In this article, we describe Web Information Retrieval (WebIR) and issues in searching the World Wide Web from a user perspective. We review different kinds of search engines and the features they offer. With the help of examples, we show how these features can be utilized to narrow down on the search results.

Introduction

The world wide web (WWW) has become so commonplace that it is an integral part of most societies today. Millions of people around the world routinely search for information and conduct transactions using the web. Without any doubt, the WWW has been one of the fastest growing and the most pervasive phenomenon in history.

The web began its existence around 1989 with a proposal called ‘Information Management: A proposal’ by Tim Berners-Lee, a researcher at CERN – the European organization for nuclear research in Switzerland. The proposal envisaged a network of remote documents linked physically using the internet and logically using hypertext. The proposal was called the ‘World Wide Web’. In response to this proposal, Berners-Lee’s boss had this remark: ‘Vague, but exciting’. A little more than ten years later, this small idea has grown to mammoth proportions beyond imagination. As of July 2001, the web contained 125,888,197 hosts (Source: Internet Software Consortium http://www.isc.org/) serving almost 1 billion users around the globe. The vagueness and excitement today lies not just in creating documents and interconnecting them over the web; but in extracting required documents from the web.

Keywords
Web searching, web information retrieval, WWW, search engines, directories.
Web Mining or Web Information Retrieval (WebIR) is the process of extracting useful information from among the petabytes\(^1\) of data that make up the WWW. WebIR has a history almost as long as the web itself. One of the first efforts towards WebIR was the creation of directories of web sites. An example is yahoo.com. Web directories manually organized contents of the web into a taxonomy of categories and sub categories. Web site owners who wished to be listed in the directory had to submit their site to the directory for perusal and inclusion.

This rudimentary approach towards WebIR was replaced by more sophisticated mechanisms like crawler based search engines. These search engines maintained an index of words and phrases and a list of all their occurrences in the web. This index was automatically generated by programs variously called as ‘spiders’ , ‘crawlers’ and ‘scooters’. These agents periodically went around the web collecting and indexing web pages.

There are many shortcomings to both the above approaches. As a result, research in WebIR is a contemporary and pertinent topic with many open questions. A number of alternative approaches are still being pursued for increasing the effectiveness of web searches. Many of them are conceptually elegant, but have shown less than satisfactory results in practice.

In this article, we provide an introduction to the topic of WebIR. We can look at WebIR along two broad perspectives: (a) the perspective of the web user, and (b) the perspective of the WebIR engine. In this article, we shall concentrate mainly on the user perspective and address the WebIR engine only briefly. Specifically we shall look at different kinds of search mechanisms that are available to the user; and how these mechanisms can be used to conduct effective web searches.

**Types of Web Users**

Web users can be broadly divided into three kinds based on their search strategies. These are: (a) a casual user searching the web for something that is loosely defined (b) a researcher look-
ing for serious research level content over the web and (c) a professional looking for business intelligence by searching the web. As we see in the subsections that follow, the search strategies of each of the above user types differ substantially.

Regardless of what a user is looking for, the web user faces a daunting task of finding relevant information over the web. Part of the complexity comes from the size of the web. Crawler based search engines can today cover only 30-40% of the web [1]. Added to this, the web is growing at a rate faster than the progress in the technology of web crawling. But the complexity of searching the web is not limited to size. The intangible nature of the web also adds to the user’s burden. Unlike a database, the web has no single logical structure. It is not possible to discern the web’s structure by simply browsing through the web. Given this, typical strategies that are adopted by the three classes of users are explored below.

Casual User

The casual user searches the web for general information to satisfy his/her curiosity. The operations of the casual user are in the form of browsing the web starting from some arbitrary location; or in the form of queries over a web search engine.

Usually for any keyword search, the user is likely to obtain vast amounts of matches. The less carefully the user’s query is defined, the more likely are the chances of the user getting inundated with information. For obtaining the desired results it is important even for the casual user to work at precisely formulating the query.

For example, suppose the user is looking for information about the history of the Vijayanagar empire. One of the first tasks of the user is to choose the correct search engine. It might seem intuitively obvious that a search engine like khoj.com or searchindia.com which are India-specific would be more suited than an international search engine like google.com.
However, a search for the term *Vijayanagar* on these search engines tells a different story. Google.com returned 9000 pages, of which the first few pages are about the Vijayanagar empire. On the other hand, khoj.com returned 4 matches where the first page is about ‘Temples of Tamilnadu’; and searchindia.com returned just 1 result which was on the ‘Vijayanagar Thermal Power Station’.

