# A perspective on Web information retrieval

Massimo Melucci (massimo.melucci@unipd.it)*
*Department of Information Engineering, University of Padua, Italy.*

David Hawking (david.hawking@csiro.au)
*CSIRO ICT Centre, Canberra, Australia.*

(Received: ……; Accepted: ……)

Since the late nineteen-nineties Web information retrieval has attracted the attention of both researchers and commercial companies and has been fueled by theoretical results, methods, experiments, and implementations. The contributions published so far form a formidable knowledge base, even though the subject is rather young and many problems are still open and worthy of further research.

A call for papers was issued in November 2003 for this special issue on Web information retrieval. Forty-seven papers were received by the end of March 2004 and were each reviewed by three independent reviewers. The five best papers, i.e. [1], [2], [3], [4] and [5] were selected for the special issue, thus providing a perspective on the subject and outlining the directions for the future. Some additional submissions will be considered, after revision, for publication in the regular issues of the journal.

Table I shows the coverage of the main topics provided by the papers selected for this issue – the central topics of each paper is highlighted with "*", whereas "+" highlights the other topics of the paper.

Table I. Topics versus papers.

| Topic | [1] | [2] | [3] | [4] | [5] |
|---|---|---|---|---|---|
| web user | | | * | | + |
| web search | * | * | * | * | * |
| topic distillation | | * | | | + |
| models | + | + | | | |
| link analysis | * | | | | |
| systems | | | + | + | + |
| performance | * | * | * | * | * |

* Correspondence address: Dipartimento di Ingegneria dell'informazione, via Gradenigo 6/a, 35131 Padova, Italy

The five papers of this special issue report on diverse topics of Web information retrieval from different yet complementary points of view. In this introduction, a summary of the main topics covered by the selected papers is presented. The reader may jump directly to the abstracts for more specific information about the content of individual papers.

Web user communities and behaviour are addressed in [3] and [5]. The former paper leverages the search behaviour of communities of anonymous previous searchers thus combining information from the current user with information from the community, whereas the latter reports on a study of portal search engines in a health domain in which many of the searchers are health "consumers".

Some functionalities for Web search are specifically studied in [3], [4] and [5], although this topic is naturally inherent in all the five papers. In the first of these, the problem of vague queries is addressed by proposing a notion of collaborative search. Using context-sensitive and personalized search techniques, the selection patterns of previous searches are exploited to personalize the results of future searches. A more general type of question answering, which is becoming even more important in the context of Web search, is addressed in [4] – instead of concentrating on factoid questions, the authors propose a methodology to extract information to answer questions similar to those typical of a Frequently Asked Questions list. In [5], a study on high quality search in the area of mental health compares general purpose and portal search engines for domain-restricted searches.

The sometimes ill-understood idea of *topic distillation* recognizes the fact that some relevant Web search results are of much greater utility to searchers than others. For example, the entry page to the BBC or CNN news site is likely to be much more useful to the searcher who types the query 'news' than would an arbitrary news report from the past. Bharat and Henzinger [7] defined topic distillation as a "process of finding quality documents on a query topic".

Topic distillation is addressed in [2] and [5], from two very different perspectives. The former reports extensive experiments aimed at devising optimal evidence combinations within the framework of the topic distillation task of the TREC Web Track. In the 2003 version of this task, quality documents are defined as entry pages of sites principally devoted to the topic, which are not subsites of higher level sites principally devoted to the topic. The authors show that their proposed decision mechanism to devise the optimal combinations of evidence is useful in improving retrieval effectiveness. By contrast, Tang et al [5] take an evidence-based medicine perspective when defining quality within the domain of health information. General-purpose and domain

specific engines are compared both on their ability to return relevant information and on the extent to which the returned documents provide advice which is in accord with best available evidence. An ideal health portal search engine would not return results which, while relevant to the query, provide dubious, misleading or harmful advice.

In the field of Web information retrieval, the issue of modeling has been faced by the researchers in two main ways – one is the adaptation of retrieval models which proved to be effective in non-Web domains, the other is the definition of new models which exploit the availability of links. Paper [2] exemplifies the first way – the authors performed the experiments using two probabilistic models, i.e. Robertson *et al.*'s Okapi model coupled with the BM25 weighing scheme, and Amati and van Rijsbergen's Divergence From Randomness framework. Paper [1] is an example of the second way as it focusses on PageRank, a well-known link-based, query-independent ranking algorithm.

