Comparing Internet search tools

Angeles Maldonado Martínez and Elena Fernández Sánchez
CINDOC-CSIC, Spain

Abstract: *The principal aim of this paper is to apply documentary criteria to the principal search tools on the Internet in order to evaluate them individually and achieve a comparative analysis of them. First, the 10 most important search tools on the Web were selected. The selection criterion is the number of links with the principal search tool's pages. So many directories and search engines are considered Web page databases and are analysed using the criteria applicable to all documentary databases. The recollection, analysis and retrieval of information resources are submitted to the control. All the criteria used are remained withdrawn in a questionnaire that has been applied to each one of the selected directories and search engines, keeping with in an individualised evaluation and comparative analysis of the collection. The questionnaire is accompanied by a system of punctuation that will permit the establishment of a ranking of the studied search tools.*

Keywords: Internet, Web, search tools, directories, search engines, evaluation

1. Introduction

The Internet has become the main source of information resources that can be found nowadays. There are important search tools to help find information in cyberspace. These tools are databases, mainly directories and search engines, that facilitate the process of locating the required information. They collect, store and manage HTML documents.

Before the emergence of the Internet many huge bibliographic databases played the same basic role in information retrieval. With the improvement of search software, scientific information has become very easy to find.

2. Selection of search engines and directories

The number of Internet search tools that allow us to navigate in cyberspace is steadily growing. Our starting point for the selection of tools to be studied was their inclusion in the Spanish Internet database Buscopio (www.buscopio.com/), which covers more than 3000 items, of which 110 were international search tools, both search engines and directories. Although the criteria of taking the most linked Web sites has no objective or intrinsic value, the simple fact of being a popular search tool on the Internet must have some meaning.

Search engines & directories	Number of linked pages
Yahoo!	1,170,599
Excite	458,239
Lycos	437,618
Webcrawler	436,184
Infoseek	356,963
Altavista	350,085
Hotbot	238,667
Nerdworld	34,229
AOLNetFind	16,192
Northern Light	15,780

Table 1: Ranking of search tools by number

These 110 addresses were examined using AltaVista (wwwq.AltaVista.digital.com). The number of links of each of the 110 URLs was determined. We took the precaution of consulting the Internet early in the morning, when the traffic is lower, to minimise the number of errors. Two different searches were conducted for each of the 110 addresses, first in the 'link' field giving the electronic address of the main Web page, followed by another search with the same address, but this time in the 'URL' field. We consider that the number resulting from subtracting one result from the other gives the linked pages of the search tool. Table 1 shows the top 10 ranking obtained.

These 10 search engines and directories were the object of the study we present here.

3. Evaluation of Internet search tools: an information scientist's perspective

3.1. Evaluation criteria

A comparison of the 10 selected Internet search tools with conventional bibliographic databases was performed.

3.1.1. Data collection and document analysis

Every 'conventional' bibliographic database has its own input format, divided into fields to include all the interesting details and data concerning each document selected for the system. Although the data and fields differ in different databases, the mere fact of being conventional bibliographic databases means that all of them have the following fields: title, author, source, location, publication year, classification, keywords, abstract, language and document type.

In the case of HTML pages, the main differences compared to conventional bibliographic database are in the source field. This field is replaced by the URL. The authors are replaced by the corporate source and the exact geographic location by the country of publication.

In databases of some importance the language is controlled, therefore the search results obtained are much better than in the case of databases with non-controlled language. The control of language usually takes place only in the following fields: classification, descriptors, country of publication, language, document type and type of corporate source.

In most conventional databases, the selection, capture and analysis of documents are done manually. This is not the

Internet Search Tools	Automatic Submission		Manual Submission								
	Meta Tag	No Meta Tag	Categories	Title	URL	Description	Key Words	Country	Language	Resources	Organi-zation
Yahoo			X	X	X	X					
Excite		X	X					X	X		
Lycos		X									
Webcrawler		X	X					X	X		
Infoseek	X										
Altavista	X										
Hotbot	X										
Nerdworld			X	X	X	X					
AOLnetFind	?	?									
Northern Light		X									

Table 2: Data collection and document analysis

case with Internet search tools; search engines carry out these processes fully automatically. Simply by introducing a URL to the system, the system is able to locate and analyse the contents of the page represented, and even go to the linked addresses. There is a clear difference between those search engines that are able to recognise META tags and get information from them and those not able to use META tags. On the other hand, directories are created in a totally manual way. To introduce a page in a directory, a submission form has to be filled in, with several compulsory fields. The form is very similar to the one used in conventional databases although the data collected could vary, depending on the policy of each system.

3.1.2. Information retrieval

A comparison between conventional bibliographic databases and Internet search tools was carried out, taking into account the basic commands used in the retrieval procedures, including the use of Boolean algebra and the logical operators (AND, OR and NOT) to incorporate multiple concepts into a search strategy, truncation of words to cover variations in word endings, parentheses to group the terms that are combined by OR, and proximity operators to search phrases.

