Information Retrieval on the Web

Jacques Savoy

Institut interfacultaire d'informatique
Université de Neuchâtel
Pierre-à-Mazel 7, 2000 Neuchâtel (Switzerland)
Jacques.Savoy@unine.ch www.unine.ch/info/

Abstract

For the information retrieval (IR) community, the Web now presents a new paradigm, while also generating new challenges and attracting growing interest from around the world. An important example of these challenges is managing huge text collections and evaluating the usefulness of hyperlinks contained within them.

Keywords**:** IR, distributed IR, Web searching.

1. Indexing and search processes

Among users looking for information on the Web, 85% submit information requests to various Internet search engines. In order to respond to these queries, search engines index each Web page, representing it by a set of weighted keywords. Thus search engines seek out any useful and pertinent Web pages through using robots or spiders that crawl through the Web. Once pages are found, the indexing process might begin by removing all very frequent and non-significant words (such as "the", "are", "of"). In a second step, a stemming procedure is applied to remove inflectional and derivational suffixes, in order to conflate word variants into the same stem or root (e.g., "thinking", "thinkers" or "thinks" may be reduced to the stem "think"). In a third stage, the pages found must then be represented by a set of weighted keywords.

Based on the TREC experiments (trec.nist.gov), the best weighting procedure takes three factors into account [SAV 01]. First, if a term appears more frequently than another, its associated weight would be increased (assuming that frequent words are more important in describing the semantic content of a document). Second, if a term appears within many pages, its weight would be decreased (assuming such words are not really helpful in discriminating between relevant and non-relevant items). Third, Web page size might be taken into account by assigning greater weights to short pages than to longer ones (usually describing more than one topic). Finally, an inverted file is updated such that for each keyword, the system can find a list of all Web pages (with an associated weight) indexed under this term. Using this structured file, a search engine would then quickly find all Web pages matching a given set of search keywords and compute a score for each retrieved page, indicating its degree of similarity with the submitted request.

Various search techniques attempt to improve their search performance by: 1) taking Web page structures into account (e.g., giving more credit to words appearing in the title field), 2) considering the distance between search keywords appearing within a page, 3) assigning appropriate weights to each search keyword, or 4) suggesting different paradigms (probabilistic, logic or language-based model) or formulas when computing the degree of similarity between a Web page and the user's query [BAE 99].

2. Distributed IR

When handling a relatively small number of documents, a single inverted file is usually sufficient to store all information representing document content. However, given the number of documents contained on the Web, this is not sufficient. Search engines must respond to queries using a distributed system whereby they dispatch queries to numerous processes, that in turn inspect their own associated inverted files. The engine sending the original query receives numerous ranked lists, each containing the retrieved items found, along with their scores. The engine must then merge these various lists and present the user with a single list.

This merging might be achieved by interleaving the result lists in a round-robin fashion. According to previous studies [POW 00], retrieval effectiveness for this type of interleaving scheme is around 40% inferior to performances achieved through using a

single inverted file. However, we know that each retrieved item has a similarity score and we assume that these values are directly comparable. This type of strategy, called raw-score merging, can be used to produce a final list by sorting the items according to the scores computed separately by each search process. Retrieval effectiveness for such merging techniques tends to be 15% less effective when compared to searches based on a single inverted file [SAV 01].

Over the last few years, various researchers have suggested a variety of merging strategies that might more effectively resolve this problem. For example, Powell *et al.* [POW 00] proposed computing a score for each inverted file based on the similarity between the user's request and the content of each inverted file. Rasolofo *et al*. [RAS 01] suggest taking the length of the retrieved list into account, thus giving more importance to those results extracted from longer lists. This leads us to questions regarding the effectiveness of commercial search engines available on the Net.

3. Evaluation of search engines

Based on 33 requests submitted to eight search engines, Gordon & Pathak [GOR 99] found that half the searches returned only one relevant item. From this study, we also discovered that the probability of the first returned item being relevant was around 15%, while the probability of the second Web page being relevant was around 3.5%.

In a more recent study, Hawking *et al.*, [HAW 01] evaluated 20 search engines using 54 requests. In this case, the precision achieved after retrieving 20 documents (or the percentage of relevant and retrieved items after inspecting the first 20 Web pages) was around 0.5 for the best search engines (Northern Light in this case) while Google showed the second best performance. Commercial search engines however are mainly concerned with response delay, which they try to keep less than two seconds.

When analyzing search results over a short period of time, Selberg & Etzioni [SEL 00] showed that search results are surprisingly unstable. When submitting the same request at a two weeks interval, the first ten retrieved pages may change widely, from around 63% for the HotBot, InfoSeek or Lycos search engines to 28% for the AltaVista engine. Given the highly dynamic nature of the Web, this phenomenon may be viewed as normal. However, some of these pages that had disappeared were found to reappear after a few days. For example, the Lycos engine found around 50% of the URLs in the top ten result list but they were not present in the top 200 of a subsequent result set and they then reappeared in the top ten of another subsequent result set.

