Muhammad Ali Norozi

The Contextual Features in Schema-Agnostic Environment

Thesis for the degree of Philosophiae Doctor

Trondheim, March 2014

Norwegian University of Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Computer and Information Science

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Computer and Information Science

ISBN 978-82-326-0120-2 (printed ver.)
ISBN 978-82-326-0121-9 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2014:98

Printed by NTNU-trykk

To Bibi Zahra, Maeda, Sani-E-Zahra and Sultana Ali Norozi.

Abstract

> *"Like your body your mind also gets tired so refresh it by wise sayings."*
>
> – ALI

RELEVANCE scoring and estimation deals with both *finding* the relevant set of answers and *ordering* them according to the degree of their relevance to the user-intent. The traditional information retrieval (IR) systems successfully find and order the relevant documents and leave them to the users, who then have to locate the relevant information embedded somewhere within the document. In contrast, estimating relevance in semi-structured retrieval means not only *retrieving* and *ordering* the relevant documents but also *locating* the relevant information within the document as well. When it comes to semi-structured retrieval, the traditional IR style retrieval is simply insufficient.

The main focus of this thesis is estimating relevance in a *schema-agnostic* environment. Here, "schema-agnostic" means that the *schema* or the *structure* exists explicitly within the documents but the user does not or need not know that schema. In such an environment, the structure is generally defined loosely, which means: (a) it can evolve over time, (b) it can constitute a large part of the data, and (c) it might exist seamlessly within the document. The natural question that comes into mind is, why is such a structure there at all? The structure in a schema-agnostic environment is there to be used by retrieval systems for several useful tasks. This thesis is about unveiling the capabilities of the *structural constructs* within semi-structured documents in schema-agnostic settings.

Structural constructs can form what we call the *structural context* of the relevant item. A structural context builds up the internal and external contextual features of a semi-structured document. These contextual features help with a series of tasks. The work presented in this thesis contributes towards understanding and utilizing the contextual features in the retrieval of *focused* information in schema-agnostic settings.

During the course of this study we have identified, implemented and experimented with several intuitive types of contextual features in semi-structured retrieval settings. *Contextualization* is the generic process of utilizing features in the structural context of the retrievable units in relevance scoring. The proposed retrieval approaches, based mainly on contextual features, exhibited notable improvements in retrieval effectiveness, during empirical analyses.

The evaluations and empirical analyses are performed in several tasks, spread across different phases of this study. The tasks are performed by looking at different aspects and challenges of the semi-structured retrieval domain. The following tasks are performed at different phases of this study: ad-hoc tasks, granulation tasks, and standard tasks offered by INitiative for the Evaluation of Xml retrieval (INEX). The contributions of this thesis are also grouped by these tasks.

Preface

> *"Often your utterances and expressions of your face leak out the secrets of your hidden thoughts."*
>
> – ALI

THIS doctoral thesis was submitted to the Norwegian University of Science and Technology (NTNU) in partial fulfillment of the requirements for the degree of *philosophiae doctor*. The work has been performed at the Department of Computer and Information Science, NTNU, Trondheim and Centrum Wiskuned & Informatica (CWI), Amsterdam, the Netherlands. The doctoral program has been supervised by Dr. Øystein Torbjørnsen (Microsoft Development Center Norway), Professor Jon Atle Gulla, and Professor Svein-Olaf Hvasshovd. At CWI, the work has been mentored by Professor Arjen P. de Vries. During the process, the candidate has been supported by the Information Access Disruptions (iAd) project and funded by the Research Council of Norway and NTNU.

Acknowledgements

“The value of each man depends upon the art and skill which he has attained.”

– ALI

Crafting the Ph.D. is like writing a book of poetry. Like poetry, you have to get in love with a certain area, and, like poetry, from that loved area, something new and creative should suddenly come into your mind. Each and every sentence is crafted like a verse in poetry. Hence, it is the best of both a science and an art. Science comes into play when you ascertain your imagination, everything else is art. The art of ideas and the art of putting those ideas into a limited number of pages, with the help of science. When you start a Ph.D., a hard piece of rock in the name of the Ph.D. topic is handed over to you. Throughout the years you have to carve and craft that hard rock, tirelessly. By the end of this tiresome journey you have to have one beautiful and monolithic *sculpture* out of that hard rock, which will automatically reflect your creativity. In this thesis, I exhibit the *sculpture* out of this work.

The thought provoking, formalized curiosity, and mind boggling yet creative experience of a Ph.D. is certainly not the work of one individual, the Ph.D. candidate. Unfortunately, the names of those great people and things who made this work possible should not be mentioned on the cover of this book, rather only in this section. I name them as follows:

First and foremost, I thank my supervisors Dr. Øystein Torbjørnsen and Prof. Jon Atle Gulla, for their endless support and encouragements. Our meetings were an important source of knowledge and experience, which helped me to discover, structure and clarify the ideas presented in this thesis. The iAD project is also to be thanked for the annual meetings and get-togethers. They were fruitful and enlightening. My deepest and heartfelt gratitude goes to my good friend and co-author Dr. Paavo Arvola for his kindness, courage and easy going personality. I also extend my warm gratitude to Prof. Arjen P. de Vries for giving me an amazing opportunity to stay and learn from the energetic group at the Centrum Wiskunde & Informatica, Amsterdam.

Finally, I would extend my deepest love and gratitude to my family and friends for their support, encouragement, and patience, which always kept me going, despite hardships. Without your enduring trust, love, and support throughout all these years, I never would have completed this work.

Muhammad Ali Norozi Trondheim 02.03.2014

Contents

List of Tables

List of Figures

Part I

Introduction

> *"If you cannot get as much things as you desire than be content with what you have."*
>
> – ALI

The first part of the thesis includes the background information for the research work done. It includes four chapters. Chapter 1 broadly introduces the research area and builds up the motivation. Chapter 2 outlines the necessary background material. And Chapter 3 summarizes the papers included in the second part, in terms of their contributions and capacity to answer the research questions. Finally Chapter 4 concludes the thesis with some pointers for future work.

CHAPTER 1

Introduction

> *"Human is a wonderful creature;*
> *they see through the layers of fat,*
> *hear through a bone and*
> *speak through a lump of flesh."*
>
> – ALI

Overview of the research area in light of the opportunities, the research scope, and the subsequent emerging research questions are presented in this chapter. In the next section, the overall research area is briefly introduced and then the motivation for this thesis is outlined. Later, a brief overview of the research context is also described. The research questions are formalized and the corresponding contributions in the form of the publications made during this research have also been identified. The chapter ends with a summary and outline of the rest of the thesis.

1.1 Relevance Estimation in Schema-agnostic Environment

In information retrieval (IR), the concept of relevance has quite a crucial and central role [17]. Relevance is multidimensional and dynamic in nature. The meaning of relevance is primarily dependent on the users' perception of information and their information needs. Relevance estimation should therefore measure the degree of closeness between the users' *perceived* information need and the information (search outcomes) provided by the IR system as an answer set. In this study, relevance or relevance estimation has been referred to as the laboratorical and *"situational"* (as named by Borlund [17]) perspectives of the complex concept of relevance. This means that the degree of relevance of the search outcomes (full-documents or parts of documents) to the information needs are measured (mathematically or statistically) in laboratory settings based on different situations or tasks.

Relevance estimation is therefore an important component in both full-text and semi-structured retrieval systems and often distinguishes their implementations. Relevance in full-text or document retrieval is the scoring of matching documents according to the users' intent and ranking them accordingly. One of the vital reasons for the high popularity of today's successful search engine solutions can be attributed to their smart relevance scoring methodologies.

The intelligent and deliberate organization of information using the *structure* apparatus in different ways, for example, markups, structural tags, annotations and so on, has been quite advantageous [4]. The existence of these structural hints in *semi-structured* data opened up new opportunities within the *schema-agnostic search environment*. "Schema-agnostic" because the schema or structure is known and operationalized by the retrieval system but unfamiliar and/or unknown to the user. The purpose of such a type of schema is twofold: (i) the user need not to remember the seemingly complex structure of the documents, while (ii) the system not only remembers it (understands and stores), but also extracts and utilizes meaning information from them.

On the one hand, semi-structured retrieval intrinsically bridges the gap between the two irreconcilable viewpoints, (i) the IR viewpoint and (ii) the databases viewpoint [31]. And on the other hand, more importantly, because of its semantic and syntactic expressiveness, it also makes the automatic processing and exchange of useful information much easier.

A central topic in information retrieval communities is to index heterogeneous data collections where there are many varieties of data, with large variations in structure: both unstructured data (text documents), structured data (e.g,. database records), and semi-structured data (e.g., XML or HTML data), all present at the same time [12, 4]. In such an environment "schema agnostic" means that the queries might not use the schema information, while the retrieval systems can possibly use the schema/(meta)-information to come up with a set of relevant and focused answers (most specific part of the document). In this study semi-structured are referred to those documents containing explicit structural information representing the logical and physical structure of the document [4] in order to systematically convey the intent of the author of the document. Almost all of the experimental analyses are applied to, but not specific to or dependent on, documents in XML form. The reason for choosing XML as the use case, in this study, is primarily driven by:

- The simplicity and flexibility of the XML standard, by definition (W3C[1]).
- Its widespread and interdisciplinary use and application [23, 1].
- Most importantly, the possibility of evaluating our retrieval methods against the gold standards – by using the *INitiative for the Evaluation of XML Retrieval* (INEX) benchmark in use since 2002 [32, 25, 84].

[1] eXtensible Markup Language (XML), a standard developed by World Wide Web Consortium (W3C) – http://www.w3c.org.

Relevance estimation and scoring approaches should therefore take into account the structural information whenever and wherever they exist and use them to improve ranking of results with or without minimal prior knowledge of the schema of the data. The schema can be automatically recognized by analysing the data sources or the data itself. Important research questions here are: (i) to find retrieval approaches and evaluation metrics which will give a perceived better result and overall performance for the end users and (ii) how to score various data with schema relative to each other? The variance and resulting inaccuracies in the document structures, vocabulary and document content dictate ranked retrieval (partial matching in contrast to exact matching) as the most meaningful search paradigm.

In a collection of documents of different types it is difficult to combine the scores of the retrievable items. For instance, in a document with no structure there will be no score from the structure features. These documents therefore might get a smaller score than documents with structure. To compensate for this, the relevance score for these documents could be boosted, but which factors should be taken into account? Can we use the inherent structure of the document (extracted automatically or semi-automatically) to increase the visibility of the contents and concepts ingrained within the document?

The main question that we ask here is: does the active use of structure in documents affect the quality of retrieval, for example by improving the *relevance* estimation? The next section briefly establishes the motivation for this work.

1.2 Motivation

It is becoming increasingly popular to publish data on the Web in the form of semi-structured documents, which is useful both for data exchange, data semantics, and ease of automation [4, 1]. The representation in Figure 1.1 if retrieved using the conventional content-only based retrieval techniques, has the following drawbacks when it comes to searching for semi-structured documents:

- Does not make use of the strengths of the self describing and explicit *structural hints*[2], ingrained within the document.
- Return reference to the entire documents and not specific fragments thereof. This is problematic, since large semi-structured documents (e.g., whole conference proceedings in one document) may contain hundreds or thousands of elements storing many pieces of information that are not necessarily related to each other.
- Small elements relevant to the users' query might be ignored because of having less textual evidence.

