Evaluation of Web Document Retrieval: A SIGIR'99 Workshop

Maristella Agosti and Massimo Melucci
Universita` di Padova
Dipartimento di Elettronica e Informatica
Via Gradenigo 6/A -- 35131 -- Padova -- Italy
{agosti,melo}@dei.unipd.it

A Framework for the Workshop

In this decade the information retrieval research community has started new relevant efforts to improve previously available evaluation techniques and methods and to address new ones. Evaluation techniques are critical to research in all areas of science and engineering, so they are for the information retrieval area. The evaluation of information retrieval systems can take different approaches, face different issues, and propose or use different methods, as it has been addressed by Harter and Hert in [1]. The two orthogonal themes they propose to consider are the ``Different IR Theoretical Perspectives'' and the ``Diversification and Hybridisation of IR Systems''. These two themes are comprehensive of many aspects and efforts that most of IR researchers believe are important for conducting future IR evaluation efforts. This framework was proposed to the thirty-three people who attended the workshop. Fourteen people were from United States, thirteen from Europe, three, two, and one people respectively were from Korea, Japan and Australia.

Different IR Theoretical Perspectives.

The different and alternative approaches need to be considered separately, because different conceptual models of IR require different and alternative models of evaluation. Some of these theoretical perspectives that have been put forward over time and presented in [1] are: the so called black-box model, where inputs are used to produce different outputs, and the evaluation approach is based on a Cranfield-like evaluation methodology; the interaction model, where IR is considered as an interactive communication process; the approach that sees IR as creativity, exploration, and discovery; the approach that considers IR as navigation which is based both on the network/hypertext paradigm together with the networked environments; and the approach where IR is considered as logical inference.

Diversification and Hybridisation of IR Systems.

The second theme to be considered is the proliferation of systems that often implement in an integrated way some of the different views and functions of IR, but which are not only IR systems. Some of those are: the World Wide Web, and a digital library. Both of them are systems which include IR functions, but they are systems that have many other different functions that are not traditionally the target or scope of an IR system.

The Focus of the Workshop

Web Functions and User's Information Needs.

Some of the possible types of information need are: Retrieving aims to locate relevant Web pages where emphasis may be placed either on recall or precision. Finding aims to locate a unknown attribute usin g a known one. Mining aims to locate novel information, associations, patterns, or rules. Exploration is retrieving information being relevant to an information need that may change at search time. Extracting is transforming an unstructur ed set of unstructured Web pages into a structured database of structured record.

By ``function'' we mean the service being provided by a software system, such as a search engine, to meet a final user's information need. The main functions be ing performed within the Web context are browsing hyper-links, querying search engine indexes, and navigating search tool directories. One of the aspects that characterizes the Web is that search engines are used not only for retrieval purposes. Indeed one may use search engines to look for a specific resource, e.g. an address or a business. In general, there is a many-to-many correspondence between function and information need. Thus, an information need can be met throu gh one or more function, or a given function can meet more than one information need. This is one of the issues of diversification and hybridisation of IR systems.

Web Data.

Another facet of diversification and hybridisation is that about the data. The issues about data should be considered whenever an evaluation framework is set up are listed in the following.

1 No collections -- we can hardly speak about collection for Web document retrieval since the Web evolves rapidly and the evolution cannot be controlled.

2 Data and structure -- in the Web context it is possible to find all different types of data, i.e. un-structured, semi-structured, or structured data.

3 Multi-lingual data -- though English is the main language used to write Web pages and to formulate queries, non-English languages are becoming more and more used.

4 Multi-media data -- the Web contains different types of document media, other that programs, such as Java applets.

5 Hyper-links -- the presence of links, though untyped 1makes Web document retrieval different since relevant information are often in linked documents.

6 Metadata -- metadata implement the semantic interoperability between heterogeneous data sources. They may be used to improve retrieval effectiveness, but they are rarely used.

Web Technology.

Web technology and standards change. Among the others, we think that the technologies which are a characterization of the Web context are the following ones.

* The Web browser with its predefined browsing functionalities is ultimately the unique interface between the final user and the Web. Additional functionalities are provided by the search engines and implemented using Web languages.

* Slow connection can make impossible an effective browsing. Indeed, to be effective, browsing or navigation should allow the final user to follow links rapidly in order to retrieve the desired information.

* Web pages are mostly implemented as HTML documents, but not only. For example, XML or Java are often used to implement Web pages and HTML is going to be legacy.