One of the reasons for the disparity is the difference in the amount of data stored in each of the search engines. Also, an Indian search engine is more likely to contain *utilitarian* information (like information about a locality called Vijayanagar) which in this case is not the desired result. On the other hand, an international search engine is more likely to archive pages about *history*. The essential point here is that the user should choose a search engine based on the *context* in which the query has to be answered.

Most search engines can search for phrases rather than single words. Phrases are usually specified by embedding them within double quotes. For example, a search on the Vijayanagar kingdom can be made more specific by a search term called ‘Vijayanagar kingdom’ rather than just *Vijayanagar*.

Some search engines like AltaVista provide case sensitive search mechanisms. In AltaVista, capital letters match only capital letters while small letters match both capital and small case letters.

AltaVista also used to provide an option to the user to refine his/her query based on the query results. The refinement process suggests words that are ‘close’ to the search term which would help the user in directing the search in right directions. The user is asked to either include or exclude other terms in addition to the original query. AltaVista has now discontinued its ‘refine’ feature. We have mentioned the feature anyway for the sake of completeness.

Other search engines like Google, provide an option for searching within the search results. By using this, the user can progres-
sively compress the search space and increase result quality.

For example, suppose the user were indeed searching for the locality called ‘Vijayanagar’ in Bangalore and not the kingdom Vijayanagar. A google query *Vijayanagar* returns pages that mainly pertain to the Vijayanagar kingdom. In order to refine the search, the user may choose to search within the query results and provide *Bangalore* as a search term. As our search shows, this indeed returns pages pertaining to the locality called Vijayanagar as the first few results.

**Researcher**

The researcher typically looks to the web for information to help in his/her research. The search is more serious than that of a casual user, and is often augmented by other activities like annotations, bookmarking, etc. on the user’s end.

Search provided by the researcher is usually more complicated than a simple keyword search. It is necessary for the researcher to be much more precise in formulating search terms. For example, suppose the researcher is looking for literature about IP tunneling in computer networks. The term ‘tunneling’ is quite general and may denote any of the following: technology for building tunnels, tunneling of data packets in the internet protocol (IP tunneling), or characterization of the movement of electrons in super conductors. A small survey over search engines over this term proves interesting. Google.com returned about 3,22,000 pages in response to the search term *tunneling*. The first three pages respectively were about: Scanning tunneling microscopy, Quantum tunneling and IP tunneling. AltaVista found about 91,500 pages, out of which the first three were respectively about: IP tunneling, Tunneling in public works department, and a page on heavy construction material. Excite.com search was similar to AltaVista in terms of the number of documents returned, and the first three results were exactly as that of AltaVista.

It is necessary for the researcher to provide enough qualifiers
It is necessary for the researcher to provide enough qualifiers and context information to the search term in order to direct the search. The search may include phrases or case sensitive searches, but clearly, much more is needed than phrases or case sensitive search. Search engines provide many other features which are much more relevant to such requirements. Some of these features are described below.

**Logic expressions:** Many search engines accept logic expressions of search terms. The user may specify keywords like AND, OR or NOT to denote logical operators. In addition, many search engines offer modalities like 'include' or 'exclude' of search terms. For example, the search term *tunneling* +IP *electrons* indicate a search on the term 'tunneling' which should return documents that not only contain the term 'tunneling' but should also contain the term 'IP' and should not contain the term 'electrons'. The above query returned about 82,900 documents from google.com out of which all of the results in the first few pages pertained to IP tunneling.

**Search scope:** Some search engines enable the user to define the search scope for the query. For example, a query term like *site:http://www.acm.org/* would search for keywords only in pages that are contained within the acm.org domain.

**Stemming:** Some search engines also provide 'stemming' queries. Stemming queries search for all stems of a given keyword. For example, an AltaVista query *electro* searches for all endings of the word 'electro' like electron, electronic, electronics, electrolyte, electrostatic, etc. The user can use a general stem as a first level query, and use facilities like searching within query results or query refines to obtain the desired information.