[2] The structure is not strict but rather agnostic—it is part of the data, it can either be employed or ignored.

```
<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/article2.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 1.1: *An example of XML representing semi-structured document.*

- Might return misleading results for queries that explicitly refer to the structure of the semi-structured documents, for example, twig like or xpath queries, querying both the Content And Structure at the same time (CAS queries at INEX).
- Misrepresentation or even ignorance of structural indices. For example, considering the structure as a part of the content only, and hence ignoring its semantic significance. Or even pruning it away from the semi-structured document, and processing the structure-less document with the traditional IR techniques.

Standard IR or document retrieval finds only a reference to the relevant documents. While in semi-structured retrieval, reference to the full document is usually not a useful answer, the granularity of the search should be refined. Within conventional IR, the areas of passage retrieval and question answering have resemblances with semi-structured information retrieval. The concept of the *logical document* [77] instead of just the document comes into play in semi-structured IR. The end users now are not interested in retrieving full documents and locating the relevant information within the documents themselves, rather they require the information to be retrieved. That is retrieval of relevant information at fine-grained granularity levels.

A structurally unaware retrieval system could seemingly be categorized as document retrieval instead of information retrieval, because, an information retrieval system by definition is the process of the retrieval of information, independent of both the *size* and the boundaries of the document. Information retrieval is closer to the focused retrieval perspective of retrieving information in contrast to documents. Focused retrieval is increasingly important [86]. Figure 1.1 shows an example of a semi-structured environment (an XML document in this case). The whole document could be returned as an answer by a document retrieval system. A structure-aware or semi-structured retrieval system should inherently retrieve the most specific and relevant parts of the document, satisfying the user's intent.

The retrieval must not only return the most specific parts of the documents but also it should take into account the degree of *relevance* of the retrieved document fragments to the query posed. And based on that, the document parts should appear in ranked outcomes. There is a shift not only in the way parts of documents are presented but also in the way they are treated by retrieval methods. The task is quite similar to that of the passage retrieval task [75], the only difference is the explicit availability of the structural constructs in the former. Hence, in focused retrieval, not only the retrieval units but also the indexing units should be purified. For an indexing unit, traditionally the $tf \times idf$ or other measures at the document granularity level are maintained, but in the focused approach, $tf_e \times ief$[3] or other elements or focused level measures should be maintained.

In this thesis, we consider structure from viewpoints within and outside of the semi-structured document. The structure within the document (internal hierarchical

[3] tf_e term frequency and ief inverse element frequency at focused granularities.

structure) provides the local context for the information. The structure outside the document (the bibliographical or external structure) provides a broader or a kind of global (in terms of reachability) context for the information. Therefore, structure internally (hierarchical) and externally (bibliographical) provides both *context* and broad *semantics* to the content. These context and semantics should possibly be used to boost or reduce the documents estimated relevance scores. Without using the structural information, the search outcomes might simply be irrelevant or misleading with regard to the queries posed.

The relative importance of the contents in different parts of a document could be learned from the intuitive structure(s) (hierarchical, bibliographical and / or others), and we hypothesize that this learning could be part of the retrieval model. The textual content or set of keywords lying in the body of a document could be less important than keywords lying in the title, generally. This entails that the importance score of each and every structural elements in the focused retrieval has an added implication for retrieval effectiveness.

In this study, we also hypothesize that the textual context of an element, *structurally*, contains traces of evidence. Utilizing this context in scoring is called *contextualization* [66, 67, 63]. The context of an XML element originating from the structural context of the contextualized element helps to lessen the effects of “biasedness” due to the sizes of the elements. We have found that contextualization improves the retrieval quality as well as the “focusedness” of the system.

In this study, the inherently *difficult* problem of processing the variable indexing units in a focused manner is addressed by the *scoring system*, while the problem of retrieving and presenting the right granularity levels as a search outcome is performed by the *selection system* [64].

1.3 Research Context

This research is directed by the Microsoft Development Centre Norway. This research is a part of the “information Access Disruptions” (iAD) project performed and co-operated by (in alphabetical order) Accenture, BI Norwegian Business School, Comoyo – Telenor, Cornell University, Dublin City University, Funcom, Induct, Microsoft® Development Centre Norway, Netview Technology, Norwegian University of Science and Technology, TIL, Uninett, University of Oslo, University of Tromsø, and Zxy Sport Tracking. The project is funded by the Research Council of Norway as a Centre for Research-based Innovation and hosted by Microsoft Development Centre Norway.

The iAD project focuses academic research on how to (cited directly from [34]):

1. *Create schema agnostic indexing services fusing structured, unstructured and multimedia content in precision, analytics and scale optimized information access services;*

2. *Develop scalable and fault-tolerant system architectures including data processing and mining platforms for capturing, cleaning, and extracting knowledge from high-speed data streams;*
3. *Develop and validate in real environments next generation infrastructure for distributed information access;*
4. *Develop extreme precision solutions for access to multimedia;*
5. *Identify disruptive processes either within information access or enabled by information access solutions that can be used as a cluster foundation for Norwegian IT innovation;*

This research has been conducted in the context of points 1 and 4 in the list of activities at iAD, shown above.

The context of this work primarily encompasses the exploration and investigation of the role and significance of the seemingly instrumental and flexible structures within semi-structured documents, in the pursuit of identifying the importance of relevant and focused information. The research work described herein, therefore, falls into the broader area of semi-structured retrieval [3] in schema-agnostic settings and more specifically into the combination of XML and information retrieval [4], or the XML IR.

Research in semi-structured retrieval has a significant bearing on the research conducted in the broader information retrieval communities. The scope of research in semi-structured retrieval, fundamentally, tends to construct a bridge between the traditional information retrieval and the traditional database research [31].

The annual TREC (Text REtrieval Conference) [92] aims at carrying out research in the more conventional information retrieval tasks, and therefore provides essential benchmarking and evaluation tool-kits for researchers in this area. Similarly, in semi-structured retrieval, the sole standardization and benchmarking initiative is also conducted yearly by the INEX (INitiative for the Evaluation of XML retrieval) workshops [32].

The research conducted in this study primarily makes use of standard document collections, gold standards, relevance assessments, as well as various relevant evaluation metrics and tool-kits offered by INEX. In addition, a couple of the studies done as a part of this work have made use of the evaluation framework from TREC and the iSearch test collection [56].

The issues addressed in this work are to a great extent related to or compared against the research work conducted under the auspices of INEX, mainly, and partially with the results of TREC. This work aims to contribute to the field of semi-structured retrieval. The research questions addressed here are enumerated in the next section and the corresponding papers addressing these research questions are listed subsequently.

1.4 Research Questions

Precisely outlining the research objectives for this study is intrinsically not very easy because of the constructive nature (both theoretical and empirical) of the discipline. The objectives have evolved and developed over time. Hence, the overall research objective (broadly) for this study can be stated as follows:

RQ: ***How to effectively use the structure within semi-structured documents as evidence in the pursuit of "further" improving retrieval effectiveness?***

From the above research problem or goal we have formulated the following four mutually disjoint research questions:

RQ1: What is the role and the significance of the structural context in the ranking of the focused items, and what kind of structural context can be "beneficially" utilized?

RQ2: How can we improve the retrieval approaches which make use of the structural context, and subsequently, how should the retrieval effectiveness of those improved strategies be evaluated?

RQ3: How to improve the retrieval of small elements in focused retrieval?

RQ4: How can we effectively utilize the scoring of multiple systems to retrieve focused results at varied granularity levels with good-enough precision (scoring)?

Secondarily, we have also formulated an efficiency related research question ($\mathbf{RQ}_{ef}$), which is not the primary focus of the thesis, but has been added to cope with the necessary efficiency bottlenecks. This research question has been addressed marginally in some of the studies in this thesis.

$\mathbf{RQ}_{ef}$**:** How *efficiently* can we carry out the semi-structured retrieval task?

The background and the subsequent opportunities identified in the earlier sections are represented in the research questions above. These research questions are based on knowledge of the area gained through the reiterative observation of the empirical results and relating and evaluating them against the state of the art approaches. The research questions will be revisited again throughout the thesis, where they will be matched against the contributions, research outcomes, and the research challenges.

1.5 Publications and Contributions

This dissertation is submitted as a paper collection. The research work conducted in three and one-half years resulted in seven selected papers which have been published in peer reviewed international venues. In Section 3.3, the papers are grouped together in a thematic structure (can be seen pictorially, in Figure 3.1). These papers are also included in full-text in Part II of this thesis.

Paper A **M.A. Norozi**, "**Relevancy in Schema Agnostic environment**", JCDL 2011, June 13–17, 2011, Ottawa Canada.

Paper B **M.A. Norozi**, "**Faster ranking using Extrapolation techniques**", International Journal of Computer Vision and Image processing, IJCVIP 2011.

Paper C **M.A. Norozi**, A. P. de Vries and P. Arvola, "**Contextualization from the Bibliographic Structure**", Task Based and Aggregated Search, Proceedings of the 34th European Conference on Information Retrieval (ECIR), Barcelona, Spain, 2012.

Paper D **M.A. Norozi**, P. Arvola and A. P. de Vries, "**Contextualization using Hyperlinks and Internal Hierarchical Structure of Wikipedia Documents**", Proceedings of the 21st International Conference on Information and Knowledge Management (CIKM), Maui, HI, US, October 2012.

Paper E **M.A. Norozi** and P. Arvola, "**Kinship Contextualization: Utilizing the Preceding and Following Structural Elements**", Proceedings of the 36th ACM SIGIR conference on Research and development in Information Retrieval, Dublin, Ireland, July 2013.

Paper F **M.A. Norozi** and P. Arvola, "**When is the Structural Context Effective?**", Proceedings of the 13th Dutch–Belgian Information Retrieval Workshop (DIR), Delft, The Netherlands, April 2013.

Paper G **M.A. Norozi**, and P. Arvola, "**Selection Fusion in Semi-structured Retrieval**", Proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM), Burlingame, CA, US 2013.

The contributions of this study are broadly related to the semi-structured retrieval and link analysis ranking methods. They are indeed applicable to a wide range of areas. For example, the research on entity extraction or entity search is a direct application of this study. Also in multimedia retrieval, the documents are generally represented in semi-structured form and searching through myriads of them is a contemporary requirement and a future need. In natural language processing, it is usually worth extracting and using the latent structures within the documents in order to detect important object relationships or features.

1.6 Summary and thesis outline

In this chapter an overview of the problems in the domain of semi-structured retrieval in schema-agnostic environments has been presented. The remainder of the thesis is devoted to introducing the theoretical background, drawing the overall picture of the thesis, and hence relating the contributions from the papers to the broader research objectives for this study. Following is a brief outline of what can be expected in the rest of the chapters:

- **Part I**

  Chapter 2 **Background**. Presents the existing work in the area, their possible opportunities and their relation to the topic of this thesis. In addition, it puts forward the evaluation framework and the way the results analysed are interpreted.

  Chapter 3 **Research Summary**. Presents the overall thematic structure of the research work conducted during the Ph.D. process. Later, there will be a brief summary of the included papers, with a retrospective understanding. It also relates the contributions as a result of the papers to the research questions.

  Chapter 4 **Concluding remarks**. Concludes the thesis with some retrospective remarks and the possibilities of future ventures.

- **Part II**

  Papers : This part comprises the included papers (Papers A–G).

CHAPTER 2

Background

"If matters get mixed up then scrutinize the cause and you will know what the effects will be."

– ALI

BACKGROUND information and the state-of-the-art motivation for this thesis are outlined in this chapter. Contemporary and past studies on semi-structured retrieval are presented in view of why they are not enough to effectively solve the imminent problems and challenges related to this study. In addition, the purpose is also to build up the necessary fundamental context that supports the second part of the thesis. From the existing studies, the main research challenges and opportunities that led to some of the key contributions of this work will also be identified. Because of the experimental nature of this study, it is also essential to provide a description of the evaluation framework, the INEX and TREC benchmarks, and the representative test collections with their associated query topics.

Section 2.1 develops the important preliminaries required before commencing part II. Section 2.3 outlines the evaluation framework, applicable to the objectives of this study. Section 2.2 introduces the contextualization model. Finally Section 2.4 argues the need for a competitive baseline system, as well as introducing different strong baseline systems to be considered and compared against each other within this study.

2.1 Semi-structured retrieval

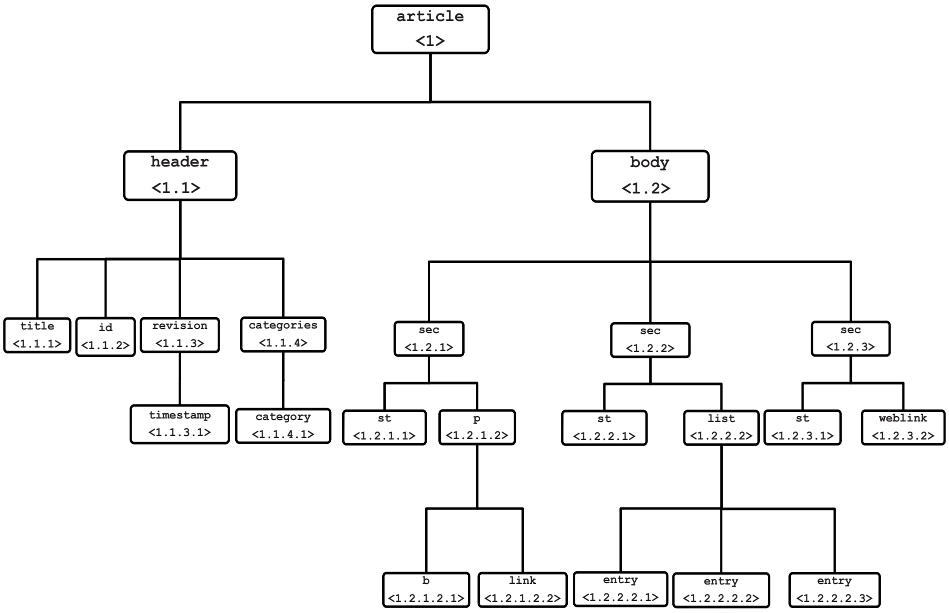


Figure 2.1: *Semi-structured data such as XML is usually represented as a Tree structure, the tree formed from the XML example in Figure 1.1*

2.1.1 Semi-structured data

Semi-structured data [2, 3], lies in between the strictly structured [27] and unstructured data. The strictly structured data belong to the database viewpoint of data, i.e., every new data entry coming in *must* comply with the existing structure and dependencies. While "unstructured" (raw) refers to the information's having, physically, *no* structural constructs defined or in-line with the data, although logically and implicitly it may have a clear structural appearance, outlining the internal semantic structure of the information. This means that the information items have no pre-defined annotations, markups or any other signs of explicit (syntactic) structure, although they might have an ingrained implicit (semantic) structure.

Semi-structured data has certain peculiar characteristics which class it into its own data category. However, it does have commonalities and differences with both its structured and unstructured counterparts. Like the structured, semi-structured data contain physical and explicit structural constructs, e.g., tags, annotations, markups or any other metadata aimed at defining loose structural boundaries for the information. Similarly, like unstructured data, the explicitly specified structure in a semi-structured document might as well be considered implicit (because the structure or schema is not strictly defined and can also be ignored), for ease of processing (indexing) and / or retrieving. Unlike the structured data, here: (i) the structure is irregular [2]—the same data might be organized in different ways, e.g., the publications in a proceedings might be marked as ⟨`papers`⟩, ⟨`articles`⟩, or none of them; and (ii) the structure can be very large and at the same time constantly and rapidly evolving as well.

In semi-structured data, the distinction between structure and content is *blurry*: the structure might be treated as an integral part of the content, or otherwise. On the one hand there are end-users' queries addressing simultaneously both the content and structure of the documents, without specifying any structural hints in their requests. On the other hand, the structure constitutes a large part of the data: it keeps on growing rapidly and the structure boundaries can be violated. Structural constructs can be kept as data in one source and as a schema in another.

The flexible structure within semi-structured data can be indicated in the form of annotations, tags or markups. Typically, the structure is specified using a markup language, and in this study the XML language is used to represent semi-structured data. Although there could be other possible representations of the data, XML is chosen because of the availability of a test collection in XML and the respective retrieval assessments. The approaches proposed in this work are generic and the findings are not specific to the XML language.

2.1.2 Internal Representation

The *internal* structure in-grained within a semi-structured document forms visually a hierarchical structure (a tree, see Figure 2.1) and it represents the sequential structure of the document. The structure is sequential (the depth first ordering of Figure 2.1) because of the way individual fragments of information follow one another in sequence, constructing a holistic picture of the concepts in the document's order.

The most common and widely understood form of representing the internal structure of a semi-structured document is the tree representation [69]. In the tree form, the node or vertex represents the structural component (a markup or an entity) and the directed edge represents the containment relationship (element–sub-element, parent–child). An edge in the internal tree representation can be denoted as a list of vertex pairs $(n_i, n_j) \in \{\text{set of edges}\}$, which implies that n_i is a parent of n_j. The following functions can also be defined based on this implication:

$$
\begin{aligned}
parent(n_j) &= \quad n_i \quad : (n_i, n_j) \in \{\text{set of edges}\} \\
children(n_i) &= \quad \{n_j\} \ : \forall n_j \ (n_i, n_j) \in \{\text{set of edges}\} \\
ancestors(n_i) &= \begin{cases} n_j & : n_j \in parent(n_i) \cup \\ & \exists n_k \in parent(n_i) \\ & s.t. \ n_j \in ancestors(n_k) \end{cases} \\
descendants(n_i) &= \begin{cases} n_j & : n_j \in children(n_i) \cup \\ & \exists n_k \in children(n_i) \\ & s.t. \ n_j \in descendants(n_k) \end{cases} \\
siblings(n_i) &= \begin{cases} n_k & : (n_i, n_j) \in \{\text{set of edges}\} \cap \\ & \exists k(n_k, n_j) \in \{\text{set of edges}\} \\ & \forall k \neq i \end{cases} \\
kinship(n_i) &= \begin{cases} n_j & : (n_i, n_j) \in \{\text{set of edges}\} \cup \\ & \exists n_k \in children(n_i) \ \cup \\ & \exists n_k \in ancestors(n_i) \ \cup \\ & \exists n_k \in descendants(n_i) \cup \\ & \exists n_k \in siblings(n_i) \\ & s.t. \ n_j \in kinship(n_k) \end{cases}
\end{aligned}
$$

A Dewey encoding [33, 82, 55] or labelling scheme can be employed to capture the internal tree structure of XML documents (as shown in Figure 2.1, second line on each node). In this way each element in the document possesses a unique index within the document, and together with the document's unique identifier, this becomes the unique identity of a particular structural component for the entire collection.

There are several ways of encoding or labelling the internal structural components within a semi-structured document: for example, pre- and post-order encoding, and other requirement-specific labelling [52, 26, 21]. Pre- and post-order encoding has been used in the TopX retrieval system—specifically designed for semi-structured documents (XML) [83]. We choose Dewey encoding for its appropriateness for our specific empirical and theoretical requirements. Notwithstanding, Dewey encoding enables us to determine more complex ancestor relationships than the other counterparts [33, 20]. This aspect of Dewey encoding, to easily discover the ancestral path or context of a structural component, enables us to apply and interpret the structural context in several intuitive and challenging settings [66, 63].

2.1.3 External Representation

In addition to the internal hierarchical dependencies, semi-structured documents also widely reference other semi-structured documents, e.g., for the completeness

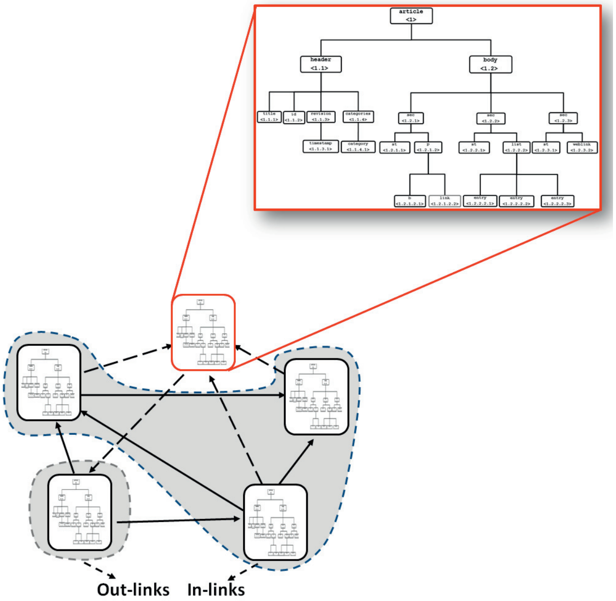


Figure 2.2: *Link structure of five semi-structured documents. Dashed lines represent the out-links (one document in the grey cloud, out-link(s) context) and the in-links (three documents in the grey cloud, in-links context) of the contextualized document (red box).*

of the concepts described within them or for any other bibliographical coherence. Figure 2.2 depicts five semi-structured documents referencing each other. The external structure or the bibliographic structure of these semi-structured documents form a directed graph. The node or vertex in this graph can be considered to be either (a) the structural component (a specific part of the document) or (b) the entire document that triggers the link (i.e., the source of the link). If option (a) is considered, the specific semantics of the link can be analysed from the surrounding structural and textual constructs. But if option (b) is chosen, the whole document can provide the contextual meaning to the link. In the graph, the directed edge represents the link from the source document d_i to the destination document d_j. A similar interpretation is possible for the destination document: it could be either a structural construct or the entire document. In this study, option (b) is deliberately chosen, both for the source and destination documents—in the studies which employ the external representation of semi-structured documents.

The bibliographic network of five documents in Figure 2.2 can be represented in matrix notation by an adjacency matrix **A** in such a way that:

$$\mathbf{A_{ij}} = \begin{cases} 1 & \text{if there is a link from document } d_i \text{ to } d_j \\ \varepsilon & \text{if there is no link from document } d_i \text{ to } d_j, \\ & \text{while there is a link from document } d_j \text{ to } d_i, \\ & \text{where, } 0 < \varepsilon \ll 1 \\ 0 & \text{otherwise} \end{cases}$$

The value ε, a very small positive value in the range [0, 1], is added to indicate the reverse edge: in order to ensure an *irreducible* and *aperiodic* matrix [50, 49]. For Figure 2.2 the corresponding adjacency matrix **A** can be constructed as follows:

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

2.1.4 Structural Components

Independent and thorough handling is required to uncover the features in the explicitly represented structure, ingrained within the hierarchical and bibliographical structure of semi-structured documents. Apart from its representational and morphological features, the structural components can be employed for several other intuitive and complex retrieval and evaluation scenarios as well [66, 73]. Structural components can also be characterized *internally* and *externally.*

Externally, the bibliographical structure of the documents can be of immense value for the semi-structured retrieval problem. These implicit sources of information, in the form of links, have been used in several problem solving activities within IR [47]. When the external structural components are considered *in isolation*, they form a strong backbone for the broader area of link analysis ranking and retrieval [90, 46]. Instead of considering them *in isolation* if we generalize them and hence consider them at a *contextual* level, they play an important role in improving the retrieval effectiveness of the structural components occurring in a *good* neighbourhood. For example, Figure 2.2 also sketches the out-links and the in-links context of the node in red. In Paper C we have investigated this issue: seeking to get an answer to the core research question as to whether or not the external neighbourhood can play a role in the effectiveness of retrieval of the relevant items.

Internally, the hierarchical structure on the one hand provides semantics to the content, and on the other hand the possibility of formulating the tree structure in such a way that the individual structural constructs support each other and form a type of context (which we refer to as the *structural context*). This broadens the scope of the meaning of each individual constructs. These internal structural associations between the structural components, in isolation and in context, have been used as a source of evidence in several intuitive settings [28, 79, 69]. The content within the structural components has also led to evidence propagation from the content to the structure [55, 68, 54]. On the other end of spectrum are the retrieval methods which simply ignore the structure, and consider each of the structural components as an independent document in itself, e.g., [80, 53, 36]. Since not all structural components possess equal status in the hierarchical structure (such as, content appearing in the ⟨`title`⟩ element possesses a stronger likelihood factor), some of the components should naturally be considered better than others. Hence, there is a need for retrieval approaches which take into account the role of those dominant or better structural components on retrieval, for example, by weighting them higher than others [85]. The main problem here is to determine the influence factor of each structural component, which requires either a manual tuning based on domain knowledge or automatic process based on either language modelling or the tree (hierarchical) topology [85, 87].

In our work, the most relevant way of treating the structural features is when the concepts from the tree and / or graph topology theory are applied to the structural dependencies (the tree and graph structures, Figures 2.1 and 2.2). This is usually done by *spread of activation* [5], i.e., the generation and propagation of evidence (activation) along the hierarchical (tree-) and bibliographical (graph-) structure of the documents [66]. *Contextualization* is a mechanism for estimating the relevance of a given structural component along with the content, with information obtainable from the structural context of the retrievable unit in question [9, 43]. We will re-visit this approach later in this chapter with more insight and elaboration.

2.1.5 Retrieval

The area of semi-structured retrieval in a schema-agnostic search environment is an interdisciplinary and widely useful field of study in IR research. Both its inception and its implications cross traditional boundaries from information retrieval communities [48, 72, 45] to relational databases [31, 93] and at the same time have widely ranging implications across digital libraries communities as well [14]. In the following, we provide an overview of existing retrieval approaches towards relevancy in semi-structured data within a schema-agnostic environment, and an account of why they are not significant enough to answer the research questions raised in this study (Section 1.4).

Broadly, there are two major viewpoints for the semi-structured retrieval problem.

- The data-centric viewpoint.
- The information-centric viewpoint.

The data-centric viewpoint, as the name suggests, makes use of techniques and perspectives from relational databases. This viewpoint does not directly qualify the schema-agnostic specifications, but is still studied here because: (i) these approaches still contribute to semi-structured retrieval, hence provide a broader background to the area; and (ii) they have indirect implications for solving the problems in this research domain.

Traditionally, semi-structured data have a more relational database oriented characterization, *structure* being the common denominator between the two. Historically also, the initial research in semi-structured retrieval originated from the database communities. Here, instead of reinventing the wheel, the existing database technologies are employed to cope with the semi-structured retrieval problem. The answer is therefore a data-centric view of semi-structured retrieval—to fit it into the relational databases perspective [31]. Specifically, here we directly utilize the capabilities of relational databases to (i) represent and (ii) retrieve semi-structured data. Inherently, from its inception, this viewpoint fell victim to one or another form of the exact matching paradigm, primarily because of the inability of relational databases (back then), to handle partial matching. (Later, some kind of partial matchings were also introduced [33].) XQuery [16], developed by the W3C Consortium inspired by the relational databases, addressed the problem of semi-structured retrieval. Unfortunately, this approach is not considered suitable for the objectives we have outlined so far, mainly because (a) they require a specific knowledge of the schema or structure beforehand (and hence are not schema-agnostic), (b) it is complicated to translate the user query into an XQuery, (c) the syntax of XQuery is by far more complicated than the syntax of standard IR systems, and (d) its nominal mechanism for ranking (which means it is inclined towards exact matching) [21].

The information-centric or the information retrieval viewpoint of the semi-structured problem falls into two broader categories: (i) the *vector space* based formulations [19] and (ii) the probabilistic or language based formulations [87, 8, 80, 70].

XSEarch [21] uses an extended *vector space model* for retrieval and ranking, and the same kind of approach was employed by Schlieder and Meuss [77]. They employed the *interconnection relationship* between the XML elements to manipulate the structure in retrieval and ranking. Their intention was to answer the fundamental question, under what conditions are the elements of an XML document semantically related? A variation of $tf \times idf$ is used in the XSEarch system, where tf correspond to the number of occurrences of a query term in the structural fragment and the *inverse leaf frequency* ilf is the number of leaves containing a query term divided by the number of leaves in the corpus (the data tree), see Equation 2.1. The $tf_e \times ilf$ score together with the interconnection relationship measure (calculated based on how close the elements are in the relationship tree) are used to determine the ranking of the answer set.

$$\begin{aligned} tf_e(k, n_l) &:= \frac{occ(k, n_l)}{max\{occ(k', n_l)|k' \in words(n_l)\}} \\ ilf(k) &:= log\left(1 + \frac{|N|}{\{|n' \in N|k \in words(n')|\}}\right) \end{aligned} \tag{2.1}$$

XRANK [33] and ObjectRank [13] are probabilistic approaches that generalize the

PageRank approach proposed by Page and Brin [71]. They consider the dataset as a tree or graph (Figures 2.1 and 2.2). Unlike PageRank, which employs a one-size-fits-all approach, XRANK recognizes that the data tree has different types of edges, namely, containment edges and hyperlink edges. ObjectRank, on the other hand, calculates both the global (PageRank) and the keyword-specific ObjectRank of each node in the *authority transfer schema / data graph.* Unlike XRANK, ObjectRank is a relational approach and is applicable only to relational databases.

The research involoving XRANK, ObjectRank, and other graph based methods, corresponds to the study of link analysis ranking (LAR) [18, 90, 47]. LAR models are employed due to the tree- and graph-theoretic formulation of the "internal" hierarchical tree and "external" bibliographical structure of semi-structured documents, as indicated before as well.

XXL [81, 82] was proposed mainly to introduce active support for ranked retrieval. In addition, ontological information or relationships have also been integrated as a basis for effective similarity searches. In the same line, XPRES [94, 95] extends the classical probabilistic model that exploits the semantic of different text parts given in semi-structured documents. Like XSEearch, XPRES extends the classical weighting measure $tf \times idf$ and call it $tf_e \times ief$ (where, *ief* stands for *inverse element frequency*).

BM25F-based (probabilistic) [74, 55] semi-structured retrieval has been introduced to score individual structural elements [36]. In this approach, each structural component is scored as if it were an independent document. This method ignores any hierarchy, i.e., the parent–child relationships (which usually contains contextual information), but rather focuses on the elements independently.

2.1.6 Challenges

In light of the background information laid down so far, there are several challenges or problems that need to be addressed in order to be able to contribute in the semi-structured retrieval domain. Generally, the problem in semi-structured retrieval arise when the *structural features* are not appropriately handled, and the challenges therefore are: (a) to properly use them in indexing and retrieval and (b) to learn from them the implicit semantics or meaning of the textual content and finally (c) apply the information gleaned to improve the retrieval and organization of the result list.

Overlap

One of the fundamental challenges that one most certainly faces when shifting the units of retrieval and indexing from documents to the structural components (logical documents), is the problem of *overlapping* result lists [42]. The same content appears several times in the results because of the overlapping or nested structural components. In the containment relationship (parent–child in the hierarchical structure), as each element is considered as an independent document, the content in the child also belongs to the parent element as well, so overlap is a natural and inevitable consequence. An intuitive yet simple approach to deal with it is to prune the overlaps (either the parent or the child) after retrieval. A slightly more advanced approach is to push the overlapping elements down in the result list [58].

A typical strategy is to select the elements on the basis of their scores so that among the overlapping elements the element with the best relevance score is selected. We call this score-based selection strategy the *first-come first-choice* (FCFC) selection principle [64]. The FCFC principle is a widely used strategy at INEX, but it is a rather straightforward technique, having no pre-defined and / or intellectual basis. For example, an intellectual basis might be one based on screen-size requirements, user preferences, or any other logical constraints.

Camps [73] and Mihajlović et. al. [58] have argued that the usefulness of a particular element can be modelled using (i) the relevance score, (ii) the size of the element, and (iii) the amount of irrelevant information it contains. Camps also argues that the semi-structured retrieval model should provide (a) a *ranking* considering the dependencies in the hierarchical yet nested structure of the documents and also (b) the *usefulness* of the retrieval elements compared to other candidate elements in the same path.

In our *selection fusion* methodology [64], we argued that the selection of an element for the result list should be an *independent* task, itself dependent on different *selection* scenarios. Consequently, relevance scoring should also be done independently of the selection criteria. Hence, we proposed that semi-structured retrieval should be executed in two individual phases: (i) the relevance scoring phase and (ii) the selection criteria. Further details are in Paper G.

Focused-ness

From its inception, the purpose of semi-structured retrieval has been primarily to retrieve (i) the relevant elements (ii) at the right granularity levels. In other words, the elements which are as *focused* as possible, while still covering the user's intent, should be a candidate answer to the user's request. In a nutshell, a semi-structured retrieval system should be able to address both (a) the scoring problem, how well a system scores the relevant elements, and (b) the selection problem, the selection of the right granularity level for the elements, i.e., of the appropriate size and the type required for the specific task(s) (as hinted earlier as well). The appropriate size or type is dependent on the relevance as well as many application- and user-specific attributes.

The right granularity level, or the focused-ness, of the result list is very much subject to different use cases and user interface scenarios [40, 66]. For example, the need for a shorter excerpt of information as a result set, is almost mandatory when a user is querying the system using a device with screen size constraints, using for example a smart phone (which is quite common, nowadays). Also, many other different use cases require the retrieval of elements at various granularities. For a snippet retrieval or fragment search, a smaller element is more suitable than a whole section or full article. For example, a user interested only in abstracts of search results, to skim through them before opening the whole document, might not be satisfied if any other selection scenario is presented. One could imagine a number of other different such use cases and user interface scenarios. At INEX, different tasks have been considered for being a representative use case of a focused retrieval task in a schema-agnostic settings [88, 51].

However, apart from the tasks, measuring the performance follows rather a one-size-fits-all principle, where system developers have to guess what is the correct granularity level appropriate for the metrics involved in the evaluation [17].

Size bias

Due to the challenge of "focused-ness" in semi-structured retrieval, there is a great variation in the result list, in terms of the sizes of textual content available in each answer. This challenge prompts the normalization of the varied sizes of the retrieved items. If not taken care of, usually these heavily skewed retrievable items, in terms of size, cause a *biasedness* in the overall results list [39]. The same observation was made in our work on *contextualization* [66] (see Section 2.2). One intuitive way to handle this issue is to simply remove the small items from the collection. But the intuitive reason to reject this is that this way of normalization simply skips or removes the information which might be of great interest to the user. In the not so extreme ends of this spectrum are the not-so-small-items, e.g., paragraphs (containing logical fragments of information, could be categorized as useful answers, if relevant). We argued in Papers D and F that usually these not-so-small-items suffer from having scant textual evidence, because they do not contain enough textual content in comparison to the larger items like sections, body, article, etc. We argued that the problem of size bias can also be handled (in addition to the propositions given by Kamps et. al. [39]) by using the structural context to alleviate the scantness of evidence in those not-so-small-items. The evidence are *accumulated*, *combined* and *propagated* into the relevance of retrievable items, from their structural context, to lessen the effects of "size bias" [65]. This way the smaller items get pushed up (in the result list) by the force of the evidence lying in their contextual features, while the larger items because of not having enough features in their context do not accordingly get enough push, hence the outcome is a normalized result list. It has also been observed at INEX that the top ranked retrieval systems (runs) are mostly those containing larger elements in their result lists, e.g., article runs [11], because the not-so-small-items did not qualify, based on the aforementioned reasons.

2.2 Contextualization

The aim here is to identify the importance and accordingly utilize the explicit structural and contextual features within and outside the documents [9, 66], to address some of the known issues and challenges within the domain of semi-structured retrieval. The purpose therefore is to expose the explicit and implicit capabilities of the internal and external contextual structures. The main research question imminent here is, and it relates to both the main research goal RQ and research question RQ1, from Section 1.4:

> **What is the role and significance of the *structure* in the retrieval of relevant information, in semi-structured settings?**

In its core, the *contextualization* method is designed to address the above research challenge, the fundamental goal of this research. It is, therefore, a mechanism designated to estimate the relevance scores of given retrievable units primarily relying on the contextual structural evidence. Hence it is essentially about exploring the features in the *structural context* of a retrievable unit [9, 66, 67]. In this line of research, the hypotheses were based on intuition and a theoretical background for the structural context[1] of an element or document (contextualized item). In addition to effectively representing the document(s), the intuitive structural contexts also hold within themselves the likelihood factor of the *contextualized* item, implicitly. These likelihood factors or evidences have to be recognized, appreciated, and integrated in the core of the retrieval model. Next, a characterization of contextualization and what it symbolizes is presented in light of the existing studies and the present study.

2.2.1 Internal Contextual Features

There could be several ways of exploring the features in the internal context of the semi-structured documents, because there could be various interpretation of the internal context. For example, some of the most intuitive ones might be as follows:

- Using the explicit hierarchical structure of documents as way to unleash the features in the internal context [66, 9].
- Using textual content in proximity as an internal context, for example a possible related work could be [15].
- Semantic analysis of the items in the neighbourhood as an internal context, for example, the XSEarch system [21], could be modified to do such handling.

Although all of the above or any other interpretation of internal context is worth exploration, this work focuses primarily on the internal contextual features extracted from the *hierarchical structure* of the documents. Below a synopsis is given of the hierarchical structure as an internal context.

[1] Nodes in the structural proximity or in kinship relationship

Hierarchical structure as internal context

The role of hierarchical structures, in one or another form, in trying to solve some core XML retrieval problem, have been studied before in different settings [9, 8, 43, 79, 57]. The aim of hierarchical or internal *structural features utilization* is to determine automatically or semi-automatically the best set of elements which can construct the structural context, in line with the definition in the previous section. The structural context can help improve the effectiveness of the retrieval. The structural context can be established using the tree relationship functions defined in Section 2.1.2 (based on the containment relationship). However, there are two quite difficult and challenging tasks: (i) determining which *types* of elements (in a tree relationship) should form the structural context and (ii) what should be their *impact* on the final relevance scores. There are some alternative approaches proposed below, but the area has not been very widely studied. This thesis is one of the very few studies that have been conducted to explore these features in schema-agnostic settings, with much more focused objectives and research goals.

Sigurbjörnsson et al. [79] argued that taking the root level only (i.e., ⟨`article`⟩ element in the example case) as a context improves the overall retrieval. Camps [73] considers the length normalization technique also as a form of contextualization, in that they used score propagation to normalize the length of the elements by the relevance scores of the elements in the context (hierarchical surroundings). Arvola et al. [8] uses a binary value to include or exclude different element types from the hierarchical context—doing both score propagation and consolidation, for a contextualization process. Ogilvie and Callan [69] utilizes the children of the element to "smooth up" the parents (smooth up tree). The smoothing up method in their hierarchical modelling is quite similar to the contextualization. In other words, they contextualize the scores of the individual keywords instead of whole elements. In the vertical contextualization approach, again by Arvola et al. [9], the impact or strength of the contextualization is adjusted with the help of different manually tuned parameters.

Instead of considering only a specific element as a context or using the children elements to smooth up the parent element or using a parameter to find the impact of each of the units in the context, we have proposed a generalized mechanism based on the Markovian random walk principle [67, 66]. Both the type and the impact factors of the elements in context are calculated automatically. The type of elements constituting the structural context is based on the tree relationship functions; the impact factor is systematically calculated using the random walk principle (Papers B and D).

The hierarchical tree structure of a document is considered as the permissible finite states: with nodes representing the states and edges representing the permissible transitions from one state to another. The Markovian random walk [50] is conducted on the resultant tree structure of the documents (represented in matrix form). Essentially, at any time step, the random walk process either (a) makes a state change along the edges (transition from parent to child) or (b) makes a transition from child to parent along the edge or (c) randomly changes the state to another state without following the edges (i.e., a random jump). This process is repeated until the expected probability of visiting a state converges to a limit, and that is where we get the *stable* state. This stable state or the eigenvector [50] of the hierarchical structure of the document forms the *impact* factor of each and every node (state) in the tree structure of the document. The impact vector is used together with the contextualization model [66] to estimate the role of the structural context in the effectiveness of element retrieval. The study in Paper B is specifically conducted to explore the performance aspects of the approaches based on the Markovian random walk formalization.

2.2.2 External Contextual Features

Similar to internal, there could be numerous aspects and interpretations of the external context. Some of the most suitable ones can be enumerated as:

- Features in the bibliographical or hyperlink structure of documents can be used in the external context [67, 66].
- External semantic (concepts) or linguistic features of the documents in the same domain could be also used as external context. For example, documents in physics domain share a set of concepts, external to the concepts already expressed within the documents.
- Semi-structured documents from similar areas can be connected together with dummy external entities (based on some external relationships of the documents). For example, documents published at a particular conference and in a particular track can be connected together with a suitable type of external entity. This way, the internal tree structure can be broadened by the external links to become a forest. Which also can be represented graphically by a hierarchical structure: the nodes in this case could be the documents and concepts and the links can symbolize the relationships between the documents and the external structures (the conceptual nodes).

The focus of this work is intentionally limited to the first point in the above list: bibliographical and hyperlink structures. In the following, we give some background information for the first point.

Bibliographical and Hyperlinks structure as external context

One of the novel contributions of this study is the introduction and application of bibliographical and hyperlink contextualization in both the traditional IR and semi-structured retrieval approaches. In the area of link analysis ranking (LAR), the *authority score* of a document in the bibliographical structure of a set of documents identifies the relevance of the document [46, 38] in isolation. On the other hand, in bibliographical contextualization the authority score only signifies the *impact* of a particular document in the external structural context. In addition to the impact factor, in external contextualization, each document also possesses some language-model based scores as well, the *basic scores*. The authority scores and the basic scores of *all* the documents (units) in the structural context and the *magnitude* of the context are taken into account in contextualization.

Contextualization in the bibliographical and hyperlinks settings has also two dimensions: (i) which type of relationships should be considered in the structural context and (ii) as discussed above, the impact and basic scores of each and every unit of structural context. The types of relationships that exist in the bibliographical structure are the $link - to$ (out-links) and $linked - by$ (in-links). And based on these two types, the structural context could be formed by (i) out-linking the relevant documents or (ii) in-linking the relevant documents or (iii) both. In addition, two different but related types of bibliographic structure are considered: *independent* and *dependent* on the query topics. Hence, in total, six possible relationship types can be formulated from the two relationship types [67]. Given these relationships, we posed the research question, can the evidences lying in the structural context surrounding the document externally be *intelligently* materialized? In Paper C we have developed a formalism that can be used to materialize and then utilize the contextual evidences in the structural context to improve the retrieval effectiveness.

The scope of the external context described above can be further broadened, for example, by going one or more levels deeper into the bibliographical structure. Instead of just considering only the direct out-links and in-links (which means one-hop, forward and backwards into the link structure), there could be double or multiple hops possible. For example, a double-hop could be, *out-links* (forward) and *in-links* (backwards) of the out-linking documents in the context and similarly for in-links. And multiple-hop would be the recursive definition of the previous statement. This way we could broaden the scope of the external context. This study only focuses on the one-hop external context. Double and multiple-hop external context studies can be an interesting future direction.

The finite permissible states with transitions in the bibliographical structure of documents (external, see Figure 2.2), can receive a similar treatment as that of the finite states in the hierarchical tree structure of a document (internal, Figure 2.1). The difference is that now the states are the documents instead of parts of the document (elements) and the transitions are the bibliographical references (out- and in-links). In the random walk process, a state change denotes the *authority* flow in the bibliographic structure of documents. A stable or an equilibrium state is established iteratively when the expected probability of visiting a state converges to a limit (the eigenvector of the adjacency matrix $\mathbf{A}$ in Section 2.1.3). This equilibrium state specifies the impact of a document in the bibliographical structure of the documents in the structural context.

Both the hierarchical and bibliographical contextualization approaches rely heavily on a large number of matrix operations. In medium scale matrix operations, there were minimal computational problems. In Paper D, we used a large semantically marked up test collection, Wikipedia (see Section 2.3.5), featuring 135 million hyperlinks with 2.66 million documents. The matrix **A** in that case became as large as $(2,668,160 \times 2,668,160)$. Handling such large scale matrices would have been a problem if we had not used the *extrapolation technique*[2] proposed in Paper B, which is about improving the performance of methods operating on very large matrices. See Paper B for details.

2.2.3 Hybrid Contextual Features

Hybrid contextualization is the process of merging together the features accumulated from the external and internal structural contexts. This contextualization approach can be performed in two steps: (i) good documents, having strong evidence in their structural context and strong basic relevance scores, are selected first, by applying bibliographical contextualization, and (ii) the most relevant parts of the documents are retrieved, based on good structural context and basic scores, applying hierarchical contextualization. Unfortunately, this approach has not been widely explored in this study. We have proposed and evaluated it briefly, in Paper D, and experimentally the approach exhibited promising prospects: with retrieval effectiveness above all other proposed methods.

Hybrid contextual features intend to explore the effects of the structural context when the capabilities of internal and external contextual features are merged together in an intelligent order. Based on the interpretation of internal and external context, identified in the previous two sections, a similar treatment is possible with the hybrid approach. We have not chosen to pursue that, because of not having enough theoretical and empirical information available now. This is one of the interesting future prospects from our study, and hence requires an independent and focused investigations.

As a retrospective note, this kind of contextualization very much resembles the specification of the *in-context* tasks offered at INEX (see Section 2.3.2). One of the future prospects can be to explore the similarities of hybrid contextualization with the in-context tasks at INEX and subsequently supplement one or both of them to perform even better.

2.3 Evaluations

Measuring the effectiveness of semi-structured retrieval models using standard IR approaches is insufficient or even inappropriate for several reasons. Some of which are:

[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 power method) and using the properties of Markov chains; $\lambda_1 = 1$ (dominant eigenvalue) [61].

(a) the retrievable units are heterogeneous, they could be a mix of full-text, sections, paragraphs or sentences of the documents,

(b) there is no notion of structural relevance in a standard IR setting,

(c) the indexing units are not the documents but rather parts of the documents, therefore the global and local parameters in IR based approaches do not make sense here,

(d) the evaluation turns a blind eye towards the *selection* (to cope with the overlap problem), it (in standard IR evaluations) only measures the relevance or scoring effectiveness [86, 10].

Hence, one has to define and delineate the evaluation framework for the semi-structured retrieval problem, specifically in schema-agnostic settings. In this section, the evaluation framework and the standard benchmark for the evaluation of semi-structured retrieval, i.e., INEX [32] used in several settings this study, are presented. Like traditional IR, the standard semi-structured information retrieval evaluation benchmark comes with standard toolkits, namely: test collections, topics, relevance assessments, tasks, tracks and metrics. The relevant settings, retrieval tasks and evaluation measures used in INEX and also applied in this study are briefly described here. A custom-defined task that is on the one hand suitable for our approaches and objectives, and on the other hand capable of simultaneously addressing both the more complex problem of 'overlap' and elements 'size bias' [66] is also defined later in this section.

2.3.1 Evaluation Model

Based on the theoretical challenges and empirical complexities, the evaluation task in semi-structured retrieval is inherently more difficult and challenging than a general IR evaluation. The experimental and evaluation model presented in Figure 2.3 forms the general framework used for the evaluation of the semi-structured retrieval task in this work. The technical nature of this research and the lack of resources prohibited us from having interactive evaluation settings with user involvement. Hence, the evaluation model has no indication of the user, in the overall picture. The evaluation model is inspired by the system-oriented information retrieval research proposed in [35].

The framework in Figure 2.3 focuses explicitly on the evaluation of semi-structured retrieval, specifically for this work. For example, it contains process nodes referring to the contextualization, random walk and selection process. It also consists of process nodes (e.g., indexer) which extract structure, content and other information from the semi-structured documents and maintain them in separate indices. Similarly, the query requests are processed and then represented. The relevance assessments comes from the evaluation benchmarking initiatives such as INEX [30] or TREC [92]. The query representations are matched against the pre- and post-processed indexing units. The pre- and post-processing of the index, among other

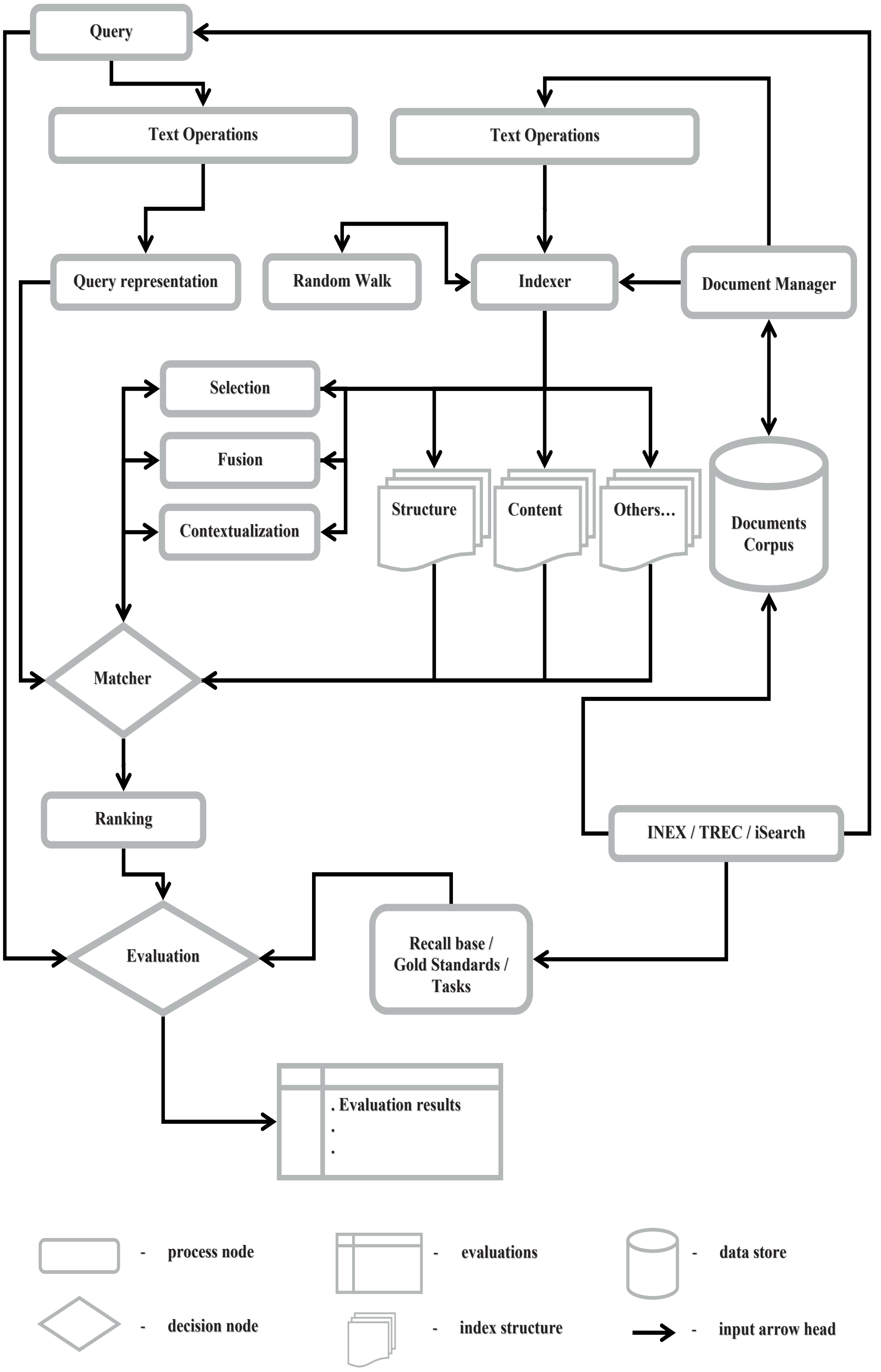


Figure 2.3: *Experimental and Evaluation Model.*

processes, includes: language pre-processing, pruning the unwanted meta-data, random walk score calculations, context aggregations, the calculation of the usefulness of each structural components, and so on. The results from the retrieval process (matching of query representation with document representations) are then reported. These reported results are then compared against the gold standards and recall base provided by INEX / TREC. The evaluation results will then expose the effectiveness of the retrieval model in question (Figure 2.3).

2.3.2 Standard Retrieval Tasks

The ad hoc track in INEX 2009 and 2010 features four different tasks, based on different assumptions. Historically, within the ad hoc track at INEX, several competitive tasks were proposed, evaluated and talked about over the years, so as to be representative of different use cases of semi-structured retrieval [51]. The results organization turned out to be the deciding factor for the definition of the different tasks at INEX [88]. The organization of results in some ways define the (semantic) interpretations of the textual results shown to the user.

INEX evaluations from 2009 and 2010 were employed in most of the studies conducted in this work. The organization of the results used at INEX 2009 and similarly in 2010 have been categorized into: (1) element (focused) retrieval tasks—thorough and focused tasks, and (2) in-context retrieval tasks—relevant in context and best in context tasks [7, 29]. The reason for mainly using the INEX 2009 evaluation metrics is because of the existence of the thorough task in 2009 which was later removed in 2010.

Element Retrieval

The primary difference between the two tasks in the category of element retrieval is the way they handle the ‘overlap’ problem. In the *focused* task, the goal is to find the best result on a path *without* overlap. This task is more user friendly, as it takes care of the repeated and nested results automatically. The *thorough* task, on the other hand, retrieves all the best results in a path, irrespective of whether they overlap or not. The goal in this task is to return an exhaustive list of items containing all the relevant elements in the collection covering the query topics.

in-Context

In the element retrieval category, the aim was to retrieve single elements that are relevant to the query topic, regardless of ‘how are they grouped?’ This means that parts of a particular relevant document might be at differently ranked positions in the result list. In the in-context tasks category, the objective is to rank the retrievable items *grouped by* documents. Hence in the *relevant-in-context* (RiC) task, all the relevant and at the same time non-overlapping elements are grouped together in the final result list. But in the *best-in-context* (BiC) task, only the best entry point *grouped by* documents are put in the result list.

2.3.3 Custom-defined task

Most of our published work are evaluated using the standard retrieval tasks offered by the INEX benchmark. In addition and also concurrently we have used a custom-defined task as well. This is because retrieving elements at various granularity levels has an unprecedented effect on the retrieval outcomes, for example, favouring retrieval systems reporting large elements in their results over systems retrieving fine-grained (focused) elements [9]. Therefore, in Paper D [66], to cope with that effect, a custom-defined task is employed: namely *granulations*. In granulations, a specific type of element is pre-selected from the collection, and hence the retrieval and evaluation is focused only on those elements. The recall base is also adjusted to take into account only those types of elements chosen in the granulation process. We have reported only two types of granulations, which are: (i) article level or full document granulation and (ii) paragraph or passage level granulation (graphically shown in Paper D, Figure D.4(a)). Of course, other types of element granulation could also be used, but for brevity and proof-of-concept, we reported only the two extreme cases. An analogous setting is also employed by Crouch et. al [24].

2.3.4 Evaluation Measures

The reason for choosing and applying the standard evaluation measures available at INEX is primarily because it is the sole initiative providing an evaluation framework for the semi-structured retrieval problem. And because of its being the sole benchmarking entity in this field, its evaluation measures and framework in general are widely acknowledged in the semi-structured research community. Lastly and more importantly for this work, it was easier to compare against the other distinguished state-of-the-art approaches (essentially the participating groups at INEX) with a great sense of confidence.

Element Retrieval

In order to appreciate the retrieval challenges within semi-structured IR, several evaluation metrics have been developed at INEX over the years. In the early days of the initiative [32], the retrieval was done regardless of the overlap problem. Eventually, the overlap problem was officially recognized [42], and that sparked the need for two logical tasks (i) with and (ii) without overlap. The non-overlapping results in the focused task are evaluated at the early precision value, i.e., interpolated Precision at 0.01 or 1% selected Recall level (iP[0.01]). The reason for using interpolated precision instead of the standard precision measure is to cope with the 'size bias' in the retrieval results. The standard precision at different rank r[3] is considered to favour shorter retrievable items than the longer ones [41].

$$
\begin{aligned}
iP[x] &= \begin{cases} \max\limits_{1 \leq r \leq |L_q|} (P[r] \wedge R[r] \geq x) & if\ x \leq R[|L_q|] \\ \\ 0 & if\ x > R[|L_q|] \end{cases} \qquad (2.2) \\
where &: \ R[|L_q|] \text{ is recall over all documents retrieved.}
\end{aligned}
$$

However, the overlapping results in the thorough task are also evaluated using the interpolated precision measure, but the one that captures the overall performance, namely, the *Mean Average interpolated Precision* measure (MAiP):

$$
\begin{aligned}
AiP &= \frac{1}{101} \times \sum_{x=0.0,0.01,\ldots,1.0} iP[x] \qquad (2.3) \\
MAiP &= \frac{1}{n} \times \sum_t AiP(t) \qquad (2.4)
\end{aligned}
$$

The *AiP* is calculated by averaging the interpolated precision scores at 101 standard recall levels $(0.0, 0.01, ..., 1.00)$.

In both the element retrieval tasks category, we have also reported our own custom-defined measure as well, the *Mean Average element Depth* (MAeD). The purpose of this measure is to get an approximation of the mean average depth of the result list, in order to get a feeling for the 'focused-ness' of the retrieval approaches [64].

in-Context

The in-context tasks to some extent resemble a special case of document retrieval. The results are grouped by documents (similar to document retrieval systems), while they are additionally enriched with the most relevant fragments of the documents. From the evaluation perspective, their retrieval performance is measured quite similarly to that of the document retrieval counterpart. The only difference is that the relevance score of the retrieved relevant document depends on the elements retrieved by the system. The measures used in the in-context tasks are based upon the generalized *non-binary* Precision (gP) and generalized Recall (gR) metrics [44] over documents. Here, the per document score at INEX is calculated using the f-score or the *harmonic mean* of precision and recall—that is, the fraction of retrieved and highlighted text in the document (text highlighted by the assessors).

$$F_\alpha = \frac{(1+\alpha^2) \times Precision \times Recall}{(\alpha^2 \times Precision) + Recall} \qquad (2.5)$$

[3] Precision at rank r is defined as the fraction of retrieved text that is relevant. Similarly Recall at rank r is defined as fraction of relevant text that is retrieved [12].

The α parameter in F_α is set to be 0.25 or 1/4, which means, in the above equation, the precision is four times as important as the recall. Hence, in the *relevant in context* task, the *Mean Average generalized Precision* (MAgP) measure is used to estimate the overall performance, where the generalized score per document is based on the retrieved highlighted text. The *best in context* task is evaluated similarly, using the MAgP measure, with the exception that the generalized score per document here is based on the distance to the assessor's best entry-point [30]. In this, the per document score is the linear discounting function of the distance d, measured in characters:

$$\begin{aligned} &\frac{n - d(x,b)}{n} \\ &for\ d < n \\ &otherwise\ 0 \end{aligned} \tag{2.6}$$

Here, n corresponds to the number of characters in the visible part of the document on the screen. At INEX, n was set to be 500, which means 500 representative and relevant characters (parts of the document) are retrieved in this task, for more details about the measures used at INEX please refer to [41].

Granulation

To evaluate the results in our custom-defined granulation task, some adjustments had to be made, both in the collection when *retrieving* and in the recall base when *evaluating*. For our two reported granulation schemes, at the article and the paragraph levels, we used the following structured queries (NEXI [89]) for retrieving articles and paragraphs respectively: (i) *//article(., about("query–expr"))* and (ii) *//p (., about("query–expr"))* [66]. In the article granulation, only the root element is allowed in the results list, while in the paragraph granulation only the small element, i.e., ⟨p⟩ element is retrieved. A similar treatment is applied to the recall base, in both cases, independently. Each item in the results list is considered as a document (in both cases), and hence the standard TREC evaluation framework [92] is applied to the tailored semi-structured retrieval approach. The standard TREC measures used and reported are the following:

- *Precision–Recall* Curve, precision values are plotted at various recall values (11-point interpolated precision–recall average curve).
- *Mean Average Precision*, to get the overall performance against all the topics, the mean of the average precision for all the topics is taken.
- *Precision@N*, precision at N retrieved results, preferably N is used in the range {5, 10, 20}.
- *rPrecision*, precision at total number of relevant results for a particular topic.
- *Bpref*, number of results judged non-relevant found before the first result judged relevant comes up.

2.3.5 Collections and Topics

In order to perform all of the above evaluation tasks defined so far, decent sized and sufficiently representative yet standard test collections are needed, to collect enough empirical and statistical proof of the validity of our proposed retrieval methods. In the Ph.D. process, the tasks of finding, indexing and retrieving the representative test collections and topics are by far the most time consuming. Thanks are due to the INEX and iSearch [56] collections for making this laborious task more convenient for the research community, and in particular for this study.

The iSearch test collection

The relatively newly released iSearch test collection comprises scientific documents from the domain of physics. One of the largest repositories, covering the main areas of physics, is arXiv.org, containing around 500,000 papers [56]. The iSearch test collection contains:

- 18,443 book records in XML.
- 291,246 metadata for articles.
- 143,571 full-text articles in PDF.
- 3,768,410 bibliographical citations among the collection.

The above results are extracted from arXiv.org, and made part of the iSearch test collection. In addition, iSearch also comes with 65 query topics with relevance assessments. Paper C used the iSearch test collection and topics for empirical analysis and evaluations. In this paper, we have used the standard TREC evaluation framework, and a document retrieval strategy, as described in the earlier section, because of the nature of the retrieval method proposed.

The Wikipedia XML collection

The most widely used, tested and analysed XML marked Wikipedia corpora has been in use at INEX and also other venues with different objectives than semi-structured ones, for years now [38, 37, 47]. The semantically marked-up Wikipedia collection [76] has been in use at INEX since 2009, and is still in use. Wikipedia has been primarily used in the ad hoc track at INEX. The large, English language, Wikipedia covers:

- 2.66 million XML marked articles.
- 50.7 Gigabytes in size.
- 135 million within collection citations—which we have extracted ourselves, not included in the INEX package.
- Approximately 32,000 unique element types.

In addition, the collection comes with 68 topics with character-wise assessments. This large semantically marked-up collection of Wikipedia documents suits our purpose for the following reasons:

1. First and foremost, it contains explicit structural components, in the form of XML and semantic markup.
2. It has a deeply nested hierarchical structure; Figure 2.1 is an example document from Wikipedia.
3. It is large and comprehensive enough for our objectives.
4. There is a huge number of citations in the collection. There are unique characteristics of the citation structure in Wikipedia compared to the citation structure of the Web [38]. Therefore, it is not characterized as a bibliographical structure but rather loosely characterized as a *hyperlink* structure. Because the links do not necessarily reflect the bibliographical semantics, two documents can reference each other without temporal ordering: A *refers* B and B *refers* A, which is not possible in a legitimate bibliographical situation.
5. It is relevant also because of the availability of the thorough task, which is needed for our contextualization models (Papers D, E).
6. There has been a large variety of runs from distinguished participants from institutions around the globe, readily available at INEX's website[4], with all supplementary information, statistics and results which can be freely used for research purposes. These submitted runs at INEX have been used in our studies to build up competitive baseline systems (see Section 2.4), both individually and based on the fusion approach.
7. As it has been in use for years now, there is a great advantage here, namely, there is quite a wide variety of research experiments that have been performed on the same collection. If we had performed our experiments on our own custom-built or synthetic test collection, then comparing the retrieval effectiveness would have been quite a problematic task in itself, and also not likely to be widely acknowledged.
8. Subsequently, with INEX and existing studies, we are able to conveniently compare our propositions with them in a straightforward and informative manner, and at the same time with a greater theoretical and statistical confidence.

In the Paper F, partially, and in Paper D exclusively, we have applied the granulation task, using the Wikipedia collection. Hence the first study we conducted on Wikipedia is primarily based on the relevance judgements from INEX, but the evaluation is done using the TREC framework instead of INEX, because of the granulation task. In Papers E and F, the INEX's element retrieval tasks was employed to observe and report the performance of the proposed methods on the Wikipedia collection. However, in the last Paper G, all the the four standard INEX retrieval

[4] http://www.inex.otago.ac.nz/

tasks (described in Section 2.3.2) were performed using, again, the Wikipedia test collection.

2.4 Baseline Systems

Comparing against not *good enough* baseline systems has already been identified as a practice that allows a lack of overall improvement to go unnoticed [6]. The main criticism that is addressed to those studies that lack such a capability, is their inability to comply with standard practices. Testing and analysing the system's performance against a weak and often or always non-standard baseline not only gives a false sense of positiveness but also hides and keeps obscure the core research challenges (to be solved) within the area. Such baseline systems prevent the contributions from getting even a remote sense of meaningfulness in a real and practical setting. The solution to this problem, as has also been argued by Armstrong et. al. (2009) [6], is: (a) to engage in the standard practices, by using standard test collections, relevance judgements, topics and (b) most importantly, compare the proposed methods with the best known results—themselves built upon the aforementioned standard practices.

Keeping the above observations in mind, in this study, a considerable amount of emphasis has been put on the effectiveness and characterizing role of the baseline systems in our evaluation processes.

2.4.1 indri - lemur

Indri, a lemur project[5], is an open source, freely available and standard search engine widely used in the IR research community. It is customizable and is largely scalable. In our Paper C, we made use of the indri baseline run, primarily employed the *#combine* operator for combining beliefs, and also applied stop-word removal and the porter stemming algorithm. The results obtained are shown in Table 2.1. The results obtained from the proposed methods are compared against the baselines and the evaluation outcomes are validated using statistical significance testing at $p < 0.05$, 2-tailed t-test (suitable for these kinds of problems). The task performed using this baseline system is done in an ad hoc document retrieval fashion.

2.4.2 Fusion

One of the strong baseline schemes (our own contribution), which has been recognized by the reviewers and the community in general, is the baseline system built upon *fusing* a set of competitive retrieval results. Data fusion techniques[6] [22,

[5] http://lemurproject.org/indri.php

[6] A process of combining two or more retrieval results in the pursuit of a resulting better set of items [78].

78, 59] proved to be the only strong and successful backbone for constructing the ambitious baseline systems used effectively in our different studies [66, 63, 64]. In the rest of this section, we will give a synopsis of those baseline systems that took part in the different retrieval tasks defined for this study.

Granulation

