ISSN 2250-3153

Methods for seeking of information on the WWW

Mr V. Veera Raghavulu, Sri B Mallikarjuna

CSE department
N. B. K. R. Institute of science and engineering college,Vidyanagar,nellore-dt., India

***Abstract*-** This paper investigates a few methods for seeking information on the WWW. The methods are:

1) Seeking information with thematic directories like Yahoo, Dmoz, Loosmart.
2) Use Search Engines like Google, AllTheWeb, AltaVista, etc.,
3) Using meta- Search Engines – Metacrawler, One2Seek, Mamma, etc.,
4) Using a few techniques for a limitation or an expansion of results.
5) Using URLs, portals, etc.

***Index Terms*-** methods, information search, information retrieval, search engines, meta-search engines, portals, intelligent agents

I. INTRODUCTION

There is ceaseless increase of the Global Internet services. One of the most used options for information search is WWW. Actually World Wide Web (WWW) arose in 1994 though Tim Berners-Lee from Laboratory of Physics, Geneva suggested hypertext system and access to documents in the Net using hyperlinks in1989. World Wide Web pages continuously increase their number.

There are more than 2 billion and seventy million pages nowadays. Specialists created and continue to create and improve instruments for fast finding of documents – directories, general and specialized, regional, national and meta searching machines. This research is devoted to information retrieval in Web and methods for its realization.

After booming development of the data base systems and systems for office automation during seventies and eighties of twentieth century, the arising and development of "information retrieval systems" is one of the most progressive directions for the information systems. "Information Retrieval" is one of the most successful areas of the natural languages processing.

According of the British computer Society Information Retrieval is purposed to ensure fast, effective and productive methods for describing, managing, searching, return and presenting of the information from CD and Internet.

Immediately after sixties were created information retrieval systems on request in text data bases of certain language. One of the founders of this direction is the Dutchman Keith van Rijsbergen.. He is working on the field of information retrieval since 1969. The group for "Information Retrieval" led by him has powerful research program based on the theory as on the experience. The purpose is to give new not used to this moment and highly effective methods for access to multimedia world. They develop models applying logistics, contingency theory, computerized linguistics etc.

A. Directories, search machines and meta-searching machines

One could differ between directories and searching machines. The directories are arranged by specialists. The information there is classified in the form of structural tree. Changing of the page or pages does not reflect automatically on the redactor's list. Open Directory Project, known as Dmoz, is Netscape property and it was realized by volunteering editors.

The Open directory on http://www.dmoz.org is just a data base.

Partners, who use the Open Directory, including AltaVista, Netscape, Lycos, HotBot etc. often posess many other data bases. The open directory Dmoz is characterised by one of the largest data bases.

It is newer than the open directory of Yahoo. Lately it characterises with changed interface for searching the information in Web or News groups. It lets also searching of FTP resources.

Search Engines are programs that let the user to search for information by key words. They create index of data bases or Internet sites on the ground of headline, key words, or full text. They suggest an interface to the user for submitting the search request. After finding the result, they generate a list of findings and let the user to click on certain hyper link. The efforts of the specialists are directed to improvement of the interface – i.e. creating of various options for the different groups of users among which in their native or preferred language.

The searching engines consist of three parts – Crawler , Spider , Robot, who travels constantly over the Web space. While coming back to pages the spider constantly search for changes in their content. The first spider is created in 1993. It is named World Wide Worm. All the spider finds goes to the second part – the catalogue. The catalogue is as a giant book. It is called also index. The third part of the Search Engine is a program, which filters millions of pages, to find the most relevant to the request. It arranges the results found.

There are also hybrid Search engines, which combine the both possibilities – associated directory and robot (for e.g. Yahoo, which belongs rather to editor directories, but submits access also to sites that are found by Google Search Engine).

Searching engines give as a result a file of hyper links to web sites, which contain the key word or phrase. According to some researchers of the retrieval process – for e.g. the site http://www.wgi.com/internet.htm, it is asserted that the No 1 way to find something in the Web is to use Search Engines. Four of every five users report that they use Search Engines daily.

According the data of the users, 80 % of Search Engines stop the search of the sites after 30 results. To answer that assertion we would announced that there are search engines which find and show much more sites. Among them they are AltaVista, Northern Light, Google, all the web. For many requests for search they produce huge number of sites. So for the specific phrase "Search Engines Optimization" AltaVista finds and shows 54 results. Many of the specialists are now directed to developments that are connected with optimization of Search Engines and the information security. Various products are created such as Search Engines Commander (a pack with revolutionary features, which lets the user to control Search Engines, to use templates, smart key words, replace etc. It is working on the back word mode).

