Contextualization using Hyperlinks and Internal Hierarchical Structure of Wikipedia Documents

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

Paavo Arvola
School of Information Sciences
University of Tampere
Tampere, Finland
paavo.arvola@uta.fi

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

ABSTRACT

Context surrounding hyperlinked semi-structured documents, externally in the form of citations and internally in the form of hierarchical structure, contains a wealth of useful but implicit evidence about a document's relevance. These rich sources of information should be exploited as contextual evidence. This paper proposes various methods of accumulating evidence from the context, and measures the effect of *contextual* evidence on retrieval effectiveness for document and focused retrieval of hyperlinked semi-structured documents.

We propose a re-weighting model to *contextualize* (a) evidence from citations in a query-independent and query-dependent fashion (based on Markovian random walks) and (b) evidence accumulated from the internal tree structure of documents. The *in-links* and *out-links* of a node in the citation graph are used as external context, while the internal document structure provides internal, within-document context. We hypothesize that documents in a *good* context (having strong contextual evidence) should be *good* candidates to be relevant to the posed query, and vice versa.

We tested several variants of contextualization and verified notable improvements in comparison with the baseline system and gold standards in the retrieval of full documents and focused elements.

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

XML retrieval, Semi-structured data, Structural indices, Schema agnostic search, Contextualization, Re-weighting, Random walks

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

1. INTRODUCTION

Focused or element retrieval addresses the possiblity to utilize the hierarchical structure of documents, and hence return the most specific (and exhaustive) text units, rather than returning only full documents. One problem with this approach is that the retrieval units have varying length in textual content, as the size of elements varies with the level in the hierarchy (see Figure 3); the leaf element or descandent elements have less textual evidences than their ancestors. This scant textual evidence makes matching those small text units, such as paragraphs, hard. As a consequence, although they are what the users (might) require, they are considered less relevant by the focused retrieval systems, only because they have too few textual content, hence too little evidence to be ranked higher for the posed user query. Fortunately, this scant textual evidence can be alleviated significantly by a method called *Contextualization* [16].

Contextualization is a mechanism to estimate the relevance of a given structural text or document unit with information obtainable from - besides the unit itself - the surrounding structural text or document units, i.e., from the context of the unit [16]. With contextualization, we assume that context of a retrievable unit gives hints about the relevance of the retrievable unit (can be document or element retrieval). Hence, it is expected in contextualization that context of a retrievable unit gives hints about the relevance of the retrievable unit.

In this study, we incorporate the idea of random walk together with contextualization on citation structure of documents and internal hierarchical structure of XML document. The approach is inspired by the random surfer model of [5, 10] over XML documents and relational databases respectively, as well as the contextualization model for XML retrieval developed by Arvola et al. [4]. The hypothesis is that contextualization together with random surfer (or walk) model will improve search effectiveness over considering retrieval units in isolation.

Until recently, the importance of contextualization (based on hierarchical relationships of element) has been studied in several settings [1, 2, 4, 19, 22, 23, 25]. Even in a schema-agnostic environment, it has been found that by contextualizing the scores of the surrounding components, such as, parents, ancestors or siblings in the scoring function of the element itself, the overall precision and recall of the focused retrieval system improves [4]. In document retrieval, the hyperlink structure of documents (i.e., inlinks and out-

links) 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. Contextualization based on the bibliographic structure of scientific documents has been shown a promising direction in [22].

The models proposed in this research paper are experimentally evaluated using the semantically annotated Wikipedia XML Collection from INEX [26], both at the granularity of a document (document retrieval) and at the XML element level (focused retrieval). We have applied several variants of contextualization, and the results are in-line with the proposed theory about the effectiveness of contextualization. The results obtained, on both document (article level) and focused retrieval (paragraph level) tasks, exhibit clear improvements over a strong and competitive baseline system – itself based on data fusion over all INEX 2009 submitted runs (see Section 3), and already achieving a performance higher than any INEX 2009 official run.

Summarizing, the contributions of this study include:

- Contextualization of the citation structure of hyperlinked documents, with random walks as a theoretically sound foundation (Section 2.1).
- Contextualization of the hierarchical structure of documents, using the same random walk model (Section 2.2).
- Developing a competitive focused retrieval system baseline based on data fusion and constructing a test setting for evaluating the retrieval of small textual units, i.e., paragraphs (Section 3).
- Experimental validation (Section 4) of the ideas proposed, using citation (Section 2.1), hierarchical (Section 2.2) and hybrid contextualization (Section 2.3) within the random walk framework.
- Evaluation of the use of citation and hierarchical information on the large semantically annotated Wikipedia XML corpora [3, 8, 11, 13, 26] (Section 4.1).

Section 5 concludes and highlights future work.

2. CONTEXTUALIZATION MODELS

Contextualization is a method of exploring the features in the context of a retrievable unit [4]. 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 (i.e., *contextualizing* documents) consists of other documents which point-to or are pointed-to by the document in question (*contextualized* document, P2), see Figure 1. The context of an element in focused retrieval and in this study consists of all the ancestors of the element in question. We use random walks to induce a similarity structure over the documents based on their bibliographic relationships, and over the elements based on the containment and reverse-containment relationships (element, sub-element and vice versa). Hence, these relationships affect the weight each contextualizing document or element 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 or element is re-enforced by the weighted scores of the contextualizing documents or elements.

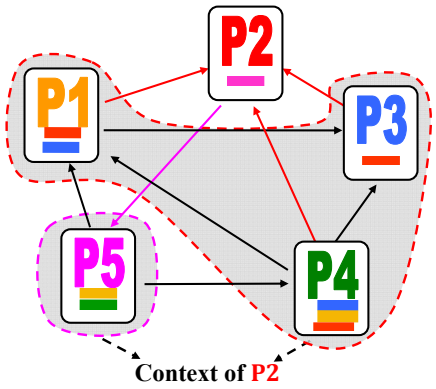


Figure 1: Citation structure of 5 documents and context of P2

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

In Section 2.1 we will explain the idea of contextualization based on citation structure, in Section 2.2 we elaborate on contextualization based on the internal hierarchical structure of XML document (see XML document in Figure 2) and in Section 2.3 we present a contextualization model based on first the citation contextualization and then hierarchical contextualization.