Rank and *score* based data fusion techniques are effectively applied to document retrieval tasks [22, 78, 59]. In this task, these fusion techniques have been modified and tested to be used and serve their purpose in the semi-structured IR settings. Thus, to carry out a fusion for semi-structured result lists, we had to somehow consider each result item as an individual document. In the granulation task, Paper D, the *reciprocal rank fusion*[7] [22] turned out to be quite effective. For each of the candidate retrieval result lists, every result item (elements in our case) is given a score based on its rank position, per query topic (the details on how we do the calculation can be found in Paper D). The retrieval effectiveness of this fusion technique can be seen in Table 2.1 (Reciprocal_Fusion row). The statistical significance test done in this case are at $p < 0.01$ and $p < 0.05$ (which means at a significance level of 0.99 and 0.95 respectively), 1-tailed t-test.

A total of 173 runs were officially submitted at INEX 2009 by participants around the globe [30]. Out of 173, we used 159 for the fusion, as 13 were not element runs, they contained ranges of fragments of *file offset lengths* (FOLs) as retrievable units. In addition, we had to make a deliberate choice of removing 61 noisy runs, having a large number of elements lacking in the document collection. Finally, a total of 98 runs were fused together, as described above.

In the full document granulation task, the fusion baseline run outperformed all the INEX 2009 officially reported runs, with a $MAP = 0.4141$. While in the paragraph granulation task, only the technical university of Queensland (qtau) run outperformed the fusion baseline [66].

INEX tasks

In a quest to construct an even better baseline system strategy, a further in-depth analysis of the properties of fusion techniques was essential. In this process, a data fusion scheme based on the sum of the normalized similarity scores, namely *CombSUM* [78], was found to be of interest. For the designated INEX element retrieval tasks, the CombSUM fusion technique gave us the best overall performance, see Paper E.

Out of 98 runs to be fused, unfortunately 56 of them have not reported any real scores with their runs, rather only the rank ordering was given. For those runs, we had to calculate artificial relevance scores for each result list item. The reciprocal rank of each item is used as its relevance score. We will refer to the

[7] A fusion technique based on the *rank* and the social voting system.

final fusion baseline run as the CombSUM_Reciprocal fusion. In the Paper G, we give a detailed synopsis of the fusion techniques in semi-structured retrieval, and their role in the construction of the independent *selection* and the *scoring* systems. Table 2.1 provides an overview of the performance by this fusion baseline (CombSUM_Reciprocal row). For validation, similar types of statistical significance testing have been performed for these tasks, as well as that of the granulation task.

2.4.3 Individual INEX runs

In addition to the competitive fusion baseline methods described above, we were also able to compare the results from our approaches to individual top-*k* official INEX submitted runs as well, task-wise [66, 63, 64]. The performance overview of some of the top individual INEX submitted runs used in this study as baseline system are depicted in Table 2.1.

Table 2.1: *A representative set of baseline systems – task wise, used in different studies (Part II).*

Run ID	**Task**	**Papers**	**MAiP**	**MAgP**	**MAP**	**iP[0.01]**	**P5**	**rPrec**
indri	ad hoc	C	–	–	.0803	–	.1938	.1041
Reciprocal_Fusion	$gran_{article}$	D	–	–	.4141	–	.6618	–
UAmsT	$gran_{article}$	D	–	–	.3578	–	.6500	–
Reciprocal_Fusion	$gran_{para}$	D	–	–	.2189	–	.4500	.3479
QTau	$gran_{para}$	D	–	–	.2286	–	.5324	.2779
CombSUM_Reciprocal	thor-foc	EG	.3396	–	–	.7273	–	–
UWFerBM25F2	thor-foc	EG	.1854	–	–	.6333	–	–
I09LIP6Okapi	thor-foc	EG	.3000	–	–	.6141	–	–
UJM_15525	thor-foc	EG	.2890	–	–	.6060	–	–
UamsFSecs2dbi100CA	thor-foc	EG	.1928	–	–	.5997	–	–
BM25BOTrangeFOC	thor-foc	EG	.2912	–	–	.5992	–	–
Spirix09R001	thor-foc	EG	.2865	–	–	.5903	–	–
LIG-2009-focused-1F	thor-foc	EG	.2702	–	–	.5853	–	–
BM25AncestorBIC	bic	G	–	.1706	–	–	–	–
BM25AncestorRIC	ric	G	–	.1865	–	–	–	–
UamsRSCMACMdbi100	ric	G	–	.1773	–	–	–	–
UamsFSsec2dbi100CA	foc	G	–	–	–	.5997	–	–
UAmsIN09article	thor	G	.2818	–	–	–	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.5 Summary

In order to be able to comprehend the concepts and contributions of the papers provided in part II of this thesis, the background information was presented in this chapter as the result of past, present and also partially future works, regarding semi-structured retrieval within schema-agnostic settings. The main research challenges arising from the existing body of work were also outlined and related to the studies conducted in this thesis. We have also had a closer look into the internal and external contextual features of semi-structured documents and briefly discussed how they can benefit the retrieval of focused and relevant elements. From the theoretical background, the evaluation environment was defined in terms of the different tasks performed at different stages of this study and how did they help solve the core challenges within the area. Also, the methods of evaluation of the experimental settings used in this study were described in light of the different evaluation measures used. Next, the importance of representative and standard test collections, query topics and benchmarking were also highlighted. Finally, the need for a competitive baseline system was emphasized, and the different ambitious and competitive baseline systems used throughout this study were introduced.

CHAPTER 3

Research Summary

"So long as fortune is favouring you,
your defects will remain covered."

– ALI

Bond between the coordinated research work that constructed the outlook of this thesis will be portrayed in this chapter. Accordingly, the summary and relation between the main findings of this work and the state-of-the-art research problems will also be characterized in light of how they support one another. To do that, first, Sections 3.1 and 3.2 will briefly sketch the overall research process, research themes, and the publications. Section 3.3 outlines a more detailed, chronological and thematic account of the included papers in the perspective of *what* they intend to achieve, theoretically and experimentally, and *how* do they relate to the topic of this thesis.

3.1 Formalities

The work described in this thesis was completed during a three and one-half years Ph.D. program, which was supported and funded both by the iAd Centre, which is financed by the Research Council of Norway, and NTNU. The Ph.D. period included one-half year of teaching duties as well. The teaching duties were carried out in the following courses:

- TDT4186 Operating Systems,
- TDT4290 Customer Driven Project,
- IT3709 Intelligent User Interfaces and
- TDT4190 Distributed Systems.

In addition to the teaching duties, the following courses were successfully completed, required as a part of the Ph.D. program at NTNU:

- DT8116 Web Data Mining.
- IT8802 Advanced Information Retrieval.
- DT8108 Topics in Information Science.
- TDT4215 Web Intelligence.
- DT8114 Ph.D. Seminar in Computer and Information Science.

Four of the courses included final exams, in addition to the obligatory practical assignments throughout the courses. The work done in the period 01.11.2011—06.06.2012 was carried out at the Centrum Wiskunde & Informatica (CWI), Amsterdam, under the supervision of Professor Arjen P. de Vries, and the rest of the work (01.12.2009—08.07.2013) was carried out at the Department of Computer and Information Sciences, IME, NTNU, Trondheim, except for conferences, workshop attendance, and vacations.

3.2 Publications and Research Themes

Table 3.1: *Overview of the papers.*

Paper	Title	Ref.	Incl.
O	Extrapolation to Speed-up Query-dependent Link Analysis Ranking Algorithms.	[60]	–
A	Relevancy in Schema Agnostic environment.	[62]	✓
B	Faster ranking using Extrapolation techniques.	[61]	✓
C	Contextualization from the Bibliographic Structure.	[67]	✓
D	Contextualization using Hyperlinks and Internal Hierarchical Structure of Wikipedia Documents.	[66]	✓
E	Kinship Contextualization: Utilizing the Preceding and Following Structural Elements.	[63]	✓
F	When is the Structural Context Effective?	[65]	✓
G	Selection Fusion in Semi-structured Retrieval.	[64]	✓

The most natural characterization of the papers in Table 3.1 (chronologically ordered) can be based on the tasks performed in different phases of this study. This classification is driven by the empirical nature of this work. The tasks form the thematic and at the same time hierarchical structure for this study (from one phase to another and within one phase from one state to another, shown pictorially in Figure 3.1)—which, coincidently, resembles the structure of the intended data. Hence, we broadly categorize the papers in the following way:

I. *Ad hoc task* – The retrieval model is based on standard document retrieval techniques and the evaluation is performed in an ad hoc fashion, which in our case means that the evaluation is based upon the TREC benchmark.

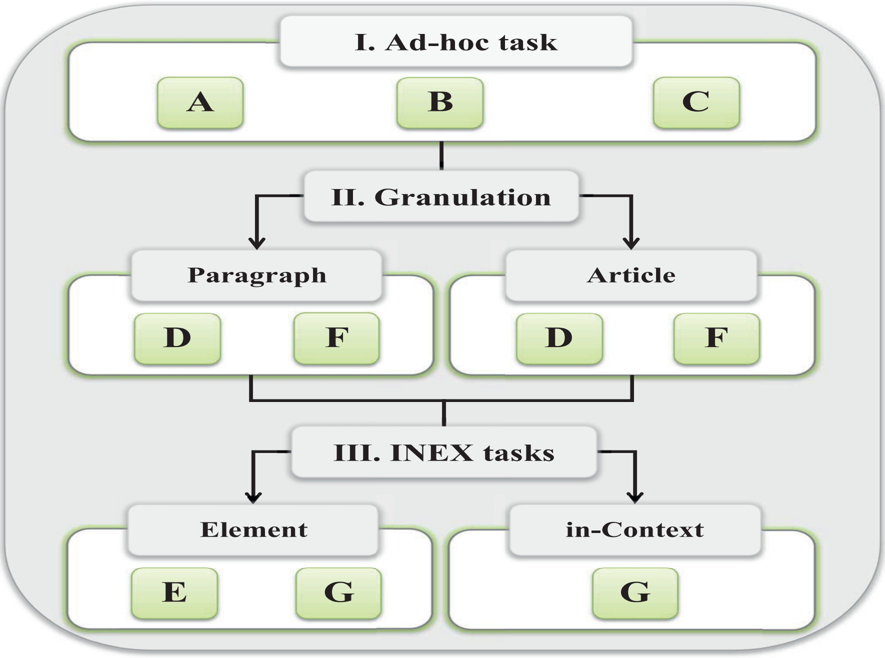


Figure 3.1: *Overview of the papers grouped by tasks (logical themes).*

II. *Granulation tasks* – A custom defined task in order to retrieve and evaluate using the granulation principle. This task includes two subtasks, (i) *article* and (i) *paragraph* granulation.

III. *INEX tasks* – These tasks include using the requirements and standards set by the benchmarking body (INEX). They are composed of four INEX defined sub-tasks, which are then classified into *element* and *in-context* tasks. The tasks are: *focused* and *thorough*, which count as element retrieval tasks, while *relevant in context* and *best in context* tasks count as in-context tasks. The details are given in Section 2.3.2.

Contributions from the included seven papers in this thesis, together with the above categorization, form the overall picture of the thesis as given in Figure 3.1. The letters in Figure 3.1 correspond to the letters assigned to the papers in Table 3.1 and also Part II of the thesis. The letters appearing in more than one theme signify that the study in that paper is performed within all those tasks, for example, the experimental evaluations in Paper D are performed within paragraph and article granulation tasks.

These tasks, by definition, are intended to address a set of designated challenges and issues within the area, as indicated in Section 2.3.2. *Structural context* has an important role to play in all three tasks and in addressing the inherent challenges and issues they pose. We, therefore, took a step by step approach, based on the hierarchical structure depicted in Figure 3.1, towards solving or at least engaging in the issues and challenges from each category that form the overall research challenges of the study. Each step involved engaging in (from top to bottom) one particular layer from the overall structure. Below, a synopsis of each of the layers is presented in relation to the outcomes of this work.

3.2.1 Ad hoc task

In ad hoc task, the *efficiency* question (not the direct focus of this thesis) has been primarily addressed using the novel *extrapolation technique* [61] proposed in Paper B. This technique has been used in the later studies [66, 63] to ensure the computational feasibility of the large scale matrix operations, while Paper A provided the theoretical background and prospects of the area. Consequently, inspired by the link analysis ranking approach from Paper B and the developed theoretical background, the *contextualization* model has been proposed in Paper C. Here, the structural context is assembled from the *external contextual features*—for example, the *bibliographical structure* of the document. Evidences were systematically collected from the structural context, to be later *combined* and subsequently *made* part of the relevance score of a document—which empirically revealed an improved retrieval *effectiveness*.

3.2.2 Granulation

From the broader ad hoc task, *granulation* tasks were formulated, where the aim was to focus on some pre- and / or post-defined structure (element) type and conduct the evaluations and analyses on those retrievable item types only. While there exist other rational granulation possibilities two representative and mutually disjoint cases are considered in Papers D and F: (i) article, as one end of the granulation spectrum and (ii) paragraph, at the other end of the spectrum. The *article* granulation was chosen to observe and analyse the disposition of the larger (-est) structural constructs in addressing the retrieval challenges within schema-agnostic settings, in order to get an answer to the core question, what is the role of “larger size” in retrieval effectiveness in the contextualization process, in terms of the amount of enclosed textual content within a retrievable element. Paper D argues that since larger structural constructs usually constitute most of the *textual evidence* within a semi-structured document, they are generally the top ranked results, retrieved by most of the semi-structured retrieval systems at INEX [9]. At the other end of the spectrum are structural constructs containing *not enough* textual evidence, the *paragraph* elements. In paragraph granulation, the aim is twofold: first to observe why the *not-large-enough* elements usually are ignored by the retrieval systems and then attempt to alleviate the problem (‘size biasedness’, Section 2.1.6) in such a way that the smaller structural constructs can also play a role in satisfying the information needs. The *internal* hierarchical contextualization model does exactly that; (i) *boosts* the relevance of smaller structural constructs appearing in a good structural context; and (ii) *setbacks* the relevance of larger structural constructs in not so good structural context.

3.2.3 INEX tasks

We took the next step to supplement and upgrade the findings from the granulation tasks phase by addressing the subsequent challenges within the standard INEX task framework. The ad hoc track at INEX features four different tasks. The primary aim in all four tasks, in response to the *overlap* challenge (described in Section 2.1.6), is to expose the significance of the *results list organization* in satisfying the user's needs. The *focused* and *thorough* tasks have an element retrieval strategy for organizing the result list. While the *relevant in context* and *best in context* tasks are inclined towards a document-retrieval oriented organization strategy. Just like the granulation tasks, the categorization within the INEX ad hoc track are also based on two extreme cases, namely, (i) the element and (ii) the document retrieval cases. In contrast to granulation, element retrieval based tasks at INEX differ in the way they handle the overlap problem. The aim of the *thorough* task is to try to retrieve (cover) as much relevant element as possible from the collection, which means allowing the result list to contain overlapping elements. On the other hand, the *focus* task is destined to retrieve a non-overlapping list of focused answers. Paper E contributes towards an element retrieval approach of presenting a focused and comprehensive list of answers, employing structural context, which primarily involves the elements in the kinship relationship function. In this way we were able to address both the "focused-ness" and "overlap" issues (from Section 2.1.6) simultaneously. In this study, we hypothesized that the structural context gathered from the elements in kinship relationship in the hierarchical structure of the document actually enhances the retrieval of the focused items. Experimental results have validated the hypothesis, by using the standard evaluation benchmarks, query topics and test collections from the INEX initiative.

The last paper in the list, (Paper G), proposes a methodology which is capable of addressing the "overlap", "focused-ness", and "size bias" challenges at the same time. We argued in the paper that the construction process of a semi-structured retrieval system should involve two individual and independent development phases. The first phase should concentrate only on the *scoring* or the relevance modelling, while the other phase should be focused on the different *selection* scenarios, based on the required use case(s) or any other user or system defined constraints. We hypothesized that the *scoring system* should comprehensively and independently gather a focused list of results, and similarly, the *selection system* should manage the different result list organization issues. The scoring system is indirectly responsible for taking care of the 'focused-ness' and 'size bias' issues, while the selection system is obliged to take care of the 'overlap' issue, based on different scenarios. The methodology is tested in all four retrieval scenarios within the ad hoc track at INEX, and the empirical results exhibited a strong statistical indication of improved effectiveness as a result of *selection fusion*.

In the next section, we take each of the papers (in alphabetical order) from Table 3.1 and broadly analyse their individual contributions which form the overall picture of the thesis.

3.3 Summary of the Included Papers

This section presents a descriptive account of what the included papers are intended to achieve theoretically and how they relate to: (a) one another, (b) the thesis topic, and (c) the research questions raised in Section 1.4, and hence to what degree they tend to answer them. All the studies are conducted within the evaluation framework presented in Figure 2.3.

3.3.1 Paper A

Relevancy in Schema Agnostic Environment.
Muhammad Ali Norozi
Bulletin of IEEE Technical Committee on Digital Libraries
In conjunction with the ACM/IEEE Joint Conference on Digital Libraries;
Volume 8.(1) pages 1–6, IEEE 2011.

Abstract: *Relevance* is an important component in full-text search and often distinguishes the implementations. Relevancy is used to score matching documents and rank them according to the users intent. One of the reasons for the high popularity of Google is its good relevancy originally based on the PageRank algorithm.

The emergence of *semi-structured* data as a standard for data representation opened up new areas which could be related to both the database and information retrieval communities. Although the information retrieval and database viewpoints were, until quite recently irreconcilable, semi-structured retrieval helped to bridge the gap. This work is about exploring relevancy in semi-structured retrieval both in isolation and as a bridge between the database and information retrieval communities.

Research process

Just after starting the Ph.D. study, the most essential first step was to engage in understanding and internalizing the different aspects of the chosen Ph.D. topic, theoretically. This paper was intended to present a theoretical overview of the thesis area. From the knowledge attained at the early stages of the Ph.D., a fairly thorough literature review in the area has been brought together in this paper, in terms of the challenges in the area, opportunities and possible research problems and directions from the research topic. The paper puts an emphasis on "how to actively employ the structure in the relevancy sub-system in a schema-agnostic environment." Broadly two points of view on schema-agnostic searches have been argued in this work, in the wake of the existing studies: (i) the relational viewpoint and (ii) the native or / and more flexible viewpoint. The relational viewpoint tends to fit the flexible (schema-agnostic) semi-structured data within the mature and not very flexible (strictly structured) domain of relational databases. While the native viewpoint is inclined towards the traditional IR style treatment of semi-structured data, which means to ignore or only marginally use the structure. The paper concludes with pointers to expected research opportunities and directions which served as a good starting point for the forthcoming contributions. The paper also offers a preliminary list of research questions which has been further refined during the course of the Ph.D., and subsequently included in Section 1.4 in this thesis. In addition, Chapter 2, the research background, has also been partially inspired by the discussion of the state-of-the-art in this paper.

Roles of the authors

The primary author, the candidate, did all the work.

Retrospective view

The paper featured a preliminary survey of the area, with a decent set of existing ideas and opportunities. It would have been probably more likely to have had wider implications if this paper had been written at a bit later point of time in the Ph.D. process, after having obtained a more thorough and deeper understanding of the area, which could have led to its being published in a venue with an higher impact factor, in order to reach a wider audience.

3.3.2 Paper B

Faster Ranking using Extrapolation Techniques.
Muhammad Ali Norozi
International Journal of Computer Vision and Image Processing;
Volume 1.(3) pages 35–52, IGI Global 2011.

Abstract: *Extrapolations* are simple and unique techniques in linear algebra that require little additional infrastructure that needs to be incorporated in the existing query-dependent *Link Analysis Ranking* (LAR) algorithms. Extrapolation in LAR settings relies on the *prior knowledge* of the (iterative) process that created the existing data points (iterates) to compute the new (improved) data point, which periodically leads to the desired solution much faster than the original method. In this study we present the novel approaches using extrapolation techniques to speed-up the convergence of the query-dependent

iterative methods like HITS and SALSA, *link analysis* based ranking methods, where hyperlink structures are used to determine relative *importance* of a document in the network of inter-connections. We work within the hubs and authorities framework defined in HITS and SALSA and propose the use of different *Extrapolation techniques* for faster ranking. Hence we come up with a novel improvement in algorithms like HITS, SALSA and their descendants (e.g., Exponentiated and Randomized HITS) using the Extrapolation techniques. With the proposed approaches it is possible to accelerate the iterative ranking algorithms in terms of reducing the number of iterations and therefore uncovered a much faster rate of convergence. In the experiments, in the concluding part of the article, we even got much better results than the theoretically predicted assertion. The results present a speed-up to the order of 3–19 times better than the original algorithms.

Research process

In addressing some of the challenges already identified in Paper A, and also in light of the earlier observations of the author, this paper formed the most plausible bridge to the future opportunities from the existing background knowledge. The primary research goal for this work was to study the behaviours of the iterative algorithms (linear algebra [50]) within *link analysis ranking* (LAR) domain, which have been applied over the years to simplify and often resolve some of the core IR challenges. In addition, when it comes to semi-structured retrieval, the LAR approaches have not only played a vital role in solving some of the issues here [13], but also opened up new research opportunities in the area [67]. The goal of this paper was to establish a strong foundation from which it would have been quite likely, at later phases of the study, to further develop the LAR based approach proposed by Guo et al. [33] (XRANK system) and the structure based approach by Arvola et al. [9]. The role of LAR based approaches here is to make computationally feasible the iterative operations on very large matrices. As has been discussed in Papers D and C, matrix operations have a fundamental role to play in identifying the importance or impact of each node (vertex) in the structural context of a document or element. A technique such as *extrapolation* accelerates the performance of the iterative ranking algorithms (having a Markov chain random process), to solve the system of linear equations representing the retrieval problem. Without some such capability, the goals achieved later in this thesis would not have been even feasible. Operating on a matrix of order 2.6 million × 2.6 million (Paper D) requires an extensive amount of computational capabilities, which were not available at that point of time, therefore accelerating the performance was the only practical and logical choice. Hence, this paper only, in contrast to other papers, addresses the *efficiency* issue (research question RQ_{ef}), which is spread across the overall picture of the thesis. Although there were other efficiency measures also applied during the course of experimental analyses, but they were not reported or even mentioned in the paper; for example, maintaining data-structures for storing a large matrix (i) column-wise and also (ii) row-wise, which means storing one matrix twice, was experimentally observed to be a very efficient alternative, given that the system has enough memory available. These aspects of the semi-structured retrieval are kept for future efficiency oriented studies.

Retrospective view

An ad hoc experimental and evaluation style of analysis was chosen in this paper. Test collections, recall bases, and other evaluation statistics are used from another related study [90, 91] in the area. The results from this study have also been effectively applied to the other standard test collections as well (Papers D and C). The implications of this study might have been even wider and more comprehensive if the findings from the subsequent papers that employed the extrapolation methods to the standard semi-structured test collections were also officially included and reported in this study, maybe at a later point when they were available.

3.3.3 Paper C

Contextualization from the Bibliographic Structure.
Muhammad Ali Norozi, Arjen P. de Vries, and Paavo Arvola
CEUR Workshop Proceedings
In Birger Larsen, Christina Lioma and Arjen P. de Vries editors.
Proceedings of Task Based and Aggregated Search workshop,
in conjunction with the 34th European Conference on Information Retrieval (ECIR); Volume 1. pages 9–13, ACM 2012.

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).

Research process

The idea for this study became the initial plan to carry out external research work at the Centrum Wiskunde & Informatica (CWI), Amsterdam. After the first round of discussions and feedback with Professor Arjen de Vries, it was decided to do a more thorough study in the area than planned. Consequently, a wider theoretical and insightful state-of-the-art study helped broaden the perspective for the overall idea. In the literature study at that time, there were considered a broad range of ideas within semi-structured retrieval and the link analysis ranking methods. The research direction for this work is still based on the initial idea but now backed with a more thorough research background. The preliminary hypothesis for this study was to analyse the effects of *external context* on improving retrieval of the relevant documents. With the theoretical settings ready, the next milestone was having an experimental setting capable enough and flexible enough to take care of both the external (at that point) and internal contexts (later). We have tested several alternatives; for example, Lucene[1], Terrier[2], and PF-TIJAH[3] [53]. After an initial round of investigations of the technical specification of these open-source retrieval systems, the decision was taken to build a custom-designed semi-structured retrieval system for our own specialized research purposes. The amount of time required to customize the already mature retrieval systems was anticipated to be more than developing a precise and customizable retrieval system from scratch which can effectively serve our flexibility requirements. Based on these findings, the next step was to have a standard and representative test collection and an evaluation framework. The iSearch test collection [56] offered both the packages for this study, (i) a test collection and (ii) supporting evaluation data to compare with. We concluded that experiments performed on the iSearch citation structure was one of the novelties in this study. In addition, the *Indri* retrieval system was used as the baseline to compare our results against. Indri had already been used for the same collection in different settings [56]. The paper addressed research question RQ2 from the list of research questions in this thesis (Section 1.4). The empirical and theoretical observations made in this paper were originally done on a broader scale than actually reported. For example, the effects of publication date on impact scores and also a wide range of experimental results, unfortunately, could not be added to the paper, because of space limitations. However, the results which were part of the paper provide a sufficient and necessary empirical overview of the proposed methods.

Roles of the authors

The primary author did most of the work. This paper is co-authored with Dr. Paavo Arvola from the University of Tampere and Professor Arjen P. de Vries

[1] http://lucene.apache.org/core/

[2] http://terrier.org/

[3] PF-TIJAH, a widely used and originally developed at the information systems research group at CWI.

from the Centrum Wiskunde & Informatica, Amsterdam, the Netherlands. The role of the candidate was: the initiator of the idea; setting up the experimental environment; evaluations and analyses; and writing most of the paper. Dr. Arvola helped with feedback/discussions during the development of the idea and helped during the write up. Professor de Vries helped with improving the text and by giving fruitful feedback/discussions during the development of the idea and during the writing process as well.

Retrospective view

Retrospectively, a more renowned publication channel, with more flexible space requirements (number of pages), would have been a better choice for this kind of study. The potential for improvements was observed to be not so large in this collection. As argued in the paper as well, this was primarily due to the bibliographical incompleteness of the collection. There were quite a large number of bibliographical links which referred to documents outside the iSearch collection [67]. Having expressed these limitations, it would have been wise to also apply the ideas in this paper to a different test collection with a rather steady and more importantly, a semantically inclined link structure[4] [38], together with the iSearch test collection.

3.3.4 Paper D

Contextualization using Hyperlinks and
Internal Hierarchical Structure of Wikipedia Documents.
Muhammad Ali Norozi, Paavo Arvola, and Arjen P. de Vries
In CIKM'12, Proceedings of the 21st ACM International Conference on
Information and Knowledge Management.
ISBN 978-1-4503-1156-4. pages 734–743,
ACM Press 2012.

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

[4] Links or references supplementing or augmenting the logical incompleteness or semantic relationships of the concepts discussed in the *source* document with that of the *target* document.

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.

Research process

This research was also performed at CWI, Amsterdam. With the experimental environment almost already in place, the initial idea of this work was to *generalize* the use of the *structural context* both way from external context to the internal context, their combination in one or another way. Hence, from the existing experimental framework, the next rational step was to make the retrieval system also capable of indexing the internal structure of the semi-structured documents. An indexing structure which provides a straightforward access to the structural constructs within the documents. Dewey encoding or labelling is chosen to provide such a capability. Here, we hypothesized that the context, *external* and *internal*, can be used to deduce the retrieval effectiveness of a retrievable unit (document or element) in granulation. This study also involved operating with medium to large size matrices, hence we have used the outcomes from Paper B to make the process sufficiently efficient. This study tend to primarily target the research questions RQ1 and RQ2 and marginally RQ3 as well. In addition, the ideas presented in this paper are applied to the widely used and semantically marked Wikipedia collection at INEX [76]. The study provides an exploratory investigation into the role and importance of internal and external context, in the contextualization process, and in improving and unveiling the retrieval effectiveness of small and large retrievable units, using the respective granulation tasks.

Roles of the authors

The primary author did most of the work. The role of the candidate was: the initiator of the idea; the experimental set-up; evaluations and analyses; and writing most of the paper. Dr. Arvola helped with setting up the baseline system—which in this paper was based on the fusion technique. In addition, he helped with feedback/discussions during the development of the idea and during the write up. Professor de Vries helped with the improving the text and by giving fruitful feedback/discussions during the development of the idea and during the writing process as well.

Retrospective view

The review and acceptance of this paper in such a reputable venue was overwhelming. However, there is still some room for improvement, for example as one of the reviewers felt, there is a need for a better explanation of the baseline system, which in this paper was based on the fusion of the collection of runs. In addition, concerning the *scalability* issue, we have not discussed which efficiency parameters and measures we had adopted to accomplish the overall retrieval tasks. Some of the scalability concern were, namely: (i) the representation issue of a large matrix of order within the range (2.6 million × 2.6 million) and (ii) to carry out operations on such a large matrix iteratively, with limited resources. Scalability issues in these types of settings should be addressed independently in isolation, not together with the effectiveness studies.

3.3.5 Paper E

Kinship Contextualization:
Utilizing the Preceding and Following Structural Elements.
Muhammad Ali Norozi and Paavo Arvola
Proceedings of the 36th ACM SIGIR conference on
Research and development in Information Retrieval, pages 837–840,
ACM Press 2013.

Abstract: The textual context of an element, *structurally*, contains traces of evidences. Utilizing this context in scoring is called contextualization. In this study we hypothesize that the context of an XML-element originating from its *preceding* and *following* elements in the sequential ordering of a document improves the quality of retrieval. In the tree form of the document's structure, *kinship* contextualization means, contextualization based on the horizontal and vertical elements in the *kinship tree,* or elements in closer to a wider structural kinship. We have tested several variants of kinship contextualization and verified notable improvements in comparison with the baseline system and gold standards in the retrieval of focused elements.

Research process

Paper D provided a competitive baseline which offered a cutting-edge target to measure against. In this study we invested some more efforts to get an even better baseline system or at least a different baseline system which can offer a competitive target. As in the earlier papers, we have used the test collection and topics from INEX, in addition we have also applied their evaluation measures and toolkits. The *sub-tree of interest*, the structural context, is chosen from the elements in the kinship of the relevant information, to be *kinship contextualized.* The hypothesis therefore was, does the *preceding* and the *following* textual content wrapped within the structural boundaries (elements in kinship) improve the retrieval of focused elements? RQ2 and RQ3 were the focal-point research questions in this study.

Roles of the authors

The role of primary author was: the initiator of the idea; setting up the experimental environment; the evaluations and analyses; and writing most of the paper. Dr. Paavo Arvola helped with feedback / discussion during the development of the idea and helped during the writing process.

Retrospective view

The main limitation of this paper is that it is too precise and might be abrupt, and because of the limited space there is not enough description of the background information with sufficient detail. But from the perspective of this thesis, the paper sits well as it is since there is sufficient background material both in the introductory part of this thesis and the papers before it. The reviewers also suggested doing experiments with different types of fusion methods as a baseline. One future extension could be to compare the approach in this paper (graph or structure based) with language-modelling based approaches. These approaches tend to include the context (based on linguistic features, textual and / or semantic, proximity features) into relevance scoring, from the sequential ordering of the documents.

3.3.6 Paper F

When is the Structural Context Effective?
Muhammad Ali Norozi and Paavo Arvola
CEUR Workshop Proceedings.
Proceedings of the 13th Dutch-Belgian Workshop on Information Retrieval;
ACM 2013.

Abstract: Structural context surrounding the relevant information is intuitively and empirically considered important in information retrieval. Utilizing this context in scoring has improved the retrieval effectiveness. In this study we will objectively look into the significance of the *structural context* in contextualization process, and try to answer the core question of under which circumstances do we need to deal with the such types of context?

Research process

The occurrence of the *worst-case scenario*[5] is inevitable within IR settings in general and in semi-structured retrieval in particular. Apart from the practical vulner-

[5] We define worst-case as; a targeted practice of eliciting favourable rankings, by designing document such that the ranking system performs badly.

ability concerns, theoretically it is also wise to study and possibly take preventive actions against malicious use cases when designing new approaches. Along this line of thought, the first thing that has to be defined is the worst-case itself. The definition of worst-case situation in contextualization with structural context could be: "how does the contextualization model behave when the structural context (accidentally or intentionally) becomes the *non-context* (misleading context)?" In this paper, we have taken the non-context as the context of an irrelevant document or element. Subsequently, we have hypothesized that under which circumstances the structural context is helping and / or misleading the retrieval of relevant items. In this paper, the research direction falls under RQ1. The paragraph and document granulation tasks are employed to quantify the experimental behaviour with the worst-case scenario as the primary focus. In addition, query term probabilities are also quantized, to see what the distribution of the any or all query words along the collection is. Apparently, at the document level, the probabilities were found to be higher while at the element or paragraph level, the probabilities were comparatively quite low. From the experimental results, we drew the conclusion that in the worst-case, the contextualization process would get as good results as the basic scoring method—the baseline. Hence, it does not hurt the retrieval effectiveness, even if the structural context is muddled.

Roles of the authors

Same as in Paper E.

Retrospective view

While the idea seemed theoretically, logically and sequentially meaningful (in the holistic picture of the thesis), the experimental evaluations were not considered to be enough. On the contrary, the initial purpose of the paper was more a theoretical approach towards the core question of the effectiveness of the structural context. Nevertheless, as a reflection on this work, a more detailed and thorough empirical analysis with a wider set of implications is required to reinforce the theoretical propositions in this work. However, in the perspective of this thesis, this work supplements the puzzle in the overall picture. The concepts should therefore be easily comprehensible in the context of the other related papers.

3.3.7 Paper G

Selection Fusion in Semi-structured Retrieval.
Muhammad Ali Norozi and Paavo Arvola
In CIKM'13, Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, Burlingame, CA, USA, October 27 – November 01, 2013, ACM Press 2013.

Abstract: Semi-structured retrieval aims at providing focused answers to the user's queries. A successful retrieval experience in semi-structured environment would mean a satisfactory combination of (a) matching or scoring and (b) selection of appropriate and focused fragments of the text. The need to retrieve items of different sizes arises today with users querying the retrieval systems with varied use case, user interface and screen-size requirements. Which means that different selection scenario serve different requirements and constraints. Hence we propose, a novel type of fusion; the *selection fusion*—a fusion methodology which fuses an all-purpose and comprehensive ranking of elements with a specific selection scheme, and also enables evaluation of the ranking in many selection perspectives. With the standard Wikipedia XML test collection, we are able to demonstrate that a strong and competitive baseline ranking system improves retrieval quality irrespective of the selection criteria. Our baseline ranking system is based on data fusion over the official submitted runs at INEX 2009.

Research process

The theoretical flow of ideas in this thesis took us further towards a methodological paper. We strongly felt the need for the introduction of a new methodology in semi-structured retrieval for the following established reasons from the paper:

- It is hard or nearly impossible to design a retrieval strategy (one-size-fit-all) which can target all or different *selection* scenarios.
- Relevance scoring in semi-structured retrieval should be carried out independently of the selection scheme, which might later be able to serve a number of different selection use-case scenarios.

Based on the above brief motivational background, the paper builds on the hypothesis that the selection system and the scoring system are two independent and disjoint system development phases in the semi-structured retrieval system development process. *Selection fusion* methodology is a fusion scheme which brings together these two independent systems into one improved retrieval experience. It can be used in a lot of different use-case scenarios, based on different selection schemes. The paper primarily focusses on the RQ4 question, but the research questions RQ1, RQ2 and RQ3 are also indirectly addressed as well.

Roles of the authors

Same as in Paper E.

Retrospective view

The limitations and shortcoming of the paper are primarily to be attributed to the use of only one test collection. Hence, the methodology could as well be tested with other realistic (or real-world) and representative test collections.

3.4 Summary and Overview

This chapter summarizes the research work done in this study. The tasks were performed in different phases, spread across several research papers. The phases were categorized based on the retrieval tasks performed at each stage. The research outcomes were also later outlined in light of the research objectives and the challenges. We have also described the research process and the retrospective analyses of each of the papers. To summarize, Table 3.2 sketches the overview of the papers in relation to the research questions in Section 1.4, they address.

Table 3.2: *Overview of papers (Table 3.1) times the research questions (Section 1.4).*

Papers	**Research Questions**					
	RQ	RQ1	RQ2	RQ3	RQ4	RQ$_{ef}$
A	×					
B	(×)					×
C			×			
D		×	×	(×)		
E			×	×		
F		×				
G		(×)	(×)	(×)	×	

CHAPTER 4

Concluding Remarks

"If two opposite theories are propagated one will be wrong."

– ALI

LESSONS learnt and the findings from this thesis are summarised in this chapter. Section 4.1 outlines the conclusions grouped by tasks executed during each phase of this study. Section 4.2 gives an overview of the contributions in relation to the research questions and challenges put together earlier in this thesis. Finally, Section 4.3 lays down some future prospects as a direct consequence of this work.

4.1 Conclusions

The coordinated and directed effort in this thesis revolved around the primary research premise expressed in Section 1.4, which is:

> ***How to effectively use the structure within semi-structured documents as evidence in the pursuit of "further" improving retrieval effectiveness?***

Hence, the overall target of this thesis was to discover and unveil the role of the structural evidence (originating from the structural context) in the retrieval of semi-structured items in a schema-agnostic environment. To achieve this, we started with a more general approach and at every subsequent stage of this study we came closer and closer to accomplishing this goal. We began with identifying and recognizing the retrieval opportunities in the graph-based perspective of the *external structural context* of the documents. Later we indulged in a more specific and thorough handling of the structural context, originating both from within and outside the documents. Characteristically, the different stages of research correspond to the standard and tailored retrieval tasks performed at various points in this study, referring to Figure 3.1, grouped by ad hoc, granulations and INEX tasks. Below we provide a treatment of the objectives and contributions of this study from two perspectives: (i) each of the retrieval tasks are revisited, and (ii) the designated research questions from Section 1.4, addressed throughout this study (spread through different research papers).

4.1.1 Retrieval Tasks

Here we align the retrieval tasks with the objectives of this study. In the next sub-section the contributions are discussed in relation to the research questions.

Ad hoc Task

In the ad hoc retrieval task the objectives were a bit general and broad because we worked on that at an early stage of this Ph.D. The aim was to focus on exploring ideas, list the most common strategies for semi-structured retrieval, and hence figure out what could be possible further research directions in this area. The experiments conducted at that stage were mainly evaluated in an ad hoc fashion. The research questions are aligned with the contributions from papers within ad hoc task in the Section 4.1.2.

Granulations

Based on the theoretical and experimental background built in the earlier stage, more focused objectives characterize this stage of the development of the work. The experimental settings were further expanded and hence we were able to deal with a large quantity of data and operating on a wide range of the features within the data. Physically, there are varied amounts of textual content available in different structural constructs within a semi-structured document. It was an intuitive next step, at this stage of the research, to study the impact of the structural constructs, comprising varied textual content sizes, on the retrieval effectiveness. The objectives in this retrieval task were to analyse and materialize the role of the structural constructs and use them in different retrieval settings. The contributions in this custom built task, in relation to the established research questions, are gathered in Section 4.1.2.

INEX

Referring to the hierarchical classification of the work in this thesis in Figure 3.1, the next logical and more focused objective was to perform the retrieval tasks within the framework of the standard benchmarking initiative, INEX. With slightly better background and research maturity, here we engaged in more hardcore and standard focused or element retrieval challenges and objectives. The implication of the work from this task has theoretical, experimental, and methodological significance. The take away messages from this task can also be grouped together by the research questions and papers, in the next subsection.

4.1.2 Research Questions Revisited

The contributions of this study spread across different papers are here grouped together by the research questions they address. Some of the papers address multiple research questions, while most of the research questions are addressed in more than one paper.

RQ: The most common research strategies in semi-structured retrieval, this comes under the overall research goal, as expressed in the earlier section. The following papers have directly or indirectly addressed this research question:

Paper A This paper presents a thorough survey of the area. In addition, it also helps *formulate* the preliminary research questions and also directs us towards possible research opportunities in the area, and hence is a step towards the aforementioned research goal.

Paper B Based on the prospects provided in Paper A, a study was conducted of the applicability of the link analysis ranking algorithms to our research goals. The contributions of the paper were primarily on the efficiency level, however it also uncovers some of the important features of iterative algorithms, which later turned out to be quite crucial.

Paper C *Contextualization* based on the bibliographical structure was a novel retrieval strategy, which characteristically addresses the main research goal of this thesis. The approach was experimentally evaluated in ad hoc settings. It was generic enough to serve other purposes as well, later in this study.

RQ1: What is the role and the significance of the structural context in the ranking of the focused items, and what kind of structural context can be "beneficially" utilized?

Paper D The main contribution of this paper is the discourse around the argument on the relevance and significance of structural context in the retrieval of small and large elements. Both article (large) and paragraph (small) granulation are featured in this paper.

Paper F The contribution of this paper can be associated to a rather negative interpretation of the structural context from RQ1, namely, what is the role of *misleading* structural context on semi-structured retrieval effectiveness? And also theoretically we argued for the usefulness of the structural context; i.e., *when* and under *which* circumstances is structural context beneficial.

Paper G Here we explore the distinctive role of the structural context in coordination with the fusion techniques by providing the capability of a strong and independent relevance scoring system for the semi-structured retrieval system development process.

RQ2: How can we improve the retrieval approaches which make use of the structural context, and subsequently, how should the retrieval effectiveness of those improved strategies be evaluated?

Paper C This paper proposes a model which can exploit both the features in the structural context of documents and the features in the graph-based interpretation of the biblio*graphical* structure of documents for the ad hoc retrieval task. In ad hoc retrieval settings, the purpose of such a model was to understand the role of structure and its empirical significance for retrieval effectiveness.

Paper D The contextualization model together with the random walk principle satisfy the specification of such an approach (RQ2). This approach actively makes use of the structural context. The other novelty of this paper is the competitive baseline system and the evaluation framework based on granulation principles.

Paper E The contribution of this paper is towards building up the structural context from the subtree of interest (obtained as a result of kinship relationships) using the INEX evaluation framework and the element retrieval strategy.

Paper G The structural context can be used as an integral part of the methodology, which enables the retrieval of items on diverse granularity levels. In this paper we have unveiled the role and importance of the structural context in helping to build a strong scoring system.

RQ3: How to improve the retrieval of small elements in focused retrieval?

Paper D This paper shows how the relatively small retrievable items or the items generally *ignored by* the retrieval system have a greater chance to improve the retrieval, primarily with the support of: (a) the evidence accumulated from the structural context and (b) a retrieval model capable of using such information in its relevance scoring mechanisms.

Paper E Kinship contextualization helps in reducing the ‘size bias’ problem and therefore enables semi-structured retrieval systems to effectively retrieve items without being *size* conscious. This finding is based on the experimental and statistical evidence gathered in the paper.

Paper G This paper addressed RQ3 as follows: by proposing (i) a relevance scoring system which delivers a comprehensive and highly relevant set of answers and (ii) selection criteria that allow the selection of small items. In experimental analyses, we were able to retrieve items of all sizes irrespective of the amount of content available within them.

RQ4: How can we effectively utilize the scoring of multiple systems to retrieve focused results at varied granularity levels with good-enough precision (scoring)?

Paper G The primary focus of this paper was to achieve a retrieval strategy capable of retrieving focused items at different and representative granularity levels. The *selection fusion* is a novel and unique methodology which proposes and empirically exhibits such a retrieval experience. Experimental evaluations validated this hypothesis in the paper by systematically performing the four standard and designated tasks and benchmarking framework at INEX.

RQ$_{ef}$: How *efficiently* can we carry out the semi-structured retrieval task?

Paper B The novelty of this work was to make the iterative algorithms perform in a cost-efficient manner. The *extrapolation* technique was the primary contribution. Even though this research question cannot be directly associated with the overall research goal of this thesis, the outcomes have played a fundamental role in the experimental settings and evaluations later in the study.

4.2 Contributions and Overview

Based on the overview (Figure 3.1), the following set of concise and precise contributions can be asserted in outline form, paperwise and oriented to the research questions:

C1: Theoretical background of the area.

C2: Extrapolation techniques and the inclusion of LAR based techniques in semi-structured retrieval. Although this latter proposition was not clearly specified in any of the papers, it is indirectly used in most of the work.

C4: Contextualization using the structural context on the bibliographical structure of documents together with the Markovian random walk principle.

C5: Contextualization of the hierarchical and hyperlink structures of documents in granulation selection and evaluation scenarios.

C6: Worst-case analyses of the use of the structural context and the situation under which structural context might be beneficial.

C7: Selection fusion—a methodology leading to a general semi-structured retrieval system, capable of serving a wide range of use case scenarios.

In light of the contributions outlined above, the conclusions in Section 4.1, and the overview of the papers and research questions in Table 3.2, Table 4.1 draws an overall picture of the thesis.

Table 4.1: *Overview of the thesis. Research questions versus the contributions*

Research questions	Papers	**Contributions**						
		C1	C2	C3	C4	C5	C6	C7
RQ	A,B	×	×					
RQ1	D,F,G				×		×	×
RQ2	C,D,E,G			×	×	×		×
RQ3	D,E,G				×	×		×
RQ4	G							×
RQ$_{ef}$	B,C,D,E,F,G		×	×	×	×	×	×

4.3 Future Work

This study, just like other scientific studies, certainly is not the end of this research topic. Despite the depth and the breadth of the research contributions, there are still quite a few interesting and challenging issues open for further investigation. During the course of this thesis, we have also argued for some of the possible future works as well. In this section, we sketch an overview of the possible future opportunities based on the understanding of the area acquired during the Ph.D. process.

The laboratory settings in this work focused on a non-interactive research style, where there is no direct user involvement. A research on the retrieval usefulness of the structural features of semi-structured documents in an interactive style of evaluation and experiment, with direct user involvement, would be a worthwhile direction for the future. In addition, in [63, 9] we have argued for the importance of the evaluation framework and metrics. If not chosen carefully, the evaluation metrics at times suffer from certain element size or type bias. These "biasedness" might lead to an overall skewness in the top ranked result lists. Here, in particular, a semi-automatic or even an automatic interactive evaluation approach with direct user involvement can be an important and interesting research direction for future work. The existing and available data collections either lack this capability or are not comprehensive and large enough to have wider implications.

In addition, the efficiency question in this study is also only indirectly addressed. The computationally and spatially expensive operations might be the utmost performance barrier in the real world scenario if they are not handled with sufficient care and technical understanding. Computationally, in this study the query throughput and the time for indexing are not sufficient or practical enough to be used in an interactive and real-time setting. Most of the operations are performed asynchronously and offline. Moreover, spatially the processes were found to be memory intensive. For example, we needed a large enough main memory and disk space to do the operations on large-scale matrices. The experiments in this study were run using hexa-core, 3.2Ghz processor with 12GB of main memory and 2TB of disk space. Running the experiments on a computer with a lower specification was not possible. Hence, the operations were generally found to be computationally and spatially expensive. A space–time trade-off decision is required here, based on the available resources, use cases, and real world situations. Therefore, a possible line of future research lies in the bottleneck efforts to study the computational-cost and/or -benefits of these kinds of approaches towards semi-structured retrieval. The core research question here would be: how to cost effectively perform or improve the performance of these expensive approaches? Maybe treat them in a distributed and/or parallel scalable setting.

In the retrieval tasks, such as the granulation task, experimenting with other types or sizes of granularity would be possible. As argued earlier in the thesis, hybrid contextualization could be further developed and evaluated to be used as a possible organizational or retrieval strategy within the in-context related tasks at INEX.

Finally, one of the core challenges in semi-structured information retrieval is the problem and challenge of setting up a laboratory environment suitable for the research goals that need to be achieved. A state-of-the-art experimental environment which can offer: (a) efficient evaluations, (b) customizability, and (c) suitability for generic research goals in semi-structured IR, is a necessary research tool. And in the same line of thought, lies the creation or selection of a representative documents corpora, query topics and real world retrieval scenarios and / or challenges. A test collection should be (a) representative, (b) widely used and reported, and (c) have relevance assessments. Therefore a test collection employed for research that ignores any of these properties might exhibit and support misleading or even false results and hypotheses in experimental evaluations.

Bibliography

[1] Gartner Survey Shows XML Usage Reaches 86 Percent in Systems Integration Projects Using Web Services. `http://www.information-management.com/news/6699-1.html`, May 2003.

[2] S. Abiteboul. Querying semi-structured data. *Database Theory-ICDT'97*, pages 1–18, 1997.

[3] S. Abiteboul. *Querying semi-structured data.* Springer, 1997.

[4] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: from relations to semistructured data and XML.* Morgan Kaufmann Pub, 2000.

[5] J. R. Anderson and P. L. Pirolli. Spread of activation. *Journal of experimental psychology. Learning, memory, and cognition*, 10(4):791–798, 1984.

[6] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 601–610. ACM, 2009.

[7] 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.

[8] 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.

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

[10] P. Arvola. *The Role of Context in Matching and Evaluation of XML Information Retrieval.* PhD thesis, University of Tampere., 2011.

[11] P. Arvola, J. Kekäläinen, and M. Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13(5):460–484, 2010.

[12] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval.* Addison-Wesley Harlow, England, 1999.

[13] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-based keyword search in databases. In *Proceedings of the Thirtieth interna-*

tional conference on Very Large DataBases-Volume 30, pages 564–575. VLDB Endowment, 2004.

[14] D. Bamman, A. Babeu, and G. Crane. Transferring structural markup across translations using multilingual alignment and projection. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 11–20. ACM, 2010.

[15] M. Barg and R. K. Wong. Structural proximity searching for large collections of semi-structured data. In *Proceedings of the tenth international conference on Information and knowledge management (CIKM)*, pages 175–182. ACM, 2001.

[16] S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie, J. Siméon, and M. Stefanescu. XQuery 1.0: An XML query language. *W3C working draft*, 12, 2003.

[17] P. Borlund. The concept of relevance in IR. *Journal of the American Society for information Science and Technology*, 54(10):913–925, 2003.

[18] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, 2005.

[19] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer. Searching XML documents via XML fragments. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 151–158. ACM, 2003.

[20] V. Christophides, D. Plexousakis, M. Scholl, and S. Tourtounis. On labeling schemes for the semantic web. In *Proceedings of the 12th international conference on World Wide Web*, pages 544–555. ACM, 2003.

[21] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv. XSEarch: A semantic search engine for XML. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, page 56. VLDB Endowment, 2003.

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

[23] J. Cox. Survey Showing XML use Growing Fast in Enterprises. `http://www.networkworld.com/news/2001/0226xml.html`, February 2001.

[24] 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.

[25] L. Denoyer and P. Gallinari. The Wikipedia XML corpus. In *ACM SIGIR Forum*, volume 40, pages 64–69. ACM, 2006.

[26] M. Duong and Y. Zhang. LSDX: a new labelling scheme for dynamically updating XML data. In *Proceedings of the 16th Australasian database conference-Volume 39*, pages 185–193. Australian Computer Society, Inc., 2005.

[27] R. Elmasri and S. Navathe. *Fundamentals of database systems*. Addison Wesley, 2009.

[28] N. Fuhr and K. Großjohann. XIRQL: A query language for information retrieval in XML documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 172–180. ACM, 2001.

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

[30] S. Geva, J. Kamps, M. Lethonen, R. Schenkel, J. A. Thom, and A. Trotman. Overview of the INEX 2009 ad-hoc track. In *Focused Retrieval and Evaluation*, pages 4–25. Springer, 2010.

[31] G. Gou and R. Chirkova. Efficiently Querying Large XML Data Repositories: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(10):1381–1403, October 2007.

[32] N. Gövert and G. Kazai. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In *INEX Workshop*, pages 1–17. Citeseer, 2002.

[33] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked keyword search over XML documents. In *Proceedings of the 29th ACM SIGMOD international conference on Management of data*, page 27. ACM, 2003.

[34] iAD Team. iAD Information Access Disruptions Centre. `http://www.iad-centre.no/about.html`.

[35] P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer, 2005.

[36] K. Itakura and C. Clarke. A framework for BM25F-based XML retrieval. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 843–844. ACM, 2010.

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

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

[39] J. Kamps, M. De Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 80–87. ACM, 2004.

[40] J. Kamps, M. Koolen, and M. Lalmas. Locating relevant text within XML documents. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 847–848. ACM, 2008.

[41] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 evaluation measures. In *Focused access to XML documents*, pages 24–33. Springer, 2008.

[42] G. Kazai, M. Lalmas, and A. P. de Vries. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 72–79. ACM, 2004.

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

[44] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[45] J. Kim, X. Xue, and W. B. Croft. A Probabilistic Retrieval Model for Semistructured Data. *Collections*, pages 228–239, 2009.

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

[47] M. H. A. Koolen. *The meaning of structure: the value of link evidence for information retrieval*. PhD thesis, SIKS, the Dutch Research School for Information and Knowledge Systems., 2011.

[48] M. Lalmas. XML Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–111, January 2009.

[49] A. Langville and C. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[50] D. Lay. *Linear algebra and its applications*. Addison-Wesley Reading, Mass, 1994.

[51] M. Lehtonen, N. Pharo, and A. Trotman. A Taxonomy for XML retrieval use cases. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 413–422. Springer, 2007.

[52] Q. Li and B. Moon. Indexing and querying XML data for regular path expressions. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 361–370, 2001.

[53] J. List, V. Mihajlović, G. Ramírez, A. P. de Vries, D. Hiemstra, and H. E. Blok. TIJAH: Embracing IR methods in XML databases. *Information Retrieval*, 8(4):547–570, 2005.

[54] D. Liu, C. Wan, L. Chen, and X. Liu. Automatically weighting tags in XML collection. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1289–1292. ACM, 2010.

[55] W. Lu, S. Robertson, and A. MacFarlane. Field-weighted XML retrieval based on BM25. *Advances in XML Information Retrieval and Evaluation. Fourth*

Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), pages 161–171, 2006.

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

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

[58] V. Mihajlović, G. Ramírez, T. Westerveld, D. Hiemstra, H. E. Blok, and A. P. de Vries. TIJAH scratches INEX 2005: Vague element selection, image search, overlap, and relevance feedback. In *Advances in XML Information Retrieval and Evaluation. Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, pages 72–87. Springer, 2006.

[59] M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the 11th international conference on Information and knowledge management*, pages 538–548. ACM, 2002.

[60] M. A. Norozi. Extrapolation to speed-up query-dependent link analysis ranking algorithms. In *Proceedings of the 8th International Conference on Frontiers of Information Technology - FIT '10*, page 2. ACM, 2010.

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

[62] M. A. Norozi. Relevancy in Schema Agnostic Environment. *Bulletin of IEEE Technical Committee on Digital Libraries*, 8(1):1–6, 2011.

[63] M. A. Norozi and P. Arvola. Kinship Contextualization: Utilizing the Preceding and Following Structural Elements. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 837–840. ACM, 2013.

[64] M. A. Norozi and P. Arvola. Selection Fusion in Semi-structured Retrieval. In *Proceedings of the 22nd ACM international conference on Information and knowledge management (CIKM)*. ACM, 2013.

[65] M. A. Norozi and P. Arvola. When is the Structural Context Effective? In *Proceedings of the 13th Dutch-Belgian Information Retrieval workshop (DIR)*, 2013.

[66] M. A. Norozi, P. Arvola, and A. P. de Vries. Contextualization using hyperlinks and internal hierarchical structure of Wikipedia documents. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM)*, pages 734–743. ACM, 2012.

[67] M. A. Norozi, A. P. de Vries, and P. Arvola. Contextualization from the Bibliographic Structure. In *Proceedings of the Task-Based and Aggregated Search workshop (TBAS2012), in conjunction with ECIR 2012*, pages 9–13. ACM, 2012.

[68] P. Ogilvie and J. Callan. Using language models for flat text queries in XML retrieval. In *Proceedings of INEX 2003 Workshop, Dagstuhl, Germany*, pages 12–18, 2003.

[69] P. Ogilvie and J. Callan. Hierarchical language models for XML component retrieval. In *Advances in XML Information Retrieval*, pages 224–237. Springer, 2005.

[70] P. Ogilvie and J. Callan. Parameter estimation for a simple hierarchical generative model for XML retrieval. In *Advances in XML Information Retrieval and Evaluation*, pages 211–224. Springer, 2006.

[71] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[72] D. Petkova, W. B. Croft, and Y. Diao. Refining Keyword Queries for XML Retrieval by Combining Content and Structure The Problem of Adding Structure to Keyword Queries. pages 662–669, 2009.

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

[74] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM, 2004.

[75] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58. ACM, 1993.

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

[77] T. Schlieder and H. Meuss. Querying and ranking XML documents. *Journal of the American Society for Information Science and Technology*, 53(6):489–503, 2002.

[78] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The 2nd TREC*. Citeseer, 1994.

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

[80] B. Sigurbjörnsson and J. Kamps. The effect of structured queries and selective indexing on XML retrieval. In *Advances in XML Information Retrieval and Evaluation. Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, pages 104–118. Springer, 2006.

[81] A. Theobald and G. Weikum. Adding relevance to XML. *The World Wide Web and Databases*, pages 105–124, 2001.

[82] A. Theobald and G. Weikum. The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking. *Advances in Database Technology – EDBT 2002*, pages 311–340, 2002.

[83] M. Theobald, H. Bast, D. Majumdar, R. Schenkel, and G. Weikum. TopX: efficient and versatile top-k query processing for semistructured data. *The VLDB Journal*, 17(1):81–115, 2008.

[84] A. Trotman, M. del Rocio Gomez Crisostomo, and M. Lalmas. Visualizing the Problems with the INEX Topics. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 826–826. ACM, 2009.

[85] A. Trotman. Choosing document structure weights. *Information processing & management*, 41(2):243–264, 2005.

[86] A. Trotman, S. Geva, and J. Kamps. Report on the SIGIR 2007 workshop on focused retrieval. In *ACM SIGIR Forum*, volume 41, pages 97–103. ACM, 2007.

[87] A. Trotman and R. A. O'Keefe. Identifying and ranking relevant document elements. In *INEX 2003 Workshop Proceedings*, pages 149–154, 2003.

[88] A. Trotman, N. Pharo, and M. Lehtonen. XML-IR users and use cases. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 400–412. Springer, 2007.

[89] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In *Advances in XML Information Retrieval*, pages 16–40. Springer, 2005.

[90] P. Tsaparas. *Link Analysis Ranking*. PhD thesis, University of Toronto, 2004.

[91] P. Tsaparas. Using non-linear dynamical systems for web searching and ranking. *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 59–70, 2004.

[92] E. Voorhees, D. K. Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

[93] J. Wang and J. X. Yu. Answering Tree Pattern Queries Using Views : a Revisit. In *Proceedings of the 14th International Conference on Extending Database Technology*, page 4. ACM, 2011.

[94] J. Wolff, H. Florke, and A. Cremers. XPRES: A ranking approach to retrieval on structured documents. Technical report, Citeseer, 1999.

[95] J. Wolff, H. Florke, and A. Cremers. Searching and browsing collections of structural information. In *adl*, page 141. Published by the IEEE Computer Society, 2000.

Part II

Selected Papers

"Contentment is the capital
which will never diminish."

– ALI

This part contains the papers that constitute the main body of the thesis. There are seven papers included, titled alphabetically as Papers A–G. All of the papers have been accepted and published in peer-reviewed international venues. The format of the text, tables and figures has been altered from the original publisher, in order to improve readability.

PAPER A

Relevancy in Schema Agnostic Environment

Muhammad Ali Norozi.

Appeared at the ACM/IEEE Joint Conference on Digital Libraries, Ottawa, Canada, 2011

Abstract: *Relevance* is an important component in full-text search and often distinguishes the implementations. Relevancy is used to score matching documents and rank them according to the users intent. One of the reasons of the high popularity of Google is its good relevancy originally based on the PageRank algorithm.

The emergence of *semi-structured* data as a standard for data representation opened up new areas which could be related to both the database and information retrieval communities. Although the information retrieval and database viewpoints were, until quite recently irreconcilable, semi-structured retrieval helped to bridge the gap. This work is about exploring relevancy in semi-structured retrieval both in isolation and as bridge between database and information retrieval communities.

A.1 Introduction

Users on the web are expanding from being active information *consumers* to becoming active information *producers* [18, 17], which leads to an unprecedented growth of information. With such a boom in information retrieval and the information explosion, there is an ever-increasing demand for accessibility, coverage, quick responses, and relevant results from relatively vague and loose queries. The huge collection of information inherently entails a loss of performance and efficiency, because it takes time to process (index, cluster, etc), retrieve (query) and keep the information up-to-date in the huge repositories. Thus there is a growing concern about the usability and the interaction time between the user and Information Retrieval (IR) systems. A trade-off between the quality of the results and query response time is mostly considered as an option. In such challenging settings a user query must yield meaningful, manageable and most importantly "*relevant*" set of results from IR systems.

A central topic in the iAD (Information Access Disruptions [10]) project is to index data collections where there is a big variety of data, both unstructured data (text documents), structured data (e.g. database records) and semi-structured data (e.g. XML or HTML data), all present at the same time. In such a setting we want to focus on *relevancy* in a *schema agnostic systems*. Here "schema agnostic" only means that the queries need not to use the schema information, while the search engine can possibly use the schema (meta)-/information to come up with a set of relevant documents. Relevancy calculations should take into account structure information whenever it exists and use this to improve ranking of results without or with minimal prior knowledge of the schema of the data. Schema can be automatically recognized by analysing data sources or the data itself. Important research questions are to find metrics which are giving a perceived better result for the end user and how to score various data with schema relative to each other. Variance and resulting inaccuracies in the document structures, vocabulary and document content dictate ranked retrieval as the only meaningful search paradigm.

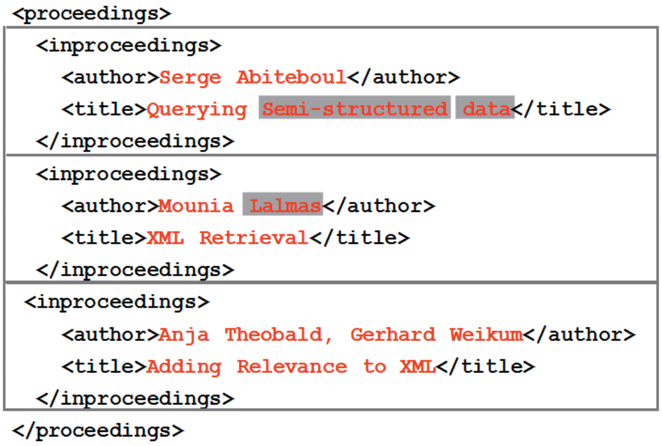
To calculate relevancy many different metrics are being used. Examples on metrics are $tf \times idf$ score [2, 17], static weight for document, proximity of terms, freshness of document. The metrics are weighted relative to each other and combined into a final score. Different users can have different weightings based on their preferences. Structured and semi-structured documents will make it possible to find new metric such as, weighting of scopes and coexistence of terms in scopes.

In a collection of documents of different types it is difficult to combine the scores. In a document with no structure there will be no score from structure metrics. These documents therefore might get less score than documents with structure. To compensate for this we can boost score for these documents but how much should this be boosted? Can we use the inherent structure of document (extracted automatically or semi-automatically) to increase the visibility of contents and concepts ingrained in the document?

As in legacy search engines (library systems or digital libraries) and web search, schema agnostic search environment must also define a measure of relevance or merit for each response. In such an environment (schema agnostic) relevancy subsystem must efficiently generate only a few responses that have the greatest relevance scores in that particular setting.

Different ways of calculating relevancy scores will be researched in schema agnostic environment (XML or semi-structured dataset) as a main focus of this study.

A.2 Motivation