Meta searching engines do not support data bases with addresses of the sites. They support data bases of searching engines. When using meta searching engines, the search is accomplished in parallel with several basic searching machines and directories. In some cases, near the results the means of search is cited. Some search machines for meta search of the information are developed to the level of capacity to find doubled page addresses, obtained as a result of different engines of search and directories.

Some search machines the means used is indicated near the result of search. Some search machines for meta search are developed to such level of capacity and they can find out the doubled addresses of the pages, obtained as a result of various search machines. Among them are: Copernic (which can be downloaded for free), One2Seek, which gives absolutely relevant and non repeated results, Metacrawler etc.

II. METHODS FOR INFORMATION RETRIEVAL

The basic classification of the ways for information search includes searching of the information by key words and expressions, option for thematic search (by categories) and combined search. If the subject to be searched does not exists as a category or sub category, the best thing is to choose respective category or sub category or to make combined search.

Forwarding of an enquiry for searching by more than one key word or use of an expression containing logical operator restricts the search and increases the relevancy of the findings.
On http://www.ibscorporation.com/search_engines.htm one can see an interesting classification of Search Engines.

Every strategy is realized by various approaches. They may be used also for comparing of the possibilities of the separate means of information search in the Web.

III. INFORMATION RETRIEVAL THROUGH DIRECTORIES

Look smart directory suggest an option for search in 10 categories and 8 directories such as United Kingdom, Canada, Hong Kong etc. Certain interesting directory is Work & Money. If its sub directory Industries is selected, then in a consequence one may reach sites of the companies that produce specific products – for example - Industries>Manufacturing>Food & Beverage>Food Companies>Dairy Products>Cheese Companiesr. Looksmart suggests also Looksmart Centres – Careers, Real Estate – all 15. The center for insurance is also interesting - with information for all sorts of insurance. Even pictures are available for search. So for "Imae" it finds 2005 objects. Webcrawler also submits very rich information due to the help of its partnership with Looksmart. Webcrawler offers more than 1.5 million web sites and 100 000 organized categories of the sites in its own directory.

Yahoo contains 14 categories. With categorical search one can view interesting sites according the respective themes and sub themes. So for Regional>Regions>Europe>Government one can see the sits of international organizations, government etc. Global symbols are admitted to be used. If the search is by the first symbol and there is inserted a global symbol after it, then interesting results can be obtained.
118654 categories and 820577 sites for S*.
Interesting sites in this query may be found in Home>Science, Finance etc.
It permits an access to 3500 companies accounts for free. There are 30 categories and 2361 sites found for "Quotes".

There are 135 resources found for the expression "Analysis for Portfolio"+free. Respectively for the expression "Analysis for Portfolio" AND "free" Hot bot finds 1000 sites. For the last expression Northern Light finds 364 positions in 196 sources. With Yahoo one can find out interesting resources on the field of students' working with computer system, with Windows"98 etc. . For the expression image(*.jpj OR *.gif) Yahoo finds 129 thousand web pages.

The directory NBCI (http://www.nbci.com) suggest an access to 20 group categories. Graphs and indexes for Stock market exchange are available for browsing. NBCi admits use of global symbols in searching by key words, permits sensible search, restricts the search by language and date.
One of the newest directories is this of http://www.hyperseek.com/.
Search with demoz –http://www.demoz.com.

It is possible to find easy for trademark. The user should determine what type of trademark is searched – Federal, Canadian, European. The site suggests also web hosting service only for 14.95 $ monthly. They offer the pack SmartSuite Web Hosting, which is appropriate for everyone who starts electronic commerce. The site suggests to start http://www.register.com/businessresurce/, where one can trace for example 14 steps for building of a successful website.

Dmoz.org, which data base includes 2 million and 700 thousand sites. more than 387 thousand categories organised in more than 16 group categories in 64 languages, if one chooses World->Bulgarian, gives an access to 875 Bulgarian sites may be obtained which are grouped in 16 themes. So if one choose education, then they could be available to be browsed.

Certain interest represents the way Search Engines arrange the results found after specific query. Usually on the top of the list the most relevant results are put. The agents follow rules the most important of which are the place and frequency of the findings of the keyword in the site.