4. Other challenges in Web searching

When implementing effective search engines for the Web, we have encountered a variety of problems. First of all, given the huge number of pages available on the Web, search engines can only index a fraction of all the available information. Lawrence & Lee Giles [LAW 99] estimated that the coverage of various search engines varies widely, with Northern Light having the greatest coverage, meaning 16% of the Web, and Lycos having an estimated coverage of around 2.5%. Moreover, the overlap between the different search engines remains low, such that combining the results produced by various search engines greatly improves Web coverage, and thus providing proper justification for using metasearch engines.

Secondly, given the huge amount of available information, search engines operate by creating numerous inverted files to store keywords associated with each Web page. Third, some Web pages retrieved cannot be accessed by users (e.g., the famous error 404 "Page not found") because the corresponding page has been moved or removed. Fourth, the content of Web pages may change over time or the corresponding information may become out-of-date (e.g., when searching in newspapers or financial sites). Finally, retrieving a page does not mean that its content is credible. This leads to the conclusion that finding authoritative sources on the Web is a challenging problem. Readers interested in discussions about search engines strategies and practices may consult www.SearchEngineWatch.com.

5. Web users and their queries

Resolving various technical problems is only one aspect of creating adequate search engines. The other involves information on the users, their needs and their habits. Recently, various studies analyzed current Web users and their requests [JAN 00],

[SPI 01]. The users are tending toward greater simplicity, including the writing of shorter queries (average query length is around 2.4 words, representing around 28% of submitted requests of one keyword and 32% for two search keywords). Web users' work sessions tend to be shorter (with 54% of the sessions being only one query), and they also tend to view fewer pages that results from each query (35% of the users examined only one page of the results provided by a search engine).

When writing a request, users frequently enter personal names, make spelling errors or use non-English words. From an analysis of search topics, users seem to be looking for Web sites on commerce, travel and employment (24.4% of the queries), people & places (20.3%), computers & Internet (10.9%), health & science (7.8%), sex & pornography (7.5%), and entertainment & recreation (7.5%).

Queries submitted containing Boolean operators represent the minority (around 6%) and these operators are often expressed with mistakes. Also scarce is the use of relevance feedback (usually through the button labeled "More like this"), but this practice seems to increase over time.

6. Why is it so difficult to find the right Web page?

Automatic retrieval of information by computers can be viewed as a complex task, especially given the underlying ambiguity of all natural languages. On the one hand, authors and users frequently write different words or expressions when referring to the same concept ("accident" may be expressed as "event", "incident", "situation", "problem", "difficulty", "unfortunate situation", "what happened last week", etc.) [FUR 87]. On the other hand, specific terms may have different (and sometimes contradictory) meanings and interpretations (e.g., polysemy relative to the word "lead" in "environment Canada plays a lead role …", "lead pollution" and "lead mining"). Moreover, additional linguistic phenomena including anaphora, ellipses, pronominal references, spelling errors etc. tend to render the process of indexing and matching requests to documents an imprecise, incomplete and uncertain exercise. Thus a computer cannot infer that a logical string match would always mean a match relative to a word's true sense.

7. Link-based retrieval

Assuming that links between pages can provide useful semantic information about document relationships, various retrieval strategies were suggested that might take them into account. To do so, two retrieval models have recently been suggested.

1.7.1. Kleinberg's algorithm

In this scheme [KLE 98], a Web page pointing to many other information sources must be viewed as a "good" hub while a document with many Web pages pointing to it is a "good" authority. Likewise, a document that points to many "good" authorities is an even better hub while a Web page pointed to by many "good" hubs is an even better authority.

To compute these values for a given page D_i after c+1 iterations, the formulas for the hub and authority scores $H^{c+1}(D_i)$ and $A^{c+1}(D_i)$ are:

$$A^{c+1}(D_i) = \sum H^{c}(D_j), \text{ for each } D_j \text{ parent of } D_i$$

$$H^{c+1}(D_i) = \sum A^{c}(D_j), \text{ for each } D_j \text{ child of } D_i$$

which is computed for the k best-ranked documents retrieved by a classical search model, and this set of pages is increased with their children and parents. The hub and authority scores were updated for five iterations (because the ranking did not change after this point), and a normalization procedure (dividing each score by the sum of all square values) can be applied after each step.