```
<proceedings>
  <inproceedings>
    <author>Serge Abiteboul</author>
    <title>Querying Semi-structured data</title>
  </inproceedings>
  <inproceedings>
    <author>Mounia Lalmas</author>
    <title>XML Retrieval</title>
  </inproceedings>
 <inproceedings>
    <author>Anja Theobald, Gerhard Weikum</author>
    <title>Adding Relevance to XML</title>
  </inproceedings>
</proceedings>
```

Figure A.1: *Find papers by author "Lalmas" on the topic of "semi-structured data"*

It is becoming increasingly popular to publish data on the Web in the form of semi-structured documents which is useful both for data exchange and data semantics. The representation in Figure A.1 if retrieved using the existing Text-based or conventional search engine based on traditional IR techniques, have two main drawbacks when it comes to searching for semi-structured documents:

- It is not possible to pose queries that explicitly refer to structure of the documents, e.g., twig like queries.
- Search engines return references (i.e. links) to documents and not specific fragments thereof. This is problematic, since large semi-structured documents may contain thousands of elements storing many pieces of information that are not necessarily related to each other.

Since a reference to whole semi-structured document is usually not a useful answer, the granularity of the search should be refined. The concept of the *logical document* [23] instead of just document comes here; the users are now not interested in the document but the most specific part of the document i.e., the logical document.

Figure A.1 shows an example scenario, the overall document will be returned as relevant by a text-based search engine. But the document is not relevant to the posed query. In this query the user is interested in the documents authored by "Lalmas". The above document should not be retrieved at all by a structure-aware search engine. The set of answers in the set of retrieved documents should be semantically related, i.e., the set of the answer nodes are meaningful fragment of the semi-structured documents. For example, a paper and an author should be in the answer set only if the paper was written by this author.

The retrieval must not only return the most specific part of the documents but also it should take into account the degree of *relevance* of the retrieved document fragments with the posed query. And based on that the documents should appear in the ranked outcomes. The document in Figure A.1 should appear lower down in the ranked outcomes for the given query.

Structure provides both *context* and *semantics* to the content as seen from the motivation scenario discussed above. This context and semantics should possibly be used to boost or reduce the documents relevancy scores. And without using the structural information, the search outcomes would simply be irrelevant or misleading to the posed queries.

Secondly, from the structure the importance of content in different parts of document could be learned. Text or set of keywords lying in body of a document could be less important than keywords lying in the title.

A.3 State of the art

Semi-structured retrieval research is an interdisciplinary field of study. Both its inception and its implication crosses traditional boundaries from information retrieval community [13, 21, 12] to relational databases [8, 28] and at the same time having a wide range implications across digital libraries communities [4]. In the following, we provide an overview of existing approaches towards relevancy in semi-structured data or schema-agnostic environment, and an overview of why they are not significant enough to answer the research questions raised (Section A.5) in this study.

Broadly, there are two major approaches towards semi-structured retrieval problem. Hence, we look into the existing work from these two perspectives:

- The relational approach towards semi-structured retrieval
- The native approach - conventional information retrieval approach

The relational approaches tend to make use of techniques from already mature area, the relational databases. Instead of considering the semi-structured retrieval problem in isolation, it is considered in the relational databases space. The benefit of adopting such an approach is that you don't have to reinvent the wheels. There are tools and techniques which have been in use for years and a lot of work have already been done on them over the years. So the answer to the semi-structured retrieval is to mold it in the relational databases space, and hence by doing that we retain the efficiency and strength of an already mature area [8]. Specifically, the relational approach directly utilizes relational databases to represent and retrieve semi-structured data, which enables to use all important capabilities of relational databases. XPath and XQuery, developed by W3C Consortium inspired by relational approaches but not necessarily using the capabilities of relational databases, address the problem of semi-structured retrieval. They are not suitable for our purpose mainly because (a) they require a thorough knowledge of schema or structure beforehand (b) complicated to translate the user query into an XQuery (c) syntax of XQuery is by far more complicated than syntax of standard IR system and (d) nominal mechanism for ranking.

Apart from limitations discussed before, it is not always as easy to adopt to an existing and mature framework. The inherent peculiarities of the semi-structured retrieval prohibits making use of some of the main strengths of the relational databases. The saving due to significantly reducing system re-engineering costs in the semi-structured environment is less than reinventing the wheel in the specialized storage and query processing systems tailored for semi-structured settings. Hence the native approach is to consider the semi-structured retrieval in its own particular settings, from scratch to further improve semi-structured retrieval.

XRANK [9] proposed by Guo et.al generalizes the idea initially proposed by Page and Brin [19]. Like Google's PageRank, XRank consider the dataset as a Tree or Graph (see Figure A.2). Unlike PageRank which consider one-size fit all approach, XRANK advocates that the data tree has different type of edges namely containment edge and hyperlink edge. Random Surfer in the XRANK instead of following just the hyperlinks also visits the containment edges (CE) (elements, sub-elements), hyperlink-edges (HE) and reverse containment edges (CE^{-1}) (sub-elements, elements). Like PageRank, XRANK is also calculated offline independent of any query. Equation A.1 taken from [9] summarizes the random surfer model of XRANK.

$$\begin{aligned} e(v) &= \frac{1-d_1-d_2-d_3}{N_d \times N_{de}(v)} + d_1 \sum_{(u,v)\in HE} \frac{e(u)}{N_h(u)} \\ & \quad + d_2 \sum_{(u,v)\in CE} \frac{e(u)}{N_c(u)} + d_3 \sum_{(u,v)\in CE^{-1}} e(u) \qquad \text{(A.1)} \\ & \quad e(v) \text{ is ElemRank of } v \end{aligned}$$

Structural information is mainly used in the calculation of the *ElemRank*, and to answer the queries that have structural dependencies (i.e., structural queries). A more in-depth study is required to observe experimentally the effects of structure on the quality of *relevancy* calculation. In case of PageRank, it was claimed based on intuition and based on the theoretical interpretation of *Markov chain model*, to mimic the users' behaviour on the Web, but in case of XRANK, the random surfer model given in Equation A.1, also resembles users' behaviour? There still is room to explore the structural benefits ingrained in the semi-structured retrieval, a more active use of structure throughout retrieval processes.

Just like XRANK, ObjectRank [3] inspired by Google's PageRank, more actively utilizes the link structure of the semi-structured data. It calculates both global (PageRank) and keyword-specific ObjectRank of each node in the *authority transfer schema / data graph*. Unlike XRANK, ObjectRank is a relational approach and applicable only on Databases.

The work in XRANK, ObjectRank and other graph based methods correspond to the study of Link Analysis Ranking [5, 17, 27]. The motivation is based on intuition as a semi-structured document form a tree structure in most of the cases and a complete graph with cycles in specific cases (as can be seen in Figure A.2).

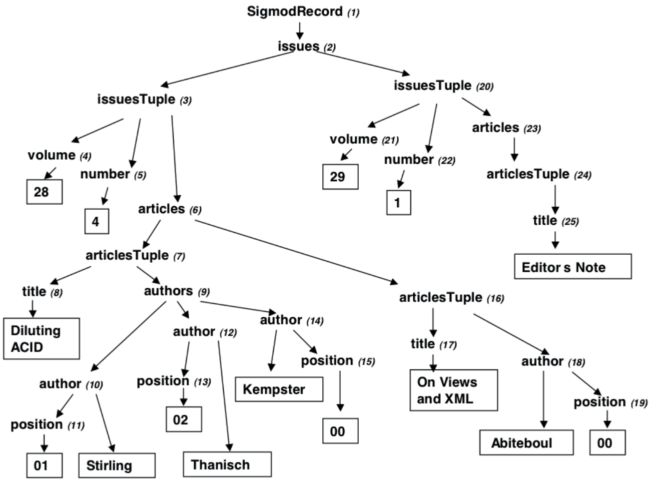


Figure A.2: *Semi-structured data, XML, is represented mostly as a Tree structure (Figure taken from [6])*

XSEarch [6] on the other hand instead of using the link structure of the Tree representation of the semi-structured data, uses extended *Vector Space Model* for retrieval and ranking, and the same kind of approach was employed by Schlieder and Meuss [23]. They make use of the *interconnection relationship* among the XML elements to use the structure in retrieval and ranking. By doing that, they tried to answer the question that under what conditions elements of a XML document are semantically related.

Again, relevancy scoring is not effected directly by the structural constraints. Rather the structural elements gets user-defined weights (manual process), instead of structural elements lying in the heart of relevancy scoring. Under what conditions the semantic constructs in the document, i.e., the structural elements could be used automatically or semi-automatically to purify the relevancy scoring?

A variation of $tf \times idf$ is used for relevancy scoring in XSEarch, where tf correspond to number of occurrences of a query term in a fragment and *Inverse Leaf Frequency* ilf: number of leaves containing a query term divided by number of leaves in the corpus (the data tree), see Equation A.2. The $tf \times ilf$ score together with the interconnection relationship measure (calculated based on how close the elements are in the relationship tree) are used to determine the ranking of the answer.

$$\begin{aligned} tf(k, n_l) &:= \frac{occ(k, n_l)}{max\{occ(k', n_l)|k' \in words(n_l)\}} \\ ilf(k) &:= log\left(1 + \frac{|N|}{\{|n' \in N|k \in words(n')|\}}\right) \end{aligned} \tag{A.2}$$

XXL [24, 25] was mainly proposed to introduce active support for ranked retrieval. In addition, ontological information or relationship has also been integrated as a basis for effective similarity search. In the same line XPRES [29, 30] extends the classical probabilistic model, that exploits the semantic of different text part given in semi-structured document. Like XSEearch, XPRES extends the classical weighting measure $tf \times idf$ and call it $tf \times ief$ (*ief: Inverse Element Frequency*).

BM25F-based [22, 16] XML retrieval has recently been introduced to score individual XML elements [11]. In this approach each XML element is scored as if it were an independent document. This method ignores hierarchy i.e., the parent-child relationships (which usually contains the contextual information), but rather focuses on the elements independently.

To sum up, this section has presented some of the methods that employ structure in the document to somehow improve or purify the retrieval. They have formed a good basis and background for this study and at the same time provided the prospects for possible future work. We believe that a more active and exclusive use of structure in the semi-structured documents would be a worthy contribution in the field of semi-structured retrieval.

A.4 Issues and challenges

The main challenge in the schema-agnostic environment (as discussed from existing work) is that: how the implicit and explicit structure of the document helps to improve the semantics of the retrieval, i.e., improved relevancy? The other challenges as also identified by [1, 3, 15] are:

- The structure is irregular, inconsistent and possibly inaccurate, the same piece of information can be structured in different ways.
- The structure is implicit and is part of the data.
- The schema could be very large. And it keeps on evolving rapidly, and hence the distinction between the schema and data keeps on blurring.

- Differentiating between semantically meaningful constructs and semantically meaningless.
- The two dimensional view of the proximity.
    1. Result specificity: more specific results higher than less specific results. One dimension of result proximity.
    2. Keyword proximity: another dimension of result proximity.
- Users require the most specific answer (part of the document only) instead of the whole document as the answer.
- What constitute an *indexing unit*?
- Partial matching elements that do not meet the structural constraints perfectly should be ranked lower and should not be omitted from search outcomes.

In this context, returning a set of *relevant* and notion of ranking at the finest granularity of semi-structured documents (e.g., in case of XML, it is XML element), is a challenging task in itself. Few of the above challenges have already been addressed in existing work as identified in the Section A.3, but there combination as a whole would be interesting future work and core of this study. From these challenges and issues we have formulated the research questions for this research.

A.5 Research Questions

The overall research can be stated as a number of research questions:

1. How semantics in the document, i.e., the structure, could possibly be used to understand the content in the document and possibly use it to improve retrieval?
2. How should the structure extracted from the semi-structured document be represented? Which type of index structures provide better or worse results?
3. The support for full-text (keyword) and structured query, using the structure to boost the relevancy scores of the documents in either cases.
4. How would the proximity be impacted by the structure in the documents? Does the parent-child relationship add to the conventional proximity measure?
5. How to accommodate variation in the data, as distinction between the schema and data is getting blurred?

The requirements and challenges described in the previous sections are represented in the research questions above. These questions are based on current knowledge of the area, hence it could as well be extended or further purified later based on increased understanding of the subject area.

A.6 Expected Results and Contributions

We would like to use capabilities of different algorithms utilizing different index structures in isolation and together with one another to see their impact on the overall retrieval in general and ranking in particular. At the moment, we are in the processs of implementing different index structures (Dewey inverted index [9] and its different flavours) and use them together with the existing state-of-art methods for example, XRANK's ranking algorithm (*ElemRank*, see Equation A.1) and index structures (interconnection index) used in XSearch system. And by doing that we would like to measure the impact of different indexing schemes on the search outcomes. We would evaluate the effectiveness of our approach with the evaluation metrics and standard datasets from the INitiative for the Evaluation of XML Retrieval (INEX) [7, 26].

One of the previous contributions by the author [18] together with XRANK and / or ObjectRank could be a valuable contribution. As identified in Section A.3 XSEarch lacks the capability to automatically weight the XML elements, together with the recent work by Liu et. al [14] could possibly be a positive contribution to XSEarch system. $BM25F$ could also possibly be improved by incorporating structural construct in the algorithm. At the moment $BM25F$ scores XML elements independently without considering the context surrounding it, i.e., the structure.

A.7 Goals achieved so far and further plan

As the research candidate is quite early stage of the study, the most of work performed so far is around specifying the research questions, planing, taking courses, performing literature review and getting a field overview. It is expected to write a self-contained search engine from scratch, or maybe customize some of the open-source solutions available such as, *Lucene* [1] / *Solr*[2] or maybe *Nutch*[3]. Alternatively, *Terrier*[4] search engine could also be customized to fit the purpose of this study.

According to the plan and after the experimental setup, the candidate is suppose to start testing the initial ideas described in the last section and proceed with the research question 1 (Section A.5).

A.8 The Methodology

In this section we discuss the research design and methodology and its appropriateness for this study. The process of collecting, recording, and analysing data is quite

[1] The Apache *Lucene*^{TM} project: http://lucene.apache.org/

[2] http://lucene.apache.org/solr/

[3] http://lucene.apache.org/nutch/

[4] http://terrier.org/

crucial in this study, because of its innovative and technical nature. An account of the assumptions of the study have also been discussed briefly.

A.8.1 The research paradigm and its rationale

The study will be conducted within the quantitative paradigm. There are two major reasons for selecting the quantitative paradigm: firstly, this research demands an in-depth study. Secondly, we can explore the problem in different experimental settings possibly using the standard practices, metrics and evaluation framework from the state of the art, INEX, TREC and the likes.

A.8.2 Experimental Research

Experimental research is the demand of my research topic primarily because it is a collection of research designs which use the manipulation and controlled testing to understand causal processes. Generally one or more variables and heuristics are manipulated to determine the effect on a dependent variable, which could be thresholds, performance and throughput bottlenecks. Thus Experimental research is a systematic and scientific approach to research in which the researcher manipulates one or more variables, and controls and measures any change in other variables [20].

Generally the experimental research is used when:

- There is a time and performance priority in the causal relationships.
- There is a consistency in a causal relationship.
- The magnitude of the correlation is great.

In semi-structured or schema agnostic retrieval environment we will actively monitor the influence of different approaches towards retrieval in the state-of-the-art settings. Through the experimental research, we intend to empirically evaluate the feasibility of the different approaches chosen as described in previous sections, for the problem of semi-structured retrieval. Both the relational and the native approaches as described in Section A.3 towards semi-structured retrieval will be experimentally compared and parametrized using state of the art evaluation techniques mainly relying on the metrics and dataset from INEX [7, 26].

A.9 Conclusion and Implications

In this paper an overview of the problems in the semi-structured retrieval or schema-agnostic search has been presented. In the state-of-the-art section a number of possible solutions have been discussed, along with their shortcomings. Given the existing work, we still think that there is a lot that need to be done in the semi-structured retrieval, specifically there is not much done in the ranking or relevancy in schema-agnostic environment. How to actively employ the structure in the relevancy subsystem? There is still enough room for introduction, innovation and improvements in the ranking of semi-structured dataset. We think that the research outcomes from this study will be quite beneficial in future. The preliminary research questions proposed aimed to develop a combination of methods to better answer the full-text and structured queries to semi-structured data.

The contributions from this study will be applicable to a wide variety of areas. For example the research on entity extraction or entity search is a direct application of this study. Also in multimedia retrieval mostly the documents are represented in semi-structured form and searching through myriad of them is a contemporary requirement and future need. In natural language processing it is usually worth to extract and use the latent-structures in the documents in order to detect important objects or features.

This Ph.D. study is expected to finish by August 2013, with a Ph.D. defence in October / November 2013.

A.10 References

[1] S. Abiteboul. Querying semi-structured data. *Database Theory-ICDT'97*, pages 1–18, 1997.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*. Addison-Wesley Harlow, England, 1999.

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

[4] D. Bamman, A. Babeu, and G. Crane. Transferring structural markup across translations using multilingual alignment and projection. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 11–20. ACM, 2010.

[5] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, 2005.

[6] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv. XSEarch: A semantic search engine for XML. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, page 56. VLDB Endowment, 2003.

[7] L. Denoyer and P. Gallinari. The wikipedia xml corpus. In *ACM SIGIR Forum*, volume 40, pages 64–69. ACM, 2006.

[8] G. Gou and R. Chirkova. Efficiently Querying Large XML Data Repositories: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(10):1381–1403, Oct. 2007.

[9] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram. XRANK: Ranked keyword search over XML documents. In *Proceedings of the 29th ACM SIGMOD international conference on Management of data*, page 27. ACM, 2003.

[10] iAD Team. iad information access disruptions centre. http://www.iad-centre.no/about.html.

[11] K. Itakura and C. Clarke. A framework for BM25F-based XML retrieval. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 843–844. ACM, 2010.

[12] J. Kim, X. Xue, and W. B. Croft. A Probabilistic Retrieval Model for Semistructured Data. *Collections*, pages 228–239, 2009.

[13] M. Lalmas. XML Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–111, Jan. 2009.

[14] D. Liu, C. Wan, L. Chen, and X. Liu. Automatically weighting tags in XML collection. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1289–1292. ACM, 2010.

[15] S. Liu, Q. Zou, and W. Chu. Configurable indexing and ranking for XML information retrieval. *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 88, 2004.

[16] W. Lu, S. Robertson, and A. MacFarlane. Field-weighted XML retrieval based on BM25. *Advances in XML Information Retrieval and Evaluation*, pages 161–171, 2006.

[17] M. A. Norozi. Information Retrieval Models and Relevancy Ranking. Master's thesis, Centre of Mathematics for Application, University of Oslo, 2008.

[18] M. A. Norozi. Extrapolation to speed-up query-dependent link analysis ranking algorithms. In *Proceedings of the 8th International Conference on Frontiers of Information Technology - FIT '10*, page 2. ACM, 2010.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1998.

[20] K. Peffers, T. Tuunanen, C. Gengler, M. Rossi, W. Hui, V. Virtanen, and J. Bragge. The Design Science research process: a model for producing and presenting information systems research. In *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006)*, pages 83–106, 2006.

[21] D. Petkova, W. B. Croft, and Y. Diao. Refining Keyword Queries for XML Retrieval by Combining Content and Structure The Problem of Adding Structure to Keyword Queries. pages 662–669, 2009.

[22] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM, 2004.

[23] T. Schlieder and H. Meuss. Querying and ranking XML documents. *Journal of the American Society for Information Science and Technology*, 53(6):489–503, 2002.

[24] A. Theobald and G. Weikum. Adding relevance to XML. *The World Wide Web and Databases*, pages 105–124, 2001.

[25] A. Theobald and G. Weikum. The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking. *Advances in Database TechnologyâĂŤEDBT 2002*, pages 311–340, 2002.

[26] A. Trotman, M. del Rocio Gomez Crisostomo, and M. Lalmas. Visualizing the Problems with the INEX Topics. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 826–826. ACM, 2009.

[27] P. Tsaparas. *Link Analysis Ranking*. PhD thesis, University of Toronto, 2004.

[28] J. Wang and J. X. Yu. Answering Tree Pattern Queries Using Views : a Revisit. In *Proceedings of the 14th International Conference on Extending Database Technology*, page 4. ACM, 2011.

[29] J. Wolff, H. Florke, and A. Cremers. XPRES: A ranking approach to retrieval on structured documents. Technical report, Citeseer, 1999.

[30] J. Wolff, H. Florke, and A. Cremers. Searching and browsing collections of structural information. In *adl*, page 141. Published by the IEEE Computer Society, 2000.

PAPER B

Faster ranking using Extrapolation techniques

Muhammad Ali Norozi.

IGI Global: International Journal of Computer Vision and Image Processing (IJCVIP) 2011.

Abstract: *Extrapolations* are simple and unique techniques in linear algebra that require little additional infrastructure that needs to be incorporated in the existing query-dependent *Link Analysis Ranking* (LAR) algorithms. Extrapolations in LAR settings relies on the *prior knowledge* of the (iterative) process that created the existing data points (iterates) to compute the new (improved) data point, which periodically leads to the desired solution much faster than the original method. In this study we present the novel approaches using extrapolation techniques to speed-up the convergence of the query-dependent iterative methods like HITS and SALSA, *link analysis* based ranking methods, where hyperlink structures are used to determine relative *importance* of a document in the network of inter-connections. We work within the hubs and authorities framework defined in HITS and SALSA and propose the use of different *Extrapolation techniques* for faster ranking. Hence we come up with a novel improvement in algorithms like HITS, SALSA and their descendants (e.g., Exponentiated and Randomized HITS) using the Extrapolation techniques. With the proposed approaches it is possible to accelerate the iterative ranking algorithms in terms of reducing the number of iterations and therefore uncovered a much faster rate of convergence. In the experiments, in the concluding part of the article, we even got much better results than the theoretically predicted assertion. The results present a speed-up to the order of $3 - 19$ times better than the original algorithms.

B.1 Introduction

Users on the web are expanding from being active information *consumers* to becoming active information *producers*. With such a boom in information retrieval and information explosion, there is an ever-increasing demand for access, coverage, quick responses, and relevant results. The huge collection of information inherently entails a loss of performance and efficiency, because it takes time to process (index, cluster, etc), retrieve (query) and keep the information up-to-date in the huge repositories. Thus there is a growing concern about the usability and the interaction time between the user and Information Retrieval (IR) systems. A trade-off between the quality of the results and query response time is mostly considered as an option. In such challenging settings a user query must yield meaningful, manageable and most importantly *"relevant"* set of results from IR systems in a reasonable time.

The apparent ease with which the users click from document to documents provides a rich source of information which could be used to understand *what* and *where* to find the important documents. The semi-structured and diverse collections of documents are held together by the billions of annotated connections called *hyperlinks*. Analyzing these myriad interconnections between the documents forms the basis for Link Analysis Ranking (LAR). These analyses will help us identify the proximity and relevance of documents amongst each other. And enables us to find out the social or informational organization of the documents (the sociology of information). We can utilize the contextual exposition of documents to deduce the importance or *popularity* of the documents in the network, by using the core graph theory concepts and techniques.

The citation structures of the documents contain a wealth of useful but "implicit information". Through citation structure hundreds and millions of documents can be pulled together into a network of knowledge. Foremost such a structure represents the users' behaviours and needs. Users on the Web usually discover most relevant and valuable information through the recommendations and references from a good source of information.

One of the main concerns in the link based ranking methods is the convergence to a "good solution" or an equilibrium state. An equilibrium state is a state where system under certain presumptions can declare the set of good results corresponding to user query. Most of the link analysis based ranking models are iterative in nature, they iteratively move towards the required equilibrium state (the good solution). Convergence is a central phenomenon in iterative algorithms [19]. In linear algebra, the iterative methods are employed when direct methods would be prohibitively expensive and in some cases impossible even with best possible computing power to find out the actual solution. Essentially the iterative methods such as "power method" [19] provide an approximation to the true solution starting from a *seed* value. This work deals with the convergence properties and behaviour of the famous query-dependent LAR algorithms [17, 22, 3, 28, 1].

The major contribution of this work is the improvements primarily in the convergence behaviour of the query dependent LAR algorithms using “careful periodic" applications of extrapolation step during iterations. We have distinctively applied extrapolation techniques to query-dependent LAR algorithms, which was not done before. The parameters are manipulated extensively in the empirical evaluation and hence extracted a very novel performance gain due to extrapolation. We concluded the article with an extensive experimental evaluation.

In the study by Kamvar *et al.*, [16] they have found an improvement of order 3 at-most due to extrapolation, in PageRank algorithm. By applying extrapolation carefully in the query-dependent algorithms, improvements of order in range $(3-19)$ have been discovered in this work, see Table B.2 and Appendix in [24].

The document is therefore organized as follows; Next section defines the theoretical background and the preliminaries of the problem at hand. Motivation for this paper is also presented in the same section and later the novel idea of *extrapolation* to speed the rate of convergence is explained in a reasonable detail in the Section *Extrapolation.* In the *Experimental Evaluations* section the idea proposed in earlier sections are empirically assessed. In the last section we conclude the study with important results and possible future work.

B.2 Theoretical Background

B.2.1 Link Analysis Ranking (LAR)

The presence of (hyper-) link information clearly augmented a great deal to the characterization of the *informative content* present in the documents. LAR approaches are intended to resolve some of the intrinsic weaknesses of the content-based Information Retrieval (IR) models. Through the analyses of network of the documents (due to citation structure) LAR approaches bring in a whole new horizon to information retrieval space. The essence of LAR therefore is that the “overall information" of a hyperlink database of documents is not composed of only static “textual information", but also another, the “hyper" information.

Link Analysis Ranking is the next step from just content-analyses. It involves analyses and understanding of a very huge and jumbled network(s) of documents. From such a huge and massive network extracting useful information is quite a challenging and difficult task. The challenge is not just because of size of network, but also because of its diversity and unpredictability. The huge network(s) of documents hence forms the core of link analysis ranking.

The resultant hyperlinked graph of the network of document will be given as an input to the LAR algorithms. This graph is encoded in an adjacency matrix $\mathbf{A}$, where $\mathbf{A}[i, j] = 1$ if there is a link from node i to node j and 0 otherwise (see also [19]). The LAR algorithm iteratively operate on the hyperlinked graph (the adjacency matrix $\mathbf{A}$) and returns the n-dimensional rank vector $\vec{x}$ with weights computed for each node in the graph, where x_i is the weight of i^{th} node. The weights are actually the *probabilities of relevance* of each document to the user query. LAR algorithms are thus meant to discover *authoritative* documents through analyzing the hyperlink graph [28, 2].

The two pioneer LAR algorithms PageRank [26] and HITS [17] are query-independent and query-dependent respectively. They were followed by substantial amount of research [11, 13, 10, 6, 27, 5, 14, 8, 16, 25], to name just a few.

B.2.2 Extrapolation

Extrapolation is the process of constructing new data points outside a discrete set of *known* data points. It is similar to the process of *Interpolation*, which constructs new points between known points, but its results are often less meaningful, and are subject to greater uncertainty. Interpolation is a specific case of curve fitting, in which the function must go exactly through the data points. In case of convergence, Extrapolation techniques can be employed to accelerate the convergence by using the known data points (values from successive iterates) to construct new data points (principal eigenvector(s)). Techniques for accelerating the *convergent series* are often applied in *numerical analysis*, where they are used to improve the speed of numerical integration, and other well-known series [21, 19].

B.3 Extrapolation Techniques to accelerate the Convergence

Extrapolation techniques are novel as they offer new ways of taking into consideration important properties of the iterative method for effectively accelerating the computation of the query dependent family of algorithms. Faster convergence and efficient computational speed in *query dependent* algorithms are quite crucial, because they operate on *query time*. For example, for a large matrix representing a network of documents referencing each other through hyper-links, it is fairly expensive to compute the operation $\vec{x}^k = \mathbf{A}\vec{x}^{k-1}$ several times as $k \to \infty$.

Extrapolation techniques were previously used by Kamvar *et al.*, [16], specifically tailored to the PageRank problem. In this study it is employed to the query-dependent counterparts such as HITS, its improvements and SALSA, and therefore we came up with more in-depth analyses of their convergence behaviours (see Section B.4).

B.3.1 Fixed Point Iteration

Extrapolation techniques in LAR stems from another popular method in numerical linear algebra called *fixed point iteration*. For a given function f defined on *real*

Algorithm 1 The HubAvg Algorithm

