INTERNET Challenges for Information Retrieval[1].

Jeremy Ellman MARI Computer Systems
Wansbeck Business Park
Ashington NE 63 8QZ
Email : Jeremy.Ellman@mari.co.uk.
and
John Tait School of Computing & Information Systems
University of Sunderland, UK.

Abstract

This paper examines the problem of Internet resource discovery from the point of view of Information retrieval. It firstly discusses the traditional Information Retrieval task, and then considers current approaches to Internet Information Retrieval. These approaches give rise to problems such as Information Burial, which are defined and discussed. Finally, a model is proposed that partially circumvents some of these problems.

1) Introduction

The Internet World Wide Web has experienced exponential growth since its inception in 1993 (Berners-Lee et al 1994). It represents (amongst other things) a huge information resource (Schwartz et al 1993, Obraczka et al. 1993) and dynamic collection of documents. This presents several opportunities and problems to "traditional Information Retrieval".

This paper concentrates on the issue of textual information retrieval. Whilst it is possible to find graphics, video, and audio files on the Internet, multimedia information retrieval is generally at an early stage. Consequently there are fewer possibilities to exploit that work within the Internet networked environment.

This paper is structured as follows: Firstly we examine the nature of the traditional Information Retrieval task. Then we consider current approaches to Internet Information Retrieval. Next problems that the Internet raises for the application of Information Retrieval concepts are considered. Finally, a model of Internet Information Retrieval is then proposed that partially circumvents these problems.

[1] In Proc. BCS Information Retrieval Specialist Group 1996

2) The Information Retrieval Task.

This section crudely summarises the nature of the Information Retrieval (IR) task (Salton 1989 and many others) as it is traditionally understood. Its purpose it to highlight several assumptions that are intrinsic to IR that are not immediately applicable to Internet resource discovery. Once these concepts are identified, they can be replaced with broader, more robust approaches.

In general the traditional approach to IR might be summarised as follows:-

① There a large collection of documents
② A searchable index is created from this collection
③ A user has a fixed requirement for information
④ By querying the index the document that contains the Information can be found.

2.1) Implicit Assumptions

There are several assumptions implicit in the above view of IR. Although they are obvious, it is worth raising them, since I shall argue below that they do not always hold when applying IR to the web.

2.2) The Index Assumption

The role of the Index is of course key to Information Retrieval. It is built by processing the document collection, and storing key information about each document. Thus, when a user's query references some more or less particular information, they can be sure that an appropriate index entry will be found. This is a natural consequence of the implicit assumption that the index has referenced the entire document collection.

2.3) Vocabulary Problem

The next implicit assumption concerns what I will call the vocabulary problem. This problem arises when the user’s query does not reference the exact keyword in the index (or something morphologically related to it), but has used any other term (eg a synonym, a more descriptive phrase, or a general rather than technical term). In this case the query will not locate the most appropriate document. Whilst solutions have been proposed for simple variants of this problem (eg Chen 1994), there is an implicit expectation that the user understands the domain well enough to phrase his query using appropriate terminology.

2.4) Relevance Feedback

The third implicit assumption concerns relevance feedback. It has been commonly observed that retrieval performance is enhanced when, after a first retrieval attempt, the user indicates the most relevant documents and his query is repeated. The assumption here is that user requirements are stable. Many authors believe that this is rarely the case in practice, and indeed, with respect to IR, Croft (1995a) has observed that users often use relevance feedback to browse a document collection. This could be the result of the user acquiring new terminology as a result of the retrieval process, or better understanding the type of information he is trying to locate.

3) Information Retrieval on the Web

There are two possible ways of applying Information Retrieval to the Web: Within sites, and between sites (where a site is a node on the Internet that runs a web server). Applying IR within sites has been shown (Croft 1995b) to be highly effective. That is, the IR assumptions generally hold, and users are able to quickly locate documents by query to an indexed collection.

Examples of IR systems that are used to index within sites include INQUERY (Croft 1995)[8][2] Harvest [7], WAIS, SWISH [16], and several others. Between them they apply the full gamut of techniques, such as TFIDF, Boolean query operations distances heuristics, synonymy, relevance feedback (see Salton 1989 for descriptions of these terms). Since these systems conform to the usual IR paradigm, they shall not be considered further.