2.1 Citation Contextualization

There are enough empirical and intuitive support 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) [13, 17, 20]. 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, 4] 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 [17]. 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 [20].

The bibliographic network of documents (for example, Figure 1) 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 Figure 1 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]:

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

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 [17].

2.1.1 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 quite huge for the citation structure of Wikipedia collection ($2,668,160 \times 2,668,160$, see Section 4.1). 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 per topic output of fusion run, 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., inlinking and outlinking documents from root-set R_q respectively.

2.1.2 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., $tf \cdot idf$). Documents occurring more than one times in the resultset, will get the basic score as the mean of the basic scores of all the occurences (which we observed in experiments after testing with the other options, like sum, best and worst basic scores).
- f is a parameter which determines the weight of the context in the overall scoring
- 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 retrieved documents, 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.

2.1.3 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 correspond to the non-zero cells of column 2 in the adjacency matrix $\mathbf{A}$.

Section 4 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 2.1.1) and
2. *second* based on query dependent random walk on adjacency matrix $\mathbf{A}$ (the base-set).

We have experimented with both of the approaches, see Section 4. In addition to the two variants, a third variant combines the query independent and query dependent random walk into a combination function:

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

[1]Finding the dominant Eigenvector of the system of linear equations, corresponding to the dominant eigenvalue, which is 1 in this case [20].

```
<article xmlns:xlink="http://www.w3.org/1999/xlink/">
 <header><title>Wiki markup</title><id>42</id>
   <revision>
    <timestamp>2006-10-05 14:22</timestamp>
   </revision>
   <categories>
    <category>Markup languages</category>
   </categories>
 </header>
 <body>
  <section><st>Introduction</st>
   <p><b>Wiki markup</b> is used in
    <link xlink:href="../Wi/Wikipedia.xml"
    xlink:type="simple">Wikipedia</link>.</p>
  </section>
  <section><st>Language Components</st>
   <list>
    <entry>tables</entry>
    <entry>lists</entry>
    <entry>and a lot more</entry>
   </list>
  </section>
  <section><st>See also</st>
   <weblink xlink:href="htt://www.wikipedia.org">
   www.wikipedia.org</weblink>
  </section>
 </body>
</article>
```

Figure 2: XML document

where

- $g_{qi}^{k}(y)$ is the contextualization vector which gives the authority weight of the contextualizing documents of x based on query independent walk.
- $g_{qd}^{k}(y)$ 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.

2.1.4 *Context for a better content description*

The existence of inlinks for contextualized document is certainly a positive indication, but outlinks also happen to occur in the contextualized document's context. By linking to another document, the author implicitly includes the outlinking document in its document 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 corresponds to the non-zero cells of row 2 in the adjacency matrix $\mathbf{A}$.

2.2 Hierarchical Contextualization

Hierarchical contextualization model has been studied before in different settings in XML retrieval [1, 4, 16, 19, 25, 27]. In hierarchical contextualization we tend to utilize the intrinsic structure within the XML document. The representation of documents in XML aims to follow the established structure of documents, i.e., an academic book is typically composed of ⟨**chapters**⟩, ⟨**sections**⟩, ⟨**subsections**⟩ etc., tags. This organization of document gives an intuitive starting point for manipulating text passages at the established hierarchy levels of text documents.

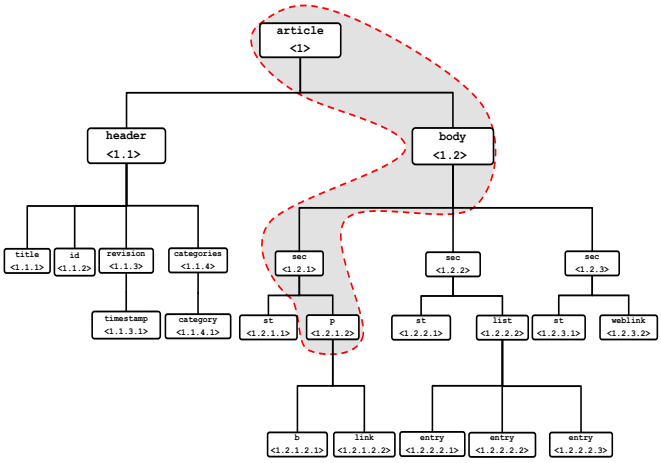


Figure 3: XML Graph of Figure 2 with context of element ⟨1.2.1.2⟩ (dewey encoding)

With contextualization on hierarchical structure of documents we aim to rank higher an element in a good context than an identical element in a not so good context within the document. In Figure 2 the ⟨**article**⟩, ⟨**section**⟩ and ⟨**subsection**⟩ form different levels of context for a paragraph ⟨**p**⟩. Hence the paragraph can be viewed in context of ⟨**subsection**⟩, ⟨**section**⟩ or the ⟨**article**⟩. While the root element ⟨**article**⟩ possesses no context.

In hierarchical contextualization the weight of the element is modified by the basic weights of its contextualizing elements. Each element in the context of the contextualized element, should possess an impact factor. An higher impact factor shows the importance of the contextualizing element and vice versa. The role and relation of contextualizing element are operationalized by giving the element a contextualizing weight. A contextualization vector is defined to capture the impact factor of each contextualizing element, and this contextualization vector is represented by a g function, in a similar way as it is defined in citation contextualization.