```
$func\ \ a^k \Leftarrow HubAvg(\mathbf{A}, a^{(0)}, \epsilon);$
$\mathbf{A} : \{adjacency\ matrix\ formed\ from\ the\ base - set\} S_q$
$\mathbf{A} \Leftarrow \mathbf{A_j} / RowNorm_j \{\forall j\}$
$a^{(0)} : \{set\ the\ seed\ values\ of\ the\ authority\ vector\}$
$h^{(0)}) : \{set\ the\ seed\ values\ of\ the\ hub\ vector\}$
while not converged do
    $\mathbb{I} :\ a^k \Leftarrow \mathbf{A^T A} a^{k-1}$
    $\mathbb{O} :\ h^k \Leftarrow \mathbf{A A^T} h^{k-1}$
    $\{Periodically.\}$
    $a^k \Leftarrow$ QuadraticExtrapolation $(a^{k-3}, a^{k-2}, a^{k-1}, a^k)$
    {Or}
    $a^k \Leftarrow$ PowerExtrapolation $(a^{k-d}, a^k, c, d)$
    {Or}
    $a^k \Leftarrow$ AitkenExtrapolation $(a^{k-2}, a^{k-1}, a^k)$
    $a^{k'} \Leftarrow a^k\ \{Normalize\}$
    $h^{k'} \Leftarrow h^k\ \{Normalize\}$
    $k \Leftarrow k + 1$
    {Compute the convergence}
end while
```

Algorithm 2 Quadratic Extrapolation

