Contextualization from the Bibliographic Structure

Muhammad Ali Norozi
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
mnorozi@idi.ntnu.no

Arjen P. de Vries
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
arjen@cwi.nl

Paavo Arvola
Department of Computer Sciences
University of Tampere
Tampere, Finland
paavo.arvola@uta.fi

ABSTRACT

Bibliographic or citation structure in a document contains a wealth of useful but implicit information. This rich source of information should be exploited not only to understand *what* and *where* to find the important documents, but also as a contextual evidence surrounding the important and not so important documents. This paper measures the effects of *contextual* evidences accumulated from the bibliographic structure of documents on retrieval effectiveness.

We propose a re-weighting model to *contextualize* bibliographic evidences in a query-independent and query-dependent fashion (based on Markovian random walks). The *in-links* and *out-links* of a node in the citation graph could be used as a context. Here we hypothesize that the document in a *good* context (having strong contextual evidences) should be a *good* candidate to be relevant to the posed query and vice versa.

The proposed models are experimentally evaluated using the *i*Search Collection and assessed using standard evaluation methodologies. We have tested several variants of contextualization, and the results are significantly better than the baseline (indri run).

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models*; H.3.4 [**Information Storage and Retrieval**]: System and Software—*performance evaluation*; H.2.1 [**Database Management**]: Logical Design—*data models*; E.1 [**Data**]: Data structures—*trees*; E.5 [**Data**]: Files—*organization/structure*

Keywords

Contextualization, Re-weighting, Random walks

General Terms

Measurement, Performance, Design, Experimentation

ECIR '2012 Barcelona, Spain

1. INTRODUCTION

Documents' bibliographic structure (i.e., inlinks and outlinks) provides both a wider *context* and a wider *semantics* to the content. This far-reaching context and semantics should possibly be used to boost or reduce the documents retrieval scores. Without using the structural information (citations graph), the search system would simply ignore the documents containing a wealth of implicit information in its context as irrelevant to the query topic in question.

Until recently, the importance of contextualization has been studied in several settings by [1, 2, 9, 7, 10] in a schema-agnostic environment. It has been found that by contextualizing the scores of the surrounding components, elements or parents (ancestors) or siblings in the scoring function of the element itself, the overall precision and recall of the focused retrieval system improves [2].

In this study we incorporate the idea of random walk together with contextualization on bibliographic structure of documents, inspired by the random surfer model of [4, 3] over XML documents and relational databases respectively. The hypothesis is that this would improve the search effectiveness in aggregated search.

Shortly, the contributions of this study include:

- The introduction of contextualization with random walk as a theoretically sound model (Section 2).
- Experimental validation of the ideas proposed using query-independent/-dependent random walk with inlinks and outlinks contextualization(Section 3).
- Evaluated the use of bibliographic information on (a subset of) the *i*Search Collection [6] (Section 3.1).

Section 4 concludes and highlights future work.

2. CONTEXTUALIZATION MODEL

Contextualization is a method exploring the features in the context of a retrievable unit [2]. In document retrieval, in turn, this means combining the evidences from a document and its context using different but plausible combination functions. The context of a document consists of other documents which point-to or are pointed-to (*contextualizing* documents) by the document in question (*contextualized* document, P2), see Figure 1(a). We use random walks to induce a similarity structure over the documents based on their bibliographic relationships. Hence, these relationships affect the weight each contextualizing document has in contextualization. A contextualization model is a re-scoring scheme, where the basic score, usually obtained from a fulltext retrieval model, of a contextualized document is re-enforced by the weighted scores of the contextualizing

documents.

The premise is that *good context* (identified by random walk and contextualization) provides evidence that a document is a good candidate for a posed query and therefore documents should be contextualized by their bibliographically similar documents. Good context is an *evidence* that should be used to deduce that a document is a good candidate for the posed query.