The important research question here is: which types of element context help to improve retrieval effectiveness? More specifically which types of context serves our purpose, which is, to boost the ranking of contextualized element in good context and vice versa. Sigurbjörnsson et al. (2004) [27] argued that by taking the root level only (i.e., ⟨**article**⟩ element in the example case) as a context improves the overall retrieval. Camps (2007) [25] later also found that the use of article as a contextual information clearly helps to improve retrieval effectiveness. Arvola et al. (2005) [1] uses a binary value to include or exclude different element types in hierarchy from the context. Ogilvie and Callan (2005) [23] utilizes the children of the element to smooth up the parents (smooth up tree). The smoothing up method in their hierarchical modeling is quite similar to contextualization. In it they contextualize the scores of individual keywords instead of whole elements. In the vertical contextualization approach again by Arvola et al. (2011) [4] the impact or

strength of the contextualization is adjusted with a help of different parameters. Instead of considering only a specific element as a context or using the children to smooth up the parent element or using a parameter to find the impact of each of the units in the context, we propose a generalized mechanism based on the Markovian Random walk principle.

The tree-structure of the XML document is considered as a graph. Myriad of random surfers traverse the XML graphs. In particular, at any time step a random surfer is found at an element and either (a) makes a next move to the sub-element of the existing element by traversing the containment edge, or (b) makes a move to the parent-element of the existing element, or (c) jumps randomly to another element in the XML graph. As the time goes on, the expected percentage of surfer at each node converges to a limit the dominant eigenvector of the XML graph. This limit provides the impact or strength of each element in the context of the contextualized element in the form of g function. We consider all the ancestors of the contextualized element in contextualization; where the contextualization vector g identifies the importance of each of the unit of context (see Equation 4).

Contextualization model formulated in this way, is independent of the basic weighting scheme of the elements and it could be applied on the top of any query language and retrieval systems. We have applied the contextualization model on the top of the baseline system which is the result of fusion from the INEX 2009 offically submitted runs by the participants (see Section 3.2).

In the experiments we evaluated the retrieval effectiveness at different granularity levels. We mainly tested, retrieval effectiveness at article level (⟨article⟩ element), and at paragraph level (⟨p⟩ element); a brief intuition is explained in Section 3.3. The most improvements in retrieval are observed when ⟨p⟩ elements are retrieved. The primary reason is because paragraph has the most context (hierarchical depth) and most specific element in context (see Figure 3).

2.2.1 Combination Function

The re-ranking function based on the random walk principle described earlier can be formally defined as follows:

$$CR(x, f, C_x, g^k) = BS(x) + f \cdot \frac{\sum_{y \in C_x} BS(y) \cdot g^k(y)}{\sum_{y \in C_x} g^k(y)} \quad (4)$$

where

- $BS(x)$ is the basic score of contextualized element x (text-based score, e.g., $tf \cdot ief$)
- f is a parameter which determines the weight of the context in the overall scoring.
- C_x is the context surrounding the contextualizing element x, i.e., $C_x \subseteq ancestors(x)$, $\subseteq$, because only the context containing the query terms are considered.
- $g^k(y)$ is the generalized contextualization vector based on random walk, which gives the authority weight of y, the contextualizing elements (ancestors) of x in XML graph.

2.3 Hybrid Contextualization

Hybrid or twofold contextualization is when the externally accumulated evidences re-enforce the evidence accumulated from within the hyperlinked and hierarchical XML documents. In this approach we first select the best documents based on the citation contextualization (Section 2.1) and later retrieve the most relevant and most specific context from the XML hierarchy using the hierarchical contextualization. The re-ranking functions are the same as before, first we use the re-ranking function, Equation 2 and later we use Equation 4 for better contextualization.

Contextualization with the hybrid approach provided the most benefit in the retrieval effectiveness, based on our empirical studies (see Section 4).

3. TEST BED AND BASELINE SYSTEM

In order to study the effect of contextualization on focused and element levels, we need a suitable baseline and a test bed with adequate evaluation methods. Next, in Section 3.1 we introduce the test bed, then in Section 3.2 a baseline system based on data fusion is introduced and examined briefly in Section 3.4 with the evaluation procedure of Section 3.3.

3.1 Test collection

The outcome of the present study relates to the Initiative of Evaluation for XML retrieval INEX [11] and the test bed provided by it. INEX is a forum for the evaluation of XML and focused retrieval offering a test collection with topics and corresponding relevance assessments, as well as various evaluation metrics. Aside evaluating element retrieval, passage retrieval evaluation is also supported in INEX. In this study we use the data provided by the 2009 INEX ad-hoc track. The track has 68 topics with character-wise relevance assessments, and the test collection, English Wikipedia, covers around 2.66 million XML marked articles and 50.7 Gigabytes of XML marked data [26].

This large, semantically marked-up, Wikipedia collection has been used in INEX since 2009 and is still in use. The reason for using the INEX 2009 test topics (instead of 2010) is the larger variety of elements in the participants' results. This is mainly because of the existence of the thorough task, where elements are retrieved regardless of overlap, i.e., in the results a section and its sub elements, paragraphs, may be retrieved within the same results [11]. The large variety of elements is a necessity for a data fusion of results, which our baseline system is based on.

3.2 Baseline System

Contextualization is independent of basic scoring method, thus we are able to implement the baseline system quite freely. In this study, we use a fusion run as our baseline system for which 159 element runs out of total 173 runs from the INEX 2009 participants was used. The remaining 13 were not element runs, i.e., they contained ranges of fragments or file-offset-lengths (FOL) as retrievable units and were omitted from the fusion. In addition, in order to avoid noise, we made a decision to remove 61 runs having an extensive number of non-existing elements. Thus, a total of 98 runs from the participants of all tasks (best-in-context, relevant-in-context, focused, fetch and browse) of the ad-hoc track were used in fusion.