Indexing machines identify all pages, which contain the search terms of the query, while "search engines" identify small part of these data. Indexing machines ignore hyper links while some search machines as "Page and Brins search engine" base the search on Internet hypertext structure.

Indexing machines first identify pages which contain the users terms for search and then they apply a function of arranging the

pages in order to identify the relevant pages. Various indexes use various characteristic features as for example frequency of the findings, preferences for the text of the headline, presence of meta tags, statistics of the previous visits, relevancy for the search terms. The context of the terms in the query also could be included, but this requires involvement of the user.

There are intelligent agents are used too in order to support the Internet search. Meta- search engines combine the results of the index machines, through combining of their ranking functions. In order to show the importance of the information contained in, except of citations, they use hyper links. They are used as a reference, recommendation to certain site.

The authors include hyper links to their sites. The new searching machines use hyper links to identify the collection of the connected documents under the given topic. Google expands the use of the hyper link structure relying on the text for identifying of the relevant pages.

Search Engines check if the key words appear in the headline or on the top of the text as first paragraph. They accept the fact that the page is relevant, when the key words are at the beginning. They also analyse how many times the key words appear in respect of other words and in case of the more often appearance they increase the relevancy.

The Length of the meta tags is another criteria which plays important role when arranging the site in the result. The recommendable length is 150 symbols, where some search machines as HotBot accept length only to 250 symbols. Some search machines increase a little the relevancy of certain sites due to advertising motivation. Exite for example uses the popularity of the links as a part of the method of arranging. Infoseek and HotBot also increase a little of the relevancy of the sites which contain key words in the meta-tags. But Lycos does not read meta tags at all. There are many examples with this searching machine when the sites are highly ranked without meta tags. Search Engines could decrease the level of ranking of the site if it determines the Search Engines as a slam. According to researchers on http://www.make-it-online.com/sf.html some Search Engines as Excite, give priority to the sites with key words in their title tag and also with key words which are repeated often within the whole content. In the same research it is pointed out that AltaVista ranks better. Some search engines and directories can arrange the results according the date– AltaVista, HotBot, iWON, MSN Search, NBCi, Northern Light, Yahoo. Some make also clustering of the sites– AltaVista, HotBot, iWON, MSN Search, Northern Light, Yahoo.

IV. SEARCH WITH USE OF ADVANCED SEARCH

Almost all of the basic directories and Search Engines suggest Advanced or Power Search, where the user enter his query in a special form. The interface in most of the cases is easy for the user.

By Yahoo one may find for "informatics" OR "Computers Architecture" with Advanced Search 11 categories and 215 sites. Again Yahoo with Advanced Search for “Quotes Insurance” gives 450 web pages.

Excite divides the sites of different languages in its own data bases. It gives better options for search by other languages by Advanced Search. While for the expression “Travail йtudiants" OR "travail pur d'йtudiants"-.ca by usual search there is no one site to be found with Excite unless one use the complex search – then a list of 484 sites is obtained.

Through this way of search the information in the complex expression divides into three parts– for the first part the option is included that it is good to contain, for the second that , it must contain for .ca that all sites which has the fourth part in their address should be excluded. (Canada). Besides a restriction is introduced for language of the search – French. Other ways of using of the options are available too. Among the sites found there are also Belgium sites.

A. Information retrieval by meta-search machines

According of some researchers the meta-search machines take less time for searching information in every data base as they search in parallel in several data bases of basic Search Engines or directories. In some cases they give only 10 % of the content of the data base, but this in most of the cases is enough to satisfy the surfing user. Some researchers recommend if the results obtained are too much then first to view these sites that are obtained by AltaVista, Northern Light, InfoSeek. Recommended metasearch machines are: WebCrawler, Thunderstone, DirectHit, WhatUseek. , Savvysearch.com, Metager.de, DogPile.

Among the best meta search engines, which find unic paes is also Ixquick (http://www.Ixquick.com), which is capable to translate the query for search to other languages.. So it gives for “Portals" OR "Vortals" an announcement for 47 unic top-ten sites. Meta search engine Ixquick shows near every site with several asterisks the relevancy of the site.

So for our query for portals or vertical portals, it put on the highest place (five stars) the site : http://www.tradeworlds.com/.