```
$func\ \ a^k \Leftarrow QuadraticExtrapolation(a^{k-3}, a^{k-2}, a^{k-1}, a^k)$
$y^{k-2} \Leftarrow a^{k-2} - a^{k-3};$
$y^{k-1} \Leftarrow a^{k-1} - a^{k-3};$
$y^k \Leftarrow a^k - a^{k-3};$
$Y \Leftarrow \left(y^{k-2} y^{k-1}\right);$
$\gamma_3 \Leftarrow \mathbf{1};$
$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \Leftarrow -Y^\dagger * y^k;$
$\gamma_0 \Leftarrow -(\gamma_1 + \gamma_2 + \gamma_3);$
$\beta_0 \Leftarrow \gamma_1 + \gamma_2 + \gamma_3;$
$\beta_1 \Leftarrow \gamma_2 + \gamma_3;$
$\beta_2 \Leftarrow \gamma_3;$
$a^k \Leftarrow \beta_0 * a^{k-2} + \beta_1 * a^{k-1} + \beta_2 * a^k;$
```

numbers and a given initial point x_0 in the domain of f, the fixed point iteration is:

$$x_{k+1} = f(x_k), k = 0, 1, 2, \dots \tag{B.1}$$

The series $x_0, x_1, \dots$ are expected to converge to x. If the function f is continuous, then x is a *fixed point* of f, i.e., $x = f(x)$.

Equation (B.1) is the standard fixed point iteration. Now consider the standard LAR problem, we will get a correspondence with fixed point iteration, i.e:

$$\vec{x}^{(k)} = \mathbf{A}\vec{x}^{(k-1)} \tag{B.2}$$

B.3.2 Aitken Δ^2 Extrapolation

Let us consider f in equation (B.1) as an iterative numerical process, then the intermediate iterates of the linear convergent series, x_i, x_{i+1} and x_{i+2} can be used to extrapolate the fixed point x. This three-point extrapolation scheme is well known as Aitken Δ^2 extrapolation [7].

Aitken's Δ^2 (three-points) extrapolation can be used to speed up the convergence of any sequence that is *linearly convergent*[1] [7]. Aitken Δ^2 extrapolation is oldest and most popular extrapolation technique. It forms the basis for other extrapolation techniques. It has also been used to speed-up the convergence of power method for faster computation of PageRank [16].

In LAR, Aitken acceleration computes the principal eigenvector of the Markov matrix in *one step*, under the assumption that the power iteration estimate $\vec{x}^{(k-2)}$ can be expressed as the *linear combination* of the first two eigenvectors, $\vec{u}_1$ and $\vec{u}_2$.

$$\vec{x}^{(k-2)} = \vec{u}_1 + \alpha_2\ \vec{u}_2 \tag{B.3}$$

where $\vec{u}_1$ is the principal eigenvector and $\vec{u}_2$ is the second eigenvector of Markov matrix.

Equation (B.3) shows that from the nonprincipal eigenvectors (the values of $\vec{x}^{(k-2)}$ from successive iterates), we can extrapolate the value of the principal eigenvector $\vec{u}_1$. The previous values calculated in the successive iterates could be used to extrapolate the new value (the new data point outside the known data points), the principal eigenvector. This way we could accelerate the rate of convergence of the already convergent series produced by the query-dependent LAR algorithm.

The Extrapolation step when applied *periodically*, enables us to subtract off the estimates of the nonprincipal eigenvectors from the current iterates $\vec{x}^{(k)}$. For the derivation of Aitken acceleration and the empirical proof that it can extrapolate the principal eigenvector for power method see [16, 7]. Aitken extrapolation technique is crucial primarily because the subsequent extrapolation techniques build upon

[1] A sequence $\{x_i\}$ is said to converge linearly to x^* if there is constant $1 > c > 0$ such that $||x_{i+1} - x^*|| \leq c||x_i - x^*||$ or alternatively $||x_{i+1} - x_i|| \leq c||x_i - x_{i-1}||$

the ideas advocated in this technique. It serves to provide a general *premise* for extrapolation. It is therefore essential to have a sound appreciation of this technique to comprehend the newer more sophisticated techniques of extrapolation used for accelerating convergence.

In a nutshell we use the *priori knowledge* (which we acquire from the prior iterates of an LAR algorithm) as a *basis* to extrapolate the new and better value (the principal eigenvector). We use the assumption that the new iterate(s) can be expressed as a linear combination of the *last few* iterates. With some changes to this basic assumption various extrapolation techniques can be formulated (for example, *Quadratic Extrapolation* assumes that last three iterates $\vec{x}^{(k-3)}$, $\vec{x}^{(k-2)}$ and $\vec{x}^{(k-1)}$ together with current iterate $\vec{x}^{(k)}$ can be used to express the new and improved iterate value, see Equation B.4). In case of Aitken Extrapolation we are using three-points $\vec{x}^{(k-2)}$, $\vec{x}^{(k-1)}$ and $\vec{x}^{(k)}$ to extrapolate the next point $\vec{u}_1$.

The extrapolation methods are different from standard fast eigensolvers, which mostly relies on the matrix factorization and/or matrix inversion. The extrapolation methods that we study here rely upon the fact that the principal (first) eigenvalue of the Markov matrix is, $\lambda_1 = 1$ [19], in order to find an approximation to the principal eigenvector. This information can be used to compute the estimates of the nonprincipal eigenvectors during the iterations. Through the *estimates* computed during the successive iterates of power method, we expect to extrapolate the value of the principal eigenvector. Specifically, we ignore the non-dominant eigenvectors corresponding to the negligibly small values of non-dominant eigenvalues ($<< 1$ or ≈ 0).

Algorithm 1 and 2 depicts the apparent elegance of extrapolation on improvement of HITS algorithm, the HubAvg algorithm [28, 24]. In Algorithm 1 we are periodically applying the extrapolation step (lines 9 to 14). Also observe in Algorithm 2, we are only using prior knowledge (the values of intermediate iterates) to extrapolate the new and improved value (the expected principle eigenvector). The most expensive operation in Algorithm 2 is the solution of the overdetermined system of linear equation (line 7).

What is crucial here is to identify *theoretically* and *empirically* that the extrapolation methods accelerate the convergence, and the computed value is actually the principal eigenvector. In Section B.3.3 we have provided theoretical proof for the importance of *Quadratic extrapolation* and its capability to extrapolate the principal eigenvector of the Markov matrix. In Section B.4 we have provided experimental evidences of the capabilities of Extrapolation to improve the convergent sequence in query-dependent LAR algorithms.

The assumption in equation B.4 is not in contrast to reality rather it is used to form the much stronger relation later in the derivation. Of course, the matrix **A** can have more than 3 eigenvectors. In the next section a relation based on this assertion has been formulated (Algorithm 2 is written based on that assertion). Empirically the Quadratic extrapolation derived from this assumption provides much better rate of convergence than the original algorithms (see Figure B.2).

B.3.3 Quadratic Extrapolation

Like Aitken Δ^2 Extrapolation, *Quadratic Extrapolation technique* also uses the idea of taking the *linear combination* of last few iterates. Unlike Aitken, Quadratic extrapolation uses the *first three* (instead of first two) eigenvectors of Markov matrix to express the iterate ($\vec{x}^{(k-3)}$). Therefore, it assumes that the iterate $\vec{x}^{(k-3)}$ can be expressed as the *linear combination* of first three eigenvectors ($\vec{u}_1$, $\vec{u}_2$ and $\vec{u}_3$) of Markov matrix.

These assumptions in Quadratic extrapolation enable us to approximate the principal eigenvector in *closed form*[2] using iterates $\vec{x}^{(k-3)}$, $\vec{x}^{(k-2)}$, $\vec{x}^{(k-1)}$ and $\vec{x}^{(k)}$.

Formulation

From the theory of Power method, we know that the seed vector $\vec{x}^{(0)}$ can be expressed as the linear combination of *all* the eigenvectors of the Markov matrix (see [19]). Thus,

$$\vec{x}^{(0)} = \vec{u}_1 + \alpha_2 \vec{u}_1 + \ldots + \alpha_m \vec{u}_m$$

While in the specific settings of Quadratic Extrapolation it is assumed that the Markov matrix $\mathbf{A}$ in equation (B.2) has only 3 eigenvectors. Based on this assumption the iterate $\vec{x}^{(k-3)}$ can be expressed as linear combination of these 3 eigenvectors.

The quadratic extrapolation can be formulated from this premise that the matrix $\mathbf{A}$ has only 3 eigenvectors and therefore we can approximate the iterate $\vec{x}^{(k-3)}$ as:

$$\vec{x}^{(k-3)} = \vec{u_1} + \alpha_2 \vec{u_2} + \alpha_3 \vec{u_3} \tag{B.4}$$

The assumption is not in contrast to reality rather it is used to form the much stronger relation later in the derivation. Of course, the matrix $\mathbf{A}$ can have more than 3 eigenvectors. Later in the section we will form a relation based on this assertion. In the experimental analyses in Section B.4 we have also provided empirical results verifying the validity of this assumption. It is shown in the experiments that Quadratic extrapolation derived from this assumption provides much better rate of convergence than the original algorithms (see Appendix in [24]).

Now we are in a position to derive the required model using equation (B.4). From the assumption (equation (B.4)), the characteristic polynomial $p_A(\lambda)$ of the Markov matrix $\mathbf{A}$ can now be written as:

$$p_A(\lambda) = \gamma_0 + \gamma_1 \lambda + \gamma_2 \lambda^2 + \gamma_3 \lambda^3 \tag{B.5}$$

[2] An *equation* or *system of equations* is said to have a *closed-form solution* if, and only if, at least one solution can be expressed analytically in terms of a bounded number of certain “well-known" functions [19].

Since the Markov matrix $\mathbf{A}$ is stochastic, we know from the theory of Markov chain that the first eigenvalue of $\mathbf{A}$ is $\lambda_1 = 1$ [9, 19]. Thus:

$$p_A(\lambda = 1) = 0 \Rightarrow \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 = 0 \tag{B.6}$$

According to *Cayley-Hamilton theorem* [9] any matrix $\mathbf{A}$ satisfies it's own characteristic polynomial, i.e., $p_A(\mathbf{A}) = 0$. Therefore multiplying $\vec{x}^{(k-3)}$ with the characteristic polynomial, we have:

$$p_A(\mathbf{A})\vec{x}^{(k-3)} = [\gamma_0\mathbf{I} + \gamma_1\mathbf{A} + \gamma_2\mathbf{A}^2 + \gamma_3\mathbf{A}^3]\vec{x}^{(k-3)} = 0 \tag{B.7}$$

This can be simplified as:

$$\gamma_0\vec{x}^{(k-3)} + \gamma_1\vec{x}^{(k-2)} + \gamma_2\vec{x}^{(k-1)} + \gamma_3\vec{x}^{(k)} = 0 \tag{B.8}$$

Since we knew from the power iterations:

$$\vec{x}^{(k-2)} = \mathbf{A}\vec{x}^{(k-3)} \ldots \quad \vec{x}^{(k)} = \mathbf{A}\vec{x}^{(k-1)}$$

From the above equations and equation (B.8), after simple steps we have:

$$\vec{x}^{(k-3)}(-\gamma_1 - \gamma_2 - \gamma_3) + \gamma_1\vec{x}^{(k-2)} + \gamma_2\vec{x}^{(k-1)} + \gamma_3\vec{x}^{(k)} = 0 \tag{B.9}$$

This can be further simplified as:

$$(\vec{x}^{(k-2)} - \vec{x}^{(k-3)})\gamma_1 + (\vec{x}^{(k-1)} - \vec{x}^{(k-3)})\gamma_2 + (\vec{x}^{(k)} - \vec{x}^{(k-3)})\gamma_3 = 0 \tag{B.10}$$

Define the following:

$$\begin{aligned} \vec{y}^{(k-2)} &= \vec{x}^{(k-2)} - \vec{x}^{(k-3)} && \text{(B.11)} \\ \vec{y}^{(k-1)} &= \vec{x}^{(k-1)} - \vec{x}^{(k-3)} && \text{(B.12)} \\ \vec{y}^{(k)} &= \vec{x}^{(k)} - \vec{x}^{(k-3)} && \text{(B.13)} \end{aligned}$$

Inserting equation (B.13) in equation (B.10) gives:

$$\vec{y}^{(k-2)}\gamma_1 + \vec{y}^{(k-1)}\gamma_2 + \vec{y}^{(k)}\gamma_3 = 0 \tag{B.14}$$

and

$$\left(\vec{y}^{(k-2)} \quad , \quad \vec{y}^{(k-1)} \quad , \quad \vec{y}^{(k)}\right)\vec{\gamma} = 0 \tag{B.15}$$

For the solution of the above system we don't want to have the trivial solution $\gamma = 0$, thus we constrain the leading term of the characteristic polynomial γ_3 as:

$$\gamma_3 = 1 \tag{B.16}$$

After substituting the value of γ_3, equation (B.15) can be written as:

$$\begin{pmatrix} \vec{y}^{(k-2)} & \vec{y}^{(k-1)} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = -\vec{y}^{(k)} \tag{B.17}$$

Hence we have an *overdetermined* system of linear equations:

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = -\mathbf{Y}^{\dagger}\vec{y}^{(k)} \tag{B.18}$$

Here $\mathbf{Y}^{\dagger}$ is the *pseudoinverse* of the matrix shown in the left side of the equation (B.17), $(\vec{y}^{(k-2)},\ \vec{y}^{(k-1)})$. From the equation (B.5) and above equations we can therefore find the coefficient of the characteristic polynomial $q_A(\lambda)$.

We may divide the characteristic polynomial with $(\lambda-1)$ to get $q_A(\lambda) = p_A(\lambda)/(\lambda-1)$. Hence we have now:

$$q_A(\lambda) = \frac{(\gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2 + \gamma_3\lambda^3)}{(\lambda - 1)} = \beta_0 + \beta_1\lambda + \beta_2\lambda^2 \tag{B.19}$$

By polynomial division and after some simple algebraic operations we get the values for beta (as depicted in Algorithm 1 as well):

$$\beta_0 = \gamma_1 + \gamma_2 + \gamma_3 \tag{B.20}$$

$$\beta_1 = \gamma_2 + \gamma_3 \tag{B.21}$$

$$\beta_2 = \gamma_3 \tag{B.22}$$

By Cayley-Hamilton theorem, for any vector $\vec{z}$ in $\mathbb{R}^n$ we also have:

$$q_A(\mathbf{A})\vec{z} = \vec{u}_1 \tag{B.23}$$

where $\vec{u}_1$ is the principal eigenvector of matrix $\mathbf{A}$ corresponding to eigenvalue $\lambda_1 = 1$. Thus by letting $\vec{z} = \vec{x}^{(k-2)}$:

$$\vec{u}_1 = q_A(\mathbf{A})\vec{x}^{(k-2)} \tag{B.24}$$

$$= [\beta_0 + \beta_1\mathbf{A} + \beta_2\mathbf{A}^2]\vec{x}^{(k-2)} \tag{B.25}$$

$$= \beta_0\vec{x}^{(k-2)} + \beta_1\mathbf{A}\vec{x}^{(k-2)} + \beta_2\mathbf{A}^2\vec{x}^{(k-2)} \tag{B.26}$$

Using the power iterations in above equations:

$$\vec{x}^{(k-2)} = \mathbf{A}\vec{x}^{(k-3)} \dots \quad \vec{x}^{(k)} = \mathbf{A}\vec{x}^{(k-1)}$$

Thus we get the closed form solution for $\vec{u}_1$, as:

$$\vec{u}_1 = \beta_0\vec{x}^{(k-2)} + \beta_1\vec{x}^{(k-1)} + \beta_2\vec{x}^{(k)} \tag{B.27}$$

The equation (B.27) together with equations (B.20) - (B.22) and equation (B.18) can be used to implement the Quadratic extrapolation. Hence together they will help to provide an approximation to the principal eigenvector of the Markov matrix $\mathbf{A}$. The above derivation steps are inspired from the work in [16, 7].

Discussion

In the above formulation in equation (B.18) we have to solve an *overdetermined system* of linear equations:

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = -\mathbf{Y}^{\dagger}\vec{y}^{(k)}$$

The above overdetermined system can be solved through any *least-square method*, for example, through the *QR* factorization of *Gram-Schmidt algorithm* [9, 19]. The Quadratic extrapolation technique can be further optimized by applying a better solver to the overdetermined system above.

The Quadratic Extrapolation improves convergence much better than the original rate of convergence of the algorithm, based on the empirical results (see Section B.4 and Appendix in [24]). In the slow convergent series, the Quadratic Extrapolation is proved to be an effective technique. For example, if the second eigenvalue of the Markov matrix is close to 1, i.e., $\lambda_2 \to 1$, theoretically and empirically the convergence of power method tends to slow down. In such a situation the slow converging sequence of Power method can be accelerated radically by Quadratic Extrapolation.

The important thing in Quadratic Extrapolation is that it should be applied periodically. Once the parameters for Extrapolation (such as $\vec{x}^{(k-3)}$, $\vec{x}^{(k-2)}$, $\vec{x}^{(k-1)}$ and $\vec{x}^{(k)}$) are ready we could either apply Quadratic extrapolation step immediately or apply it at any other instance with appropriate values. It doesn't necessarily need to be applied too often to achieve maximum benefit. Experiments in Section B.4 reveal interesting insights about the potential of Quadratic Extrapolation in various settings. By manipulating the periodic application of Quadratic extrapolation we can administer the convergence behaviour of an algorithm.

Theoretically, Quadratic extrapolation technique is used to *subtract off* the errors in the current iterate along the direction of the second and third eigenvectors, as mathematically represented by equation (B.4). By doing that it enhances the convergence for the future application of the power method. The approximate principal eigenvector as a result of Extrapolation step serves as a good approximation for the further iterates, which help to converge much faster.

In the next section we will explore another interesting technique for extrapolation, where some important properties of the Markov matrix are exploited to make a more generic and cleaner formulation of Extrapolation.

B.3.4 Power ($\mathbf{A}^d$) Extrapolation

Based on the ideas initially put forward in Aitken Extrapolation and Quadratic Extrapolation discussed in previous sections Haveliwala et al., [13] construct another interesting formulation of extrapolation. Similar to both Aitken and Quadratic extrapolation, here also by subtracting off the errors along several nonprincipal eigenvectors from the current iterates, it is intended to accelerate the rate of convergence of Power method. Not just relying on the values of successive iterates but other important properties of Markov matrix, i.e., the *nonprincipal eigenvalues* could be exploited to accelerate the convergence.

In linear algebra, finding the nonprincipal eigenvalues of a Markov matrix is a problem in itself. Calculating nonprincipal eigenvalues of Markov matrix may increase the computational overheads, instead of providing any improvements. Thus apparently the idea of using the nonprincipal eigenvalues for acceleration may not seem conducive in general. But in a study by Haveliwala and Kamvar [12], they discovered interesting insights about the nonprincipal eigenvalues of the Markov matrix in PageRank algorithm. They have proved that the modulus of second eigenvalue of the Markov matrix (or the *Google matrix*) is given by the damping factor 'c' in PageRank algorithm (where $\alpha = c$, in the original formulation in [26]). Thus if the row stochastic matrix $\mathbf{S}$ in PageRank formulation ([26]) has at least two irreducible closed subsets, then the second eigenvalue of $\mathbf{S}$ is given by:

$$\lambda_2 = c$$

Note that the webgraph can have many eigenvalues with modulus of 'c' (i.e., one of c, $-c$, ci, and $-ci$). These eigenvalues of the webgraph has been exploited in the power extrapolation to approximate the principal eigenvector of the hyperlink matrix $\mathbf{A}$ corresponding to the webgraph. In the next section we will briefly formulate the use of second eigenvalue for accelerating the convergent series of power method, and in Section B.4 we will also present experimental findings for this approach.

Formulation

In Power Extrapolation the iterate $\vec{x}^{(k-2)}$ can be represented as linear combination of three eigenvectors ($\vec{u}_1, \vec{u}_2$ and $\vec{u}_3$) of Markov matrix. Making use of the same assumption as it was in *Quadratic Extrapolation*. Thus:

$$\vec{x}^{(k-2)} = \vec{u_1} + \alpha_2\vec{u_2} + \alpha_3\vec{u_3} \tag{B.28}$$

The nonprincipal eigenvalues corresponding to nonprincipal eigenvectors $\vec{u_2}$ and $\vec{u_3}$ are 'c' and '$-c$' respectively, according to the results in [12]. From Power Iterations we have:

$$\vec{x}^{(k)} = \mathbf{A}^2\vec{x}^{(k-2)}$$

$$(since \quad \vec{x}^{(k)} = \mathbf{A}\ (\mathbf{A}\vec{x}^{(k-2)}) \quad as \quad \vec{x}^{(k-1)} = \mathbf{A}\vec{x}^{(k-2)})$$

Substituting the above relations in equation (B.28) we get:

$$\vec{x}^{(k)} = \mathbf{A}^2(\vec{u_1} + \alpha_2\vec{u_2} + \alpha_3\vec{u_3}) \tag{B.29}$$

$$\vec{x}^{(k)} = \vec{u_1} + \alpha_2\lambda_2^2\vec{u_2} + \alpha_3\lambda_3^2\vec{u_3} \tag{B.30}$$

We can replace $\lambda_2 = c$ and $\lambda_3 = -c$ in equation (B.30), thus we have:

$$\vec{x}^{(k)} = \vec{u_1} + c^2(\alpha_2\vec{u_2} + \alpha_3\vec{u_3}) \tag{B.31}$$

$$\vec{x}^{(k)} = \vec{u_1} + c^2(\vec{x}^{(k-2)} - \vec{u_1}) \tag{B.32}$$

From equation (B.32) the closed form of the approximated principal eigenvector $\vec{u}_1$ will be:

$$\vec{u_1} = \frac{\vec{x}^{(k)} - c^2\vec{x}^{(k-2)}}{1 - c^2} \tag{B.33}$$

The above derivation leads to $\mathbf{A}^2$ Extrapolation which subtract off error along the eigenspaces corresponding to eigenvalues c and $-c$. There is a general derivation as well to the case where the eigenvalues of modulus of c given by $c \times d_i$, where d_i are the d^{th} root of unity, are used to form a generalized closed form based on variable d [13]. For example, for $d = 4$ the nonprincipal eigenvalues of modulus of c are given by c, ci, and $-ci$, which means 4^{th} roots of unity.

The generalized case will have the closed form of the principal eigenvector $\vec{u_1}$ as:

$$\vec{u_1} = \frac{\vec{x}^{(k)} - c^d\vec{x}^{(k-d)}}{1 - c^d} \tag{B.34}$$

For details about the derivation of equation (B.34) see [13].

The implementation of $\mathbf{A}^d$ Extrapolation is much more simpler than the Quadratic Extrapolation, just to implement equation (B.34). Theoretically the overhead due to $\mathbf{A}^d$ extrapolation is negligible since it is applied only *once*. The convergence is also found to be similar to the Quadratic Extrapolation, but the wallclock-speedup is higher in $\mathbf{A}^d$ extrapolation [13]. In contrast to the findings by Haveliwala et al., our results show that the convergence behaviour due to power extrapolation in query-dependent algorithms are not comparable to that of Quadratic extrapolation in the empirical settings, see [24].

Discussion

In the case of PageRank the *Eigengap* $(1 - |\lambda_2|)$ for the Markov matrix $\mathbf{A}$ is given exactly by the teleportation probability $1 - c$, in accordance to the findings by Haveliwala et al., discussed above. Theoretically, if the second eigenvalue λ_2 is close to 1, then the convergence of the power method will be slow. Because of the fact that convergence of the power method depends on $|\lambda_2|/|\lambda_1|$ factor, and k must be fairly large before $(|\lambda_2|/|\lambda_1|)^k$ converge to $\mathbf{0}$ [19, 9]. In PageRank reducing the factor $|\lambda_2|/|\lambda_1|$ correspond to the reduction of the damping factor c, because we can only change the numerator which is $|\lambda_2|$. But reduction of damping factor means increase in teleportation $(1 - c)$ and hence increasing the chance for the spammers to inflict the rankings. However a high teleportation probability $((|\lambda_2|/|\lambda_1|) \to 1)$ constitutes slow convergence of power method. Under such circumstances it is highly rewarding to accelerate the slow convergence of power method. The methods of Extrapolation discussed above supposedly provide a faster convergence even with high teleportation probabilities.

The $\mathbf{A}^d$ Extrapolation is based on second eigenvalue of the Markov matrix. In HITS and other query-dependent algorithms such as SALSA, the power extrapolation cannot be directly applicable. The second eigenvalue of the Markov matrix should be calculated independently. It is not a trivial problem to compute the nonprincipal eigenvalues of Markov matrix in general case. The teleportation scheme cannot be directly employed in the query-dependent algorithms, unless we incorporate the damping factor the same way as it was done in PageRank [26, 18].

There is a phenomenon called *Deflation* in linear algebra which can be used to compute the nonprincipal eigenvalues of the Markov matrix. Deflation is a technique of reducing the dimension of Markov matrix corresponding to its dominant eigenvalue. Once we compute the dominant eigenvalue, the Markov matrix can be reduced to one lower dimension (by subtracting the column corresponding to the dominant eigenvalue). We can now compute the dominant eigenvalue of the reduced dimension matrix, which will be the second eigenvalue of the original Markov matrix. Repeat this process until all the eigenvalues are computed. Finding the nonprincipal eigenvalues of Markov matrix is therefore not an easy task. The main techniques for dimensionality reduction are principal component analysis (PCA) and the famous singular value decomposition [19] .

In [24] we have discussed one way to incorporate the damping factor into the query-dependent HITS and SALSA too. After that modification it becomes then possible to apply power extrapolation under the same assumption as we had applied for PageRank algorithm.

B.3.5 Insights into Extrapolation

Extrapolation is one of the effective techniques in numerical analysis, but its use for acceleration of convergence in power method is novel. There are a lot of things to explore in the topic of Extrapolation. The studies so far just provide a definition level insight into the topic. There could be a lot of different and useful insights into much dynamic aspects of Extrapolation.

A new premise

The Extrapolation methods described in the previous sections are built upon the *premise* constructed about the initial function (equations (B.3), (B.4) and (B.28)). The common thread among all the extrapolations techniques is the initial assumption. In case of Quadratic Extrapolation we assumed that the matrix $\mathbf{A}$ has 3 eigenvectors, and expressed the current iterate as the linear combination of these 3 eigenvectors. A quadratic function illustrates a much closer representation to the reality in this case, and hence provides much better convergence.

One of the possible prospects in Extrapolation could be to start with another new *premise*. A much deeper understanding of Markov matrix and the insight into properties of LAR algorithms in question could be valuable to form a new premise(s). It could be possible to also consider the *personalization* (next section) factor in the construction of the new premise in Extrapolation. A possible future work from this study could be to explore extrapolation independently with the focus on formation of a new premise.

Extrapolation parameters

From experiments in the Section B.4 we found that from the behaviour of the *convergence graph* we could manipulate the extrapolations parameters to control the convergence. In this regard the number of times we apply extrapolation step is quite crucial. If extrapolation is applied more often than required it might increase the overhead instead of improving. We will discuss in detail about the exploitations of the parameters in extrapolation during experimental analyses in the next chapter.

Hybrid Extrapolation technique

We might as well think of extrapolation in a 'hybrid' environment. That is, considering the properties of different extrapolation techniques and depending on the behaviour of the graph we could apply different extrapolation techniques at different instances during iterations of the power method. We will experiment with the hybrid of Quadratic Extrapolation and Power Extrapolation to experimentally examine this approach. But it also requires an independent study to come up with a possible *framework* which could be used to exploit different extrapolation techniques in a hybrid environment in order to achieve a much controlled convergence in power method. And to observe more closely the dependences and dynamics of different extrapolation techniques applied simultaneously. It could turn out to be a novel approach towards active use of extrapolation. And again, the factor of personalization can also be employed in the hybrid extrapolation scheme.

The findings discussed in this section just depict limited implications of extrapolation. There are quite a lot of other possibilities too for further innovation in the field of Extrapolation.

B.4 Experimental Evaluations

In this section the focus is on the empirical evaluations of the extrapolation technique discussed in the Section B.3. We will specifically observe the effectiveness extrapolation on the query-dependent LAR algorithms, such as; HITS, SALSA and their improvements. For brevity, only the results of HITS and HubAvg algorithm

$$\begin{array}{lll}
\mathbf{1} & : & \mathbf{182}, \mathbf{183}, \mathbf{12}, \mathbf{-1} \\
\vdots & & \\
5 & : & 325, 326, 327, 328, 329, -1 \\
\vdots & & \\
51 & : & 1296, 1297, 694, 707, 715, 789, 502, -1 \\
\vdots & & \\
3403 & : & 3405, 3406, -1
\end{array}$$

Figure B.1: *Inverted file for query "amustement park"*

has been discussed here, but for more comprehensive understanding see [24]. Specifically, we will observe the peculiarities of Extrapolation techniques in improving the rate of convergence and hence a faster ranking.

We have primarily relied on the dataset used in [3]. The dataset is gathered using the prescriptions of Kleinberg [17] (as described in the section below) and is stored in an inverted file format, see Figure B.1. The effectiveness of the findings of this study is compared with the findings described in the work by Borodin *et al.*, and also with the findings in both [23, 22].

B.4.1 Experimental setup

The algorithms that are tested, operate on a collection of pages that is created following the guidelines of Kleinberg [17]. Search engine Google is queried for each of the queries shown in Table B.2 (when a query consists of more than one word, we put the '+' symbol in front, so as to ensure that all pages contain the query terms). The first 200 pages returned by Google form the *Root-set* as prescribed by Kleinberg. For each page in the Root-set, all the *out-links* are stored of that page, and the first 50 *in-links*, in the order they are returned by Google. One way of obtaining in-links is to use Google queries of form $link : url$ (e.g., $link : www.fastsearch.com$).

Every page is first assigned a '*docid*' (document id), and from the pages' *docids* an inverted file will be generated corresponding to each query. An inverted file (for query "amusement parks") looks like Figure B.1. This means that *docid* **1** the first row (boldface), contains link to *docids* $182, 183$ and 12, and -1 indicate end of list (out-links). Using the inverted file as an input to a *script* (such as, a bash-script, or a python code or a matlab code), we could convert the inverted file to an adjacency matrix **A** form. Where the entries of **A** corresponding to *docid*-**1** will be; $\mathbf{A}(1, 182) = 1$, $\mathbf{A}(1, 183) = 1$ & $\mathbf{A}(1, 12) = 1$ and rest of the entries in 1^{st} row of matrix **A** will be; $\mathbf{A}(1, j) = 0$, where $j \notin (182, 183, 12)$. The resultant adjacency matrix **A** corresponding to each query given in Table B.2, can be given as an input to the LAR algorithms.

B.4.2 The Queries

There are some standard set of queries appeared in the literature and used in previous works [3, 17, 20]. The choices of queries are driven by the fact that they become a representative of the whole Web. Therefore it is expected that through their representativeness they unveil the implicit properties of the algorithms, e.g., HITS support of tightly knit communities [28, 17]. Every query represents a topic on web. Webgraph can be considered as a set of clusters of *strongly connected nodes*, each cluster theoretically represents some topic(s). In principle we want to align query topic(s) with the topic(s) on the webgraph. By testing the algorithms with different representative queries (topics), we tend to observe the behaviour of the algorithms on these topic(s), using single or multi-topic queries at a time.

There are queries where the most relevant results are not textually expressed in the most relevant documents, e.g., the phrase “search engine" doesn’t appear in the most of the search engines main pages. Thus we would have those types of queries also which are usually not expressed within the relevant documents, such as, “search engines", “automobile industries", etc.

There are also queries for which we have conflicting communities on the webgraph. For example the query “iraq war" and “abortion" can have sets of conflicting clusters of webgraph. It is interesting to observe how different algorithms treat these queries.

The queries in the Table B.1 with the statistical information will be used for the experiments; the queries are exactly the same as is used in [3].

B.4.3 Query Statistics

The query statistics provided in the [28] are used in our study (see Table B.1). The table provided here shows the analytical study of the datasets corresponding to each query. The assessment information of the dataset gives a broad picture of the neighbourhood graph (the base-set [17]). The information therefore is useful for conceptualizing and understanding the underlying structure of graph and broad picture of manifestation of algorithms. Sometimes from the assessed dataset we could predict the expected behaviours, performances or outcomes of the algorithms. For example, for the query “search engines" there are $11,659$ nodes and $292,236$ links, which means the underlying graph for this query is quite big ($11,659 \times 11,659$) and dense. There will be memory contention issues for such a big graph. It would be interesting to observe how different algorithms will react to such a big graph.

In case of PageRank, Haveliwala [11], presents a memory efficient approach that lowers the main memory requirements for the huge webgraph (using *Block-Based strategy*). In HITS usually the memory concern is not that terrible, because the graphs are usually of order $1,000 - 5,000$ nodes. Nevertheless, efficient usage of memory is a favourable property for HITS too, given the fact that it is computed at query time, and the retrieved pages could be sizable.

Table B.1: *Query Statistics*

Query	Nodes	Hubs	Authorities	Links	Avg out
abortion	3340	2299	1666	22287	9.69
affirmative action	2523	1954	4657	866	2.38
alcohol	4594	3918	1183	16671	4.25
amusement parks	3410	1893	1925	10580	5.58
architecture	7399	5302	3035	36121	6.81
armstrong	3225	2684	889	8159	9.17
automobile industries	1196	785	561	3057	3.89
basketball	6049	5033	1989	24409	4.84
blues	5354	4241	1891	24389	5.75
cheese	3266	2700	1164	11660	4.31
classical guitar	3150	2318	1350	12044	5.19
complexity	3564	2306	1951	13481	5.84
computational complexity	1075	674	591	2181	3.23
computational geometry	2292	1500	1294	8189	5.45
death penalty	4298	2659	2401	21956	8.25
genetic	5298	4293	1732	19261	4.48
geometry	4326	3164	1815	13363	4.22
globalization	4334	2809	2135	17424	8.16
gun control	2955	2011	1455	11738	5.83
iraq war	3782	2604	1860	15373	5.90
jaguar	2820	2268	936	8392	3.70
jordan	4009	3355	1061	10937	3.25
moon landing	2188	1316	1179	5597	4.25
movies	7967	6624	2573	28814	4.34
national parks	4757	3968	1260	14156	3.56
net censorship	2598	1618	1474	7888	4.87
randomized algorithms	742	502	341	1205	2.40
recipes	5243	4375	1508	18152	4.14
roswell	2790	1973	1303	8487	4.30
search engines	11659	7577	6209	292236	38.56
shakespeare	4383	3660	1247	13575	3.70
table tennis	1948	1489	803	5465	3.67
vintage cars	3460	2044	1920	12796	6.26
weather	8011	6464	2852	34672	5.36

Profoundly analyzed datasets provide valuable input for assessment or comparison of the LAR algorithms. Therefore it's very crucial to have a sample and representative datasets with statistical information available, which could be used for experimentation. We could have done our own evaluation based on the pre-labelled corpus, such as *TREC collection*[3], but due to limited time constraint, we primarily rely on the dataset provided on [29].

We refer and confide on the query statistics given in the Table B.1 during experimentations.

B.4.4 Measures

To assess the quality and accuracy of the results of an algorithm the 'precision' over $top - 15$ results has been compared with the results in [3]. Usually for the relevancy ranking algorithms the measures used to assess the algorithms are 'precision' and 'recall'. But the results of the text based search are expected to have high recall, and therefore only the precision over $top - 15$ results are considered, and also

[3] **T**ext **RE**trieval **C**onference(TREC) is the primary benchmark for information retrieval.

considering the behaviour web user, only the accuracy of the $top-15$ results or the first page of results are important.

B.4.5 Convergence

In general an algorithm should declare convergence once the value of $Residual_i = Rank_{k+1} - Rank_k$ stabilizes. In most of the literature L_1 norm or residual of the authority weights of two successive iterates is used to detect convergence (see $y-axis$ of Figure B.2). Hence, for measuring the convergence in all of the algorithms the following well-known measure is employed:

$$\delta_k = ||\mathbf{A}\vec{x}^{(k)} - \vec{x}^{(k)}||_p \quad p = 1 \tag{B.35}$$

In linear algebra, equation (B.35) is generally used as an indicator of convergence for most of the iterative algorithms. In almost all of the experiments, L_1 norm in Equation (B.35) has been compared against $\epsilon \in (10^{-16} - 10^{-5})$.

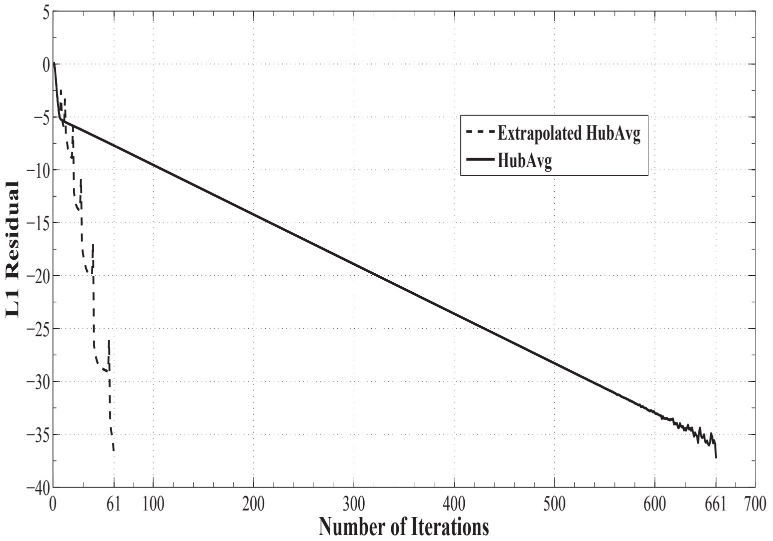
There are other possible ways also to measure convergence, for instance in [16] *Kendall's-tau rank correlation (KDist)* measure is used, to see if the residual, L_1 norm is a good measure of convergence. Haveliwala [11] suggests to use *induced orderings*, rather than residuals, by looking at the ordering of the pages *induced* by the rank vector to measure convergence. With induced ordering, PageRank vector converges in as few as 10 iterations in comparison with convergence with L_1 residual which takes about 50+ iterations.

We choose L_1 norm in Equation (B.35) as a measure of convergence for the sake of simplicity.

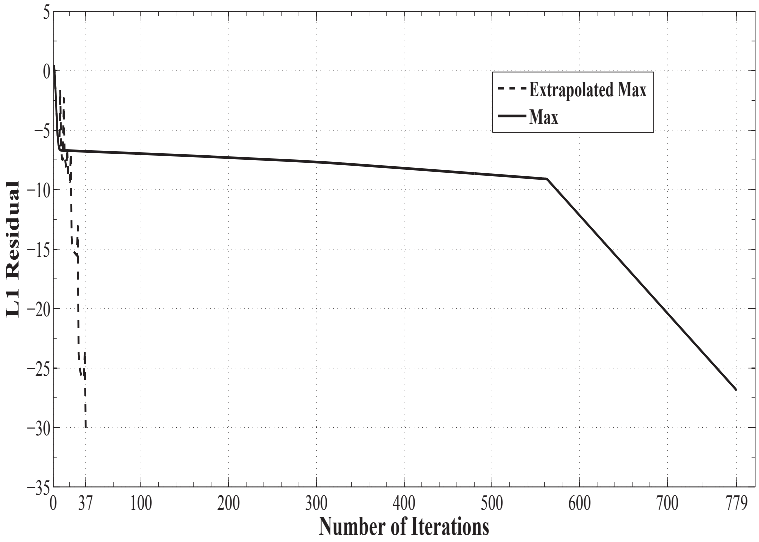
B.4.6 Experimental Results

We have exclusively looked into convergence property of the query-dependent LAR algorithms. The application of a single extrapolation step is considered to be equivalent to 0.5 times or less the cost of an iteration of power method (32% of cost of an iteration [13]), e.g., Figure B.2(b) extrapolation step is applied 6 times only while the improvement is surprising (the original algorithm stablizes after 779 iterations while extrapolated version in just 37 iterations). Hence the effects of extrapolation step is not that severe. The improvements in number of iterations is interpreted as almost equivalently to the improvements in time, e.g., 2 times improvement in number of iterations is treated as $1.8 - 2.0$ times speedup in the wall-clock time.

Extrapolation techniques are therefore very effective, the extrapolated algorithms in our experiments yielded a net speedup of over 3 (see Table B.2), the speedup could be even more significant in practice; for example we even got a speedup of 19 on our dataset depending on careful application of *extrapolation step* (see Figure B.2(a) for query “basketball").



(a) Query “computational complexity"



(b) Query “basketball"

Figure B.2: *The convergence graphs. The spikes in each graph shows the point where extrapolation step is applied*

Table B.2 provides a comprehensive overview of all the results for each of 34 queries that we used. In the table:

–itr, *is the number of iterations*
–ext, *is the number of times extrapolation applied*
–E, *is the Extrapolated version of the algorithms*, *and*
–N *refers to normal version*

Table B.2: *Results of the experiments with* ***Extrapolation***

		HITS		HubAvg				HITS		HubAvg	
Queries	#	**E**	**N**	**E**	**N**	**Queries**	#	**E**	**N**	**E**	**N**
abortion	itr	12	17	43	106	globalization	itr	17	24	52	139
	ext	3		8			ext	3		5	
affirmative action	itr	397	2529	71	199	gun control	itr	43	149	100	448
	ext	3		10			ext	4		11	
alcohol	itr	21	38	41	135	iraq war	itr	146	736	39	70
	ext	2		6			ext	14		5	
amusement parks	itr	26	57	34	61	jaguar	itr	22	31	131	879
	ext	5		8			ext	4		6	
architecture	itr	45	98	155	817	jordan	itr	30	78	62	131
	ext	4		17			ext	3		6	
armstrong	itr	30	68	26	44	moon landing	itr	34	79	59	178
	ext	5		4			ext	2		4	
automobile industries	itr	50	175	64	217	movies	itr	131	568	41	72
	ext	9		10			ext	10		4	
basketball	itr	20	27	49	119	national parks	itr	17	21	33	95
	ext	3		3			ext	2		4	
blues	itr	29	52	48	163	net censorship	itr	35	76	284	1048
	ext	6		11			ext	3		18	
cheese	itr	17	24	44	87	randomized algorithms	itr	63	193	80	239
	ext	3		7			ext	13		9	
classical guitar	itr	43	160	38	88	recipes	itr	90	397	53	113
	ext	10		8			ext	12		7	
complexity	itr	21	32	101	321	roswell	itr	100	341	175	538
	ext	3		11			ext	13		11	
computational complexity	itr	53	144	61	661	search engines	itr	9	13	23	41
	ext	5		9			ext	2		3	
computational geometry	itr	28	58	44	98	shakespeare	itr	29	72	64	270
	ext	6		6			ext	5		6	
death penalty	itr	12	18	36	70	table tennis	itr	26	45	42	114
	ext	1		5			ext	6		9	
genetic	itr	26	40	43	91	vintage cars	itr	35	60	91	587
	ext	3		7			ext	3		7	
geometry	itr	25	45	45	108	weather	itr	26	53	32	54
	ext	5		6			ext	6		4	
Median	itr	**29.0**	**59.0**	**48.5**	**125.0**	**Average**	itr	**50.2**	**191.7**	**67.8**	**247.1**
	ext	**4.0**		**7.0**			ext	**5.3**		**7.5**	

The extrapolation columns in the Table B.2 indicate the best performance in terms of number of iterations that we got as a result of tweaking the parameters. Overall we have applied extrapolation steps **8** times on an average to get rapid convergences. So, the overhead of net application of extrapolation is very less, almost 3 – 4 iterations of LAR algorithm. On average the algorithms converge after just **46** as a result of extrapolation in comparison to the average **170** iterations of the original algorithms. A net average speedup of order **5.78** in all the algorithms presents a very good reason for the usefulness of extrapolation techniques.

The quadratic extrapolation technique should be applied *periodically* to subtract off the errors in the current iterate (along the direction of the second and third eigenvectors). It improves convergence only when it is applied *carefully*. It is interesting to observe empirically that extrapolation applied too frequently doesn't really achieve any further benefits. Because by doing that we are not allowing the iterative power method (see Algorithm 1) to use the new computed iterate to annihilate error components of the iterate in directions along the eigenvectors with *small* eigenvalues. Quadratic Extrapolation step leaves error components primarily along the smaller eigenvectors, which the power method is better equipped to eliminate. Thus we need to allow power method to eliminate errors instead of applying extrapolation step frequently and accumulating the errors after every application. That is why it is very much important to apply the right number of extrapolation steps to gain the required improvements.

Note that Extrapolation technique can also be applied in conjunction with other acceleration techniques, such as *BlockRank* [15], or other iterative algorithms e.g., *Gauss-Seidel*, *Successive Over Relaxation*, *Conjugate Gradient* or any other methods [9, 18]. When used in conjunction with any other methods, we might expect more insights about the effectiveness of Extrapolation, both in terms of time and convergence.

We have also had a limited evaluation of the hybrid implementation of extrapolation technique, see [24] for details.

B.5 Conclusions and Future Work

The speedup in convergence due to extrapolation came first as a surprise. There were some interesting observations that came out as a result of the experiments. It is observed that a careful application of extrapolation can improve convergence inevitably. Therefore, it matters *when*, *where* and *how many times* during iterations you apply extrapolation step to gain the required acceleration. The importance of extrapolation in accelerating the convergence is hence remarkable. We have also tested *hybrid* extrapolation technique, where the effects of extrapolation based on different techniques (Aitken Δ^2, Power [13] and Quadratic extrapolation [24]) applied together have been observed in different settings.

Quadratic Extrapolation in query dependent LAR algorithm improves convergence much better than the original rate of convergence of the algorithm, based on the empirical results. In the slow convergent series, the Quadratic Extrapolation is proved to be an effective technique. For example, if the second eigenvalue of the Markov matrix is close to 1, i.e., $\lambda_2 \rightarrow 1$, theoretically and empirically the convergence of power method tends to slow down [4, 19]. In such a situation the slow converging sequence of Power method can be accelerated radically by Quadratic Extrapolation (see Section B.4). As a result of this study, it is possible to observe the convergence graph much more closely from another perspective. For example, use any other convergence measure instead of just L_1 norm. The question is; from the convergence behaviours, is it possible to automate *when*, *where* and *how* many times extrapolation should be applied to gain certain acceleration? Also the more active use of *induced ordering* [11] to measure convergence together with extrapolation could be a possible future work. *Hybrid approach* of extrapolation could also be further observed to formulate a better framework for extrapolation, which could possibly be used for *personalization* too, apart from just accelerating the convergences.

B.6 References

[1] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. *ACM Transactions on Internet Technology*, 5(1):92–128, 2005.

[2] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web. *Proceedings of the 10th international conference on World Wide Web*, pages 415–429, 2001.

[3] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, 2005.

[4] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

[5] M. Brinkmeier. PageRank revisited. *ACM Transactions on Internet Technology (TOIT)*, 6(3):282–301, 2006.

[6] K. Bryan and T. Leise. The $25,000,000,000 Eigenvector: The Linear Algebra behind Google. *SIAM Review*, 48(3):569–81, 2006.

[7] J. Cioslowski. Why does the Aitken extrapolation often help to attain convergence in self-consistent field calculations? *The Journal of Chemical Physics*, 89:2126, 1988.

[8] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pages 291–299, 1999.

[9] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[10] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, 2003.

[11] T. Haveliwala et al. Efficient computation of PageRank. *Stanford University*, 8090:1998–31, 1999.

[12] T. Haveliwala and S. Kamvar. The second eigenvalue of the Google matrix. *A Stanford University Technical Report http://dbpubs.stanford.edu*, 8090:2003–20, 2003.

[13] T. Haveliwala, S. Kamvar, D. Klein, C. Manning, and G. Golub. Computing PageRank using Power Extrapolation. Technical report, Tech. Rep. 2003-45, Stanford University, July 2003, 2003.

[14] G. Jeh and J. Widom. Scaling Personalized Web Search. *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[15] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank. *Preprint, March*, 2003.

[16] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating PageRank computations. *Proceedings of the 12th international conference on World Wide Web*, pages 261–270, 2003.

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

[18] A. Langville and C. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[19] D. Lay. *Linear algebra and its applications*. Addison-Wesley Reading, Mass, 1994.

[20] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*, 33(1-6):387–401, 2000.

[21] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[22] M. Najork. Comparing the effectiveness of HITS and SALSA. *Proceedings of the sixteenth ACM Conference on information and knowledge management*, pages 157–164, 2007.

[23] M. Najork, H. Zaragoza, and M. Taylor. HITS on the Web: How does it compare? *Proceedings of the 30th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 471–478, 2007.

[24] M. A. Norozi. Information Retrieval Models and Relevancy Ranking. Master's thesis, University of Oslo, 2008.

[25] M. A. Norozi. Extrapolation to speed-up query-dependent link analysis ranking algorithms. *Proceedings of the 8th International Conference on Frontiers of Information Technology*, page 2, 2010.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1998.

[27] D. Rafiei and A. Mendelzon. What is this page known for? Computing Web page reputations. *Computer Networks*, 33(1-6):823–835, 2000.

[28] P. Tsaparas. *Link Analysis Ranking*. PhD thesis, University of Toronto, 2004.

[29] P. Tsaparas. Link analysis ranking - experiments. `http://www.cs.toronto.edu/~tsap/experiments/thesis/`, 2004.

PAPER C

Contextualization from the Bibliographic Structure

Muhammad Ali Norozi, Arjen P. de Vries and Paavo Arvola.

Appeared in Task Based and Aggregated Search Workshop, Proceedings of the 34th European Conference on Information Retrieval (ECIR), Barcelona, Spain, 2012.

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).

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

Section C.4 concludes and highlights future work.

C.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 C.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.

C.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 information 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 probabilities 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 equilibrium 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 C.1(a)) can be represented in matrix notation by adjacency matrix **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 C.1. For the Figure C.1(a) the corresponding adjacency matrix **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]:

$$g^k \quad = \quad \mathbf{A^T}\ \mathbf{A} g^{k-1} \tag{C.1}$$

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 C.1 [5].

C.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 **A** becomes huge in this case ($342,279 \times 342,279$, see Section C.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 C.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 **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 **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(see Figure C.1(b)).

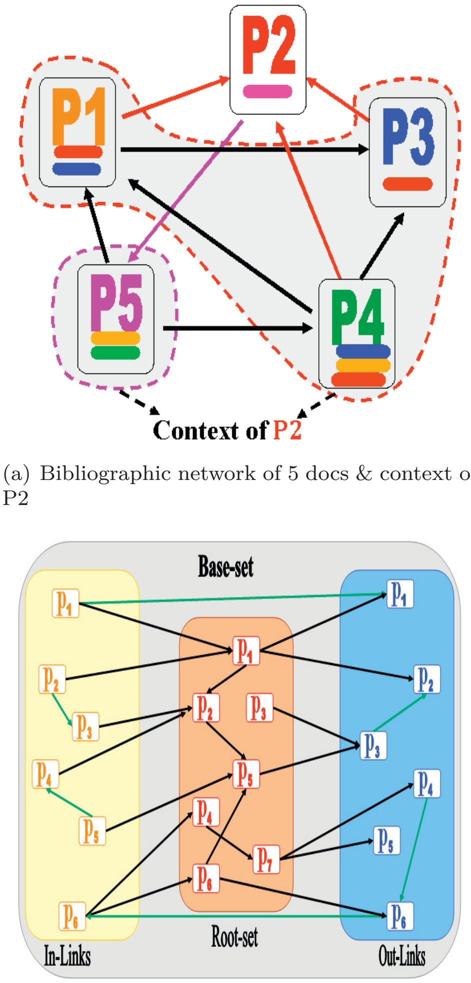
C.2.3 Combination function

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:

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

where

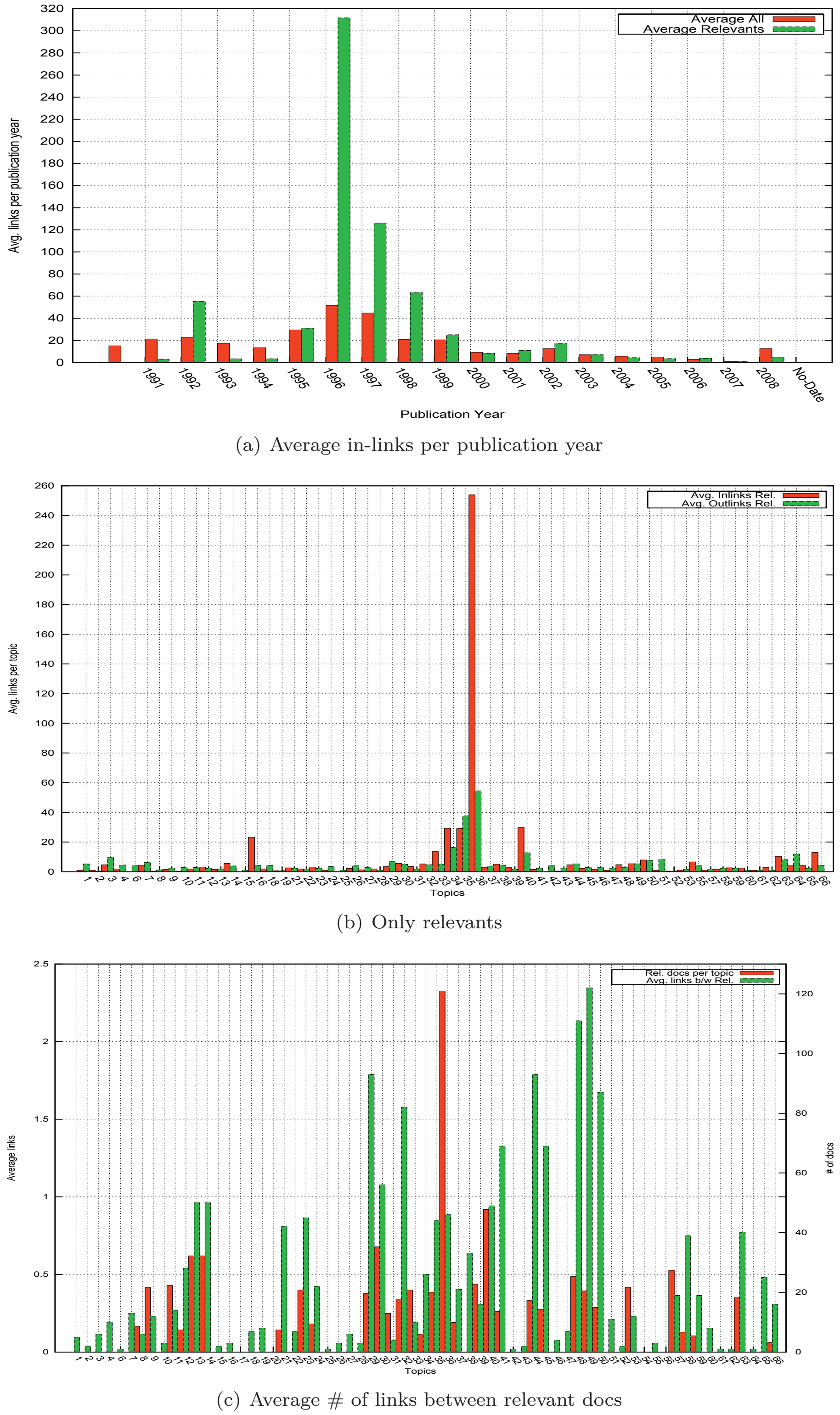
[1] Finding the dominant Eigenvector of the system of linear equations, corresponding to the dominant eigenvalue, which is 1 in this case [8].



(a) Bibliographic network of 5 docs & context of P2

(b) The Base-set

Figure C.1: *Bibliographic information and relevant retrieved*



(a) Average in-links per publication year

(b) Only relevants

(c) Average # of links between relevant docs

Figure C.2: *Average Number of links per topic*

- $BS(x)$ 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
- 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 C.2, as discussed in forthcoming Sections below.

C.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 C.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 C.1(a) correspond to the non-zero cells of column 2 in the adjacency matrix $\mathbf{A}$.

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

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

$$\begin{aligned} CR(x, f, C_x, g^k_{qi}, g^k_{qd}) = (1-f) \cdot BS(x) + f \cdot \alpha \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k_{qi}(y)}{\sum\limits_{y \in C_x} g^k_{qi}(y)} + \\ f \cdot (1-\alpha) \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k_{qd}(y)}{\sum\limits_{y \in C_x} g^k_{qd}(y)} \end{aligned} \tag{C.3}$$

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.

C.2.5 Context for a better content description

Given the bibliographic structure of the iSearch collection, Figure C.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 C.2 and C.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 C.1(a) corresponds to the non-zero cells of row 2 in the adjacency matrix $\mathbf{A}$.

C.3 Experimental Evaluation

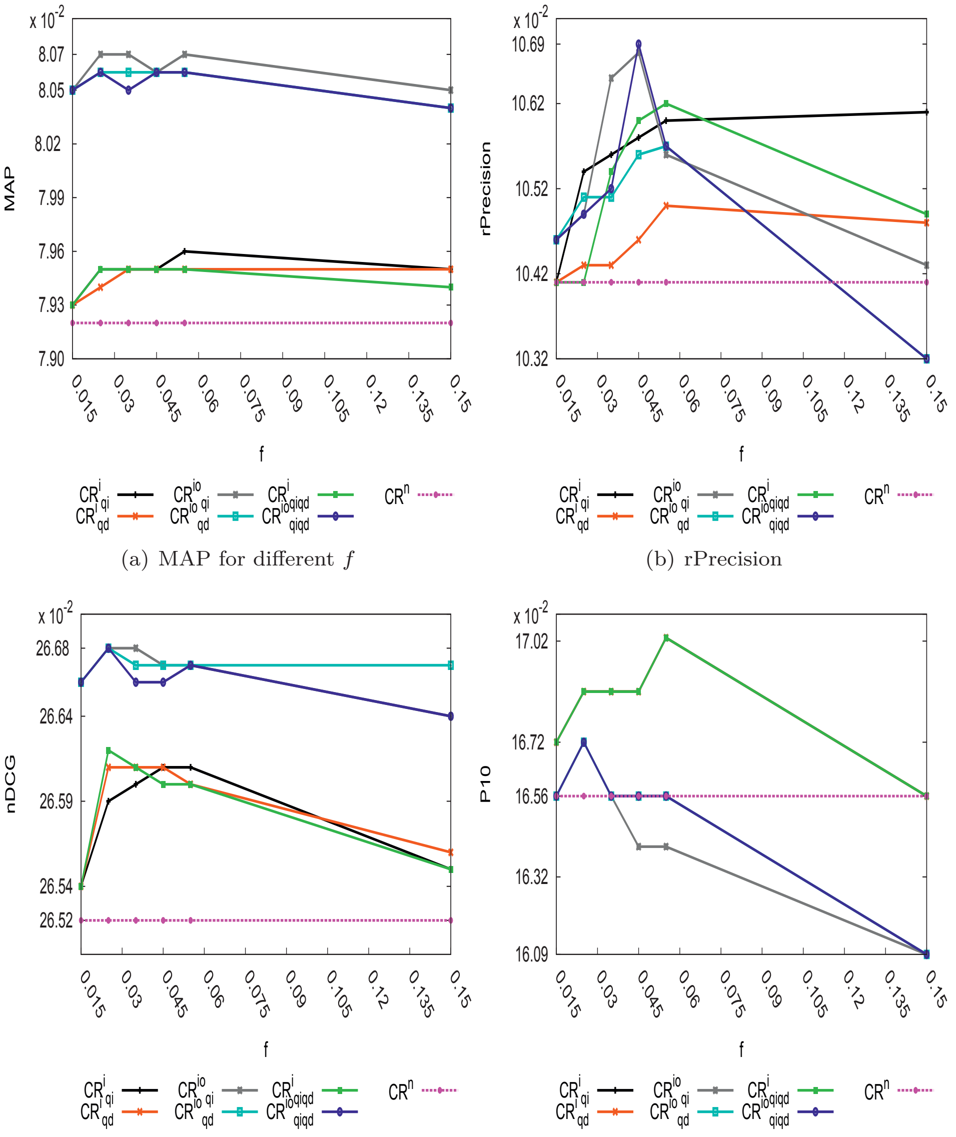
C.3.1 Experimental Settings

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 limitations 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].

[2] Citebase is created by Tim Brody from Univerity of Southampton, UK, http://citebase.org



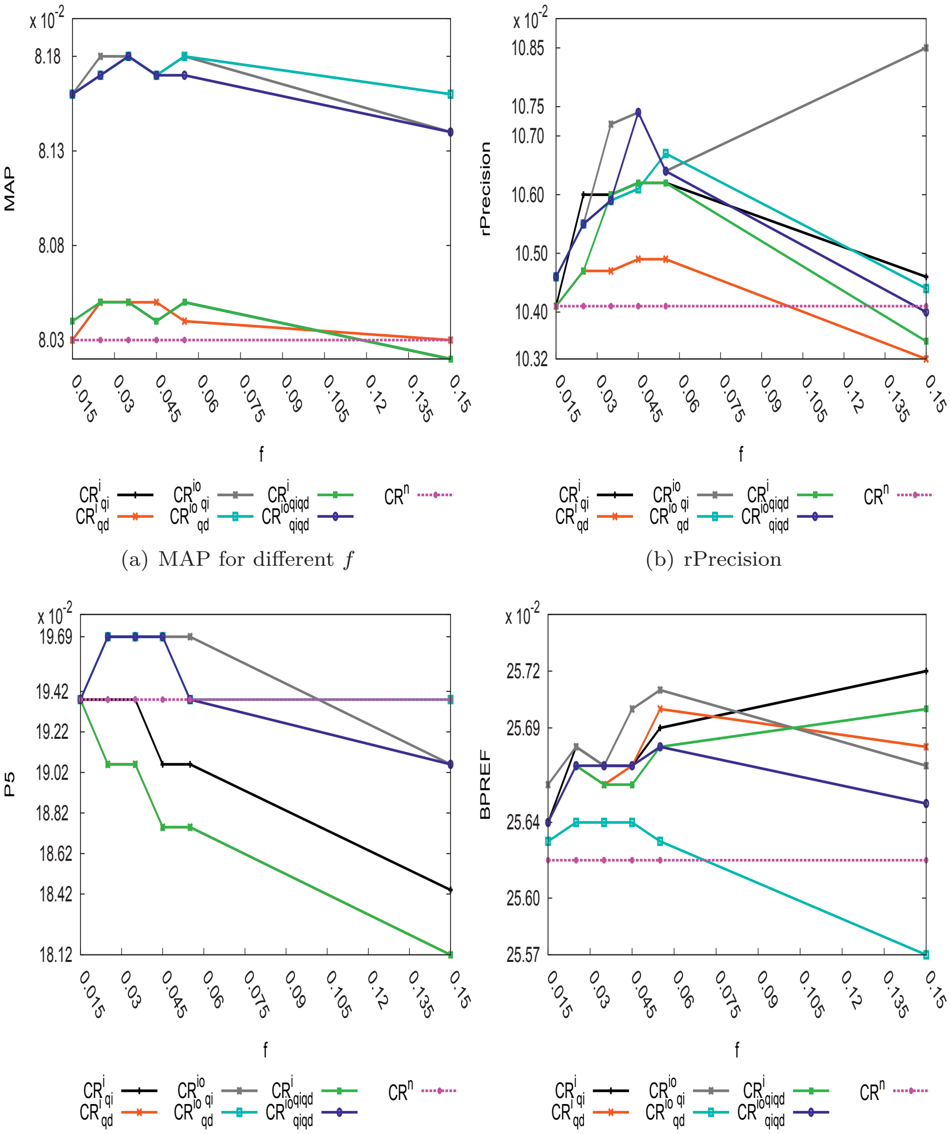
(a) MAP for different *f*

(b) rPrecision

(c) nDCG

(d) P@10

Figure C.3: *Trends for different measures @1500*



(a) MAP for different f

(b) rPrecision

(c) P@5

(d) BPREF

Figure C.4: *Trends for different measures @8k*

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 C.1 and C.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.

C.3.2 Results

We have tested seven different retrieval methods based on the propositions (see Section C.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 C.1) seperately, we tuned the following parameters and report the best performance: (i) the contextualization force f from Equation C.2 ($f \in \{0.015, 0.025, 0.035, 0.045, 0.055, 0.15\}$); (ii) the α parameter from Equation C.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 C.3, i.e., runs involving both query- independent and -dependent walks. These optimal values for f and α are obtained training with the iSearch collection. Figures C.3 and C.4 illustrate the behaviour of the methods as we change mainly the f parameter, from Equations C.2 and C.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 C.1

and C.2). As can be visually observed, the proposed methods out-perform the baseline, CR^n, in almost all the figures.

Table C.1 and C.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 C.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△	**.1060▲**	.2661△	.2330△	.1906	**.1703▲**	–
CR^i_{qd}	.055	.0795△	**.1050▲**	**.2661▲**	.2330△	.1938	**.1703▲**	–
CR^i_{qiqd}	.025-.055	.0795△	.1063△	**.2662▲**	.2329△	.1938	**.1703▲**	.2-.7
CR^{io}_{qi}	.035-.055	.0807△	**.1068▲**	.2668△	.2326△	.1938	.1656	–
CR^{io}_{qd}	.025-.055	.0806△	**.1057▲**	.2668△	.2326△	.1938	.1672△	–
CR^{io}_{qiqd}	.035-.055	.0807△	**.1069▲**	.2667△	.2325△	.1938	.1656	.2-.7

Table C.1: *Ret. performance @1500 ▲= stat. significance at $p < 0.05$ (2-tailed t-test). △= 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△	.1062△	.2878△	.2569△	.1906	**.1703▲**	–
CR^i_{qd}	.055	.0804△	.1049△	.2878△	.2570△	.1875	**.1703▲**	–
CR^i_{qiqd}	.055	.0805△	.1062△	.2878△	.2569△	.1875	**.1703▲**	.2-.7
CR^{io}_{qi}	.035-.055	**.0818**△	**.1074▲**	**.2890**△	.2571△	.1969△	.1625	–
CR^{io}_{qd}	.025-.055	**.0818**△	**.1067▲**	.2889△	.2563△	.1969△	.1625	–
CR^{io}_{qiqd}	.035-.055	**.0818**△	**.1074▲**	**.2890**△	.2571△	.1969△	.1625	.2-.7

Table C.2: *Retrieval performance @8k*

C.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.

C.5 Acknowledgements

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

C.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.

PAPER D

Contextualization using Hyperlinks and Internal Hierarchical Structure of Wikipedia Documents

Muhammad Ali Norozi, Paavo Arvola and Arjen P. de Vries.

Appeared at the 21st International Conference on Information and Knowledge Management (CIKM), Maui, Hawaii, USA, 2012.

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.

D.1 Introduction

Focused or element retrieval addresses the possibility 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 D.3); the leaf element or descendant 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 [10, 5] 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, 23, 19, 25, 22]. 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 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. 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 Wiki-pedia 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 D.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 D.2.1).
- Contextualization of the hierarchical structure of documents, using the same random walk model (Section D.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 D.3).
- Experimental validation (Section D.4) of the ideas proposed, using citation (Section D.2.1), hierarchical (Section D.2.2) and hybrid contextualization (Section D.2.3) within the random walk framework.
- Evaluation of the use of citation and hierarchical information on the large semantically annotated Wikipedia XML corpora [26, 13, 3, 8, 11] (Section D.4.1).

Section D.5 concludes and highlights future work.

D.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 D.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.

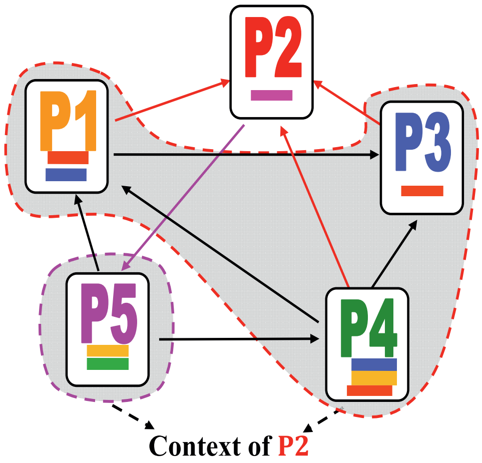


Figure D.1: *Citation structure of 5 documents and context of P2*

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

D.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) [17, 20, 13]. 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 information 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 probabilities 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 equilibrium 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 D.1) can be represented in matrix notation by adjacency matrix **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 D.1. For Figure D.1 the corresponding adjacency matrix **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]:

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

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 D.1 [17].

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 **A** becomes quite huge for the citation structure of Wikipedia collection $(2,668,160 \times 2,668,160$, see Section D.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 D.1.

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

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.

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:

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

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 D.2, as discussed in forthcoming Sections below.

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 D.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 D.1 correspond to the non-zero cells of column 2 in the adjacency matrix $\mathbf{A}$.

Section D.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 D.2.1) 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 D.4. In addition to the two variants, a third variant combines the query independent and query dependent random walk into a combination function:

$$CR(x, f, C_x, g^k_{qi}, g^k_{qd}) = (1-f) \cdot BS(x) + f \cdot \alpha \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k_{qi}(y)}{\sum\limits_{y \in C_x} g^k_{qi}(y)} + f \cdot (1-\alpha) \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k_{qd}(y)}{\sum\limits_{y \in C_x} g^k_{qd}(y)} \tag{D.3}$$

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.

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 D.2 and D.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 D.1 corresponds to the non-zero cells of row 2 in the adjacency matrix $\mathbf{A}$.

D.2.2 Hierarchical Contextualization

Hierarchical contextualization model has been studied before in different settings in XML retrieval [4, 1, 16, 25, 27, 19]. 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.

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 D.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.

