An Evaluation of Relevancy Ranking Techniques used by Internet Search Engines

Jonathan Back

The Author

Jonathan Back is now a full-time Ph.D. research student at Loughborough University, Loughborough, Leicestershire, LE11 3TU. His current research involves investigating how to capture 'user relevance feedback' in order to improve Web retrieval performance. E –mail: j.back@lboro.ac.uk

The Occasion

The 1998/99 Undergraduate Prize was awarded to Jonathan for his final year project, "An Evaluation of Relevancy Ranking Techniques used by Internet Search Engines".

Abstract

A relevancy ranking algorithm aims to sort retrieved information resources so that those most likely to be relevant are shown first. Experimentation reveals that ranking techniques employed by Internet search engines do not facilitate effective retrieval. The methodology adopted allows for a comparative evaluation of rival search engines. Additionally, different search techniques utilised by experiment participants were analysed. Results show that if more than four search terms are used, the accuracy of a relevancy ranking will increase.

Acknowledgements

I would like to thank Ron Summers, my undergraduate project supervisor, for recommending me for the LIRG prize.

Introduction

There is a plethora of search engines available: some attempt to index all information available on the Internet, while others offer specialised functionality. Typically, a search event begins when a user submits an information request (query) to a search engine comprising one or more keyword(s) or phrase(s). Some search engines can process syntactically valid question queries by using Natural Language Processing techniques or Question Template Matching. Queries can also be supplemented by the use of Boolean connectors between keywords; more advanced functionality can be achieved via the use of search menus.

An advanced search menu enables additional filters to be applied to the search. These filters can be used to restrict the search by: last update, geographic publishing region or language of publication. It is also sometimes possible to configure the search to find spelling and stem variants of search terms - this process being known as fuzzification.

At the present time, the search engine is the most popular tool for retrieving information resources from the Internet. However, many users are aware that fundamental problems exist that are common to all types of search engine:

- In an environment where resources keep moving and increasing (Koehler 1999) maintaining a fresh index is difficult. Problems associated with invalid search results being returned to a user can be attributed to the difficulties of indexing.

- Relevancy ranking is an automated technique that is used to arrange results so that those most likely to be relevant to a user's query are displayed first. For a typical search, hundreds of thousands of information resources will contain keywords that have been included within a user's query. It is infeasible for the user to evaluate the relevancy of all these resources manually. Therefore the success of a search is always bottlenecked by relevancy ranking techniques.

This paper reports an investigation into the effectiveness of relevancy ranking techniques used by AltaVista™, Excite™, Infoseek™, Lycos™ and Webcrawler™.

Relevancy Ranking

The determination of relevancy is fundamental to the success of the information search undertaken. A ranking is calculated using relevancy scores. When a search takes place, a relevancy score is assigned to all indexed resources within the database of a search engine. Most search engines assign relevancy scores based on the techniques shown in Table 1. Sullivan (1999, January) suggests that other factors such as link popularity, Web site reviews and the use of Meta tags are also used to increase the relevancy score.

A range of more advanced information retrieval techniques exist. These techniques are used by some search engines to improve retrieval effectiveness enabling superior relevancy rankings. Instant thesaurus and concept retrieval are processes that expand search terms used within a query. They try to maximise possible interpretations of an information need. Natural Language Processing and Question Template Matching aim to minimise the intrinsic difficulties of query formulation.

TABLE 1. A Simplification of Fundamental Ranking Techniques

Technique	Application
Density	Query term match / TIR
Boolean AND	More matching terms = HR
Boolean OR	Any query terms matching / TIR
Boolean NOT	More matching terms = LR
Proximity	More query terms found close together in resource = HR
Location weight	Query term match in title = HR
Term weight	Rare query terms matching / TIR

KEY: TIR: Terms in resource. HR: Higher ranking assigned.
LR: Lower ranking assigned.