Meta search engine Mamma.com finds 16 highly relevant sites for “Information Systems" OR "Information Technology" OR "Informatics". It shows the first 10, before every address in brackets which search engine is used to find the given site. For “arbait” they are 16 sites found.

The use of every meta search engine depends on the following factors: To which search engines it sends the queries of the users, of the number of the basic search engines and directories, of the fact if the user can select the right word, which Boolein operator they imply, which syntax they use; how to present the results – if they are unite in a common list or they are presented in separate. Meta search machines are useful at search of scientific, terms, phrases in brackets etc.

Metacrawler (http://www.metacrwler.com) gives access to 18 groups of categories. Metacrawler suggests information for search of busines information – for companies, for indexes, for the ways of calculating of the indexes, for the rates of shares of the companies, leading media companies as Forbes, Fortune etc.

If during the category search one choose Insurance Quotes, and on the next level – Auto Insurance, it is possible to choose hyper link to 53 companies in this sphere..

Among meta- search engines highly appreciated by many specialists is CNET Search.com, which gives an access to 800 search engines and its biggest advantage is the access to the channels – using a specialised sites.

Meta- search engine One2Seek (http://www.one2seek.com) is capable to do three kinds of search: - simple, basic and complex.

B. Intelligent agents Information retrieval

The intelligent agents are systems which use specialised servers, which Interpret “the behaviour” of the agent and they

communicate with other servers. There are various kinds of these. “Mobile agents” are autonomous intelligent programs, which move through the Net finding and interacting with services “on the user's behalf”. They have genuine navigation autonomy. They could be able to fulfil there performance on every kind of computer. They use mobile code systems as Java and Java virtual engine.

On http://agents.umbc.edu/Applications_and_Software/ one may see hyper links to various sites of various agents. If we select Software an option appears to make a choice among angents for applications for electronic commerce. One of these, ShopBot is a marketig agent, which has an address: http://www.clickthebutton.com/. It is decisive marketing assistant which works on the background and compares prices for millions of of retailers and hundreds of stores and shops.

On http://webrom.net/search_and_find/ it is possible to select among hpper links to sites of intelligent agents. The user has an option to register himself on this address in order to receive free information for the market status within a period from two weeks for up to 10 companies.

The marketing agent for price comparing PriceComparison.net (www.deatime.com), is capapable to compare the models and prices of various suppliers of various goods, for example Desktop Computers.

According to some researches the marketing through conventional devices is four times more expensive than the marketing through Internet..

The query for “Search Engines Features" OR "Meta* Search Engines" gives 14214 findings. For the same query meta- engine Beaucoup gives 15 sites. (The first results are found by Look Smart и MSN Search). The search is done in parallel by 10 search machines and directories.

Some robots fulfil also the functions of creating and support of the mirror images of sites. When the software archives are for example on many sites this means that more of the users from a certain country or region have faster access and to be able to find and copy for example certain freeware programs.

The challenges before mobile agents are able for browsing on slides of the presentation which gives the expectation for the moving agents, factors, they depend on – unpredictable surrounding, low speed or busy net links, data sources are often doubled by many computers.

For use of other technics and strategies among which is use of Regional and National search engines eventually will be revealed in the next materials.

V. CONCLUSION

There is ceaseless increase of the Global Internet services. One of the most used options for information search is WWW. There are more than 2 billion and seventy million pages nowadays. Specialists created and continue to create and improve instruments for fast finding of documents – directories, general and specialized, regional, national and meta searching machines. This research is devoted to information retrieval in Web and methods for its realization. They develop models applying logistics, contingency theory, computerized linguistics etc.

REFERENCES

[1] D.J. Reifer, Web Development: Estimating Quick–to-Market Software, IEEE Software, Vol. 17, No. 6, Pages 57 - 64, 2000.

[2] Thomas A. Powell, The Complete Reference HTML & XHTML, Fourth Edition (Tata Mcgraw Hill (TMH) New Delhi, 2005).

[3] H. M. Deitel, P. J. Deitel & A. B. Goldberg, Internet & World Wide Web How to Program (Pearson Prentice-Hall, New Delhi, India, Third Edition 2006)

[4] D. Lowe, Web Engineering or Web Gardening?, WebNet Journal, Vol. 1, No. 1 January-March 1999.

AUTHORS

First Author – V. Veera Raghavulu (Asst.Professor in CSE department), N. B. K. R. Institute of science and engineering college,Vidyanagar,nellore-dt.