling, Indexing and Searching as shown in Fig. 1.1. It may be noted that the Web crawler is the first stage that downloads Web documents which are indexed by the indexer for later use by searching module, with a feedback from other stages. Thus, it is a cascade model comprising of crawling, indexing, and searching modules.

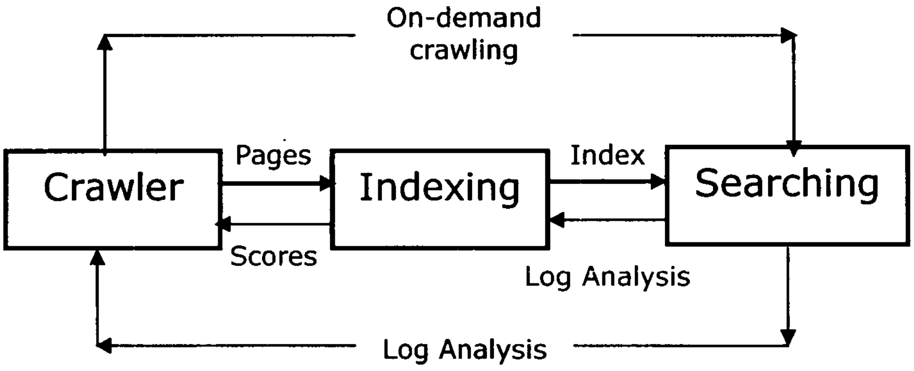

Fig. 1.1 Cyclic architecture for search engines

Besides simple indexing, the indexing module helps the Web crawler by providing information about the ranking of pages to make crawler more selective in downloading of high ranking pages. The searching process, through log file analysis or other techniques, optimizes both indexer and the crawler by providing information about *active set* of pages which are actually seen and sought by users. Finally, if required, on needfelt basis, Web crawler could provide on-demand crawling services for search engines.

1.3 MOTIVATION

The Open Web is a unique document collection owing to the properties that are listed as follows:

- Huge size,
- Dynamic,
- self-organized,
- Hyperlinked, and
- Low precision.

Huge Size The size of the web is so huge that it can't be measured accurately. In January 2004, it has been estimated that WWW contains about 10 billion pages, with an average page size of 500KB [10, 11]. The early exponential growth of the Web has slowed recently, but it is still the largest document collection in existence.

Broadly, WWW contains two type of Web i.e. Surface Web and Hidden Web [4]. Surface Web is the part of the Web which is searched and indexed by the typical search engines and also called Publicly Indexable Web (PIW) whereas the Hidden Web also called Invisible or Deep Web contains very large and wide range of publicly accessible databases that is not searched or indexed by the search engines. To access Hidden Web, a user must request for information from a particular database through a search interface. The hidden Web dynamically generates the relevant pages for the user without keeping any copy. As a result, typical search engines cannot find these dynamic pages. BrightPlanet [16] sells access to the Hidden Web that is estimated to contain over 92,000TB of data spread over 550 billion pages.

Dynamic nature WWW contains both static and dynamic collection [12] of documents. Static documents are like books in a bookshelf that don't change their content, but dynamic web pages are changed very frequently. A study by Junghoo Cho and Hector Garcia-Molina [13] in 2000 reported that 40% of all webpages in their dataset changed within a week, and 23% of the .com pages changed daily. A recent study by [14] reported that 35% of all webpages changed during their study. It has been found that pages in larger in size are changed more frequently and more extensively than smaller size webpages.

Self-Organization Traditional document collections were usually collected and categorized often by highly paid professionals. However, anyone can post a webpage on the Web. There are no standards and no gatekeepers policing content, structure, and format. The information is also volatile; there are rapid updates, broken links, and file disappearances. A U.S. study reporting [15] on "link rot" suggested that up to 50% of URLs cited in articles in two information technology journals were inaccessible within four years. The data is heterogeneous, existing in multiple formats, languages, and alphabets and this volatile, heterogeneous data is posted multiple times. In addition, there is no editorial review process to find the errors, falsehoods, and invalid statements. Moreover, this self-organization opens the door for spammers who capitalize on the mercantile potential offered by the Web. *Spammers* were the name originally given to those who send advertising emails to thousands of users within a second. With web search and online retailing, web spamming includes those using deceiving webpage creation techniques to rank highly in web search listings for particular queries. Spammers usually use minuscule text font, hidden text (white on a white background), and misleading metatag descriptions to fool early web search engines [27, 30, 37] (like those using the Boolean technique of traditional information retrieval). The self-organization of the Web also means that webpages are created for different purposes. Some pages are aimed at users who are shopping, others at users who are researching. In fact, search engines must be able to answer many types of queries, such as transactional queries, navigational queries, and informational queries.