2.1 Random Walk for context materialization

There are enough empirical and intuitive proof for the premise that a good document in citation graph is good because it contains references to alot of good documents, and more importantly, a good document is good if it is contained in a good document as a reference (recursive definition) [5, 8]. But here, the question is, can the evidences, lying loosely in the context surrounding the contextualized document, be intelligently materialized? Fortunately, the answer is yes, later in the section we will show a formalism that can be used to materialize and then utilize the contextual evidences for improving retrieval effectiveness.

Previous work [1, 2] presents a contextualization model where a binary vector represents the relevant context (a part of) a document. Here, we extend that work to use probabilistic infomation derived from a random walk over the citation structure. A random walk on the citation structure of the documents independent or dependent of a query topic will populate the contextualization vector with the probablisties that indicate *authority* of a document in the network of citations.

An alternative way to conceive the intuition behind the random walk model here is, to consider that authority and relevance information flows in the bibliographic structure of documents in the same fashion as that of the HITS model [5]. The authority flows in the bibliographic structure of documents until an equilibirium is established which specifies that a document is authoritative if it is referenced by authoritative documents [8].

The bibliographic network of documents (for example, Figure 1(a)) can be represented in matrix notation by adjacency matrix $\mathbf{A}$ such that:

$$\mathbf{A_{ij}} = \begin{cases} 1 & \text{if there is a link from page } P_i \text{ to } P_j \\ \varepsilon & \text{if } \mathbf{A_{ij}} = 0 \text{ and there is a link from page } P_j \text{ to } P_i, \\ & 0 < \varepsilon \ll 1 \\ 0 & \text{otherwise} \end{cases}$$

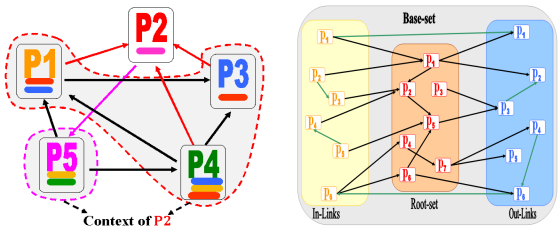
The reverse edge ε, very small value, is added to ensure a unique solution to the system of linear Equations 1. For the Figure 1(a) the corresponding adjacency matrix $\mathbf{A}$ can be:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & \varepsilon & \varepsilon \\ \varepsilon & 0 & \varepsilon & \varepsilon & 1 \\ \varepsilon & 1 & 0 & \varepsilon & 0 \\ 1 & 1 & 1 & 0 & \varepsilon \\ 1 & \varepsilon & 0 & 1 & 0 \end{pmatrix}$$

The random walk probabilities are then obtained by iteratively solving the following system of linear equations[1]:

$$g^k = \mathbf{A^T}\ \mathbf{A} g^{k-1} \quad (1)$$

[1]Finding the dominant Eigenvector of the system of linear equations, corresponding to the dominant eigenvalue, which is 1 in this case [8].



(a) Bibliographic network of 5 docs & context of P2 (b) The Base-set

Figure 1: Bibliographic information and relevant retrieved

Here g^k is the proposed contextualization vector, and k is the number of iterations. The matrix $\mathbf{A^T}\ \mathbf{A}$ constructed this way would lead to a *unique* solution to the system of linear Equations 1 [5].

2.2 Query independent and query-dependent walks

A *query independent random walk* is conducted on the entire bibliographic structure of the documents, irrespective of any query. This walk primarily captures the authoritativeness of documents in the collection. The adjacency matrix $\mathbf{A}$ becomes huge in this case ($342,279 \times 342,279$, see Section 3.1). The contextualization vector g^k depicts the scores of each document in the massive citation graph for the entire collection iteratively calculated using Equation 1.

A *query dependent random walk* is conducted on the rather smaller subset of the citation graph, corresponding to a specific query topic in question. Adjacency matrix $\mathbf{A}$ is in this case considerably smaller then the query-independent walk. The contextualization vector g^k depicts the stationary distribution of random walk (scores of documents) specific to a query. The focused subgraph can be constructed from the output of text-based search engine (indri in our case) which can be used to iteratively produce set of documents that are most likely considered to be relevant to the query topic. The Base-set S_q (which is used to form $\mathbf{A}$) can be obtained by growing query results (Root-set R_q); which includes any document that pointed to by a document in Root-set R_q, and any document that points to a document in R_q, i.e., in-linking and outlinking documents from root-set R_q respectively(see Figure 1(b)).

