

Using Citations for Ranking in Digital Libraries

Birger Larsen and Peter Ingwersen

Department of Information Studies, Royal School of Library and Information Science, Denmark

{blar,pi}@db.dk

Categories and Subject Descriptors

H.3.3. Information Search and Retrieval; H.3.7. Digital Libraries.

General Terms: Algorithms, Experimentation.

Keywords: Information Retrieval, Citation Indexing.

1. INTRODUCTION

Citations drawn from the list of references of scholarly documents may have many uses in digital libraries, including reference linking [5]. In this poster we present the main results from an experiment designed to exploit the potential power of bibliographical citations for information retrieval in digital libraries. The work is based on the first author's PhD thesis [3].

2. SEARCHING BY CITATION LINKS

The approach tested, the so-called boomerang effect [4] (labelled BE in the following), exploits citation indexing for retrieval in a way similar to CiteSeer and Google Scholar, but with one important difference: Where Google Scholar uses the *number* of citations received as part of the ranking, and CiteSeer mainly focuses on linking backward and forward from *individual* documents, the BE exploits the actual *citation links* for ranking the documents, i.e., to retrieve and rank documents that contain citations to documents on the subject searched for. The citation links themselves have been shown capable of producing high performance in traditional online databases (see, e.g., [6]). Such forward chaining, where documents citing known seed documents are retrieved, is a unique capability of citation indexes, but the main weakness of this retrieval strategy is apparent: the user has to be able to supply a known seed document of relevance [4]. The BE attempts to eliminate this requirement by identifying seed documents automatically. This is done by selecting frequently occurring citations, extracted from top-ranked documents that are retrieved by a regular keyword based search. These extracted citations are then used in a query against the citation index, resulting in a list of documents ranked by how often they contain these citations (see [3,4] for details).

3. EXPERIMENTAL SETUP

The BE was tested in a laboratory experiment using the INEX 2002 test collection [2]. This consists of 12,107 full text XML documents from the journals of the IEEE Computer Society, and includes 22 topics and relevance assessments. Results for two BE variables are reported here: Var-1 is the number of documents from which citations were extracted; Var-2 is the top-n percent of the extracted citations (which were weighted by a simple scheme, see [4]). Mean Average Precision (MAP) was calculated for the first 100 retrieved documents for each combination of Var-1 and Var-2 (Table 1, 27 runs). The keyword based InQuery search engine [1] was used for implementing the BE, and for creating a baseline run (in which InQuery was used without modifications).

4. RESULTS

The baseline run had a MAP of 0.084, which is significantly better than any of the BE runs in Table 1 ($p=0.05$, Friedman test). This is perhaps to be expected as the BE runs do not include keywords in the final query, only citation links. At its best the BE achieved a MAP up to two thirds of that of the baseline, which may be regarded as remarkable in its own right because the baseline system is optimised for keyword based retrieval. Of the two variables in the BE, the number of documents from which citations were extracted (Var-1) had the greatest effect on results. Optimal performance with the INEX2002 test collection was accomplished when citations were extracted from around 32 documents (there were, however, few statistical differences to nearby values of Var-1).

Table 1. Mean Average Precision of the boomerang effect over values of Var-1 (documents from which citations were extracted), and Var-2 (top-n percent of extracted citations)

Var-2	Var-1								
	2	4	8	16	32	64	128	256	512
25%	0.038	0.040	0.045	0.048	0.050	0.049	0.044	0.037	0.031
50%	0.040	0.045	0.049	0.052	0.053	0.051	0.045	0.036	0.031
75%	0.042	0.047	0.052	0.054	0.055	0.051	0.045	0.036	0.030

Significant differences, Var-1: 16, 32, 64, 128 > 256, 512 and 16, 32, 64 > 2. Significant differences, Var-2: none. ($p = 0.05$, Friedman test).

5. DISCUSSION

The initial experiment reported here does not outperform keyword based retrieval, but points to citation links as a potentially unique and hitherto untapped beneficial feature that might be included in search engine ranking algorithms for digital libraries. Further developments of this idea may include utilisation of unique bibliometric properties of citation links, e.g., citation age and frequency information from the network of citations.

6. REFERENCES

- [1] Callan, J. P., Croft, W. B. and Harding, S. M. (1992): The INQUERY retrieval system. In: Tjoa, A. M. and Ramos, I. eds. *Proceedings of DEXA-92*, p. 78-83.
- [2] Fuhr, N. et al. eds. (2003): *Proceedings of INEX2002*. (<http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf>)
- [3] Larsen, B. (2004): *References and citations in automatic indexing and retrieval systems : experiments with the boomerang effect*. Copenhagen: Royal School of LIS. 297 p. (<http://www.db.dk/blar>)
- [4] Larsen, B. and Ingwersen, P. (2002): The boomerang effect : retrieving scientific documents via the network of references and citations. In: Beaulieu, M. et al. eds. *Procs. SIGIR 2002*, 397-398.
- [5] Mischo, W. H., Habing, T. G. and Cole, T. W. (2002): Integration of simultaneous searching and reference linking across bibliographic resources on the web. In: Marchionini, G. ed. *Proceedings of JCDL 2002*, 119-125.
- [6] Pao, M. L. (1993): Term and citation retrieval - a field-study. *Information Processing & Management*, 29(1), 95-112.