Hyperlinked documents This linking feature, the foundation of Vannevar Bush's memex, is the saving grace for web search engines. Hyperlinks [37] make the new national pastime of surfing possible. But much more importantly, they make focused, effective searching a reality. Web search engines exploit the additional information available in the Web's extensive link structure to improve the quality of their search results. However, the advantages resulting from the link structure of the Web did not come without negative side effects. The most interesting side effects concern those sneaky spammers. Spammers soon identified the link analysis employed by major search engines, and immediately set to work on link spamming. Link spammers carefully craft hyperlinking strategies in the hope of increasing traffic to their pages. This has created an entertaining game of cat and mouse between the search engines and the spammers, which many, the authors included, enjoy spectating.

Low Precision An additional challenge in information retrieval from the Web is its low precision. As users does not look beyond the first 10 or 20 documents retrieved. Therefore, user's ability to look at documents does not increasing as rapidly as the accessible information on the Web. This means that search engine precision must also increase as rapidly as the number of documents on the Web is increasing. Another problem related to web search engines is about their performance measurements and comparison. While traditional search engines are compared by running tests on familiar, well-studied, controlled collections, this is not feasible for new search engines.

1.4 HIDDEN WEB (Deep or Invisible Web)

In the recent years, the exponential growth of the information technology has led to large amount of information available electronically and the largest repository of this information is the WWW. As the size of the Web continues to grow, searching it for the useful and relevant information has become more difficult. Broadly, WWW can be classified into two parts: *Surface Web* and *Deep Web*. The *Surface Web* is that portion of the World Wide Web that is indexed by conventional search engines. It is also known as *Publically Indexable Web (PIW).*

On the other hand, a large amount of information on the web today is available only through search interfaces (forms) [22, 26, 32, 57]. Hence a lot of data is remaining invisible to the users. For example, if a user wants to search information about some flight, then in order to get the required information, he/she must go to airline site and fill the details in a search form. As a result he/she gets the details of the flights available. These types of pages are often referred to as hidden pages and the part of the Web that is not reachable by this way is called Hidden Web. Typical, Search engines cannot discover and index such pages as they have no static links. However, according to recent studies [28, 33, 67], the content provided by many Hidden Web sites is often of very high quality in term of relevance and therefore, very valuable to many users.

BrightPlanet has quantified the size and relevancy of the deep Web in a study based on data collected between March 13 and 30, 2000 [4]. Their findings include:

- Public information on the deep Web is currently 400 to 550 times larger than the commonly defined WWW.
- The deep Web contains 7,500 terabytes of information compared to nineteen terabytes of information in the surface Web.

- The deep Web contains nearly 550 billion individual documents compared to the one billion of the surface Web.
- More than 200,000 deep Web sites presently exist.
- Sixty of the largest deep-Web sites collectively contain about 750 terabytes of information -- sufficient by themselves to exceed the size of the surface Web forty times.
- On an average, deep Web sites receive fifty per cent greater monthly traffic than surface sites and the typical deep Web site is not well known to the Internet surfers.
- The deep Web is the largest growing category of new information on the Internet.
- Deep Web sites tend to be narrower, with deeper content, than conventional surface sites.
- Total quality content of the deep Web is 1,000 to 2,000 times greater than that of the surface Web.
- Deep Web content is highly relevant to every information need, market, and domain.
- More than half of the deep Web content resides in topic-specific databases.
- A full ninety-five per cent of the deep Web is publicly accessible information without any fees or subscriptions.

It is estimated [4] that larger search engines like *Google* and *Northern Light* each index about 16% of the surface web and internet users, therefore, searching only 0.03% or one in 3000 of the pages available to them. Thus, there is need to search surface as well as hidden web resources for comprehensive information retrieval.

1.5 CHALLENGES OF HIDDEN WEB CRAWLING

Crawlers are programs that traverse the Web and automatically download pages for search engines. Conventional crawlers today rely mainly on the hyper-links on the Publically Indexable Web to discover and download pages. Due to the lack of links pointing to the Hidden-Web pages, the current search engines cannot index the Hidden-Web pages. The major challenges in designing a Web crawler to crawl the hidden web are listed as follows:

- **Identification of *Entry Points* for Hidden Web.**

  The hidden Web resources allow the users to access the underlying information by querying through their query interfaces containing attributes that tend to describe the information accessible through them. For example, the query interface of a source like amazon.com contains attributes such as author, title, ISBN etc. Infact, these query interfaces acts as the entry points to the hidden or invisible web and therefore, become potential candidates for possible extraction of hidden data by specially designed crawler. Hence, to implement an effective Hidden-Web crawler, automatic identification of the search forms to generate *meaningful* queries is needed so that it can discover and download the Hidden-Web pages.