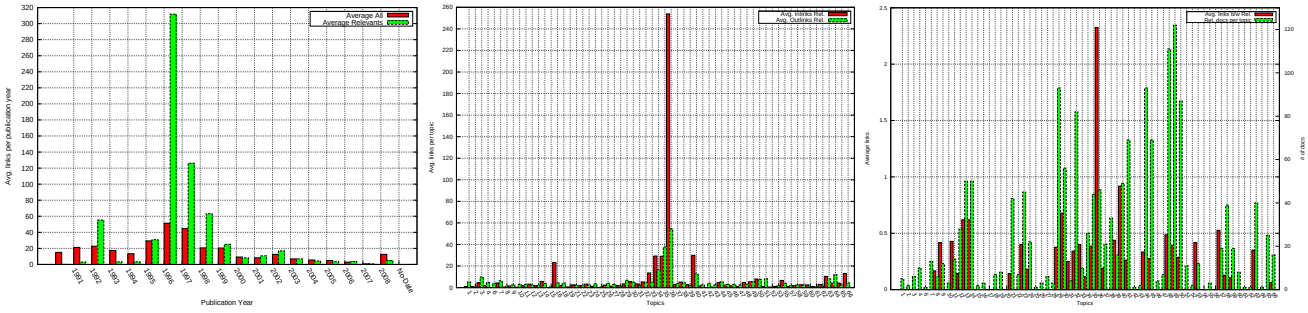
2.3 Combination function

We now give a tailored re-ranking function CR, which allows the contextualizing scores to be added to the basic scores. The function can be formally defined as follows:

$$CR(x, f, C_x, g^k) = (1-f) \cdot BS(x) + f \cdot \frac{\sum_{y \in C_x} BS(y) \cdot g^k(y)}{\sum_{y \in C_x} g^k(y)} \quad (2)$$

where

- $BS(x)$ is the basic score of contextualized document x (text-based score, e.g., $tf \cdot idf$)
- f is a parameter which determines the weight of the context in the overall scoring



(a) Average in-links per publication year

(b) Only relevants

(c) Average # of links between relevant docs

Figure 2: Average Number of links per topic

- C_x is the context surrounding the contextualizing document x, i.e., $C_x \subseteq (inlinks(x) \cup outlinks(x))$, $\subseteq$, because we are only considering the set of inlinks and / or outlinks of x in the top$-k$ retrieved documents($k \in$ 1500 and $8k$), not all the inlinks and outlinks of x.
- $g^k(y)$ is the contextualization vector which gives the authority weight of y, the contextualizing documents of x.

We can have several variants of the combination function of Equation 2, as discussed in forthcoming Sections below.

2.4 Context as the authority

Do documents cited a lot, or documents containing more in-links or authoritative documents form a good context? Let's assume that the context function C_x in Equation 2 only contextualize based on the in-links. In this case the argument would be: $C_x \subseteq inlinks(x)$. The set C_x only contains the in-links of the contextualizing document. The inlinks of a document x corresponds to its column in the adjacency matrix $\mathbf{A}$. For example, the inlinks of document $P2$ in the Figure 1(a) correspond to the non-zero cells of column 2 in the adjacency matrix $\mathbf{A}$.

Section 3 presents experiments with two variants of contextualization:

1. *first* based on random walk conducted on query independent adjacency matrix $\mathbf{A}$ (the entire bibliographic graph, see Section 2.2) and
2. *second* based on query dependent random walk on adjacency matrix $\mathbf{A}$ (the base-set, see Figure 1(b)).