In a second phase, the comparison took into account advanced search facilities such as refining the search, searching by limited subject, searching in specific fields to narrow the results, viewing the indexes to discover more search terms and viewing the controlled vocabulary.

To compare the Internet search tools, a system of score assignment was established. The databases are given a score of 1 or 0, depending on whether they meet the different basic and additional criteria. There are two exceptions: when truncation is compulsory and when the search tool classifies only part of its pages, the score is 0.5.

3.2. Results

3.2.1. Data collection and document analysis (Table 2)

Out of the 10 search tools studied, only two, Yahoo! and Nerdworld, are directories. Three of the eight search engines (Infoseek, AltaVista and HotBot) are able to recognise META tags; four (Excite, Lycos, Webcrawler and Northern Light) neither recognise nor reject META tags. We were unable to determine AOLNetFind's abilities in this respect. Excite and Webcrawler have something of an intermediate position, asking the searcher for category, country and language in their submission form.

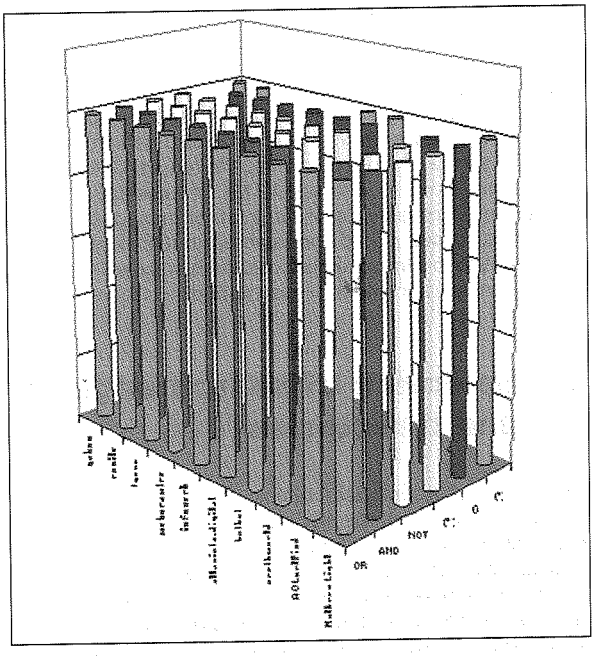

Figure 1: Basic Search Commands

The directories, Yahoo! and Nerdworld, ask the user for the same data collected in their entry forms: title, URL, description and category of the page that is to be included in the system.

We were not able to score the search tools in this respect, as both directories and search engines have different ways of entering and analysing Web pages.

From our information scientist perspective, we think that the manual system, i.e. the one used by directories, is safer because it guarantees the same information from all of the HTML pages. As a second step in importance, we regard the capability of the search tools to use the content of the META tags very favourably.

3.2.2. Information retrieval (Table 3)

The results of the evaluation of the Internet search tools with regard to the basic criteria of information retrieval are shown in Figure 1. There were four criteria. The maximum score that may be obtained is 6. Nearly all of the search tools examined met the four basic criteria, except for Nerdworld, which

Internet Search Tools	Search Tools						Refine Search	Categories	Field Search									Brow-sing	Control Lang-uage	SCORE
	OR	AND	AND	(")	( )	*			T	Url	D	KW	C	L	Dt	R	O			
Yahoo	X	X	X	X		X		X	X	X					X					9
Excite	X	X	X	X	X	X		X*						X			X	L	X	10,5
Lycos	X	X	X	X	X	X*	X	X*	X	X				X				L		11
Webcrwler	X	X	X	X	X	X*		x*												6
Infoseek	X	X	X	X	X		X		X	X			X					KW,C		10,5
Altavista	X	X	X	X	X	X		X*	X	X				X	X			KW,L		11,5
Hotbot	X	X	X	X	X	X		X*					X	X	X			KW,L		10,5
Nerdworld	X							X												2
AOLnetfind	X	X	X	X	X			X*												5,5
Northern Light	X	X	X	X	X	X			X	X			X	X	X		X	C,L,O		13
Total	10	9	9	9	8	6	2	5,5	5	5	0	0	3	5	4	0	2	6	1	

(") = phrase; () = parenthesis; * = truncation; T = title; D = description; KW = key words; C = country; L = language; Dt = date; R = resource; O = organization

Table 3: Information retrieval

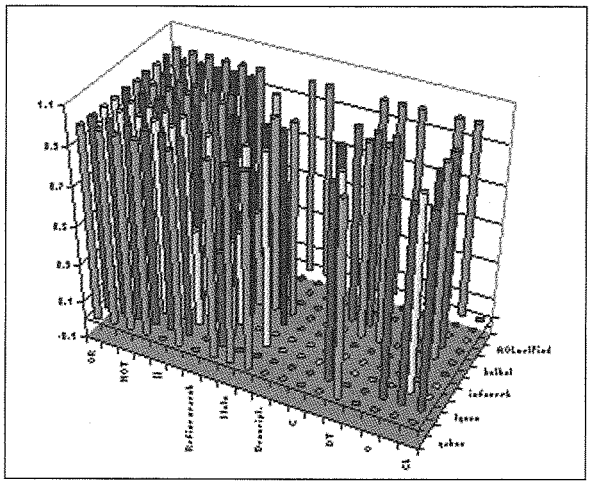

Figure 2: Basic Search Commands plus advanced search faciilities

Search tools	Scores
Northern Light	13
Altavista	11,5
Lycos	11
Excite	10,5
Infoseek	10,5
Hotbot	10,5
Yahoo!	9
Webcrawler	6
AOLNetFind	5,5
Nerdworld	2

