

The Boomerang Effect: Retrieving Scientific Documents via the Network of References and Citations

Birger Larsen and Peter Ingwersen

Department of Information Studies, Royal School of Library and Information Science
Birketinget 6, DK-2300 Copenhagen S, Denmark

{blar,pi}@db.dk

1. INTRODUCTION

In Information Retrieval (IR) research interest has increased in extending the traditional bag-of-words approach to accommodate other features of documents than text words and exploit, for example, the document structure [e.g., 1] and links in and between documents [e.g., 5] to improve retrieval performance.

The theory of polyrepresentation [2, 3] provides a theoretical background for how to exploit such features. In summary, the theory hypothesises that *overlaps* between different cognitive representations of both users' information needs as well as documents can be exploited for reducing the uncertainties inherent in IR, and thereby improve the performance of IR systems. Two or more different cognitive representations pointing at the same documents is regarded as multi-evidence of those documents being relevant, and suggests to apply a principle of 'intentional redundancy' [2] with the purpose of reducing the uncertainties by placing emphasis on overlaps between representations. Better results are expected when cognitively unlike representations are used, e.g., the document title (made by the author) vs. intellectually assigned descriptors from indexers.

Based on the theory of polyrepresentation we propose the so-called Boomerang effect as a method for retrieving scientific documents, with representations generated from 1) the structure of these documents as composed by the author(s) (i.e. title, abstract, introduction, methodology, (sub)sections, conclusion etc.), and 2) links, citations, intellectually assigned descriptors, etc. generated by other cognitive agents. A special matching function is used where bibliographic references *from* documents and citations *to* documents are used as index terms for matching documents with the query, resulting in an expansion of the initially retrieved set. The bibliometric aspects of the Boomerang effect were introduced in [4] along with detailed results from a pre-experiment testing the approach. In this paper we present the Boomerang effect from an IR perspective along with the main results from the pre-experiment, and develop the method from a purely Boolean approach to a best match approach.

2. THE BOOMERANG EFFECT¹

The theory of polyrepresentation from IR is extended with methods from bibliometrics and scientometrics to generate the Boomerang effect. Studies from the operational online IR community have shown that using citation search strategies can yield high performance [e.g., 6], but these strategies typically require a user to select intellectually one or more seed documents at least a few years old that are relevant for the information need, and may have been cited subsequently – a situation not readily

applicable to most IR research settings. The purpose of the Boomerang effect is to exploit the potentially high performance offered by incorporating link or citation information, and at the same time allow for queries formulated in natural language.

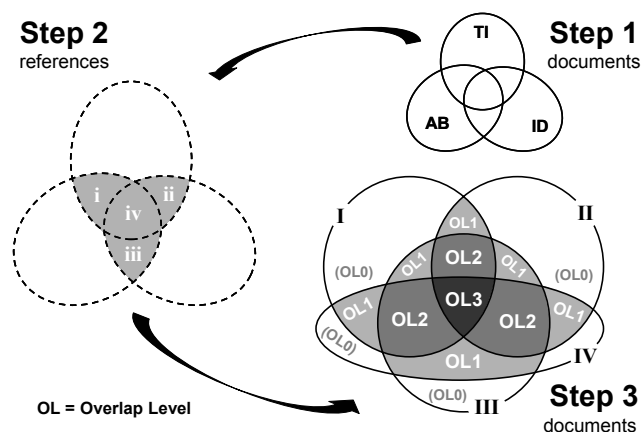


Figure 1: Example of the Boomerang effect with 3 representations. At step 3, 4 sets of documents (I-IV) citing the references contained in the overlaps at step 2 (i-iv) are retrieved, using the references as query terms. (Modified from [4])

In short, several sets of documents are retrieved using natural language queries in different cognitive representations (Fig. 1, step 1). From the retrieved sets of documents the references are extracted into a *pool* for each set, and distinct overlaps between these pools of references are identified (overlaps i-iv in Fig. 1, step 2). Following the theory of polyrepresentation and the principle of intentional redundancy only the references in the overlaps are used in the following, where, for each overlap, the set of documents that cite the references in the overlaps at step 2 are retrieved (Fig. 1, step 3). Finally, the overlaps between these citing documents at step 3 are identified and used to order the documents in an overlap structure. The result of the Boomerang effect is the documents retrieved in step 3, and the overlap structure generated in step 3, where the documents are partially ordered in a number of overlap levels (OL). The hypothesis in the Boomerang effect is that the computed precision at OL3 will be greater than at OL2, which will be greater than at OL1 etc. The number of possible OLs in step 3 is determined by the number of sets in step 1 and the number of non-empty overlaps in step 2. Note that additional documents not present in step 1 may be retrieved in step 3, because they (co-)cite the same references.

¹ We use the boomerang as an analogy because of the loop made back in time and past research, returning to yield (ideally) high precision (See Fig. 1).