We have experimented with both of the approaches, see Section 3. In addition to the two variants, a third variant combines the query independent and query dependent random walk into a combination function:

$$CR(x, f, C_x, g^k_{qi}, g^k_{qd}) = (1-f) \cdot BS(x) + f \cdot \alpha \cdot \frac{\sum_{y \in C_x} BS(y) \cdot g^k_{qi}(y)}{\sum_{y \in C_x} g^k_{qi}(y)} + f \cdot (1-\alpha) \cdot \frac{\sum_{y \in C_x} BS(y) \cdot g^k_{qd}(y)}{\sum_{y \in C_x} g^k_{qd}(y)} \quad (3)$$

where

- $g^k_{qi}(y)$ is the contextualization vector which gives the authority weight of the contextualizing documents of x based on query independent walk.
- $g^k_{qd}(y)$ is the contextualization vector which gives the authority weight of the contextualizing documents of x based on query dependent walk.
- α is the parameter moderating the share of contextualization from query independent and query dependent.

2.5 Context for a better content description

Given the bibliographic structure of the iSearch collection, Figure 2 shows that the numbers of inlinks in the documents are not very stable along year and along the query topics. The existence of inlinks for contextualized document is certainly a positive indication, but outlinks also happen to occur in the contextualized document's context. Inlinks together with outlinks provide a much wider context for the contextualized document. Combination functions, Equations 2 and 3 remain the same, only the interpretation of the contextualization function changes now to: $C_x \subseteq (inlinks(x) \cup outlinks(x))$. The set C_x now contains the inlinks and outlinks of the contextualizing document, containing the query term. The outlinks of a document x correspond to its row in the adjacency matrix $\mathbf{A}$. For example, the outlinks of document $P2$ in the Figure 1(a) corresponds to the non-zero cells of row 2 in the adjacency matrix $\mathbf{A}$.

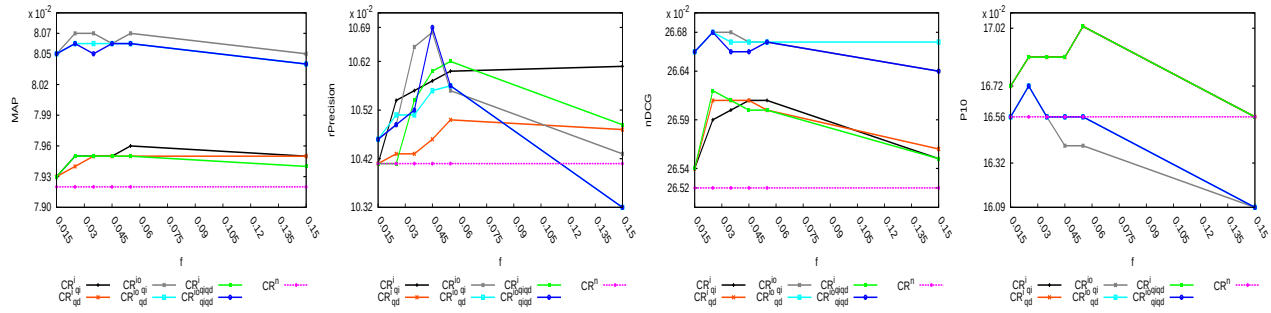
3. EXPERIMENTAL EVALUATION

3.1 Experimental Settings

The proposed approaches are evaluated using the newly released *i*Search test collection, consisting of 65 queries with relevance assessments. The collection contains 18,443 book records in XML (BK), 291,246 metadata of articles are in XML (PN) as well as 143,571 full text articles are in PDF (PF). The query set is provided with a description of the information need, task, background, ideal answer and a few keywords. We have used the keywords as query text for our experiments, because that resulted in the highest effectiveness with our baseline system.