Workshop Presentations

Nick Craswell, Peter Bailey and David Hawking asked themselves if ``Is it fair to evaluate Web systems using TREC ad hoc methods''. In other words, the question they posed is if the Cranfield evaluation model is still adequate for Web document retrieval. A preliminary answer to this question was given by the authors at the beginning of their paper by stressing that ``systems of TREC VLC Track participants were more effective than live Web systems''. The authors provide a possible reason for that -- Web and TREC ad hoc systems solve different problems. Then, they analyse into detail four hypotheses which may explain the difference in performance:

* The importance of hyper-links. Web search engines retrieve hyper-documents, while TREC systems retrieve ``flat'' documents. TREC ad hoc evaluation scheme do not consider links between documents and a document is judged as non-relevant though it is linked to relevant documents.

* Different topics. As stressed in Section 1, the authors highlight that there are many information needs which reflect in different topics being expressed as very short sentences, perhaps without any structure.

* Document quality. Within TREC a document is judged about its relevance to a topic without considering its quality. On the contrary, search engines can retrieve high quality non-relevant documents, and miss low quality relevant documents if they use quality as retrieval criterion.

* Duplicates. The presence of many copies of the same document induce a bias into retrieval effectiveness measures. Indeed, systems that retrieve many copies of a non-relevant document appear to be less effective.

Roberto Zamparelli addressed one of the workshop topics which constitutes one of the issues mentioned in Section 1. The main idea is

summarized in the title: ``Metadata Do It Yourself''. Zamparelli argued that automatic techniques to describe the semantic content of Web pages have some limits. Automatic extraction of keywords looses many important information such as non-text data, document structure and layout, and ``details'' improving the document quality. Information propagation is sometimes used by the systems to increase the evidence about the relevance of a document or a site if the latter has many in-going links from another document or site. The use of links to propagate evidence may lead to circularity -- the more a site is judged as relevant, the higher the number of in-going links, the more the site is judged as relevant. The author proposed a semi-automatic tools supporting authors to add metadata to their own documents which is based on the iteration of manual editing supported by automatic processes, such as automatic keyword extraction and relevance feedback.

Annelise M. Pejtersen, Mark Dunlop and Raya Fidel presented ``A use centred framework for evaluation of the Web'' for identifying the ``experimental boundary conditions''. They identified some issues making Web evaluation rather difficult -- lack of understanding on how search engines work, user disorientation and cognitive overload in a complex of links, the evolution of the Web may make user understanding and learning more difficult, absence of functionalities to support decision tasks which is then left to the user.

Mildrid Ljosland presented an ``evaluation of Web search engines and the search for better ranking algorithms''. The paper was in two parts: The first part reported on a case study consisting in two comparisons. A comparison was among three search engines in order to test their effectiveness, while another comparison aimed to study the size of the set of documents indexed by a search engine. The second part was a review of the possible ways for improving ranking algorithms.

The last presented paper was by Monica Landoni and Steven Bell. They presented a ``critical overview'' on the ``information retrieval techniques for evaluating search engines''. They envisage a collaboration between the IR and the Web communities in order to take advantage of the synergy with the experience in evaluation of the former. The paper proposes a framework for evaluating search engines. The starting point of the proposed guidelines is the assessment of the scenario, i.e. of the context within which evaluation has to be carried out. Then, it is necessary to select the measures being most adequate to fulfil research objectives. Collecting information about the search tools is the next step and aims to have data about functionalities and databases. The experiment is then defined in order to select queries, relevance model, judges and documents.

Some Reflections

The final part of the workshop was devoted to the discussion on the topics arisen from the presentations. In particular, comments concentrated on the following issues expressing diverse positions:

* The Cranfield model may be the most adequate one if user studies are hard and expensive to do, and links are investigated in order to employ the information they provided about content and

topology. Indeed, hyper-links have to be considered within any Web document retrieval evaluation model since they may express relevance relationships between documents.

* More work should be done to investigate what ``Web user's information need'' means. The difficulty of understanding what a Web user's information need is depends on the fact that there is no average users. Logs can not always be used to infer what the user wants.

* Web queries are rather different from the TREC topics. The latter are well-formulated query expressing a well defined information need. The former are on the contrary ill defined expression of ill defined information needs.

Acknowledgments

The workshop organizers thank Fred Gey for the help and the support, and all the attendees for their active participation.

References

[1] S.P Harter and C.A. Hert. Evaluation of information retrieval systems: Approaches, issues, and methods. In Annual Review of Information Science and Technology (ARIST), volume 32, chapter 1, pages 3--94. Information Today, Inc., Medford, NJ, 1997.