In addition, much of the content a typical researcher needs, are not directly searchable by web search engines. They are hidden in databases behind web servers. The web documents that the user retrieves would be dynamically generated by querying these databases. Some examples of such databases are: CORA (cora.whizbang.com), ACM digital library (www.acm.org/dl/),
arXiv technical report archive (www.arxiv.org), etc. Such databases are also called the ‘invisible’ web. Techniques for extracting information from the invisible web are quite different from conventional web searches. The web site www.invisible-web.net maintains a directory of a large number of databases that can be searched over the web.

In recent times, there have been some innovations in search engine technology that are especially suited for a researcher. Significant among these is the ‘Vivisimo’ search engine (www.vivisimo.com). This search engine not only searches for web documents matching the query, but also *categorizes* search results into a concept hierarchy. Building the concept hierarchy and assigning a label to each grouping is done completely automatically.

A query called *tunneling* on vivisimo.com returned the following categories: Scanning tunneling (30), Tunneling protocol (13), Mining (14), Trenchless technology (11), Proxy (8), Construction (13), Quantum (16), Security (13), History (17), and Boring(9). The numbers in parenthesis indicates the number of subcategories in each of the categories. Figure 1 shows a screen shot of a Vivisimo search.

**Business User**

A business user searching the web for business intelligence typically looks for answers to questions like the following: Who are all my competitors?, What is the market potential for our new suite of products?, etc. Such a user requires more than search results from a search engine. Information required by such a user has to be extracted using inference mechanisms from results of conventional web searches.

Such a requirement is still rather lofty to expect from current web technology. Currently, the web is a collection of data and web crawlers do not infer anything from the way these data elements are linked. For the needs of the business user, WebIR needs to be intelligent. It not only has to search data, but has to
reason about the searches and perform inferences. These goals sound similar to the goals of Artificial Intelligence (AI) of the 80s. AI proved to be elegant in theory, but fell short of expectations in practice. Nevertheless, the pursuit for creating knowledge networks has not ended. Many proposals have been made in this regard. Predominant among them is the concept of a ‘semantic web’ (www.w3.org/2001/sw/) [1].

The semantic web builds upon the current WWW and establishes a knowledge web. Knowledge artifacts that are connected in this web can be used by web crawlers or agents to perform inferences and offer more sophistication in WebIR. The semantic web is based upon a markup language called the Resource Description Framework (RDF).

In addition to semantic web, various other strategies that build upon search engine technology, have also been proposed to meet
demands like that of the business user. None of them have proved revolutionary, but a few are worthy of mention. Prominent among them is the ‘meta-search engine’ that searches over the results of multiple search engines. Meta-search engines also offer a natural language interface to the user to formulate a query. An example is AskJeeves.com (www.askjeeves.com). Suppose a business user wants to survey the book publishing scene in India. A query over AskJeeves can be formulated very simply by saying ‘Who are the major book publishers in India?’ The search then proceeds in two directions: One part of the search looks for similarities between the query and other questions already stored in the search engine. Another part of the search processes the query and searches the web for documents that might possibly answer the question.

Another approach by a company called About.com utilizes human experts or guides to help the user in locating the required information. About.com consists of human experts who have expertise in more than 700 topics. They distill information about their topic from the web to make up the About search database. They also guide individual users on their more complex requirements.

Another approach that deserves mention is ‘web cartography’ as implemented in the VisIT search engine. VisIT displays logical maps of web contents rather than pages of search results. It was developed at the University of Illinois, Urbana-Champaign (www.visit.uiuc.edu/). In response to a query, VisIT shows search results in a ‘search space’, which allows the user to quickly scan hundred of hits at once. Arrows are drawn showing who is referencing whom in the search space. The user can navigate through the search space and mark areas of search results rather than individual pages.

The Search Engine

At the heart of a web search is the search engine or the WebIR engine. In this article we have concentrated mainly on the user

Suggested Reading

Box 1. The Perspective of the WebIR Engine.

WebIR engines may be classified into three types: directories, search engines and meta-search engines. Directories use (human generated) taxonomies to classify web contents and perform search. Search engines use automated crawlers to index documents from the web. Meta-search engines use results of existing search engines and perform further filtering.