Solution: *Since the search interfaces acts as the entry points for the hidden web contents, a mechanism that automatically extract domain-specific search interfaces has been proposed in this work.*

- **Large number of Domains:** The search forms are distributed widely over large number of domains like *Books, Airline, shopping* sites on the World Wide Web, therefore, it is very difficult to collect the search forms of different domains and process them simultaneously.

Solution: *A domain-assisted approach for crawling has been developed. It helps in identifying the domain and in finding the relevant information thereof for the Hidden web.*

- **Integration of search forms of the same domain.** *Let us consider the two search interfaces shown in Fig. 1.2. it may be noted that both the search interfaces belong to the same domain but they contain the fields which may however be synonyms, hypernyms or meronyms rendering interface integration a critical problem in many application domains, such as semantic web, data warehouses, e-commerce etc.*

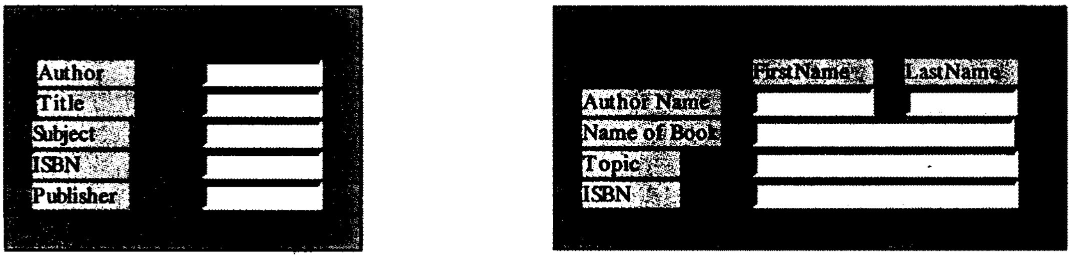

Fig. 1.2 Simple query interfaces from book domain

Thus, there is need to identify the meaning of the fields present in the search interfaces and integrate them efficiently into a Unified Search Interface.

Solution: *A novel framework for interface integration is being proposed. The proposed Interface integration involves two steps: Interface Matching and Interface Merging. The "Interface Matching" employs an "extensible domain-specific Library" for quickly identifying regions in the interface repository comprising of important mappings while "Interface Merging" merges all the interfaces into a Unified Search Interface (USI) based upon the semantic mapping identified in first step of Interface Integration by merging two interfaces at a time.*

- **Automatic Query Generation.**. Hidden Web sources store their content in searchable databases that only produce results dynamically in response to a direct request. But a direct query is a *one at a time* laborious way to search. Therefore, a mechanism need to be devised that generate the intelligent queries (to be submitted) automatically for the extraction of hidden web contents.

  Solution: *To fill the USI automatically, a Domain-specific Data Repository has been employed. It contains labels and their corresponding values. In this work, Data Extractor Engine and Search Interface Parser are used to create Domain-specific Data Repository which is further used for automatic query generation..*

1.6 ORGANIZATION OF THESIS

The topics covered in this thesis are shown in Figure 1.3. The topics are entangled, i.e., there are several relationships that make the development non-linear.

This research process has been linearized to present it in terms of chapters, but this is not the way the actual research was carried out: it was much more cyclic and iterative than the way it is presented here. The following is an outline of the contents of this thesis. The first chapters explore theoretical aspects of Hidden Web and challenges in crawling the Hidden Web:

- Chapter 2 reviews selected publications related to the Web search covered in this thesis. This chapter contains the general architecture of search engine and strengths and weaknesses of the most commonly used search engines like Google, Altavista, and Yahoo etc.
- Chapter 3 reviews the selected publications related to Web Crawling and Hidden Web. Different surface Web crawling techniques have been discussed. The need to crawl the Hidden Web has been identified and several issues of crawling the hidden web are discussed.