Table 4: Ranking of search tools

permits the use of OR as the only Boolean command, and Yahoo!, which is not able to accept parentheses. There are other differences: Infoseek does not permit the use of truncation, and Lycos and Webcrawler oblige the system to use truncation as they select all words starting with the first characters written in the search form.

Figure 2 shows the results of how the advanced search facilities are implemented by the Internet search tools under study. The only Internet search tools that incorporate the possibility of carrying out a refined search are Lycos and Infoseek. With respect to the search by categories, only the two directories Yahoo! and Nerdworld, have every page classified; both have a box to introduce a classification for every page in their submission form. Seven of the eight search engines had a small proportion of their pages organised in accordance with a subject classification. Northern Light is the only search engine not permitting a category search.

Northern Light has the facility to search by fields, offering more options than any of the other tools. Out of the nine fields considered, Northern Light offers six: title, URL, country, language, date and type of organisation. The next search tools according to the number of searchable fields are AltaVista (title, URL, language and date), Yahoo! (title, URL and date), Lycos (title, URL and language), Infoseek (title, URL and country), HotBot (language, date and geographic region), Excite (language and type of organisation). Finally, the three remaining search tools - Webcrawler, Nerdworld and AOLNetFind - do not give the chance of interrogating by fields.

We notice that the most popular fields in relation to their searchability are title, URL and language.

There is no difference in this respect, among the search engines and the directories. Of the two directories under study, Yahoo! offers five searchable fields and Nerdworld none. The search engines are also very heterogeneous in this respect.

With respect to browsing capability, Excite, Lycos, AltaVista, HotBot and Northern Light give the chance of viewing the language index. Infoseek, Northern Light and HotBot permit expansion of the country field. The three search tools that are able to use META tags, Infoseek, AltaVista and HotBot, have another feature: after having conducted a search using a single word, they present a list of compound terms that include the word and gives hints to improve the search.

Finally, the option to search using controlled vocabulary in free text is offered by every search tool examined here, with the exception of Excite. Excite claims to be capable of allowing a conceptual search to be conducted. It should be possible to ask for a term and for the system to locate pages where the term appears or pages where other conceptually related terms are present.

Table 4 shows the overall scores obtained by the search tools for their basic search commands and advanced search facilities. It shows a wide range of scores, from Northern Light with 13 to Nerdworld with only 2. A large group scored between 10.5 and 11.5 (AltaVista, Lycos, Excite, Infoseek and HotBot).

4. Conclusions

The most surprising result was to discover that there is no relation between popularity and data search capabilities among the search tools studied. Two clear examples of this finding are Northern Light and Yahoo!. Although Northern Light is the least popular in the ranking by links, it occupies first place in the ranking by possibilities of information retrieval, while Yahoo!, one of the most popular search tool, obtains a score that is not particularly high in the retrieval options.

We could not find any evidence of whether directories or search engines are better. In data collection and document analysis, directories are more reliable because they use a manual system. The same happens with the search engines that utilise META tags, although it was not clearly the case that they improved the results of searches.

From the point of view of information retrieval, both directories and search engines, even though they are databases that file HTML pages, have fewer possibilities than conventional bibliographic databases. This is due to the following factors:

- Some search tools do not have all of the search options that conventional bibliographic databases have.
- Limiting retrieval by fields is not very usual in Internet search tools. Searchers could ask for a term included in the title and URL, fields where data could be obtained automatically, in only half of the Internet search tools. Neither directories, which could obtain the data from their submission forms, nor search engines that use META tags permit searching by many fields.
- The browsing of indexes is not a common feature.
- The use of controlled vocabulary is really of no use in Internet search tools.

Finally we have to consider the value of the classification of resources in categories, and state that most of the search engines under study allow the possibility of accessing a section where some pages are classified.

Angeles Maldonado Martínez
CINDOC-CSIC
Spain

References

Hartley, R.J., Keen, E.M., Large, J.A. and Tedd, L.A. (1990) *Online Searching. Principles and Practice*, Kent, Bowker-Saur.
AltaVista: AltaVista.digital.com/
AOLNetFind: www.aol.com/netfind/
Excite: www.excite.com/
HotBot: www.HotBot.com/
Infoseek: guide.infoseek.com/
Lycos: www.lycos.com/
Nerdworld: www.nerdworld.com/
Northern Light: www.northernlight.com/
Webcrawler: www.webcrawler.com/
Yahoo!: www.yahoo.com/