The runs were fused using an acknowledged method called the reciprocal rank. The method has been found effective in document retrieval [6]. In it, every element (item) in each of the result list (candidate run) is given a score based on its ranking and the fused score for an element is the sum of their ranked scores per topic. A fusion score for an element

e is calculated as follows.

$$RRScore(e,q) = \sum_{r \in R} \frac{1}{k + rank(r,e,q)} \quad (5)$$

where

- R is the set of runs (rankings)
- and $rank(r,e,q)$ returns the rank of element e as a result of query q in run r.
- If e is not in the ranking, $rank(r,e,q)$ is not defined and the outcome of $\frac{1}{k+rank(r,e,q)}$ is 0.
- The parameter k is for tuning.

Before addressing the effectiveness of such approach as a baseline system, we introduce shortly our evaluation approach, which aims at measuring performance of very focused elements only.

3.3 Evaluation methodology

One of the key issues in semi-structured retrieval is the handling of overlap in results. A partial solution has been introduced not to accept structurally overlapping elements in the results. Still non-overlapping elements of various granularities are accepted, so that retrieval of e.g., a whole section instead of its smaller descendants separately leads to different result list than returning the descendants as individual elements. Measuring these kinds of result lists has led to numerous, typically quite complex and unintuitive metrics [9, 15, 24]. The aim of these metrics is not only to measure the matching of the text content, but also the selection of granularity level at various situations. Unfortunately, retrieving elements of various granularity levels has an uncontrolled effect on the evaluation results and has led to bizarreness in the true evaluation results, and favouring systems retrieving large elements over focused ones [4]. Thus, as a criticism, deciding the right granularity level is based on the laboratory environment (especially metrics) rather than on true user needs.

Elements low in a hierarchy are focused answers to a query and possess more context and thus supposedly benefit more on contextualization. In order to study the effect of contextualization especially on those small and focused elements, and to exclude the effect of element granularity level selection on evaluation results, we use granulation [4], where specific types of elements are pre- selected in the collection. The search is focused on those elements only. For that purpose also the underlying recall base needs to be pruned so that only those selected elements are involved (see Fig 4(a)). Obviously, a semi-structured collection can be granulated in numerous ways. In this study, we focus on two types of granulations: full document granulation and a granulation containing paragraphs ($\langle$p$\rangle$-elements) only. To put it short, the former is for document retrieval and the latter is paragraph retrieval. The paragraph level elements are very frequent in the collection (on average 274 relevant paragraphs per topic) and a list containing such elements may provide satisfactory and focused answers. It is worth mentioning that, Crouch et al. [7] had similar setting and used the paragraph as the basic index node. One obvious use case for paragraph retrieval is snippet retrieval.

In terms of structural query language NEXI (strict interpretation) [28, 29], we use the following queries $//article$ $(., about(\text{“}query-expr\text{”}))$ and $//p$ $(., about(\text{“}query-expr\text{”}))$ for full document and focused runs respectively. The “$query-expr$” stands for the title field bag-of-words query of a topic. In the full document approach only root elements (i.e. articles) are considered in the result lists and in the focused run, only elements having the name ($\langle$p$\rangle$). The corresponding runs are made by pruning the fusion results by basically taking out everything else but the lines corresponding to the structural conditions (i.e., $\langle$article$\rangle$ and $\langle$p$\rangle$). In other words, the paragraph list is a sub list of the fusion run. Corresponding recall base is made for paragraph list. The full document recall base is provided by the initiative. The fusion run contains every element retrieved by the participants. The pool was constructed from the paragraph granulation by analyzing the FOLs in the recall base against the submitted paragraphs. Out of the full set of runs used, 46 runs did contain paragraphs. So the paragraph result list is a fusion of those runs.

3.4 Thoughts of competitiveness of the baseline system

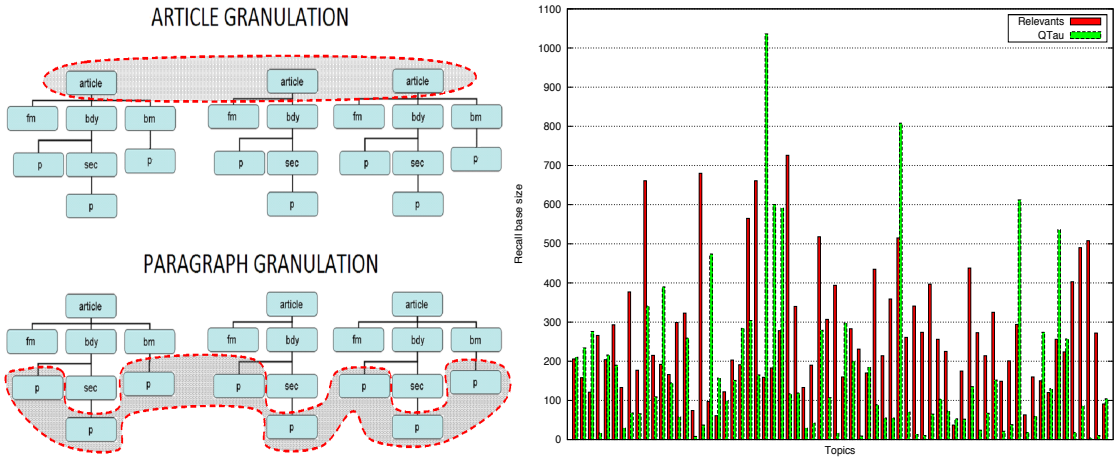
Next, we aim to give an insight of the baseline system we want to improve using contextualization in next section. In order to avoid over tuning of the baseline system, we refer only to results, which are achieved using basic values only and leave the further analysis of the data fusion of element results for later studies. Thus, our baseline system is the bare format of reciprocal rank, i.e., $k = 0$. In other words, an element at the first rank of any run yields basically the score of 1 and the second yields 0.5, third 0.33 and so on.