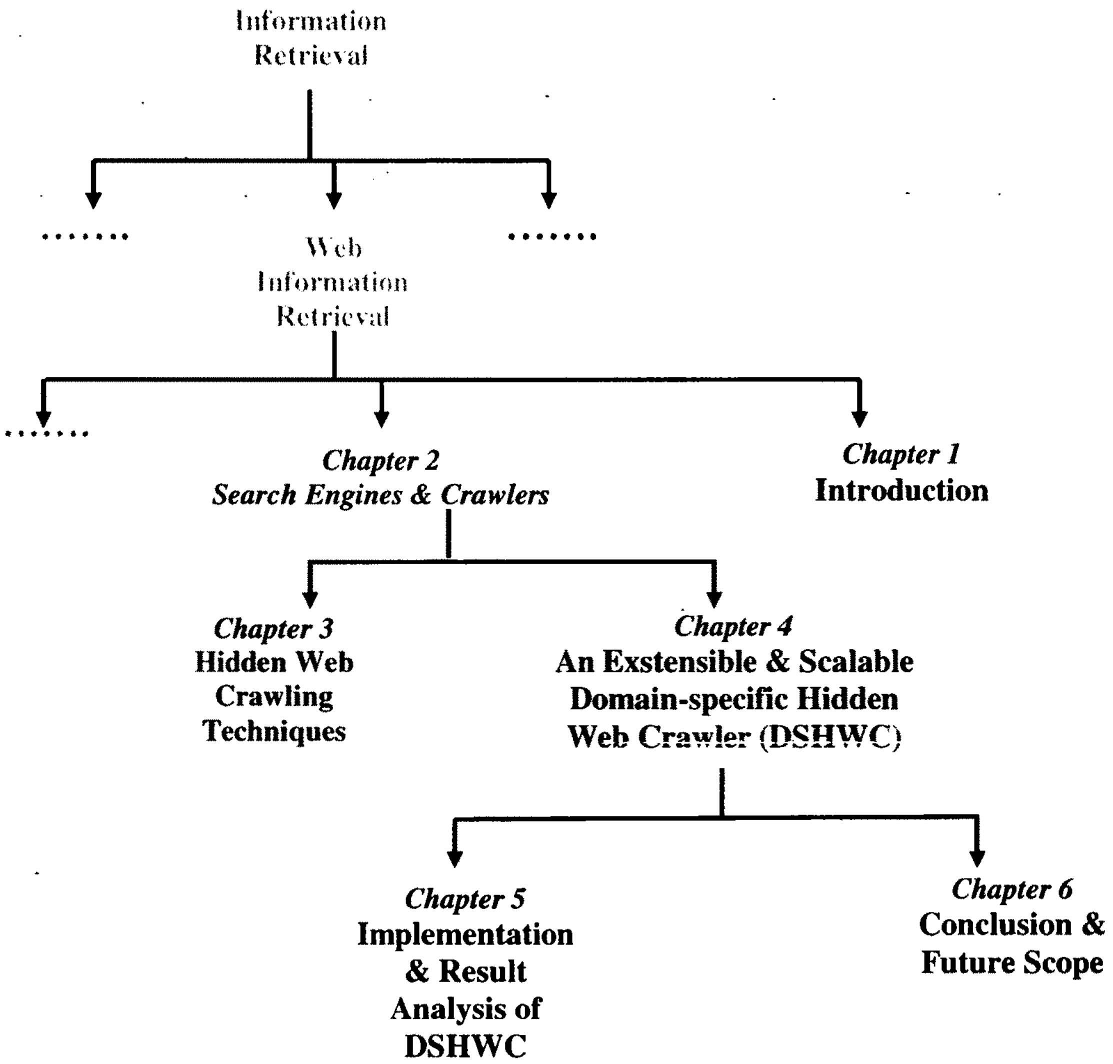
Information Retrieval
....
....
Web Information Retrieval
....
Chapter 2 Search Engines & Crawlers
Chapter 1 Introduction
Chapter 3 Hidden Web Crawling Techniques
Chapter 4 An Exstensible & Scalable Domain-specific Hidden Web Crawler (DSHWC)
Chapter 5 Implementation & Result Analysis of DSHWC
Chapter 6 Conclusion & Future Scope

Fig. 1.3 Main topics covered in this thesis. Web crawling is important in the context of Web Information Retrieval (IR)

- Chapter 4 proposes the architecture of a Domain-specific Hidden Web Crawler (DSHWC) which is able to crawl the hidden contents on the Web. The working of

DSHWC has been divided into five important phases. All the five phases and algorithms used in each phase of the hidden web crawler has been discussed in detail.

- In Chapter 5, the various important issues in implementation of Domain-specific Hidden Web Crawler (DSHWC) have been discussed. The Domain-specific Hidden Web Crawler (DSHWC) has been implemented using .NET technology. Various experiments over five domains were conducted to check the validity of DSHWC and results found are very promising. This chapter also shows the snapshots and results of proposed Domain-specific Hidden Web Crawler. The results for all the five domains have been analyzed and various standard metrics like *Precision*, *Recall* and *F-measure* were employed to measure the efficiency of DSHWC. The chapter shows the analysis of results found in the form of graphs and shows the high performance of our crawler. The analysis of the results establishes the high performance of DSHWC, the hidden web crawler.
- Finally, Chapter 6 summarizes our contributions and provides guidelines for future work in this area.
- In Appendix A and Appendix B, the results for three domains i.e. Airline, Automobile and Electronics domains are provided.
- Finally, the bibliography includes references to publications in this area.

A survey of existing search engines and crawlers is given in next chapter.