There are a variety of approaches to the Information Retrieval problem between sites. Robot approaches aim to automatically index the entire web by searching node by node. Assisted approaches apply traditional IR techniques across several manually selected sites. By contract, the Object Oriented Data Base approach relies on site operators to manually suggest what pages should be indexed, and by what criteria. Finally, web Information Retrieval agents assist the user in his search, but do not build specific indices themselves.

3.1) Robots Based Engines

Robot engines aim to exhaustively reference all web sites. They are then indexed and made publicly accessible. The problem with this approach is the exponential growth of the Web. Mauldin (1995) reports its approximate size as being 29.9 Gigabytes and 4 million URLs in April 1995. However by November 1995 Lycos [13] reports the Web size as being 9 million URLs. Whilst this is tractable it is unclear for how long it will be if the Web continues its exponential growth. Nonetheless, this approach is popular, and currently effective.

WEBCRAWLER

WEBCRAWLER [21] exhaustively searches the World Wide Web indexing key pages that it finds. From these it builds a publicly searchable index . Queries are made up of keywords combined with Boolean logical operators. WEBCRAWLER is supported by advertisers whose messages are included in the results of searches.

[2] Numbers in [] refer to URL’s (Internet web links) listed in section 7.

InfoSeek

InfoSeek[9] builds a global WWW index by carrying out limited searches of the WEB, combined with site references sent in by the site owners. InfoSeek also indexes Usenet NEWS groups.
The InfoSeek index is also publicly searchable. Queries are made up of keywords plus logical operators. Where keywords are combined, a distance separator may also be used to indicate increased term weighting.

LYCOS

LYCOS uses patented search technology developed at Carnegie Mellon University. It aims to catalogue 100% of the Internet (currently 91%) and includes 1075 million URL's. LYCOS supports a variety of techniques to determine term importance. LYCOS includes automatic document abstracting techniques top summarise contents of relevant URLs.

3.2) Robot Ethics

The general use of robots is however problematic (Koster 1994). Robots consume network bandwidth that is inevitably limited, and impose general burdens on web servers. Since it is not difficult to write a robot that does exhaustive searching, interested groups are trying to control robot development and the nuisance robots can create. These take the form of ethical guidelines for robot writers and their robots to follow. These suggest limits both in the frequency and number of accesses to remote sites, and also encourage robots to check for a sites' "robots.txt" file. This allows site owners to limit access to robots. An alternative to searching the entire web is to restrict the search to named, willing, sites. Let us call such approaches Assisted IR Engines.

3.3) Assisted IR Engines

INQUERY, HARVEST, SWISH, and several other tools are primarily Information Retrieval (IR) engines that have been adapted to Web searching. Thus they have the advantage of being "state of the art" IR engines, but the disadvantage that sites have to be explicitly indexed. WAIS is the exception to this rule as we shall see below. However WAIS, and other tools below are not popular (or effective) ways of searching the web. This is because sites need to be explicitly indexed before they can be searched. That is, they require human assistance.
These engines are most effective when used to exhaustively index one (or more) particular sites. There they allow visitors to carry out in depth searches.

INQUERY

INQUERY (Croft 1995b) is Massachusetts University's state of Art Information Retrieval engine and has been used successfully in the TREC [17] trials. (Harman 1994). It has been applied to indexing up to 100 web sites. INQUERY owes its success to the combination of evidence about relevance from the user, the document and the corpus in a probabilistic inference net model. The system uses text variants and relevance feedback to expand the users initial simple queries. This provides excellent average recall-precision performance.

WAIS

WAIS (Wide Area Information System) is particularly interesting since it was explicitly designed as an Internet resource discovery tool but before the explosive growth of the World. It is a full text information retrieval architecture that provides users with a location transparent mechanism to access information (Obraczka et al 1993). Sites are indexed by a directory server when requested, and the database then maintains a full inverted index of the referenced files. WAIS also supports server replication. In a discovery session, the user enters search terms that are assigned importance inverse to their document frequency (TFIDF). X-WAIS also supports relevance feedback (Salton and McGill 1983).
WAIS is not particularly popular as a web searching engine, possibly because it lacks coverage or because of its overly technical nature.

SWISH

SWISH (Simple Web Indexing System for Humans) is inverted file Information Retrieval engine especially designed for web sites (Udell 1995). It supports Boolean queries, and calculates document relevance. This is based on TFIDF, but it is also possible to preferentially weight terms in HTML document titles, or section headings.