This sidebar briefly describes only the search engine issues. The main issues that confront search engines are two fold: (a). how to control crawling over the web, and (b). how to rank web pages for their relevancy. Some of the techniques used are introduced here; the reader may want to consult some of the following references for more details [3, 4, 5].

Controlling web crawls: Web crawls require agent based programs to visit pages over the web and index them. Without some form of control, the crawl may go out of hand resulting in a large number of uninteresting pages. One of the techniques to determine which direction the crawling should proceed is to look for ‘hubs’ in the web. Hubs are web pages that have a large number of outgoing links. These web pages are likely to be some form of directories created by web page owners. Other methods of controlling web crawls include crawling depth first within a given domain before taking up links that point out of the domain. This ensures that the crawl finishes indexing as many pages as possible of every domain it visits.

Ranking web documents: Every search engine uses some form of ranking mechanism to rate web documents. These mechanisms are often strictly confidential in order to prevent web page owners from manipulating their pages to get them a higher ranking.

Some of the early ranking mechanisms ranked a document based on the number of words in the document that matched a given query. Subsequent variations to this approach included the following: ranking based on the position of the matched words in the document (a match near the top of the document gets a higher rank than one that matches further down in the document); ranking based on which area of the document matches are found (matches in the title or URL have a higher ranking than matches in the body).

The above mechanisms can be easily exploited by web site owners to mislead a search engine to erroneously give a higher rating to their page. A fairly tamper-proof mechanism is now used by the search engine Google for ranking web documents. This ranking is based not on the contents of a document, but on how a document is referred to by other documents over the web. A hyperlink from document A to document B is considered as an endorsement of B from A. Web documents that have a large number of links pointing to them are called ‘authorities’. They are likely to be pages of interest during a web search. Ranking of web pages may be either static or dynamic. Static web ranking assigns static ranks to web pages. In the google ranking mechanism, static ranking would mean that authorities are ranked highest. Dynamic ranking involves ranking web documents in the context of a query. Dynamic ranking slows down the search process, but would return more accurate results.

Box continued...
Clustering: Clustering entails grouping documents based on their similarity. Clustering is important to narrow down the scope of a web search. Some of the initial clustering mechanisms grouped documents based on their content. The mechanism used here is called ‘document vectorization’. A document is considered to be a bag of words. Based on the words that make up the document, the document is mapped to a point in hyperspace. Clustering is then carried out based on the distances between points in this space.

Google uses a different kind of clustering mechanism based on the linking structure of the web. Here, clustering is based on concepts of citation and cocitation. In the citation model, if document A refers to document B, then A is said to cite B and are clustered into the same group. In the cocitation model, if document A refers to documents B and C, then B and C are said to be cocited and are grouped together.

Table 1. Comparison of different search engine features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Google</th>
<th>Alta Vista</th>
<th>Lycos</th>
<th>Searchindia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-sensitive search</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Search within results</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Phrase search</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Include, exclude phrases</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Stemming</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Logic expressions</td>
<td>Limited (auto-</td>
<td>Yes (AND, OR,</td>
<td>Yes (AND, OR,</td>
<td>Limited</td>
</tr>
<tr>
<td></td>
<td>matic AND</td>
<td>NOT, NEAR)</td>
<td>NOT)</td>
<td>(use “+A or B” to</td>
</tr>
<tr>
<td></td>
<td>of search terms,</td>
<td></td>
<td></td>
<td>prefer a page containing A and B over one containing only A)</td>
</tr>
<tr>
<td></td>
<td>and OR operator)</td>
<td></td>
<td></td>
<td>url: (search within url)</td>
</tr>
<tr>
<td>Search scoping</td>
<td>Many qualifiers</td>
<td>Limited</td>
<td>Word filtering</td>
<td>mailto: (search for contributions by user with email address)</td>
</tr>
<tr>
<td></td>
<td>(site; intitle;</td>
<td>(domain:)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>link; inurl;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>etc.)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

Web Information Retrieval is a pertinent topic of research for the present day. As the web keeps growing in size, the problem of searching the web becomes only more complex. There are a number of innovative approaches that have been proposed which hold promise. However, it remains to be seen which approach finally becomes the norm, and to what extent the web is actually used by users. The kind of WebIR technology that develops over the future would determine whether the web is destined to be a large storehouse of largely unstructured data, or is actually a huge knowledge network that offers insight to people all over the world.