```xml
<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 D.2: *XML document*

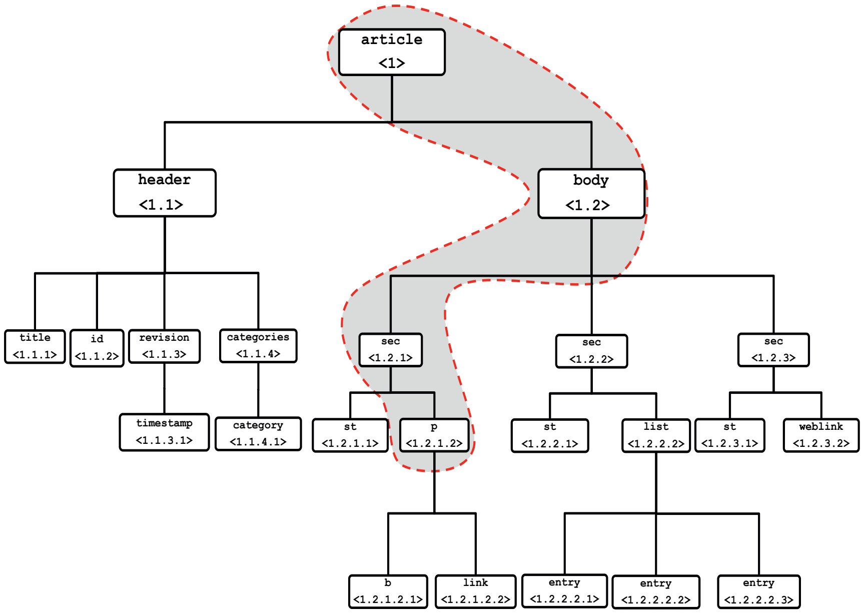


Figure D.3: *XML Graph of Figure D.2 with context of element ⟨1.2.1.2⟩ (dewey encoding)*

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 D.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 D.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 D.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 D.3).

Combination Function

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

$$CR(x, f, C_x, g^k) \quad = \quad BS(x) + f \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k(y)}{\sum\limits_{y \in C_x} g^k(y)} \tag{D.4}$$

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.

D.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 D.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 D.2 and later we use Equation D.4 for better contextualization.

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

D.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 D.3.1 we introduce the test bed, then in Section D.3.2 a baseline system based on data fusion is introduced and examined briefly in Section D.3.4 with the evaluation procedure of Section D.3.3.

D.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.

D.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.

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

where

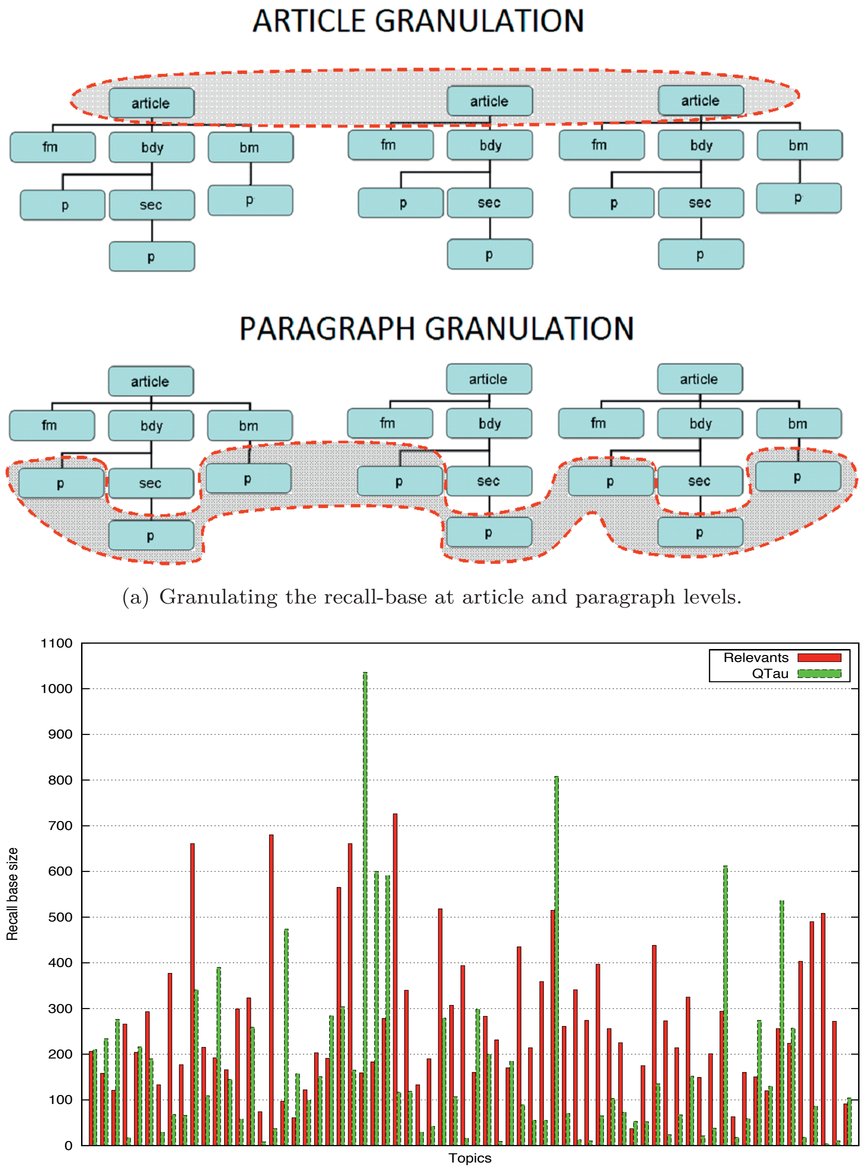
- R is the set of runs (rankings)
- and $rank(r, e, q)$ returns the rank of element e as a result of query q in run r.
- If e is not in the ranking, $rank(r, e, q)$ is not defined and the outcome of $\frac{1}{k+rank(r,e,q)}$ is 0.
- The parameter k is for tuning.

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.

D.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 D.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 (⟨p⟩-elements) only. To put it short, the former is for document retrieval and the latter is paragraph retrieval.



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

(b) Recall base sizes per topic

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

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) [29, 28], 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 (⟨p⟩). 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., ⟨article⟩ and ⟨p⟩). 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.

D.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 D.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.

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

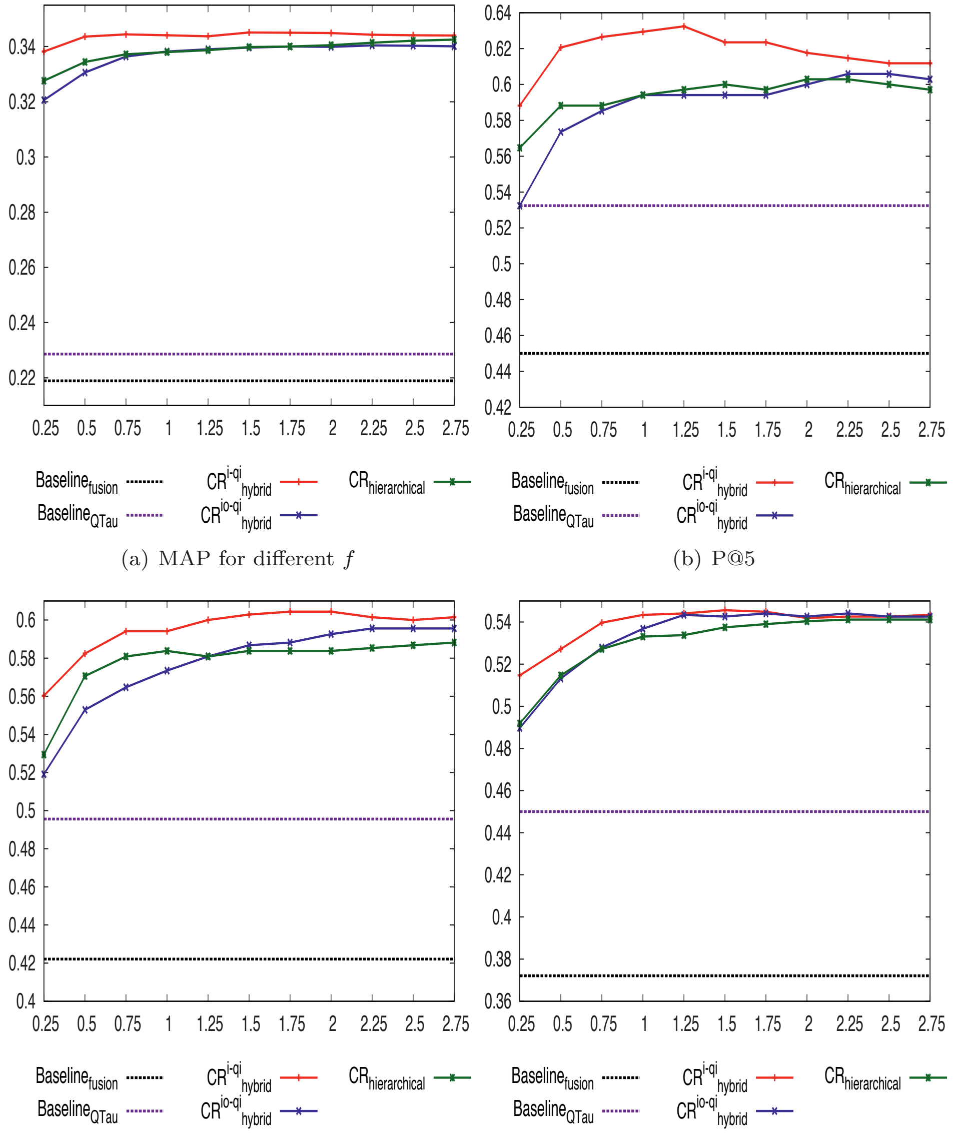
D.4.1 Experimental Settings

The proposed approaches are evaluated using the Wiki-pedia test collection, described in Section D.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 Wiki-pedia 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 D.2.1 and D.2.3). The computation of the contextualization vector g^k from Equation D.1 for the large Wikipedia collection was quite extensive, however this process is performed offline. The linear system of Equations D.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 time, we combine the iteratively computed random walk scores and the basic scores based on the proposed methods (Equation D.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 D.2 and D.1).

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



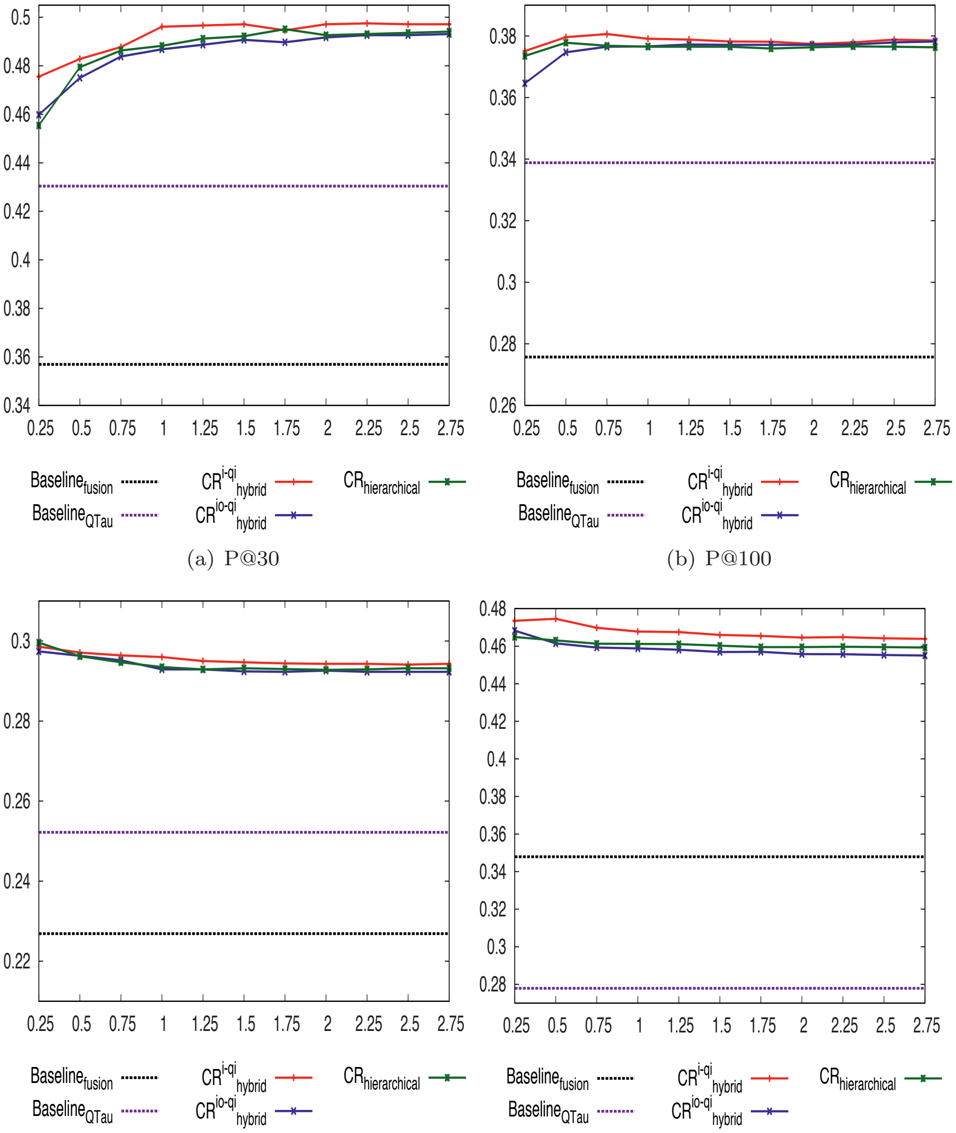
(a) MAP for different f

(b) P@5

(c) P@10

(d) P@20

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



(a) P@30

(b) P@100

(c) P@200

(d) rPrecision

Figure D.6: *Trends for different measures at different context force f for focused retrieval task (paragraph level)*

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 D.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 D.2.1. In this case also the contextualization vector g^k from Equation D.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 D.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 D.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 D.2: *Retrieval performance for document retrieval (article level).*

D.4.2 Results

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

- Baseline systems
  - Fusion run, $Baseline_{fusion}$.
  - University of Queensland run, which performed best on paragraph level, $Baseline_{QTau}$.

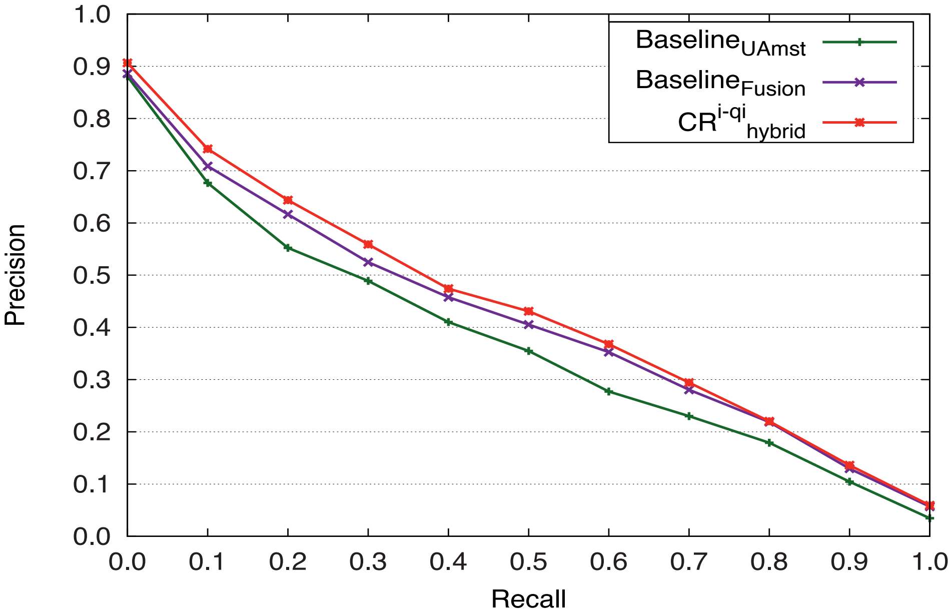


Figure D.7: *Precision - recall performance for document retrieval (article)*

  - 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 [4, 1]. While, the contextualization force has to be parametrized. For each proposed contextualization model, we tuned the contextualization force and report the values leading to best overall performance. In our parametrization process we found: (i) the optimal values of contextualization force f in citation contextualization (from Equation D.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 D.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. These 68 independent or unseen samples are then combined to produce a single or a set of estimations for the parameter f.

Figures D.6 illustrate the behaviour of the methods as we change the optimal values of f parameter, from Equations D.2, D.3 and D.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 D.7).

Table D.1 and D.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 D.6). The improvements overall are surprisingly good on both focused and document retrieval.

The best overall results among the proposed methods are obtained with CR_{hybrid}^{i-qi} 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 [13, 12] 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.

D.4.3 Discussion

Contextualization is a re-ranking model utilizing 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

[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.

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 in-line with the earlier work on use of hyperlinked and hierarchical tree (graph) structure of documents [10, 5, 12, 17] and the role of contextualization [1, 4, 23, 19, 25, 22]. 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 [10, 5, 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.

D.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 dimension 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.

D.6 Acknowledgements

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

D.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, 2011a.

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

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

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

[22] M. A. Norozi, A. P. 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.

PAPER E

Kinship Contextualization: Utilizing the Preceding and Following Structural Elements

Muhammad Ali Norozi and Paavo Arvola.

Appeared at the 36th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 2013.

Abstract: The textual context of an element, *structurally*, contains traces of evidences. Utilizing this context in scoring is called contextualization. In this study we hypothesize that the context of an XML-element originated from its *preceding* and *following* elements in the sequential ordering of a document improves the quality of retrieval. In the tree form of the document's structure, *kinship* contextualization means, contextualization based on the horizontal and vertical elements in the *kinship tree,* or elements in closer to a wider structural kinship. We have tested several variants of kinship contextualization and verified notable improvements in comparison with the baseline system and gold standards in the retrieval of focused elements.

E.1 Introduction

Contextualization [3] is a mechanism which makes possible the retrieval of items with varying length in textual content, as the size of elements varies with the level in the hierarchy (see Figure E.1); the leaf element or elements on low levels of hierarchy have potentially less textual evidences than their ancestors. The scant textual evidence in the small text units, such as paragraphs, are augmented with information obtainable from the context surrounding them.

The potential of contextualization has been revealed before in several intuitive settings [1, 2, 3, 7, 5, 8, 6]. In the existing studies, context of elements in focused retrieval has been mainly referred to the ancestor elements. In addition to the hierarchical order or the ancestor elements, documents also have an established sequential ordering (paragraph 1 comes prior to paragraph 2 and hence are siblings in the structural tree (which we refer to as the *kinship tree*, see Figure E.1), in the documents hierarchical structure). In this study the elements in the kinship tree, in the document's sequential order are considered to be the context - the kinship context. The proposed models are experimentally validated using the semantically annotated Wikipedia XML collection using INEX [9] evaluation measures. The results obtained, on focused retrieval task (INEX), exhibit clear improvements over the best submitted runs at INEX 2009, and over a strong and competitive baseline system − itself based on data fusion over all INEX 2009 submitted runs (Section E.3).

Summarizing, the contributions of this study include:

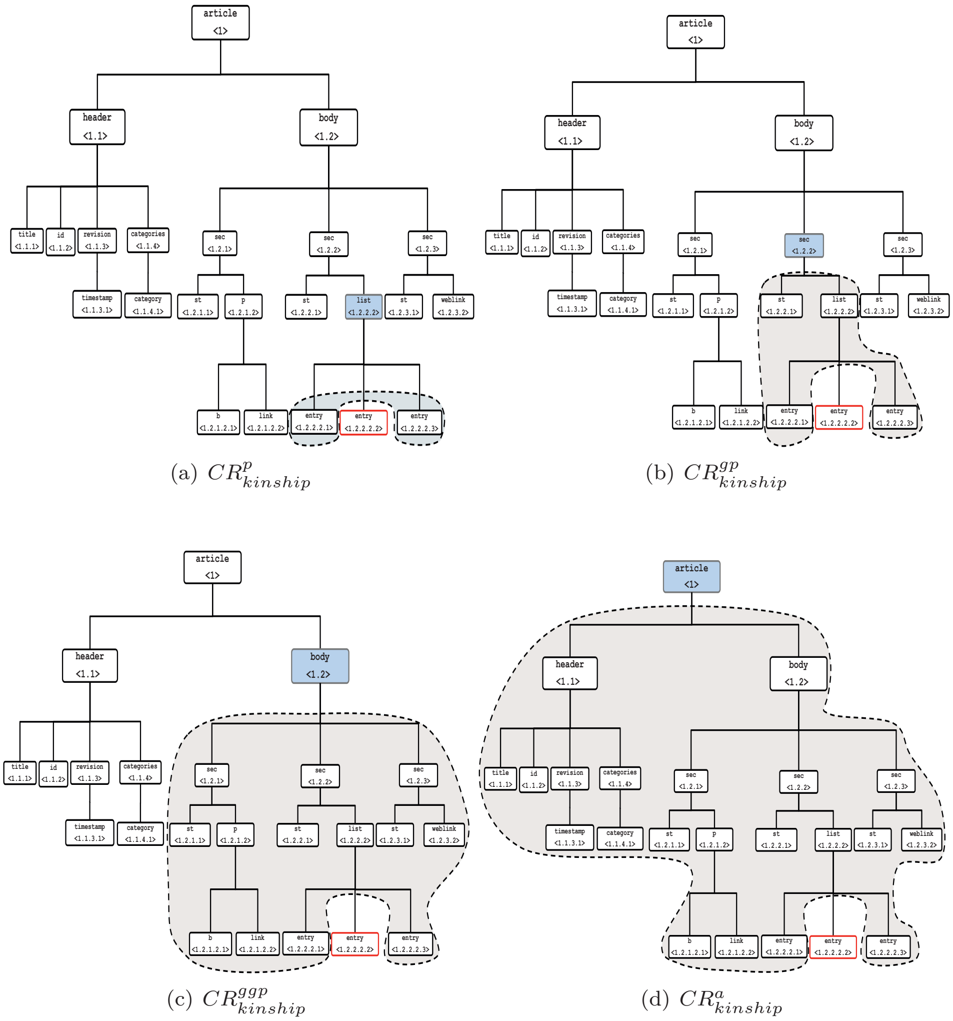
- Contextualization utilizing the nodes in *kinship* relationship (Figure E.1), in the hierarchical structure of documents, with random walks as a theoretically sound foundation (Section E.2.1).
- Developing a competitive focused retrieval system baseline based on data fusion and constructing a test setting for evaluating the retrieval of small textual units, i.e., focused retrieval (Section E.3).
- Experimental validation and evaluation (Section E.4) of the role of kinship contextualization on the large semantically annotated Wikipedia XML corpora [9] (Section E.4).

E.2 Contextualization

Contextualization is a re-scoring scheme, where the basic score, usually obtained from a full-text retrieval model, of a contextualized document or element is re-enforced by the weighted scores of the contextualizing documents or elements. We use random walks to induce a similarity structure over the documents based on the containment and reverse-containment relationships (element, sub-element and vice versa). Hence, these relationships (kinships) affect the weight each contextualizing element has in contextualization.

The premise is that *good context* (identified by random walk and contextualization model [6]) provides evidence that an element in focused retrieval is a good candidate for a posed query and therefore, the elements should be contextualized by their hierarchically similar elements in "kinship". Good context is an *evidence* that should be used to deduce that an element is a good candidate for the posed query.