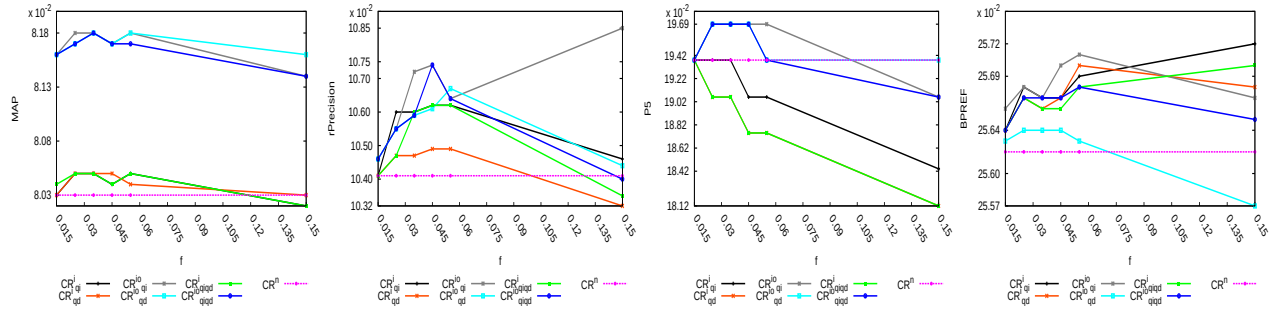
We believe that this is the first study to use the citation structure provided with the collection, based on Citebase semi-autonomous citation index[2]. There are certain limi-

[2]Citebase is created by Tim Brody from Univerity of



(a) MAP for different f (b) rPrecision (c) nDCG (d) P@10

Figure 3: Trends for different measures @1500



(a) MAP for different f (b) rPrecision (c) P@5 (d) BPREF

Figure 4: Trends for different measures @8k

tations to the citation structure extracted namely: (a) citations only covers citations among the PN and PF documents in iSearch. (b) citations has been extracted automatically.

We first evaluated our baseline system on the entire collection, and obtained a satisfactory result when compared to the related works: our Indri baseline gives a MAP of 0.1048 retrieving 1,667 relevant documents, a performance higher than earlier published results of [6].

We now define a subset of the collection that has sufficient coverage in citation structure, that we will refer to as *iSearch-Citations*. We keep only those documents that have citations (342,279 out of the original 434,817 PN and PF documents), discarding the rest from the experiments and evaluations. The baseline performance drops to a MAP of 0,0792 on this reduced data set, retrieving 974 relevant documents in the top 1,500 documents, and 1256 relevant documents retrieved in the top 8,000 documents retrieved per query. These choices are based on following reasons: (i) to widen the context, e.g., when we retrieve 1,500 documents per topic then we have a narrower context than, when 8,000 documents are retrieved per topic, (ii) to have a better coverage of the relevant documents and subsequently boost their rankings, based on their inlinks and outlinks, with the help of the proposed approaches (see next Section and Table 1 and 2).

A total of 3,768,410 citations contained in 219,242 PN documents and 123,037 PF documents. The original graded qrels contain 11,264 documents, out of which 2,878 have been assessed to be relevant. After pruning the documents without citation structure, we have 6,975 documents, of which 1,591 are relevant ones, in the modified graded qrels.

3.2 Results

We have tested seven different retrieval methods based on the propositions (see Section 2).

- No contextualization, indri run using *#combine* operator for combining beliefs and using the keywords field from queries provided, CR^n (baseline)
- Query independent - inlinks contextualization, CR^i_{qi}
- Query dependent - inlinks contextualization, CR^i_{qd}
- Query independent and dependent - inlinks contextualization, CR^i_{qiqd}
- Query independent - inlinks and outlinks contextualization, CR^{io}_{qi}
- Query dependent - inlinks and outlinks contextualization, CR^{io}_{qd}
- Query independent and dependent - inlinks and outlinks contextualization, CR^{io}_{qiqd}

For each evaluation measure (Table 1) seperately, we tuned the following parameters and report the best performance: (i) the contextualization force f from Equation 2 ($f \in \{0.015, 0.025, 0.035, 0.045, 0.055, 0.15\}$); (ii) the α parameter from Equation 3 $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$, a total of 96 runs (as each parameter contains 6 different values) per query independent and query dependent method. The α parameter is only involved in the CR^i_{qiqd} and CR^{io}_{qiqd} runs, as reflected in Equation 3, i.e., runs involving both query- independent and -dependent walks. These optimal values for f and α are obtained training with the *i*Search collection. Figures 3 and 4 illustrate the behaviour of the methods as we change