3. PRE-EXPERIMENT

Three work tasks describing an information need were constructed in co-operation with a researcher at a hospital, who subsequently assessed the relevance of the retrieved documents. A query for each work task was constructed manually and submitted to four document representations (Title, abstract, identifiers and free-text) in Science Citation Index® (SCI), which resulted in an average of 88 documents per work task. The Boomerang effect was executed as detailed above, and returned 476 documents from SCI on average per work task at OL1-4, expanding the initially retrieved set considerably. A document in step 1 will also be retrieved in step 3 if at least one of its references is contained in an overlap at step 2 – 91% of the documents were retrieved in both steps in the pre-experiment. Due to the large number of documents retrieved not all could be assessed, and two samples were drawn: Table 1 shows the precision calculated for 135 documents from step 1 also retrieved in step 3 (A), and for 143 additional documents retrieved by the Boomerang effect (B). It can be seen at each OL that precision is greater than or equal to the OL below. This holds for both samples, as well as for each individual work task (not shown).

Table 1. Precision calculated on overlap levels (OL) for sample A: 135 documents from step 1 also retrieved in step 3, and for sample B: 143 additional documents retrieved in step 3. The figures cover all three work tasks (Modified from [4])

	Sample A		Sample B	
	# documents	Precision	# documents	Precision
OL4	1	100,0 %	1	100,0 %
OL3	5	100,0 %	10	80,0 %
OL2	28	92,9 %	23	56,5 %
OL1	59	72,9 %	68	39,7 %
OL0	42	66,7 %	41	17,1 %
	135	76,3 %	143	39,2 %

4. BEST MATCH BOOMERANG EFFECT

The Boomerang effect can, as demonstrated above, retrieve documents that are potentially relevant, partially ordered in a number of levels. However, it does not rank the documents *within* a given OL because it uses a strictly Boolean approach. Below a best match Boomerang effect is proposed.

In the Boomerang effect the bibliographical references extracted from the documents in step 1 (See Fig. 1) are in essence used as *index terms*, and may as such be weighted. The frequency of occurrence of a reference in a pool in step 2 can be normalised by the total number of references in that pool. As only references occurring in the *overlaps* in step 2 are considered each reference will occur in at least two pools and may therefore have several weights assigned. In accordance with the theory of polyrepresentation, that emphasises the importance of multi-evidence, these can be *added* into one, final weight for every reference in the overlaps in step 2. In Table 2 a matrix is shown, which maps the occurrences of five references (i_{1-5}) in three pools (p_{1-3}) as an example. Note that i_5 is not assigned any weight as it occurs in one pool only. The references with associated weights can be used as weighted query terms in step 3 in an IR model that allows this, e.g., the vector space model. The result will be a list of documents with at least one of these references, ranked by how many and how heavily weighted references they contain. Using

this approach the rigidity of the Boolean sets from the pre-experiment has been broken down, and a single, continuous rank can be created by the Boomerang effect that may be compared to the performance of other best match systems.

Table 2. Example of the calculation of weights for references in the overlaps at step 2, based on the occurrence of references (i_{1-5}) in the reference pools (p_{1-3}).

	Frequency of occurrence of references in each pool					$\sum_{i=1}^5 f_{i_j}$	References weighted by the total number of references in each pool				
	i_1	i_2	i_3	i_4	i_5		i_1	i_2	i_3	i_4	i_5
p_1	2	1	1	0	0	4	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	-
p_2	1	2	0	1	0	4	$\frac{1}{4}$	$\frac{2}{4}$	0	$\frac{1}{4}$	-
p_3	2	1	2	1	1	7	$\frac{2}{7}$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{1}{7}$	-
	Final weights for i_{1-5} :						1,04	0,89	0,54	0,39	-

5. CONCLUSION AND FUTURE WORK

The potential of the Boomerang effect for yielding high precision and expanding the initial set has been indicated with data from a Boolean pre-experiment. A best match Boomerang effect is proposed with focus on weighting references based on their frequency of occurrence. Further work is presently being done on implementing and testing the best match approach with a high number of cognitive representations generated from a corpus of 12,000 scientific full text articles in XML (The INEX collection – <http://qmir.dcs.qmw.ac.uk/inex>).

6. ACKNOWLEDGMENTS

The authors wish to acknowledge the Danish branch of Dialog for generous access to their files, the helpful comments and support from the researchers at the Department for Information Studies at Tampere University, Finland, as well as the constructive comments from the three anonymous reviewers. The research reported in this paper is carried out as part of the TAPIR (Text Access Potentials for interactive Information Retrieval) research project at the Royal School of Library and Information Science.

7. REFERENCES

- [1] Chiamarella, Y. (2001): Information retrieval and structured documents. In: *Lectures on Information Retrieval: ESSIR 2000*. Berlin: Springer, p. 286-309. (LNCS; 1980).
- [2] Ingwersen, P. (1994): Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In: *Proceedings of SIGIR 1994*, p. 101-110.
- [3] Ingwersen, P. (1996): Cognitive perspectives of information-retrieval interaction - elements of a cognitive IR theory. *Journal of Documentation*, 52(1), p. 3-50.
- [4] Larsen, B. (2002): Exploiting citation overlaps for Information Retrieval - generating a boomerang effect from the network of scientific papers. *Scientometrics*, 54(2), p. 155-178.
- [5] Lee Giles, C., Bollacker, K. D. and Lawrence, S. (1998): CiteSeer: An automatic citation indexing system. In: *Proceedings of Digital Libraries 1998*, p. 89-98.
- [6] Pao, M. L. (1993): Term and citation retrieval - a field-study. *Information Processing & Management*, 29(1), p. 95-112.