1.7.2. PageRank measure

Brin & Page [BRI 98] suggest another approach called PageRank measure that first evaluated the importance of each Web page based on its citation pattern (and this computation is done independently of the current query). A Web page will have a higher score if many pages point to it. This value may increase if there are highly scoring documents pointing to it. The PageRank value of a given Web page D_i, value noted as $PR(D_i)$, having D_1, D_2, … D_m pages pointing to D_i, is computed according to the following formula:

$$PR(D_i) = (1 - d) + d \cdot [(PR(D_1) / C(D_1)) + \ldots + (PR(D_m) / C(D_m))]$$

where d is a parameter (e.g., set to 0.85 [BRI 98]) and $C(D_j)$ are the number of outgoing links for Web page D_j.

The PageRank value can be computed using an iterative procedure (e.g., five iterations). After each iteration, each PageRank value is divided by the sum of all PageRank values. Finally, as initial values, $PR(D_i)$ is set to 1/N where N indicates the number of documents in the collection.

After a classical search engine has retrieved a set of Web pages, these pages are sorted according to their PageRank values and the reranked list is presented to the user.

After evaluating each link-based retrieval schemes, we did not find any improvement over more classical IR models [SAV 01] (see also TREC-9 and TREC-10 results at trec.nist.gov). These hyperlinks can however be useful for other purposes (e.g., finding the homepage of a given person or finding micro- or macro-structure in the Internet).

Conclusion

The Web today is generating new challenges for the IR community, including the management of a huge amount of hyperlinked pages, crawling the Web in order to find appropriate Web sites to index, accessing documents written in various languages [PET 02], measuring the quality or authority of available information [KLE 98], providing precise and short answers to user requests (e.g. "Who was the first American in space?" [VOO 00]), online service location [CRA 01] (where the expected answer to the request "Quantas" is the Quantas Airlines homepage, not several Web pages about this airline company), and interactive searches looking for specific document types or Web pages in order to satisfy a particular geographical or time constraint (e.g., "Where is the nearest Chinese restaurant?" based on the fact that this request is coming from Santiago).

References

[BAE 99] Baeza-Yates, R., Ribiero-Neto, B. *Modern information retrieval*. Addison-Wesley, Reading, 1999.

[BRI 98] Brin, S., Page, L. *The anatomy of a large-scale hypertextual Web search engine*. WWW'7, 107-117, 1998.

[CRA 01] Craswell, N., Hawking, D., Robertson, S.E. *Effective site finding using link anchor information*. ACM-SIGIR'2001, 250-257.

[FUR 87] Furnas, G., Landauer, T.K., Gomez, L.M., Dumais, S.T. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 964-971, 1987.

[GOR 99] Gordon M., Pathak, P. Finding information on the world wide Web: The retrieval effectiveness of search engines. *Information Processing & Management*, 35(2), 141-180, 1999.

[HAW 01] Hawking, D., Craswell, N., Bailey, P., Griffiths, K. Measuring search engine quality. *Information Retrieval*, 4(1), 33-59, 2001.

[JAN 00] Jansen, B.J., Spink, A., Saracevic, T. Real life, real users and real needs: A study and analysis of users queries on the Web. *Information Processing & Management*, 36(2), 207-227, 2000.

[KLE 98] Kleinberg, J. *Authoritative sources in a hyperlinked environment*. ACM-SIAM Symposium on Discrete Algorithms, 668-677, 1998.

[LAW 99] Lawrence, S., Lee Giles, C. Accessibility of information on the Web. *Nature*, 400, 107-109, 1999.

[PET 02] Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds). *Results of the CLEF 2001 Cross-Language System Evaluation Campaign*, LNCS, Springer-Verlag, Berlin, 2002.

[POW 00] Powell, A.L., French, J.C., Callan, J., Connell, M., Viles, C.L. *The impact of database selection on distributed searching*. ACM-SIGIR'2000, 232-239.

[RAS 01] Rasolofo, Y., Abbaci, F., Savoy, J. *Approaches to collection selection and results merging for distributed information retrieval*. ACM-CIKM'2001, 191-198.

[SAV 01] Savoy, J., Picard, J. Retrieval effectiveness on the Web. *Information Processing & Management*, 37(4), 543-569, 2001.

[SEL 00] Selberg E., Etzioni O. *On the instabillity of Web search engines*. RIAO'2000, 223-235.

[SPI 01] Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T. Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234, 2001.

[VOO 00] Voorhees, E.M., Tice, D.T. *Building a question answering test collection*. ACM-SIGIR'2000, 200-207.

About the author

Jacques Savoy is a professor at the University of Neuchatel (Switzerland). After completing his Ph.D. thesis in 1987 (University of Fribourg, Switzerland), he became a Computer Science professor at the University of Montreal, where he began research in link-based and distributed information retrieval. His current research interests involve logic-based and cross-lingual IR models.