HARVEST

This is part of he University of Colorado’s Internet resource discovery research (Schwartz 1993). It includes in depth site searching tools, plus a tool to gather web information (GATHER).

3.4) Assisted OODB's

YAHOO

YAHOO [24] is "Yet Another Hierarchical Object Oriented database. The database is populated by site owners wishing to advertise their site. The site is indexed using the OPENDOC Information Retrieval Engine. User's may either browse YAHOO be category (eg Business, the Arts etc) or by carrying out a full search of the index. YAHOO is extremely well known as a place to publicise a new web site. This ensures the completeness of its index. The approach is interesting, since it is site owners who indicate their most important pages, and provide the keywords to classify them.

3.5) Web Information Retrieval Agents

A final approach to web IR is based on automation or assistance of the web searching task. Here there are knowledge rich approaches, (as in the Softbot, WebWatcher and Letizia) where the agent is intelligent, and knowledge poor approaches. These (CUSI, and MetaCrawler) pass queries on to full search engines, and collate the results.

SoftBot

The Softbot (Software robot, Oren and Etzioni 1994)[19] is an AI based intelligent agent to the Internet. It directs a variety of Internet tools (FTP, TELNET, ARCHIE, GOPHER etc) to achieve a user's goals. The Softbot uses AI planning techniques to select the most appropriate tools, and constructs an plan, subgoals to achieve it. These would include referencing remote servers and different tool applications where appropriate. The long term stated aim of the Softbot is to enable naive users to effectively locate, monitor and transmit information across the net.

WebWatcher

WebWatcher (Armstrong et al 1995) is an experimental information seeking assistant for Web. It is a modification of the public domain Mosaic web viewer that advises users which hyperlinks to follow next. Webwatcher incorporates machine learning methods to acquire knowledge for automatic hyperlink selection. These include TFIDF with cosine similarity measure (Salton and McGill 1993), Wordstat, a measure based on the statistics of individual words, and Winnow (Littlestone 1988) that learns a Boolean concept represented as a single linear threshold function of the instance features. A random control was also used. Whilst Winnow and Wordstat were the most effective measures, it is difficult to infer much from this since the system trials were restricted to retrieval of pages in the machine learning area.

Letizia

Letizia (Liebermann 1995) is a user interface agent that assists a user browsing the World Wide Web. As the user operates a conventional Web browser such as Netscape, the agent tracks user behaviour and attempts to anticipate items of interest by doing concurrent, autonomous exploration of links from the user's current position. The agent automates a browsing strategy consisting of a best-first search augmented by heuristics inferring user interest from browsing behaviour.

Combined Search Interfaces

METACRAWLER

METACRAWLER (Selberg and Etzioni 1995)[14] permits parallel searches of the Web by executing the same query using several of the search interfaces mentioned above: LYCOS, YAHOO, InfoSeek, INKTOMI, and OPENDOC.
Search results are cross correlated and returned as one page to the user. This is inefficient in terms of Web resources, but gives wider coverage than using any single search service.

CUSI

CUSI (Customisable User Search Interface) [5] is similar in style to METACRAWLER. except it carries out fewer parallel searches. It's results are therefore less impressive.

4) Research Issues for IR

In this section we shall consider some research questions posed for IR by the Web. We have identified the following issues:-

Index Omniscience
Information Burial
Unstable Queries
Non Uniform Information Density

We will now go on to look at these in turn.

4.1) Index Omniscience

Global Web indices violate a primary assumption discussed in Section 2 that the index has been created by analysing all possible documents. This is not possible for practical reasons. Firstly, automatic indexing robots would cause considerable annoyance when they attempt to obtain all pages at a site since it would impose a considerable work load on the site, and impede access for human users. Indexing robots therefore restrict themselves to retrieving one to ten pages per visit. Secondly, sites may be continually adding pages, amending them, or even withdrawing them. Therefore even pages that have been stored in the index may no longer be available. We could say then that a global Web index is not omniscient: It does not contain references to all possible information on the Web. This has several consequences.

Firstly, since no index is complete, even a highly focused search for an item known to be present on a the Internet may probably not locate that item since it is not in the index. This affects user confidence, since the next time that the user queries the index, and no hits are reported he may not take this as an indicator of index, rather than query deficiency.