At article level granulation, i.e., full document retrieval, the fusion run outperforms all reported official full document runs of INEX having the MAP as high as 0.4141. The best official INEX full document run yielded at the level of 0.3578 (UamsTAbi100 by the University of Amsterdam) [11]. The granulation of the run is made so that only results rows with $/article[1]$ are considered. Similarly, at paragraph level any result row ending with $/p[n]$ is considered (n is positive integer). We did the same granulation for every 46 INEX run and compared the results with ours. Early precision was used in comparison at paragraph level for two reasons. First, the granulation results in a subset of the result, so the result list may be short. Second, early precision is in line with the nature of focused retrieval.

The runs of the Technical University of Queensland (qtau) yielded the best early precision figures, especially a run called ANTbigramsThorough. Figure 4(b) represents the recall base sizes per topic at paragraph level and the number of retrieved paragraphs of the ANTbigramsThorough run. In 21 topics the number of retrieved paragraphs of the run outnumbers the number of relevant paragraphs, so a fair comparison can be made using $r-precision$ score for those topics. Accordingly, the $r-precision$ score for the run ANTbigramsThorough is 0.2779 and for the baseline fusion 0.3479. Based on these figures, we can say that the fusion approach is competitive. Next, we apply contextualization for the fusion and see if there still is room for improvements.

4. EXPERIMENTAL EVALUATION

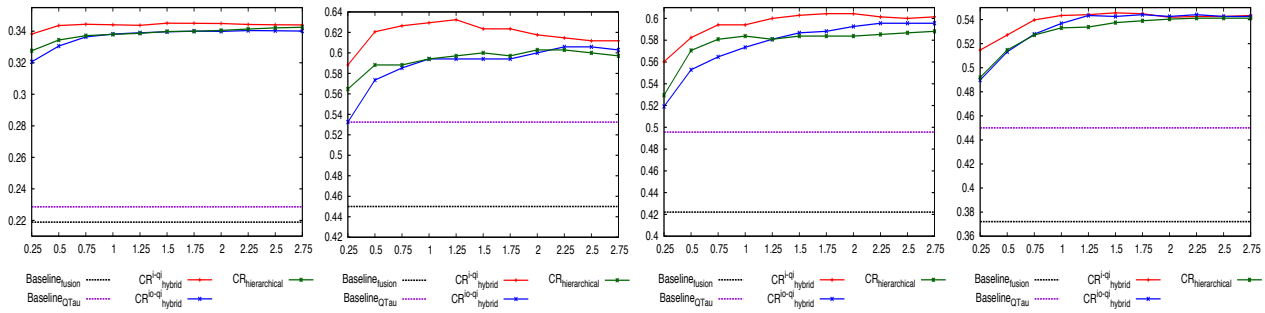
We now experimentally evaluate the propositions presented in this paper. First, we lay down the experimental settings. Later, we present some empirical evidence that our ranking models return intuitive results both on document and focused retrieval. We then evaluate the retrieval effectiveness of our models againts the competitive baseline systems



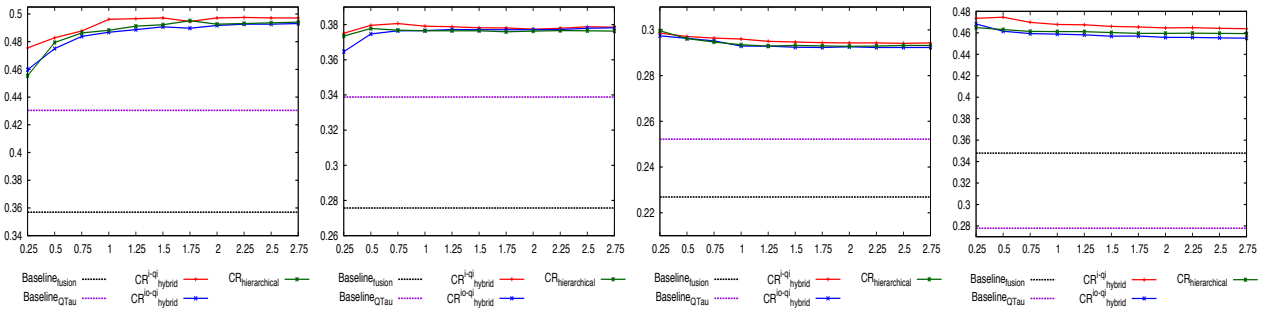
(a) Granulating the recall-base at article and paragraph levels.

(b) Recall base sizes per topic

Figure 4: Granulating overall recall-base (a) and recall-base sizes for QTau baseline (b).



(a) MAP for different f (b) P@5 (c) P@10 (d) P@20



(e) P@30 (f) P@100 (g) P@200 (h) rPrecision

Figure 5: Trends for different measures at different context force f for focused retrieval task (paragraph level)

that were introduced in Section 3. Finally, we relate the empirical evidence with the theoretical claims.

4.1 Experimental Settings

The proposed approaches are evaluated using the Wikipedia test collection, described in Section 3.1. The choice of experimenting with the Wikipedia collection is for the following reasons. First, XML documents in Wikipedia 2009 collection has a very deep internal hierarchical structure, containing overall about 32 thousand different tags [26]. Second, Wikipedia has quite a huge number of inter-document references (in the form of citations). Finally because Wikipedia collection is quite big and extensively assessed test bed used over the years at INEX [3, 11] and at other evaluation forums.

The 2.66 million semantically marked XML documents contain a total of around 135 million citations (links), which were extracted by parsing each of the documents in the collection. We use the resultant gigantic citations graph for experimentation with the citations and hybrid contextualization (Sections 2.1 and 2.3). The computation of the contextualization vector g^k from Equation 1 for the large Wikipedia collection was quite extensive, however this process is performed offline. The linear system of Equations 1 is usually solved iteratively, using the well known *Power method* [18]. The convergence of power method is accelerated using a technique called *Extrapolation* [2]. At the query

[2]Extrapolation is a technique for constructing new data points (dominant eigenvector) outside a discrete set of known data points (known values during each iteration of

time, we combine the iteratively computed random walk scores and the basic scores based on the proposed methods (Equation 3).