Methodology

A methodology is proposed that enables the quantification of the effectiveness of Internet search engine relevance ranking techniques by the application of subjective relevance judgements. It must be noted that this methodology is not suitable if the overall effectiveness of Internet search engines as an information retrieval tool is to be evaluated. This methodology does not consider factors such as user effort, response times and information visualisation that are essential in order to obtain an overall evaluation.

Traditionally, evaluating information retrieval effectiveness has involved the use of relevance as a metric. By using measures of recall and precision, such as in the Cranfield tests of the 1960s, retrieval effectiveness can be ascertained. Unfortunately, the evaluation of search engines on the Internet cannot be achieved in this way. Frické (1998), amongst others, suggests that with large information systems like the Internet recall is almost impossible to measure or estimate by all standard techniques because, in order to measure recall, all potentially relevant information resources have to be reviewed. In the case of Internet search engines, this could involve reviewing millions of Web sites, which is an impractical proposition. In any case, human relevance judgements are subjective. The use of subjective relevance judgements in order to perform a comparative analysis of rival services avoids the problems associated with compiling judgements from a number of relevance judges with different cognitive capabilities and experiences.

Experiment participants were drawn from students in the Department of Information Science and the Department of Computer Science at Loughborough University. An e-mail was sent to all students in these departments inviting them to volunteer for this study, 39 responded. Experimentation took place in January 1999. Each participant was asked to submit a query to the project web site. This site was designed to act as an interface between the participants and the search engines being evaluated.

Crucially, a participant's query was submitted to the five Internet search engines simultaneously. This enabled test conditions to be the same for each search engine, eliminating the possibility of new information sources or links becoming available during the time of the test. The *highest* ranked result from each of the five search engines was selected and returned to the participant.

The *highest* ranked result was not necessarily the result ranked at the top of the first results page. The Internet is vast; search engines cannot be expected to have indexed all resources. Problems associated with the inconsistencies of index coverage need to be avoided. Therefore, it was decided that a resource had to exist in the index of all five search engines. The *highest* ranked document that satisfied this condition was the one returned to the user. This enables an analysis of duplicate results by comparing how *high* a particular resource is ranked.

Participants were asked to quantify the level of relevancy they associated with each result returned by performing an on-line evaluation. They were asked to use their subjective judgement by assigning a graded relevancy score. Participants were able to select a relevancy score of: 100%, 75%, 50%, 25%, or 0%. Participants in the study were blinded to the search engine services - avoiding the possibility of bias towards or against a particular service.

Two measures called *fulfilment* and *quantity* are introduced to calculate the effectiveness of relevancy rankings performed by a search engine. *Fulfilment* is calculated by obtaining the ratio of the number of documents (returned by an individual Web search engine) that satisfy user information requests. During the on-line evaluation, if an experiment participant assigned a resource at 100%, satisfaction is assumed.

$$\mathit{fulfilment} = \frac{rs\,100\%}{n}$$

$rs_{100\%}$ = Number of resources retrieved that are deemed to satisfy user information requests.

Quantity can be defined as the mean relevancy value of all resources returned by an individual Web search engine.

$$\text{quantity} = \frac{\sum_{i=1}^{n} rs_{ij}}{n}$$

rs_{ij}= The % relevancy score associated with the i^{th} resource retrieved for query j.

Results and Discussion

Table 2 shows the *quantity* and *fulfilment* scores achieved by the search engines. *Optimal ranking* scores have been included within the table. These can be used to compare the relative success of search engine ranking techniques. *Optimal ranking* is calculated by using the percentage relevancy score assigned to the most relevant resource returned for each query.

TABLE 2. Ranking Evaluation Results

Search engine	Fulfilment	Quantity
Optimal ranking	59.0%	79.5%
Excite	23.1%	41.7%
Infoseek	23.1%	39.1%
WebCrawler	15.4%	43.6%
Lycos	12.8%	24.4%
AltaVista	7.7%	38.5%

Resources returned by Excite and Infoseek *fulfil* user requirements to a greater degree than resources returned by the other search engines. This suggests that relevancy ranking techniques adopted by Excite and Infoseek are more successful. However, when compared to *optimal ranking* a considerable improvement is still possible. Due to the proprietary nature of the techniques adopted it is difficult to identify the specific techniques that enabled improved performance.