Since the intention of PageRank is to distinguish higher quality or more popular pages, [1] can be seen as falling under the heading of topic distillation. However, this paper assumes the usefulness of PageRank and addresses only the question of efficient implementation. A methodology is proposed which aggregates the large page graph into a smaller host graph from which the PageRank distribution on the pages can be reconstructed. The reported experiments show that the time needed to recompute the PageRank is reduced considerably thus suggesting the possibility of implementing, for example, efficient personalized PageRank algorithms.

In [3], [4] and [5] prototype systems are introduced as proof-of-concept tools or as information retrieval services on the Web. The issue of performance is addressed in all five papers: in [1] more emphasis is put on efficiency than on effectiveness.

We invite readers to also read papers from other special issues devoted to Web information retrieval and mining which may be seen to complement this issue. In [8], the papers are a miscellanea of various aspects of Web information access: algorithms, interfaces, architectures. Aspects of information retrieval are introduced in all papers at different levels of depth. In [9], the papers are revised and extended versions of the ones presented at the 2000 WebKDD workshop held in conjunction with the ACM conference on Knowledge Discovery in Databases. There are good contributions on the use of mining algorithms for e-commerce solutions. Webometrics is an interesting area of research addressed in [10]. The papers address Web domain size, the inter-relationships between domains, the characteristics of end users, and the influence or standing of sites. In [11], six papers "report research in web retrieval and mining. Most papers apply or adapt various pre-web retrieval and

analysis techniques to other interesting and challenging web-based applications." Finally, the papers in [12] describe studies of different sizes reporting on diverse aspects of Web search.

Acknowledgements

We thank the current Editors-in-Chief and Stephen Robertson, for support and the advice. We also thank the colleagues who helped select the papers: Maristella Agosti, Gianni Amati, Einat Amitay, Ricardo Baeza-Yates, Peter Bailey, Nicholas Belkin, Andrei Broder, Jamie Callan, Yves Chiaramella, Charles Clarke, Gordon Cormack, Nick Craswell, Giorgio Di Nunzio, Susan Dumais, Norbert Fuhr, Ayse Goker, Taher Haveliwala, Peter Ingwersen, Ray Larson, Ronny Lempel, Knut Magne Risvik, Vibhu Mittal, Alistair Moffat, Nicola Orio, Iadh Ounis, Jan Pedersen, Luca Pretto, Berthier Ribeiro-Neto, Stephen Robertson, Alan Smeaton, Ross Wilkinson, Hugh Williams, and Nivio Ziviani.

## References

1. A. Broder, R. Lempel, F. Maghoul, and J. Pedersen. Efficient PageRank approximation via graph aggregation.
2. V. Plachouras, F. Cacheda, and I. Ounis. A decision mechanism for the selective combination of evidence in topic distillation.
3. B. Smyth and E. Balfe. Anonymous personalization in collaborative Web search.
4. R. Soricut and E. Brill. Automatic question answering using the Web: Beyond the factoid.
5. T.T. Tang, N. Craswell, D. Hawking, and K. Griffiths. Quality and relevance of domain-specific search: A case study in mental health.
6. K. Bharat and M.R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR'98*, pages 104–111, Melbourne, Australia, 1998.
7. *Issue on Web-based Information Systems, papers from the 10th World Wide Web conference*, volume 20(1) of *ACM Transactions on Information Systems*. ACM Press, 2002.
8. R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, editors. *Special issue on Web mining*, volume 6(1) of *Data Mining and Knowledge Discovery*. Kluwer Academic Publisher, 2002.
9. M. Thelwall and L. Vaughn, editors. *Special topic section on Webometrics*, volume 55(14) of *Journal of the American Society for Information Science and Technology*. Wiley, 2004.
10. H. Chen, editor. *Special topic section on Web retrieval and mining*, volume 54(7) of *Journal of the American Society for Information Science and Technology*. Wiley, 2003.
11. A. Spink, editor. *Special Issue on Web Research*, volume 53(2) of *Journal of the American Society for Information Science and Technology*. Wiley, 2002.