Southampton, UK, http://citebase.org

mainly the f parameter, from Equations 2 and 3, on some of the significant retrieval measures. Due to space limitations, we only report $\alpha = 0.5$ (which was one of the optimal values during training, see last column of Tables 1 and 2). As can be visually observed, the proposed methods out-perform the baseline, CR^n, in almost all the figures.

Table 1 and 2 show the overview of the retrieval performance of our approaches against the baseline at $1,500$ and $8,000$ documents retrieved per topic. All the proposed contextualization models improves the performance over baseline. The improvements are statistically significant (2-tailed t-test $p < 0.05$) on $rPrecision$, $nDCG$ and $P10$ measures. Note that, queries having no relevant results in relevance assessments (queries 5, 17, 20, 54 and 56) are not removed during evaluations and statistical significance assessments. The improvements overall are not surprisingly good because of the connectivity of the relevant documents per topic, as can be seen graphically in Figure 2(c). The preliminary per-query analyses showed a much better improvement, when we assess and evaluate one query at a time. Queries containing a wider context (such as, query 36) lay on a greater hope for the proposed approaches. Due to space limitations, we will not go in further detail about those results here.

The best overall results among the proposed methods are obtained with CR^{io}_{qi} and CR^{io}_{qiqd}, in terms of highest mean average precision values. We conclude that, context provided by in- and outlinks may indeed improve retrieval effectiveness, even though improvements are still small, but statistically significant.

				@1500				
Method	f	MAP	rPrecision	nDCG	BPREF	P5	P10	α
Baseline (CR^n)	–	.0792	.1041	.2652	.2323	.1938	.1656	–
CR^{i}_{qi}	.055	.0796$^{\triangle}$	**.1060**$^{\blacktriangle}$	.2661$^{\triangle}$	.2330$^{\triangle}$	.1906	**.1703**$^{\blacktriangle}$	–
CR^{i}_{qd}	.055	.0795$^{\triangle}$	**.1050**$^{\blacktriangle}$	**.2661**$^{\blacktriangle}$	.2330$^{\triangle}$	.1938	**.1703**$^{\blacktriangle}$	–
CR^{i}_{qiqd}	.025-.055	.0795$^{\triangle}$	.1063$^{\triangle}$	**.2662**$^{\blacktriangle}$	.2329$^{\triangle}$	.1938	**.1703**$^{\blacktriangle}$	.2-.7
CR^{io}_{qi}	.035-.055	.0807$^{\triangle}$	**.1068**$^{\blacktriangle}$	.2668$^{\triangle}$	.2326$^{\triangle}$	.1938	.1656	–
CR^{io}_{qd}	.025-.055	.0806$^{\triangle}$	**.1057**$^{\blacktriangle}$	.2668$^{\triangle}$	.2326$^{\triangle}$	.1938	.1672$^{\triangle}$	–
CR^{io}_{qiqd}	.035-.055	.0807$^{\triangle}$	**.1069**$^{\blacktriangle}$	.2667$^{\triangle}$	.2325$^{\triangle}$	.1938	.1656	.2-.7

Table 1: Ret. performance @1500 $^{\blacktriangle}$= stat. significance at p < 0.05 (2-tailed t-test). $^{\triangle}$= better than baseline