The *quantity* of relevant information returned by WebCrawler, Excite, Infoseek and AltaVista is around 40%. Lycos performed very poorly in comparison. On occasions Lycos returned very few results, suggesting that it has a high relevancy ranking threshold. A high threshold theoretically enables a user to quickly determine if relevant results exist, minimising user effort. However, in this case, relevant information was filtered out, indicating that Lycos had too much confidence in relevancy ranking techniques.

The explicitness of a query can be categorised by the number of terms used (excluding the use of Boolean connectors). Findings presented in Table 3 suggest that an increase in the number of terms used improves the relevancy of results returned. A more explicit query is more likely to find relevant information. However, it is often difficult to provide such a query if the user is not an expert in the domain being searched. This finding supports the use of information retrieval techniques that can expand a user's query.

TABLE 3. Explicitness of Query

Terms Used	Frequency	Fulfilment	Quantity
2-4	21	13.4%	32.9%
5-7	15	20.0%	42.0%
Others	3		
N	39		

Infoseek (1998, May) uses an extra search precision algorithm (ESP). ESP "improves the quality of search results for general queries. Infoseek research shows that the majority of its users routinely use general keyword (one or two-word) searches to find information and services. ESP is especially beneficial to these users because it automatically anticipates the services and information that will be most useful to them". Excite (1996) uses an intelligent concept extraction (ICE) system. ICE identifies relationships between keywords used within a query enabling retrieval by concepts instead of using traditional Boolean retrieval methods. (This feature is turned off if Boolean connectors are used by a searcher).

The use of ESP and ICE allows for query expansion. This could be a possible explanation for the improved performance associated with Infoseek and Excite.

Conclusions

The work described in this study has successfully enabled an analysis of the *effectiveness* associated with the artificial (search engine) perception of relevance. It is feasible to use this methodology for a full-scale evaluation of search engines that takes into account factors such as user effort, response times and information visualisation.

The first generation of Internet search engines involved complicated user inputs. They required the user to supply their own Boolean logic and search parameters. Over the years, advances in technology have simplified the input required from the user. Almost all commercially available search engines now cater for novice users. Current interfaces for search engines are based on speed of delivery and ease of use. The author believes that as a consequence a search is bottlenecked by the success of automated techniques. The use of an interactive intermediary during the search process would enable a user to define their information need more precisely before automated techniques are applied.

This study has discovered that search engines do not currently perform to a level acceptable to most users. However, bridging the relevancy disparity between the user and the search engine is not a simple task. Future research should investigate the feasibility of search engines that acquire intelligence by analysing user information requirements - ultimately helping to improve the relevancy of results.

Note: It must be remembered that any comparative study performed on a range of search engines is only likely to remain valid for a short period of time. Results published within this paper were based upon experimentation that took place in January 1999. This paper should not be used as a means to endorse or discredit any of the search engines evaluated.

References

EXCITE. (1996). Information retrieval technology and Intelligent Concept Extraction™ searching. *Excite', Excite Inc.* http://www.excite.com/ice/tech.html

FRICKÉ, M. (1998). Measuring recall. *Journal of Information Science, 24*, (6), 409-417.

INFOSEEK. (1998, May). Infoseek introduces "E.S.P." to dramatically improve general search results. *Infoseek', Infoseek Corporation.* http://www.ir-infoseek.com/1998_releases/esp.phtml

KOEHLER, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science, 50*, (2), 162-180.

SULLIVAN, D. (1999, January). *Search Engine Watch, Mecklermedia.* http://www.searchenginewatch.com/

Trademarks