Short title:

Empirical evaluation of Internet indexes

Please send correspondence to both authors:

Wouter Mettrop [1] and Paul Nieuwenhuysen [2]

1. CWI - Centrum voor Wiskunde en Informatica / Library, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands; Wouter.Mettrop@cwi.nl

2. Vrije Universiteit Brussel (V.U.B.), Pleinlaan 2, B-1050 Brussel, and Universitaire Instelling Antwerpen (U.I.A.), Belgium; pnieuwen@vub.ac.be

SOME EMPIRICAL RESEARCH ON THE PERFORMANCE OF INTERNET SEARCH ENGINES

Wouter Mettrop [1] and Paul Nieuwenhuysen [2]

1. CWI - Centrum voor Wiskunde en Informatica / Library, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands; Wouter.Mettrop@cwi.nl

2. Vrije Universiteit Brussel (V.U.B.), Pleinlaan 2, B-1050 Brussel, and Universitaire Instelling Antwerpen (U.I.A.), Belgium; pnieuwen@vub.ac.be

Abstract: In this paper the IRT project (Internet / Information Retrieval Tools) is described. The basic goal of IRT is to advise users of Internet search engines in retrieving information from the free public access part of the Internet. In achieving this, IRT has developed a model to evaluate search engines. This model is described in here. Evaluation criteria refer to functionality: search options, presentation characteristics and indexing characteristics (which elements of a Web document are indexed?). Also evaluated is the consistency of retrieval through search engines. This model has been tested in the period October - December 1998 on six of the major search engines. We found many differences among Internet indexes in their functionality, as well as in their consistency and reliability.

Keywords: WWW search engines, functionality, internet indexes, evaluation criteria, consistency of retrieval.

Introduction

Since March 1997 a project is running, named IRT (INTERNET / INFORMATION RETRIEVAL TOOLS) in which 11 information professionals are involved (see their names below). The main goal is to assist users of Internet search engines in retrieving information from the World Wide Web (WWW).

Starting Point

The Internet is expanding constantly. The bigger the Internet, the more important becomes the role of Internet (WWW) search engines in retrieving information. Search engines vary in size and coverage of their database, and in functionality. Many users are aware of differences in functionality features like search options and presentation characteristics, but many do not know about differences in the way engines index Web documents. Which elements of a Web document are indexed? An end-user, looking for the appropriate engine to get a result with the desired recall and precision should be aware of this.

Moreover: the "indexing behaviour" of search engines is not always consistent. Different results are obtained for the same question, even with the same engine. Ideally an end-user should also be aware of this unreliability and should be able to take it into account.

There is other research on functionality of Internet idexes (see for instance Su 1997) and on their coverage (Lawrence and Giles 1998). The Whistlestop project (Kochtanek et al. 1998) studies a few engines. The WWW site http: //searchengineswatch. internet. com/ is devoted to search engines, but not much research is mentioned there; it is primarily concerned with testing the coverage and the freshness of the greatest US search engines, which is investigated by observing how often these visit and index a few sites in the US.

However, how each search engine indexes the contents of the visited sites and how well each resulting database functions is still not clear. Clearly some investigation is needed. Also, European engines deserve more attention.

Goals of the project

The basic goal of the IRT project is to advise users of Internet search engines in retrieving information from the free public access part of the Internet. In achieving this, IRT evaluates search engines (mainly by assessing their functionality and the degree of consistency of their behaviour).

The investigated search engines

In the IRT project, 13 of the major search engines are evaluated:

AltaVista
Euroferret
Excite
HotBot
Infoseek
Lycos
MSN
Northern Light
Snap
WebCrawler
and three Dutch engines: Ilse, Search.nl and Vindex.

The six major search engines evaluated by Lawrence and Giles (1998) are among these.

Adoption of new engines in the evaluation system is possible at any time.

IRT does *not* study the "directory type engines" or "subject trees", like Yahoo!, and the ones associated with (and offered together with) the Internet indexes that we investigate.

Method of evaluation

Evaluating functionality and consistency is completely based on experiments (i.e. results of searches). IRT draws conclusions with respect to these experimental data, and not with respect to the documentation provided by the engines.

Assessing the functionality of search engines

We evaluate three categories of functionality, dealing with:

1. indexing,
2. search options, and
3. presentation of the results.

Investigation of indexing: A set has been made of relevant evaluation criteria related to indexing. A test document has been set up based on these criteria. Queries related to the criteria have been set up in order to test Internet search engines. The test document has been made available at six WWW sites in The Netherlands and in Belgium. Thus we investigate what document elements are indexed by each search engine.

Investigation of search options and presentation of results: Sets with relevant criteria / properties / features have been made. Based on these sets, the search engines are evaluated.