The second consequence of incomplete indices concerns traditional assumptions concerning term frequency. For example, the frequency of a term in a document is commonly used to indicate relevance in proportion to its occurrence in a collection. TFIDF (Salton 1989) must be less useful if the index has not sampled the entire collection, but only a sample of pages. In fact search engines typically only sample home pages, or one or two pages following in the hierarchy. Since these contain introductory material they may not refer to a term that distinguishes later parts of the collection.

4.2) Unstable Queries

IR commonly assumes that users have a well defined information requirement that can be expressed in a succinct query. Whilst there are many web users like this, there are also many seeking information in new areas. As such, they initially pose general queries, and also may learn new terminology as they retrieve different documents. Then they rephrase their requirements more specifically. This effect has also been noted by Croft (1995a), who states that relevance feedback is often used as a means of browsing a document collection.

This effect is a broader interpretation of the vocabulary problem. It is well known in IR that retrieval may fail where user's queries do not contain precisely the same terminology as key documents. Mauldin (1991) gives the example of documents that mention "connectionist approaches" as opposed to "neural networks".

The research challenge here is to assist users in phrasing queries using the correct (or most commonly used) terminology.

4.3) Information Burial

Information burial occurs when sites contain information that would not be discovered by an automated search engine. This may be because it is too deeply buried from the main index page, or because it is not stored in transparent form. This may because the information is held in a compressed postscript file for downloading. Whilst tools such as GLIMPSE (Mamber and Wu 1993) have tools to tackle this problem, they are site specific.

Information burial may also occur when sites use HTML's <ISMAP> feature. Here Information Providers may include a clickable image in their web pages. When users click on different sections of the image, the co-ordinates connect them to different URL's (links). This is useful when there are many links to a page that can be best understood graphically. Since indexing robots can not understand images, or select parts of them, information attached to images in this way is buried.

The research challenge here would be to recognize from the appropriate links, document or image titles that the page contains information to draw to the user's attention. He would then be able to select or follow relevant links.

4.4) Non Uniform Information Density

IR is usually applied to a collection of more or less interesting documents. However Web pages may contain information and also pointers to other web pages. Not all do though, and to this end, I will attempt a characterisation of WWW information density. Firstly, web sites are intentionally constructed by people as web sites. Consequently, if I were producing Web pages for my bank (or other commercial venture), I may well know about my competitors, their sites, and their expertise. However it is unlikely that I would choose to reference them. This contrasts to document collections (such as TREC), where the information content has usually been prepared for other purposes. This implies that to obtain an unbiased selection of relevant information care needs to be exercised over what links to follow.

An initial site characterisation could be as follows:-

Index Sites
Interest Group Sites
Content Sites

Index sites are characterised as only containing pointers to other sites (eg [9][13], and [14]). Their purpose is to index the complete web, or subject specific areas of it. General examples include Lycos, Yahoo and AltaVista, whereas CityNet [25] is an example of an indexing site that just contains pointers to sites with information about towns, cities etc.

Interest group sites are operated by organisations such as professional societies(eg [1][2]). They collect together content and links related to their special interest. Such sites may be inherently biased by referring for example to their own publications, or referencing members, or associates sites.

Content sites contain few links (eg [4][22]). They are used to electronically publish information in any area whatsoever and of variable quality.

The research issue here is to apply web information structure to enhance (or automate) the retrieval of information from content sites.

5) Conclusion and Future Work

The Internet introduces a large number of mostly practical problems for Information Retrieval. Users may be “less professional”, or more accurately searching for information in areas in which they are not specialists. Information is also transient, and may be augmented or disappear although it is still referenced by central indices. The retrieval process itself may also be laborious due to bandwidth constraints and slow servers. Nonetheless, it must be a proper domain for IR.

Several research challenges have been presented in this paper. They cover content recognition, the structure of the web itself , and expanding the users query to match that used in documents. Whilst this latter problem has been reported in Croft (1995b), a general approach could tackle this without the use of a document index. This truly represents a challenge to Information Retrieval[3].

[3] This area of research is under development as part-time PhD project by the first author.