In the forth coming sections we will present empirical evidence that the contextualization vector g^k together with the citation contextualization model, produces intuitive overall retrieval effectiveness (see Tables 2 and 1).

For hierarchical contextualization we index the collection and use the dewey encoding to capture the internal tree structure of the XML documents (as shown in the example, Figure 3). This way each element in the document possess a unique index within the document, and together with document's unique id, this becomes unique for the entire collection. The tree structure of XML documents are converted into a matrix, and random walk is performed on this matrix, as it is described in Section 2.1. In this case also the contextualization vector g^k from Equation 4 is computed offline for each and every XML documents in Wikipedia collection. This suggests that computing g^k vector is feasible for a reasonably large XML document collections. Again, at the query time, the scores from g^k vector and basic scores are combined to produce an overall ranking score, using Equation 4.

Focused Retrieval									
Method	f	MAP	P5	P10	P20	P30	P100	P200	rPrec
Baseline (Fusion)	–	.2189	.4500	.4221	.3721	.3569	.2757	.2269	.3479
Baseline ($QTau$)	–	.2286	.5324	.4956	.4500	.4304	.3388	.2522	.2779
$CR_{hierarchical}$	.25-2.75	**.3425**$^{\blacktriangle *}$	**.6029**$^{\blacktriangle *}$	**.5882**$^{\blacktriangle *}$	**.5412**$^{\blacktriangle *}$	**.4951**$^{\blacktriangle *}$	**.3778**$^{\blacktriangle *}$	**.2996**$^{\blacktriangle *}$	**.4649** $^{\blacktriangle *}$
$CR^{i-qi}_{citations}$	.025-1.75	.2423$^{\triangle +}$	.4912$^{\blacktriangle}$	.4500$^{\triangle}$	.3897$^{\triangle}$	.3755$^{\triangle}$	.2915$^{\triangle}$	.2465$^{\triangle}$	**.3811**$^{\triangle *}$
$CR^{io-qi}_{citations}$	.025-1.75	.2207	.4588$^{\triangle}$	.4206	.3750	.3578	.2765	.2288	**.3548**$^{\triangle *}$
CR^{i-qi}_{hybrid}	.25-2.75	**.3451**$^{\blacktriangle *}$	**.6324**$^{\blacktriangle *}$	**.6044**$^{\blacktriangle *}$	**.5456**$^{\blacktriangle *}$	**.4971**$^{\blacktriangle *}$	**.3806**$^{\blacktriangle *}$	**.2986**$^{\blacktriangle *}$	**.4746**$^{\blacktriangle *}$
CR^{io-qi}_{hybrid}	.25-2.75	**.3404**$^{\blacktriangle *}$	**.6059**$^{\blacktriangle *}$	**.5956**$^{\blacktriangle *}$	**.5441**$^{\blacktriangle *}$	**.4931**$^{\blacktriangle +}$	**.3782**$^{\blacktriangle *}$	**.2974**$^{\blacktriangle *}$	**.4615**$^{\blacktriangle *}$

Table 1: Ret. performance for focused retrieval $^{\blacktriangle *}$ = stat. significant than both the Fusion and QTau baselines runs at p < 0.01 (1-tailed t-test), and $^{\triangle +}$ = stat. significant at p < 0.05 respectively.

Document Retrieval								
Method	f	MAP	P5	P10	P20	P30	P100	P200
Baseline (Fusion)	–	.4141	.6618	.5853	.5029	.4554	.2949	.2126
Baseline ($UAmst$)	–	.3578	.6500	.5397	.4515	.3961	.2635	.1898
$CR_{hierarchical}$	.25-2.75	.4142*	.6618*	.5853*	.5029*	.4559*	.2949*	.2126*
$CR^{i-qi}_{citations}$	.025-1.75	.4186*	**.6706**$^{\triangle *}$	.5853*	**.5118**$^{\triangle *}$	**.4618**$^{\triangle *}$	.2965*	.2153*
$CR^{io-qi}_{citations}$	.025-1.75	.4159*	**.6706**$^{\triangle *}$	.5853*	.5051*	.4583*	.2951*	.2129*
CR^{i-qi}_{hybrid}	.25-2.75	**.4194**$^{\triangle *}$	**.6706**$^{\triangle *}$	.5853*	**.5125**$^{\triangle *}$	**.4608**$^{\triangle *}$	.2965*	.2148*
CR^{io-qi}_{hybrid}	.25-2.75	.4139	.6676*	.5779*	.5044*	.4549*	.2944*	.2126*

Table 2: Retrieval performance for document retrieval (article level).

4.2 Results

We have tested five different retrieval methods based on the propositions (Sections 2.1, 2.2, 2.3) and three different baseline systems (Section 3).

- Baseline systems
  - Fusion run, $Baseline_{fusion}$.
  - University of Queensland run, which performed best on paragraph level, $Baseline_{QTau}$.
  - University of Amsterdam run, which performed best on article level, $Baseline_{UAmst}$.
- Hierarchical contextualization, $CR_{hierarchical}$
- Citation contextualization
  - Query independent - inlinks context, $CR^{i-qi}_{citations}$
  - Query independent - inlinks and outlinks context, $CR^{io-qi}_{citations}$
- Hybrid Contextualization
  - Query independent - inlinks context, CR^{i-qi}_{hybrid}
  - Query independent - inlinks and outlinks context, CR^{io-qi}_{hybrid}

We did not report results on citation contextualization based on query-dependent random walk, as the preliminary experimental analysis showed not enough or desirable retrieval gains, apparently because of the definition of citations or links in the Wikipedia collection. Hence, we omit query-dependent citation contextualization from evaluations, and therefore investigate the usefulness of this approach in our future studies.