				@8K				
Method	f	MAP	rPrecision	nDCG	BPREF	P5	P10	α
Baseline (CR^n)	–	.0803	.1041	.2873	.2562	.1938	.1656	–
CR^{i}_{qi}	.055	.0805$^{\triangle}$	.1062$^{\triangle}$	.2878$^{\triangle}$	.2569$^{\triangle}$	.1906	**.1703**$^{\blacktriangle}$	–
CR^{i}_{qd}	.055	.0804$^{\triangle}$	.1049$^{\triangle}$	.2878$^{\triangle}$	.2570$^{\triangle}$	.1875	**.1703**$^{\blacktriangle}$	–
CR^{i}_{qiqd}	.055	.0805$^{\triangle}$	.1062$^{\triangle}$	.2878$^{\triangle}$	.2569$^{\triangle}$	.1875	**.1703**$^{\blacktriangle}$	.2-.7
CR^{io}_{qi}	.035-.055	**.0818**$^{\triangle}$	**.1074**$^{\blacktriangle}$	**.2890**$^{\triangle}$	.2571$^{\triangle}$	.1969$^{\triangle}$	.1625	–
CR^{io}_{qd}	.025-.055	**.0818**$^{\triangle}$	**.1067**$^{\blacktriangle}$	.2889$^{\triangle}$	.2563$^{\triangle}$	.1969$^{\triangle}$	.1625	–
CR^{io}_{qiqd}	.035-.055	**.0818**$^{\triangle}$	**.1074**$^{\blacktriangle}$	**.2890**$^{\triangle}$	.2571$^{\triangle}$	.1969$^{\triangle}$	.1625	.2-.7

Table 2: Retrieval performance @8k

4. CONCLUSIONS AND FURTHER WORK

We have presented an exploratory study into the use of context from bibliographic information to improve retrieval performance on a document retrieval task. The approach is generic and maybe applied beyond the *i*Search-Citations collection studied in this paper. The approaches proposed are particularly suited for collections with less textual evidences. The evidences are collected in a systematic way from the surrounding context of the document to be ranked. The importance of each single unit in the context is identified by the markovian random walk. Most of the proposed system are tested and found to be statistically significant against the baseline, which had a better mean average precision than the so far published results. The proposed methods both boost the rankings of the documents in good context and degrade the rankings of documents in not so good context.

The effectiveness of random walk to materialize the context was tested with six different methods. We have found that the context from in- and out-links can indeed help improve retrieval results, albeit not by a large margin. Given that the collection has not a very steady citation structure based on the amount of context present in the relevant documents (assessed), still, contextualization together with random walk is significantly plausible, both theoretically and empirically. We consider our experiments on the *iSearch-Citations* collection sufficiently promising to consider different types of evidence in future work. Specifically, we would like to investigate the effects of context derived from tweet mentions that may help improve retrieval from video collections.

5. ACKNOWLEDGEMENTS

This study was supported by Academy of Finland under grant #130482.

6. REFERENCES

[1] P. Arvola, M. Junkkari, and J. Kekäläinen. Generalized contextualization method for XML information retrieval. In *Proc. of the 14th ACM international conference on Information and knowledge management*, pages 20–27. ACM, 2005.

[2] P. Arvola, J. Kekäläinen, and M. Junkkari. Contextualization models for XML retrieval. *Info. Processing & Management*, pages 1–15, 2011.

[3] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-based keyword search in databases. In *Proc. of the 13th international conference on Very large data bases-Volume 30*, pages 564–575. VLDB Endowment, 2004.

[4] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked keyword search over XML documents. In *Proc. of the 2003 ACM SIGMOD international conference on Management of data*, pages 16–27. ACM, 2003.

[5] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[6] M. Lykke, B. Larsen, H. Lund, and P. Ingwersen. Developing a test collection for the evaluation of integrated search. *Advances in IR*, pages 627–630, 2010.

[7] Y. Mass and M. Mandelbrod. Component ranking and automatic query refinement for XML retrieval. *Advances in XML IR*, pages 1–18, 2005.

[8] M.A. Norozi. IR Models and Relevancy Ranking. Master's thesis, University of Oslo, 2008.

[9] P. Ogilvie and J. Callan. Hierarchical language models for XML component retrieval. *Advances in XML IR*, pages 269–285, 2005.

[10] G. Ramirez Camps. *Structural Features in XML Retrieval*. PhD thesis, SIKS, the Dutch Research School for Information and Knowledge Systems., 2007.