6) References

Armstrong Robert, Freitag Dayne, Joachims Thorsten, Mitchell Tom "WebWatcher: A Learning Apprentice for the World Wide Web" AAAI Spring Symposium on Information Gathering In Heterogeneous, Distributed Environments. March 1995

Berners-Lee, T,. Cailliau, R., Luotonen, A, Nielsen H.F, and Secret A, “The World Wide Web”, CACM Vol 37, 8 August 1994

Chen Hsinchun “Collaborative Systems: Solving the Vocabulary Problem” IEEE Computer May 1994

Croft W. Bruce 1995a“What Do People Want from Information Retrieval?” D-LIB Magazine (“http://www.dlib.org”), November 1995

Croft W. Bruce 1995b “Effective Text Retrieval Based on Combining Evidence from the Corpus and Users” IEEE Expert Vol 10, 6 December 1995

Eichmann David “Ethical Web Agents” Proc WWW 94

Etzioni Oren and Weld Daniel "A Softbot-Based Interface to the Internet" CACM July 1994.

Koster, Martyn 1994, World Wide Web Wanderers, Spiders and Robots, “http://web.nexor.co.uk/ mak/doc/robots/robots.html”

Liebermann Henry 1995 “Letizia: An Agent That Assists Web Brwosing” Proc WWW4

Littlestone N. "Learning quickly when irrelevant attributes abound" Machine Learning 2,4 pp 285-318

Manber Udi, and Wu Sun, 1993 “GLIMPSE: A Tool to Search Through Entire File Systems” University of Arizona Dept of Computer Science Technical Report TR 93-34

Mauldin Michael “Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing” Kluwer Academic Publishers, Dordrecht The Netherlands

Mauldin Michael "Measuring the Web with Lycos" Proc 3rd Int. World Wide Web Conference April 95 (Http://lycos.cs.cmu.edu/lycos-websize.html)

Obrazka Katia, Danzig Peter B, and Li Shih-Hao "Internet Resource Discovery Services" IEEE Computer September 1993.

Salton G. and McGill MJ "Introduction to Modern Information Retrieval" McGraw-Hill 1983

Schwartz Michael "Internet Resource Discovery at the University of Colorado" IEEE Computer September 1993

Selberg Erik and Etzioni Oren 1995 “Multi-Service Search and Comparison Using the MetaCrawler” Proc WWW4

Udell Jon, 1995 BYTE November

7) Links

1 Association for Computational Linguistics. "http://www.cs.columbia.edu/~acl/"

2 CCTA Government Information Service. "http://www.open.gov.uk/"

3 CERL Home Page. "http://portico.bl.uk/cerl/main.html"

4 CIIR Publications. "http://ciir.cs.umass.edu/info/ciirbiblo.html"

5 CUSI. "http://pubweb.nexor.co.uk/public/cusi/cusi.html"

6 D-Lib Home Page. "http://www.dlib.org/"

7 Harvest System. "http://harvest.cs.colorado.edu/brokers/cstech/query.html"

8 INQUERY Information Retrieval System. "http://cobar.cs.umass.edu/inqueryhomepage.html"

9 InfoSeek Net Search. "http://www2.infoseek.com/"

10 Information Filtering Resources. "http://www.enee.umd.edu//medlab/filter/filter.html"

11 Information Gathering Projects. "http://www.research.att.com/orgs/ssr/people/levy/sss-projects.html"

12 Information Retrieval Resources. "http://documents.cfar.umd.edu/resources/ir/"

13 Lycos, Inc. Home Page. "http://www.lycos.com/"

14 MetaCrawler Searching. "http://metacrawler.cs.washington.edu:8080/"

15 Other Digital Libraries. "http://alexandria.sdc.ucsb.edu/digital-libraries/"

16 SWISH Documentation. "http://www.eit.com/software/swish/swish.html"

17 TREC Home Page. "http://potomac.ncsl.nist.gov/TREC/"

18 The British Library and the Internet. "http://portico.bl.uk/bl-guide.html"

19 The Internet Softbot. "http://www.cs.washington.edu/research/projects/softbots/www/softbots.html"

20 UMBC Intelligent Software Agents Resources. "http://www.cs.umbc.edu/agents/"

21 WebCrawler Searching. "http://www.webcrawler.com/"

22 WordNet 1.5 on the Web. "http://www.cogsci.princeton.edu/~wn/w3wn.html"

23 World Wide Web searching tools. "http://ukoln.bath.ac.uk/BUBL/IWinship.html"

24 Yahoo "http://www.yahoo.com/"

25 CityNet"http://www.city.net/"