As defined earlier, contextualization has two general dimensions - the magnitude of contextualization (contextualization force) and the impact of each contextualizing element. The impact of each contextualizing factor is identified automatically with random walk principle, in contrast to the earlier studies [1, 4]. While, the contextualization force has to be parameterized. For each proposed contextualization model, we tuned the contextualization force and report the values leading to best overall performance. In our parameterization process we found: (i) the optimal values of contextualization force f in citation contextualization (from Equation 2) lies in: ($f \in \{0.025, 0.055, 0.10, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75\}$); (ii) and in hierarchical contextualization (from Equation 4) $f \in \{0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75\}$.

These optimal values for f are obtained by using cross-validation technique[3]. We did 68-fold cross-validation (or complete cross-validation in our case) - by randomly partitioning the collection into 68 training and test samples based on the number of assessed topics. Of the 68 samples, a single sample is retained as the validation set for testing, and remaining 67 samples are used as training set. The cross-validation process is repeated 68 times (for each fold), with each of 68 samples used exactly only once as validation set.

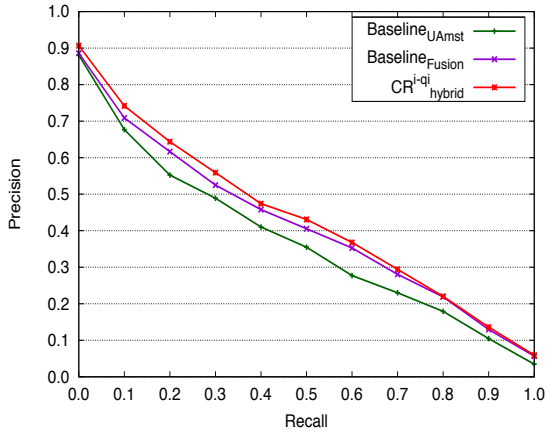


Figure 6: Precision - recall performance for document retrieval (article)

power method) and using the properties of Markov chain; $\lambda_1 = 1$ (dominant eigenvalue) [14, 21].

[3]Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

These 68 independent or unseen samples are then combined to produce a single or a set of estimations for the parameter f.

Figures 5 illustrate the behaviour of the methods as we change the optimal values of f parameter, from Equations 2, 3 and 4, on precision-oriented measures. As can be visually observed, the proposed methods out-perform notably all the baseline systems, Fusion, QTau and UAmst (Figure 6).

Table 1 and 2 show the overview of the retrieval performance of our approaches against the baselines for focused (paragraph level) and document (article level) retrieval tasks. All the proposed contextualization models improves the performance over the baselines. The improvements are statistically significant (1-tailed t-test at $p < 0.01$ and $p < 0.05$) on $rPrecision$, $P@5$, $P@10$, $P@20$, $P@30$ and so on (Figures 5). The improvements overall are surprisingly good on both focused and document retrieval.

The best overall results among the proposed methods are obtained with CR^{i-qi}_{hybrid} and $CR_{hierarchical}$, in terms of highest mean average precision, $r-precision$ and precision at N values. Documents with many and important inlinks have a higher probability of being relevant [12, 13] and hence in contextualization their role is considerable and fruitful, which is also verified in our experiments. We conclude that, context from citations, hierarchical structure of documents and their hybrid indeed improve the retrieval effectiveness, and the improvements are in-line with the theoretical anticipations.

4.3 Discussion

Contextualization is a re-ranking model utlizing the context of the relevant retrievable unit for improving the overall retrieval. We studied context from three different but related perspectives; (i) external perspective (based on citations) (ii) internal perspective (hierarchical structure) and (iii) hybrid perspective (external and internal perspective). The common thread among the three ways of contextualization is the use of the graph structure originated from the documents citation structure externally and hierarchical structure internally. We hypothesized that context gathered from graph structure of documents (from within and outside), influence the retrieval effectiveness. The experiments validated the hypothesis that utilizing the context actually enhances the retrieval of information on article and paragraph granularity levels. The results obtained in this study are inline with the earlier work on use of hyperlinked and hierarchical tree (graph) structure of documents [5, 10, 12, 17] and the role of contextualization [1, 4, 19, 22, 23, 25]. However, none of these works exploits evidence accumulated from the link structure of documents with random walk as a contextual evidence.

The authority score 'in isolation' can identify the importance of each node in the graph formed from either citations or hierarchical structure of documents. The usefulness of these authority scores in isolation (not in context) has been studied well over the years [5, 10, 17]. The novelty of this study is the utilization these useful sources of information not 'in isolation' but 'in contextualization'. That means, to use the importance score of each document or element as an impact factor for identifying how essential is the role of this document or element in context. A retrievable unit (document or element) with strong context must be boosted higher in ranking than the retrievable unit with less strong context. Extensive experimentation validated this view point.