Many differences in functionality have been found. We systematically investigate these further.

Moreover: the indexing functionality of engines is not constant. In the period October 1998 - January 1999, AltaVista, HotBot and Vindex have changed in relation to the considered criteria.

Access to the results obtained up to now is possible online through http: //www.cwi.nl /cwi /projects /IRT /colis/.

Assessing variations in the retrieval performance of search engines

We have added the time dimension to the indexing test setup. Every 29 minutes one of the test queries is sent to one of the engines. At this moment it takes about 9 days for every question to be submitted. The results (including how many times engines were unreachable) are automatically gathered, filtered and stored in a database.

Then the different results for the same queries are analysed. It is possible to express this kind of inconsistency in numbers of *documents not found but known by the engine*. In other words: some results are complete (according to the content of the engine's database at the time the query is submitted) and others are not; a result can be more or less incomplete.

The following are investigated:

1. the number of incomplete results per engine
2. the degree of this incompleteness
3. relations between incompleteness per engine and time or types of queries.

In the period October - December 1998 about 4500 test searches were performed with AltaVista, Excite, HotBot, Lycos, Vindex, WebCrawler. The number of test pages found by the search engines ranges from 1 to 6, the maximum. This, of course, is not a constant number.

Some conclusions drawn from these experiments:

1. A substantial number of results appears to be incomplete. The incompleteness varies per engine and per query. We consider only the set L of combinations (engine x query) in which the engine has found, at least one time, at least one test document. The set L contains 65% of all combinations.

   a. In the set L about 3000 experiments have been performed. Of the results 14% were incomplete. Leaving the experiments with Vindex aside, this is 5%.

   b. It appears that in 53% of the combinations (engines x query) in the set L an engine has given, at least one time, an incomplete answer.

   c. It seems that some engines always give complete answers (Excite, Lycos and Webcrawler), that others give less complete answers (Altavista and HotBot) and that some perform even less reliable (Vindex).

   d. The degree of incompleteness varies.

   e. Some combinations have only incomplete results; i.e. there are queries getting incomplete results from an engine every time they are submitted.

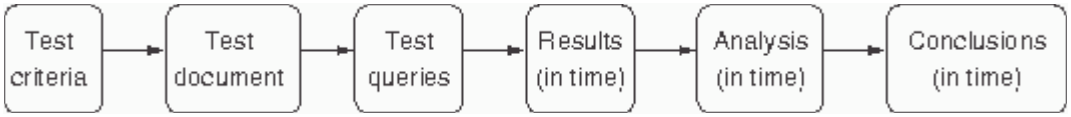
2. Sometimes an engine is unreachable (for the researchers in Amsterdam). In 27% of all combinations an engine was at least one time unreachable. Some engines are more often unreachable than others.

3. In 34% of the combinations in the set L the engine was never unreachable and gave no incomplete results.

Also these experimental results can be accessed online through http: //www.cwi.nl /cwi /projects /IRT /colis/

We plan to investigate more engines simultaneously early 1999, in the hope that this will learn us more in this area of (un)reliability and (in)consistency.

Evaluation architecture



References

Kochtanek, Thomas, Laffey, James, Ervin, Jane, Tunender, Heather, and Borwick, Jim (1998)
Project Whistlestop: an evaluation of search engines on the Web.
In: *Proceedings of the National Online Meeting. New York, 1998.* Information Today, Medford, NJ. pp. 211-221.

Lawrence, Steve, and Giles, C. Lee (1998)
Searching the World Wide Web.
Science, **280**, 3 April 1998, 98-100.

Su, Louise T. (1997)
Developing a comprehensive and systematic model of user evaluation of Web-based Search Engines.
In: *Proceedings of 18th annual National Online Meeting, New York, 1997*. Learned Information, Medford, NJ. pp. 335-344.

Sullivan, D.
Search Engine Watch. [online]
Available from http://searchengineswatch.internet.com [accessed in 1998]

Endnote

Involved in this IRT research project are (in alphabetical order):

Louise Beijer (Hogeschool van Amsterdam)
Hans de Bruin (Unilever Research Laboratorium, Vlaardingen)
Rudy Dokter (PNO Consultants, Hengelo)
Marten Hofstede (Rijks Universiteit Leiden)
Hans van der Laan (Computer and Internet consultant, Leiderdorp)
Hans de Man (JdM Documentaire Informatie, Vlaardingen)
Wouter Mettrop (CWI, Amsterdam),
Paul Nieuwenhuysen (Vrije Universiteit Brussel and Universitaire Instelling Antwerpen)
Eric Sieverts (Hogeschool van Amsterdam and Rijksuniversiteit Utrecht)
Hanneke Smulders (Infomare, Terneuzen)
Ditmer Weertman (Amsterdam)