5. CONCLUSIONS AND FURTHER WORK

We have presented an in-depth study into the use of context from citations and hierarchical structure information, in order to improve retrieval performance on document and focused retrieval tasks. To the best of our knowledge, this is the first study that takes context into account by mixing two perspectives (a) the context from the citation structure of documents, and (b) the context from the hierarchical structure of semi-structured documents. The approaches presented are generic and can be applied to different test collections and baseline systems. Evidence is collected in a systematic way, from the surrounding context of both the document itself and the element to be ranked, in document and focused retrieval respectively. In this paper, XML documents are used as a sample case of semi-structured documents. These documents have an hierarchical structure, which is often represented in a form of tree. However, the approaches could also be applicable for other generic structured (or semi-structured) test collections (e.g., Linked Data, RDF, etc.), where the structure may be represented as a general graph (with cycles). The proposed methods are particularly suited for collections that carry more types of evidence than just textual information. The importance of each single unit in the context is identified by a Markovian random walk. Most of the proposed methods are tested and found to be significantly better than the baseline system, which had an overall performance that was already better than any run submitted to INEX 2009. 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 walks to materialize the context has been evaluated in five different settings. We have found that the context from in- and out-links as well as a document's hierarchical structure can indeed improve retrieval results. Given that the citation structure of Wikipedia collection does not necessarily form a sound bibliographic semantics, because, (a) two documents can cite each other at the same time (A cites B and B cites A), without temporal ordering, (b) the link structure in Wikipedia is a (possibly weak) indicator of relevance [12] in isolation. Yet, when applying contextualization using weights obtained with the random walk principle, this information is found to be significantly plausible, both theoretically and empirically. Bibliographical structure of scientific documents could lead to even better results, as their citation structure characterizes stronger semantics, and possibly a stronger indicator of relevance. Nevertheless, we consider our experiments on the Wikipedia test 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. There are also several other venues for future work, for instance, experimenting with different granularity levels than just article and paragraph levels – identify the importance of each granularity level(s) and possibly automatically boost 'important' ones more than other 'not so important' granularity levels. The sequential document ordering, often referred to as the document order, where text passages follow each other in sequence, one after the other,

could also be considered as a second dimenion of the structural context within the random walk paradigm. Finally, graph-based methods for results list fusion may be naturally included in our current approach, where we applied random walks over result lists obtained from a separate fusion phase.

6. ACKNOWLEDGEMENTS

This study was supported by Academy of Finland under grant #140315.

7. 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 (CIKM)*, pages 20–27. ACM, 2005.

[2] P. Arvola, J. Kekäläinen, and M. Junkkari. The Effect of Contextualization at Different Granularity Levels in Content-oriented XML Retrieval. In *Proc. of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1491–1492. ACM, 2008.

[3] P. Arvola, S. Geva, J. Kamps, R. Schenkel, A. Trotman, and J. Vainio. Overview of the INEX 2010 ad-hoc track. *Comparative Evaluation of Focused Retrieval*, pages 1–32, 2011.

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

[5] 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.

[6] G. Cormack, C. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proc. of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759. ACM, 2009.

[7] C. Crouch, D. Crouch, N. Kamat, V. Malik, and A. Mone. Dynamic element retrieval in the Wikipedia collection. *Focused Access to XML Documents*, pages 70–79, 2008.

[8] S. Geva, J. Kamps, R. Schenkel, and A. Trotman. INEX 2010 Workshop Pre-proceedings. 2010.

[9] N. Gövert, N. Fuhr, M. Lalmas, and G. Kazai. Evaluating the effectiveness of content-oriented XML retrieval methods. *Information Retrieval*, 9(6): 699–722, 2006.

[10] 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.

[11] W. Huang, S. Geva, and A. Trotman. Overview of the INEX 2009 link the wiki track. *Focused Retrieval and Evaluation*, pages 312–323, 2010.

[12] J. Kamps and M. Koolen. The Importance of Link evidence in Wikipedia. *Advances in Information Retrieval*, pages 270–282, 2008.

[13] J. Kamps and M. Koolen. Is Wikipedia Link Structure Different? In *Proc. of the Second ACM International Conference on Web Search and Data Mining*, pages 232–241. ACM, 2009.

[14] Kamvar, S.D. and Haveliwala, T.H. and Manning, C.D. and Golub, G.H. Extrapolation methods for accelerating PageRank computations. *In Proc. of the 12th Int. Conf. on WWW*, pages 261–270, 2003.

[15] G. Kazai and M. Lalmas. eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval. *ACM Transactions on Information Systems (TOIS)*, 24(4):503–542, 2006.

[16] J. Kekäläinen, P. Arvola, and M. Junkkari. Contextualization. *Encyclopedia of Database Systems*, pages 174–178, 2009.

[17] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM)*, 46(5): 604–632, 1999.

[18] D. Lay. *Linear Algebra and its Applications*. Addison-Wesley Reading, Mass, 1994.

[19] Y. Mass and M. Mandelbrod. Component Ranking and Automatic Query Refinement for XML Retrieval. *Advances in XML IR*, pages 1–18, 2005.

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

[21] M. Norozi. Faster ranking using extrapolation techniques. *International Journal of Computer Vision and Image Processing (IJCVIP)*, 1(3):35–52, 2011.

[22] M. Norozi, A. de Vries, and P. Arvola. Contextualization from the Bibliographic Structure. In *Proc. of the ECIR 2012 Workshop on Task-Based and Aggregated Search (TBAS2012)*, page 9, 2012.

[23] P. Ogilvie and J. Callan. Hierarchical Language Models for XML Component Retrieval. *Advances in XML IR*, pages 269–285, 2005.

[24] B. Piwowarski and G. Dupret. Evaluation in (XML) information retrieval: expected precision-recall with user modelling (EPRUM). In *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in IR*, pages 260–267. ACM, 2006.

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

[26] Schenkel, R. and Suchanek, F.M. and Kasneci, G. YAWN: A Semantically Annotated Wikipedia XML Corpus. *Proc. of GIFachtagung für Datenbanksysteme in Business Technologie und Web BTW2007*, 103 (Btw):277–291, 2007.

[27] B. Sigurbjörnsson, J. Kamps, and M. De Rijke. An Element-based Approach to XML Retrieval. In *INEX 2003 Workshop Proc.*, pages 19–26, 2004.

[28] A. Trotman and M. Lalmas. Strict and vague interpretation of XML-retrieval queries. In *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in IR*, pages 709–710. ACM, 2006.

[29] A. Trotman and B. Sigurbjörnsson. Narrowed Extended Xpath I (NEXI). *Advances in XML IR*, pages 533–549, 2005.