E.2.1 Kinship Contextualization



(a) $CR^{p}_{kinship}$

(b) $CR^{gp}_{kinship}$

(c) $CR^{ggp}_{kinship}$

(d) $CR^{a}_{kinship}$

Figure E.1: *Kinship tree taken from example in [6], a representative XML from the Wikipedia 2009 collection.*

In this section, we will show a formalism that can be used to materialize and then utilize the contextual evidences originated from the elements in the kinship tree, in the documents sequential order, for improving the retrieval effectiveness. Use of hierarchical information as a context has been studied before in different settings in XML retrieval [3, 1, 8, 11, 5]. In hierarchical contextualization the intrinsic structure within the XML document is employed. Kinship contextualization is both horizontal (siblings) and vertical (ancestors & descendants elements) but intrinsically non-hierarchical perspective of the hierarchical information. 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`⟩, ⟨`paragraphs`⟩ and so on, structures. ⟨`chapter1`⟩ is followed by ⟨`chapter2`⟩ and within ⟨`chapter1`⟩, ⟨`section1`⟩ is followed by ⟨`section2`⟩. Elements ⟨`section1`⟩ and ⟨`section2`⟩ are siblings, and hence most likely, semantically related. The following element takes the concepts further from the preceding elements, and the preceding elements provide the basics or foundation for the following elements. Therefore, together in the document order, the *preceding* and *following* elements form a strong and connected perspective (the kinship context), surrounding the relevant information. This organization of document gives an intuitive starting point for manipulating text passages at the established hierarchy levels of text documents.

With contextualization from the preceding and following elements, we aim to rank higher an element in a good context (strong evidence in the kinship) than an identical element in a not so good context (less or no evidence in the kinship tree) within the document. In Figure E.1 the ⟨`entry`⟩ element has the preceding ⟨`entry`⟩ and following ⟨`entry`⟩ as the kinship context. Hence one element can be viewed in context of its kinships, the elements preceding and following it.

In kinship contextualization, like the other types of contextualization [6, 3], 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 Equation E.1.

The important research question here is: how far wide (siblings) and how deep (direct parent or siblings at ancestral level), the element's kinship context help to improve retrieval effectiveness? Arvola et al. (2011) [3], in their horizontal contextualization approach, used a weight array, which follows a zero centred parabola function – the impact of contextualization is adjusted with a help of different set of parameters. The weight of contextualizing element is assumed to be the function of distance, hence the weight ought to be lower the further away the contextualizing element is from the contextualized element. Instead of employing the weight function, to find the impact of each of the units in the context, we propose here, a generalized mechanism based on the *Markovian* random walk principle [6] and the kinship contextualization.

Random Walk

The tree-structure of the XML document (Figure E.1) 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. In kinship contextualization, we consider all the elements in the *kinship* of the contextualized element; where the contextualization vector g in this case, identifies the importance of each of the unit of context in kinship(Equation E.1).

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 officially submitted runs by the participants (see Section E.3).

In the experiments we evaluated the retrieval effectiveness at different granularity levels. We mainly tested, retrieval effectiveness based on the element selection in focused retrieval task (using the INEX evaluation kit); a brief intuition is explained in Section E.3.

Kinship Contextualization at different level

We have experimented with kinship contextualization at different levels of hierarchy. Kinship contextualization with elements in kinship from:

- direct parent, $CR^{p}_{kinship}$.
- grand parent, $CR^{gp}_{kinship}$.
- grand grand parent, $CR^{ggp}_{kinship}$.
- root, the ⟨`article`⟩ element, $CR^{a}_{kinship}$.

The four approaches listed above are pictorially shown in Figure E.1.

Combination Function

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

$$CR(x, f, C_x, g^k) \quad = \quad BS(x) + f \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k(y)}{\sum\limits_{y \in C_x} g^k(y)} \tag{E.1}$$

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 kinship context surrounding the contextualizing element x, i.e., $C_x \subseteq kinships(x)$, $\subseteq$, 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 (the impact) of y, the contextualizing elements (kinships) of x in XML graph. Similar interpretation is used in our earlier studies [6]

E.3 Test Settings and Fusion Baseline

We test our approach using the Wikipedia collection containing 2.66 million semantically annotated XML documents ($50,7$ Gb) and 68 related topics provided by the INEX 2009 ad-hoc track [4]. The reason for using the INEX 2009 test topics (instead of 2010) is the larger variety of elements in the participants' results which was due to the existence of the thorough task. In order to get the best possible baseline, we performed a data fusion based on sum of normalized scores (CombSUM) [10]. The element scores (for each run per topic) were normalized for the fusion as follows:

$$score_x = \frac{score_x - min(scores)}{max(scores) - min(scores)} \tag{E.2}$$

where $max(scores)$ and $min(scores)$ denote the maximal and minimal scores respectively.

We used all the 98 INEX 2009 runs delivering correct element result lists as component systems for the fusion. The 2009 runs have the largest variety in the results for the fusion in comparison to other years of the initiative. Unfortunately, most of the participants (56) did not report any real element scores, because the INEX evaluation did not require that information. For those systems an artificial score was given for each element based on their reciprocal rank, before the normalization. In other words, the first element in the result list was given a score 1, the second 1/2, third 1/3, fourth 1/4 and so on.

The focused task in INEX ad-hoc track is to retrieve most focused elements satisfying an information need without overlapping elements. An overlapping result list means that the elements in the result list may have a descendant relationship with each other and share the same text content. For instance, in Figure E.1 the entry element ⟨`1.2.2.2.1`⟩ and the ⟨`sec`⟩ element ⟨`1.2.2`⟩ are overlapping. In this study we are following the focused approach, considering a result list where only one of the overlapping elements from each branch is selected. This means that including the ⟨`sec`⟩ element in the results would mean excluding the entry element in the results or vice versa.

The fused result list contains all the elements delivered by the 98 component systems. This comprehensive result list contains overlapping elements. In order to remove the overlap, we basically selected elements having the highest score from each branch. However, many participants returned runs having full-articles only, which led to full-article bias in the fusion results. Therefore, we made a deliberate choice to exclude full-articles in the results, following a more focused retrieval strategy. The result lists were measured using the official INEX evaluation metrics and software for the focused task [4].

Contextualization and the fusion approach as scoring methods, however, do not take any stand on which elements should be selected from each branch. Thus we perform a structural fusion, where we take the element level selection from the baseline run and subsequently re-rank the elements of the baseline run. For instance (in Figure E.1) if the baseline run suggests the ⟨`body`⟩ element, we select that one, not the ⟨`list`⟩ element beneath, regardless of their mutual ranking in the full list.

E.4 Experimental Evaluation

The hierarchical structure of XML documents are captured using the dewey encoding scheme (as shown in the Figure E.1). 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 detail, in our earlier work [6]. The contextualization vector g^k from Equation E.1 is computed off-line for each and every XML document in the Wikipedia collection. This suggests that computing g^k vector is feasible for a reasonably large XML document collections. At the query time, the scores from g^k vector and the basic scores are combined to produce an overall ranking score, using Equation E.1.

We have experimented with all the four variants of kinship contextualization (Section E.2.1) and compared them against the different baseline systems, (Table E.1, sorted on *i*nterpolated *p*recision at recall 0.01, $iP[0.01]$). The runs in Table E.1 are among the best runs submitted at INEX 2009 ad-hoc track, focused retrieval task.

In the combination function given, the contextualization force has to be parametrized. For the proposed contextualization model, we tuned the contextualization force and report the values leading to best overall performance. In our parametrization process we found the optimal values of contextualization force f (from Equation E.1) lies in the range: ($f \in \{3.25, 3.50, 3.75, 4.00, 4.25, 4.50\}$). These optimal values for f are obtained by using cross-validation technique. We did 68-fold cross-validation (or complete cross-validation in our case) - by randomly partitioning the collection

Run ID	MAiP	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]
CombSUM Fusion	**.3396**	**.7577**	**.7273**	**.6539**	**.6021**
UWFerBM25F2	.1854	.6797	.6333	.5006	.4095
I09LIP6Okapi	.3000	.6244	.6141	.5823	.5290
UJM_15525	.2890	.6241	.6060	.5742	.4921
UamsFSecs2dbi100CA	.1928	.6328	.5997	.5141	.4647
BM25BOTrangeFOC	.2912	.6049	.5992	.5619	.5057
Spirix09R001	.2865	.6081	.5903	.5342	.4979
LIG-2009-focused-1F	.2702	.5861	.5853	.5431	.5055

Table E.1: *Retrieval statistics for baseline systems. CombSUM fusion run is the best (statistically significant on all measures at $p < 0.01$, 1-tailed t-test).*

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

Method	f	MAiP	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]
CombSUM Fusion	–	**.3396**	.7577	.7273	.6539	.6021
UWFerBM25F2	–	.1854	.6797	.6333	.5006	.4095
$CR^{p}_{kinship}$	3.25-4.5	.2949*	.7357*	$.6971^{+}$	.6580*	.6066*
$CR^{gp}_{kinship}$	3.25-4.5	.3034*	$.7746^{\triangle *}$	$.7308^{\triangle *}$	$.6945^{\triangle *}$	$.6457^{\triangle *}$
$CR^{ggp}_{kinship}$	3.25-4.5	.3158*	$\mathbf{.8125}^{\blacktriangle *}$	$\mathbf{.7552}^{\blacktriangle *}$	$\mathbf{.7145}^{\blacktriangle *}$	$\mathbf{.6572}^{\blacktriangle *}$
$CR^{a}_{kinship}$	3.25-4.5	.3049*	$\mathbf{.8046}^{\blacktriangle *}$	$.7490^{\triangle *}$	$.6993^{\triangle *}$	$.6499^{\triangle *}$

Table E.2: *Ret. performance for focused retrieval $^{\blacktriangle *}$ = stat. significant than both the CombSUM Fusion and UWFerBM25F2 at $p < 0.01$, and $^{\triangle +}$ = stat. significant at $p < 0.05$ respectively.*

Table E.2 and Figure E.2 show the overview of the retrieval performance of our approaches against the baselines for the focused retrieval task. The proposed contextualization model improves the performance over the baselines. The improvements are found to be statistically significant (1-tailed t-test at $p < 0.01$ and $p < 0.05$) on iP and $MAiP$ measures.

The best overall results among the proposed methods are obtained with $CR^{ggp}_{kinship}$ and $CR^{a}_{kinship}$, in terms of best $iP[0.01]$ values (early precision). The kinship context from the hierarchical structure of documents, employed in contextualization, indeed improves the retrieval effectiveness, and the improvements are in-line with theoretical anticipations.

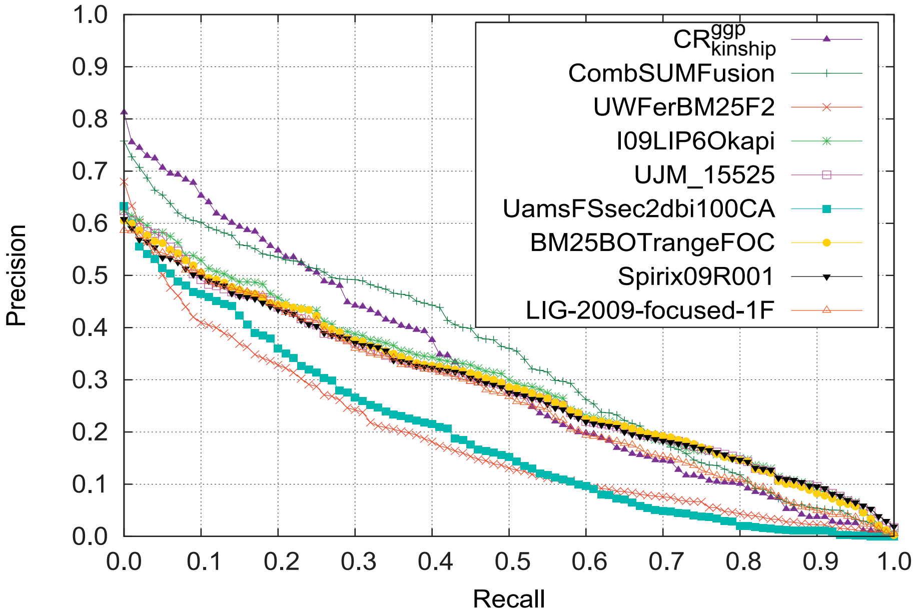


Figure E.2: *Precision - recall performance against baselines and best INEX 2009 submitted runs.*

E.5 Conclusions and Discussion

Contextualization is a re-ranking model, utilizing the context of the relevant retrievable unit, for improving the overall retrieval. We have presented an exploratory study into the use of context from elements in kinship in the hierarchical structure of information, to improve retrieval performance on focused retrieval tasks. We looked at context from document's sequential ordering, which we call the kinship context. Hence, we hypothesized that context gathered from the kinships, “horizontally" and “vertically" from the graph structure of document, influences the retrieval effectiveness. Experiments have validated the hypothesis that utilizing the kinship context this way actually enhances the retrieval of information in focused retrieval task. The results obtained are in-line with the earlier work on contextualization [1, 3, 7, 5, 8, 6]. However, none of the existing works consider the kinship context, as a source of contextual evidence.

The approaches presented are generic and can be applied to different test collections and baseline systems. Evidence are collected in a systematic way, from the surroundings, the kinship context of the element to be ranked. XML documents are used as a sample case of semi-structured documents, these documents have 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.

E.6 References

[1] P. Arvola, M. Junkkari, and J. Kekäläinen. Generalized Contextualization Method for XML Information Retrieval. In *Proc. of the 14th ACM 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 CIKM*, pages 1491–1492. ACM, 2008.

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

[4] S. Geva, J. Kamps, M. Lethonen, R. Schenkel, J. Thom, and A. Trotman. Overview of the inex 2009 ad hoc track. *Focused Retrieval and Evaluation*, pages 4–25, 2010.

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

[6] M. A. Norozi, P. Arvola, and A. P. de Vries. Contextualization using hyperlinks and internal hierarchical structure of wikipedia documents. In *Proc. of the 21st ACM CIKM*, pages 734–743. ACM, 2012.

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

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

[9] 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.

[10] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The 2nd TREC*. Citeseer, 1994.

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

PAPER F

When is the Structural Context Effective?

Muhammad Ali Norozi and Paavo Arvola.

Appeared at the 13th Dutch-Belgian Information Retrieval Workshop, Delft, The Netherlands, 2013.

Abstract: Structural context surrounding the relevant information is intuitively and empirically considered important in information retrieval. Utilizing this context in scoring has improved the retrieval effectiveness. In this study we will objectively look into the significance of the *structural context* in contextualization process, and try to answer the core question of under which circumstances do we need to deal with the such types of context?

F.1 Introduction

Document parts, referred to as elements, have both a hierarchical and a sequential relationship with each other. The hierarchical relationship is a partial order of the elements, which can be represented with a directed acyclic graph, or more precisely, a tree. In the hierarchy of a document, the upper elements form the context of the lower ones. In addition to the hierarchical order, the sequential relationship corresponds to the order of the running text. From this perspective, the context covers the surroundings of an element. An implicit chronological order of a document's text is formed, when the document is read by a user.

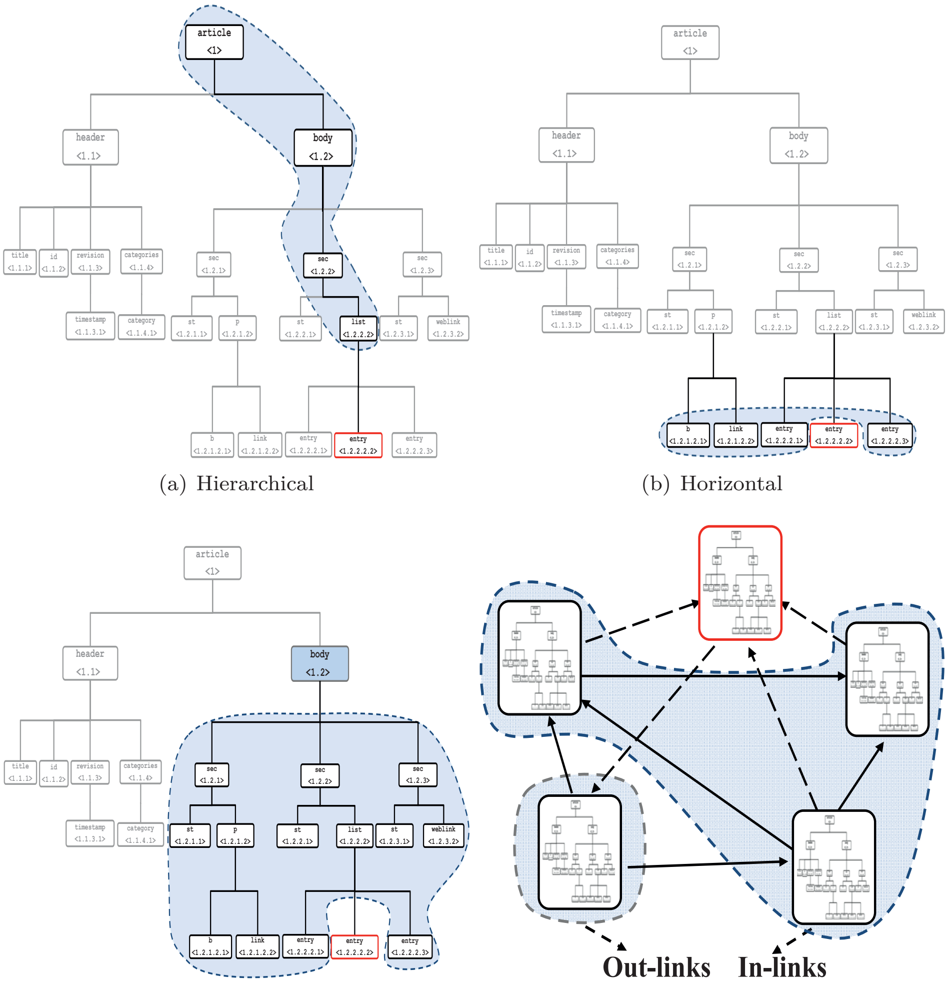
In focused retrieval, the use of context is a driving force to alleviate or "un-bias" the retrieval of items with varying length. Namely, information retrieval is based on evidence of the retrievable units at hand, and longer text units have indeed more textual evidence. This has led to a play-safe strategy where the larger elements are favoured by retrieval systems. How effective the context is to neutralize the side-effects or bias because of size or length (smaller elements with less textual evidence gets same opportunity to satisfy the users need), has been reported experimentally in many studies [1, 2, 3, 9, 6, 10, 8, 7]. The question asked here is: why the structural context is important in the retrieval of focused items? In addition, we also ask if the use of context, under certain circumstances (worst-case), would harm the retrieval. This means if the context is poor or even misleading.

F.2 Context

In semi-structured documents, context of an element covers everything in the document excluding the element itself. The surrounding items or elements of the relevant information is the *context*. The representation of the semi-structured documents aims to follow the established structure of documents, i.e., an academic book is typically composed of ⟨`chapters`⟩, ⟨`sections`⟩, ⟨`subsections`⟩ etc., structures. ⟨`chapter1`⟩ is followed by ⟨`chapter2`⟩ and within ⟨`chapter1`⟩, ⟨`section1`⟩ is followed by ⟨`section2`⟩. Elements ⟨`section1`⟩ and ⟨`section2`⟩ are siblings, and hence most likely, semantically related. The following element takes the concepts further from the preceding elements, and the preceding elements provide the basics or foundation for the following elements. Therefore, together in the document order, the *preceding* and *following* elements form a strong and connected perspective (the kinship structural context), surrounding the relevant information. Two general types of context can be distinguished based on the standard relationships. Hierarchical context, for one, refers to the ancestors, whereas horizontal refers to the preceding and following elements [3]. In existing studies, context has been referred to *externally* as the hyperlink structure of the elements as well. The context is *internal* when it is considered from within the document, and it is external when it is considered outside the document(s).

Contextualization [3] is a re-scoring model, where the basic score, usually obtained from a full-text retrieval model, of a contextualized document or element is re-enforced by the weighted scores of the contextualizing documents or elements (elements in the sub-tree of interest or structural context). In this section, we will formalize the context from in and outside the document using contextualization model.

F.2.1 Structural Context



(a) Hierarchical (b) Horizontal

(c) Kinship (d) Citation

Figure F.1: *Structural context, the sub-tree of interest, example taken from [7]*

Structural context is the *sub-tree of interest* from the hierarchical tree structure of the semi-structured document. *Internally*, in *hierarchical contextualization* [3], the intrinsic tree structure within the XML document is employed. Structural context in hierarchical or vertical contextualization is the context based on parent-child relationship in document's hierarchical structure. An element's parent or ancestors are accounted to be the structural context, while contextualizing the element itself. The sub-tree of interest is shown in Figure F.1(a). *Horizontal contextualization* [3] takes into account the sibling elements in the document's hierarchical structure as the structural context. If we visualize the document's hierarchically tree structure, horizontal structural context is horizontal, as it is based on one level (the same level as the element to be contextualized) of the tree at a time (see Figure F.1(b)). The most recent form of contextualization, the *Kinship contextualization* [7], is both horizontal (siblings) and vertical (ancestors & descendants elements) but intrinsically non-hierarchical perspective of the hierarchical information. Structural context is hence both vertical and horizontal in the document's hierarchical form, Figure F.1(c).

And *externally*, in *citation contextualization* [8], the document's hyperlink structure is taken in to account. The structural context here is based on the hyperlinks' graph of documents hyper-linking (connecting) one another in form of inlinks (indegree) and outlinks (outdegree). In this case, the sub-graphs instead of tree of interest are the out-links graph and the in-links graphs (see Figure F.1(d)).

F.2.2 Why Structural Context?

Structural context is the essential component of the Contextualization model [1]. With contextualization model, using the structural context, the aim is to rank higher an element in a good context (strong evidence in the structural context) than an identical element in a not so good context (less or no evidence in the structural context) within the document. And therefore, retrieve elements independent of their sizes. A small element, in term of size, can be viewed and hence scored in relation to its structural context, and its smaller size (which means having less evidence in total) doesn't stop it from being selected as one of the best results.

In order to cope up with the "biasness" issue (described earlier), in contextualization model, the weight of a relevant element is adjusted by the basic weights of the elements in the structural context (its contextualizing elements). In addition to basic weights, each element in the structural 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 elements in the structural context 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 Equation F.1.

F.2.3 Contextualization and Random Walks

Random walk principle is employed, for contextualization, to induce a similarity structure over the documents based on the containment and reverse-containment relationships (element, sub-element and vice versa). Hence, these relationships affect the weight each element, in the structural context, has in contextualization.

The premise is that *good structural context* (identified by random walk and the contextualization model [7]) provides evidence that an element in focused retrieval is a good candidate to satisfy the user's need and therefore, the elements should be contextualized by the elements in the sub-tree of interest. Hence, the good structural context contains a strong likelihood factor that should be used to deduce that the contextualized element is a good candidate for the posed query.

The tree-structure of the XML document (Figure F.1) is assumed to be a graph. In order for the structural context to take part in the contextualization process, each of the nodes in the sub-tree of interest should possess an impact factor. Conceptually, the impact factor is produced in the following manner: 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 number iterations), 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 structural context of the element to be contextualized, in the form of g function. All the elements, in the structural context of the contextualized element, are considered for contextualization; where the contextualization vector g identifies the importance of each of the unit of the structural context (Equation F.1).

F.2.4 Generalized Combination Functions

The generalized re-ranking combination function based on the random walk principle, which also captures the structural context, can be formally defined as follows:

$$CR(x, f, C_x, g^k) \quad = \quad (1-f) \cdot BS(x) + f \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k(y)}{\sum\limits_{y \in C_x} g^k(y)} \tag{F.1}$$

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 kinship context surrounding the contextualizing element x, i.e., $C_x \subseteq \mathit{structural\_context}(x)$, $\subseteq$, because only the structural context containing the query terms are considered.
- $g^k(y)$ is the generalized contextualization vector based on random walk, which gives the authority weight (the impact) of y, the contextualizing elements (elements in structural context) of x in the sub-tree of interest.

F.3 Effects of Contextualization on different test collections

Structural context in the contextualization framework, is independent of the basic weighting scheme of the elements and it could be applied on the top of any query language, retrieval systems and test collections. The effects of contextualization on different test collections have been observed in the existing studies. Contextualization model has been applied on the top of different and competitive baseline systems using a diverse set of test collections, e.g., semantically annotated Wikipedia collection from INEX 2009[1], IEEE collection, and iSearch scientific collection [3, 7, 8]. In order to get the best possible baseline system, a data fusion was performed based on sum of normalized scores (CombSUM) [11] and Reciprocal Ranking [4] of INEX 2009 submitted runs.

In the experimental evaluation the retrieval effectiveness at different granularity levels were observed. Mainly, retrieval effectiveness at paragraph, article and INEX's focused retrieval level selection has been observed. The approaches were evaluated using the evaluation framework provided by TREC and INEX evaluation initiatives. The reported results were shown to be promising using both TREC and INEX evaluation framework [3, 7].

The focused task in INEX ad-hoc track is to retrieve most focused elements satisfying an information need without overlapping elements. An overlapping result list means that the elements in the result list may have a descendant relationship with each other and share the same text content. For instance, in Figure F.1 the ⟨`entry`⟩ element ⟨`1.2.2.2.1`⟩ and the ⟨`sec`⟩ element ⟨`1.2.2`⟩ are overlapping. In the existing studies, in the focused retrieval task, the INEXs' focused approach is followed, considering a result list where only one of the overlapping elements from each branch is selected. This means that including the ⟨`sec`⟩ element in the results would mean excluding the entry element in the results or vice versa.

Contextualization and the fusion approach as scoring methods, however, do not take any stand on which elements should be selected from each branch. Thus a structural fusion has been performed, where the element level selection is taken from the baseline run and subsequently re-rank the elements of the baseline run.

[1] Wikipedia collection containing 2.66 million semantically annotated XML documents (50, 7 Gb) and 68 related topics provided by the INEX 2009 ad-hoc track [5].

F.3.1 Test Settings

The hierarchical structure of XML documents in the Wiki-pedia 2009 collection, are captured using the dewey encoding scheme (as shown in Figure F.1). 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 at indexing time, as it is described in detail, in our earlier work [7]. The contextualization vector g^k from Equation F.1 is computed off-line for each and every XML document in the Wikipedia collection. This suggests that computing g^k vector is feasible for a reasonably large XML document collections. At the query time, the scores from g^k vector and the basic scores are combined to produce an overall ranking score, using Equation F.1.

In the generalized combination function given (Equation F.1), the contextualization force has to be parametrized. In our earlier work [7], the contextualization force was tuned and reported the values leading to best overall performance. In the parametrization process it was found that the optimal values of contextualization force f (from Equation F.1) lies in the range, ($f \in \{.25,..., 2.50\}$). These optimal values for f are obtained by using cross-validation technique. A 68-fold[2] cross-validation (or complete cross-validation) technique has been performed - 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. These 68 independent or unseen samples are then combined to produce a single or a set of estimations for parameter f.

F.3.2 Query Term Probabilities

If a relevant element does not contain any of the query term(s), it does not match to the query. Hence, in order to retrieve such elements, some expansive methods, such as contextualization, ought to be used. It seems obvious that, in a relevant small element, the probability of occurrence of a query term is smaller than in a larger element. In order to demonstrate this lack of evidence on small elements, we calculated some posteriori probabilities for query term occurrences in a relevant document (R_d) and in a relevant paragraph (R_p, i.e., the relevant $\langle \texttt{p} \rangle$ elements from the XML graph), based on INEX 2009, 68 topics (title field) and their relevance assessments. The probabilities are calculated as the fraction of relevant elements containing any query term, or all query terms over all relevant elements of same kind. The probability of occurrence of any query term (from the query Q) in a R_p

[2] 68, because of the 68 topics from INEX 2009.

and in a R_d respectively are:

$$P\left(\bigcup_{q\in Q} q \middle| R_p\right) = 0.847, \quad P\left(\bigcup_{q\in Q} q \middle| R_d\right) = 0.995$$

This means that the probability of occurrence of none of the query terms in R_p and a R_d is 0.153 and 0.005 respectively[3]. Accordingly, the probabilities of occurrence of all the query terms in R_p and R_d, respectively are:

$$\prod_{q\in Q} P\left(q|R_p\right) = 0.127, \qquad \prod_{q\in Q} P\left(q|R_d\right) = 0.469$$

The difference in the amount of evidence at different granularity levels become even more obvious, when we draw the frequencies of the query terms in this picture. A query term occurs on average 3.4 times in a R_p and 45.4 times in a R_d.

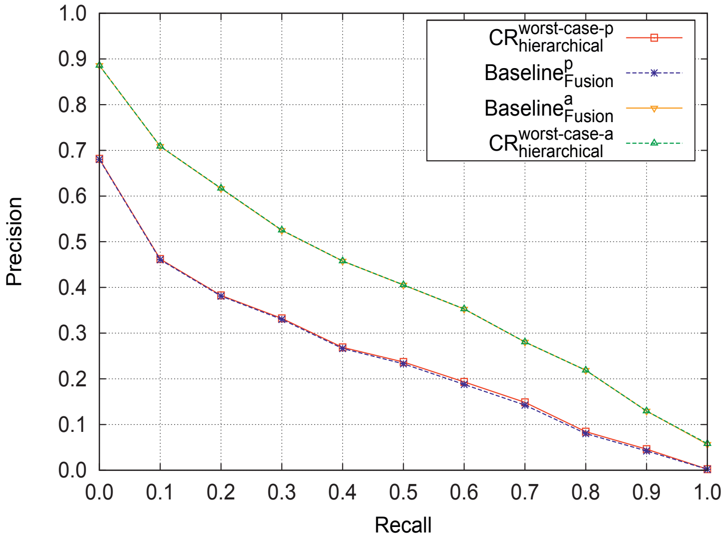


Figure F.2: *Precision - recall, worst-case scenario at article (a) and paragraph (p) granulation and the fusion baseline systems.*

F.4 Worst case analysis

Worst-case for a document d, in contextualization models, means when structural context of element x is chosen such that:

$$\begin{array}{rcl} structural\_context(x) & \notin & elements_y(d) \\ (\forall \text{ elements } y \text{ in document } d & & \text{where } x \text{ and } y \in d) \end{array} \quad \text{(F.2)}$$

[3] Test is performed without stemming or stop-word removal

The *non-structural context* (Equation F.2), should theoretically expose the worst-case effects of the contextualization model. Non-structural context is structural by definition, but physically not in the structural context of element x. How should we interpret the non-structural context, in order to experimentally visualize the worst-case scenario? Instead of taking the actual and true structural context, we randomly select the structural context from another non-relevant but retrieved document. Such a document (retrieved but not relevant) would have misleading evidence (false positive) and hence best suited for the worst-case evaluation. Randomly selecting a document with zero basic score would be trivial and not suitable for our purposes.

By applying this simplistic approach on every element to be contextualized, we can formulate the worst-case scenario. We have used the reciprocal rank fusion approach (fusing 98 INEX 2009 runs) as the baseline system, for worst-case analysis, which has been used before in our earlier work, find further details from [8]:

$$RRScore(e, q) = \sum_{r \in R} \frac{1}{k + rank(r, e, q)} \tag{F.3}$$

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.

Figure F.2 reveals the worst-case depiction of the contextualization model. Not unexpectedly, the worst-case scenario is as good as the baseline system, slightly better but not significant enough to be visible statistically. We can claim here that, when the structural context is chosen randomly (haphazardly), in the worst-case, the contextualization method will not be worse than the basic scoring method.

F.5 Conclusions and Future work

Structural context is the sub-tree of interest, utilized in conjunction with contextualization model, improves the retrieval effectiveness. We have presented an exploratory and theoretical study into the use of structural context from elements in the hierarchical structure of information, to improve retrieval performance. We looked into the structural context from document's hierarchical structure internally, and hyperlinks structure externally. We looked theoretically into the hypothesis that structural context gathered from within the document, “horizontally" and “vertically" using the hierarchical tree structure of document, and from outside, using the hyperlinks graph structure of documents referencing each other, influences the retrieval effectiveness. Worst-case experiments also support the theoretical soundness of contextualization, i.e., if we apply contextualization blindly, in the worst case, we would have as good result as the basic scoring method. The results obtained in this study are in-line with the earlier work on contextualization [1, 3, 9, 6, 10, 7]. In this study we have experimented with semi-artificial data, in the sense that we muddled the context for the worst-case analysis. However, in real data the quality of context varies as well. For example in Wikipedia there are different kinds of pages ranging from listings to topically very coherent documents. In order to get the best results in retrieval, analysing the quality and topical coherency of context would be of great benefit. The analysis of context may be topic dependent, since some queries may have contextual parts. For instance a query: “Losses Belgium in WW2", crave for answers about *Belgium* in the context of *WW2*.

F.6 References

[1] P. Arvola, M. Junkkari, and J. Kekäläinen. Generalized Contextualization Method for XML Information Retrieval. In *Proc. of the 14th ACM 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 CIKM*, pages 1491–1492. ACM, 2008.

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

[4] 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.

[5] S. Geva, J. Kamps, M. Lethonen, R. Schenkel, J. Thom, and A. Trotman. Overview of the INEX 2009 ad-hoc track. *Focused Ret. and Evaluation*, pages 4–25, 2010.

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

[7] M. A. Norozi, P. Arvola, and A. P. de Vries. Contextualization using hyperlinks and internal hierarchical structure of wikipedia documents. In *Proc. of the 21st ACM CIKM*, pages 734–743. ACM, 2012a.

[8] M. A. Norozi, A. P. 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, 2012b.

[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.

[11] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The 2nd TREC*. Citeseer, 1994.

PAPER G

Selection Fusion in Semi-Structured Retrieval

Muhammad Ali Norozi and Paavo Arvola.

Appeared at the 22nd International Conference on Information and Knowledge Management (CIKM), Burlingame, CA, USA, 2013.

Abstract: Semi-structured retrieval aims at providing focused answers to the users queries. A successful retrieval experience in semi-structured environment would mean a satisfactory combination of (a) matching or scoring and (b) selection of appropriate and focused fragments of the text. The need to retrieve items of different sizes arises today with users querying the retrieval systems with varied use case, user interface and screen-size requirements. Which means that different selection scenario serve different requirements and constraints. Hence we propose, a novel type of fusion; the *selection fusion* – a fusion methodology which fuses an all-purpose and comprehensive ranking of elements with a specific selection scheme, and also enables evaluation of the ranking in many selection perspectives. With the standard Wikipedia XML test collection, we are able to demonstrate that a strong and competitive baseline ranking system improves retrieval quality irrespective of the selection criteria. Our baseline ranking system is based on data fusion over the official submitted runs at INEX 2009.

G.1 Introduction

A relevant document is not always completely relevant; instead, the relevant information may be embedded somewhere in the document, available only in the part(s) of the document. Thus, the traditional definition of finding relevant documents applies rather to *document retrieval* than to information retrieval, which refers to finding relevant *information* from the documents [22, 11]. Document retrieval leaves the latter task to the end-users, whereas semi- structured retrieval endeavours to provide direct access to the relevant portion of the document. For semi-structured retrieval documents' logical structure needs to be denoted with a mark-up language, typically in the self-describing form, i.e., the XML. Hence XML Information Retrieval (XML IR) is a type of semi-structured retrieval [2].

Document retrieval delimits a retrieval unit in a simple way. In semi-structured retrieval, the retrievable unit has to be defined according to the user and situation specific requirements. Both the varying information needs and the varying screen-sizes of the devices set requirements for the retrievable units of the information to be returned. We refer to this as the *granularity level* of the retrieval units.

In semi-structured retrieval, the notion of selectivity has an important role to play [1, 9]. The granularity level (specific part of the document - element) at which we want to present or user would like to see the results. The users in this domain are not interested in the whole document, as a search outcome, but rather in the most specific portion of document where the relevant information can be found. Thus, there are two essential tasks in XML IR: (1) the ranking of the retrieval units according to the relevance scores and (2) the selection of the appropriate granularity level or element type [6, 29, 23]. The standard evaluation of XML retrieval forces system developers to take both of the tasks into account simultaneously. This means that a good ranking performs poorly, if the selection is not successful. Unfortunately, any of the contemporary metrics do not reveal what actually went wrong in that very common case.

Hence, there should be a method to study and develop the ranking of XML retrieval as it own case (independently) and then this ranking with different selection schemata. In this study, we suggest that ranking and selection can be successfully combined using two distinct systems for these tasks by performing a *selection fusion.* In a nutshell, in selection fusion, the following two component systems are fused together:

- Ranking / scoring system (the *structural fusion*, in this study).
- Selection scheme (INEX 2009 submitted runs, in this study).

The *ranking system*[1] ranks all elements of the collection (the comprehensive list) according to their estimated relevance, and the *selection scheme* removes all structurally inadequate or not needed elements from that comprehensive list. In the

[1] Ranking system is analogous to the scoring system, they are used interchangeably in this study.

experiments, we are able to show that a good ranking system performs well with any selection scheme. Our baseline ranking system is based on simplistic and primitive fusion methods, such as, CombSUM [30], reciprocal ranking [8] and structural contextualization [25]. Finally, the selection fusion for the ranking is done with the different runs of INEX 2009 as selection systems, one at a time.

Our scoring system, which is based on the fusion methods [30, 24, 8] and the structural contextualization [25, 26], provides a firm basis for a mechanism or a methodology which makes possible the retrieval of items with varying length and specificity, and an above average performance.

In the *structural fusion*, a relevant but not sufficiently large-enough or deep element, in a semi-structured document, is boosted by the evidence lying in its structural surroundings [6, 25]. In the opposite case, the structurally surrounding elements should degrade the importance of the non-relevant element.

The hypothesis is that a sufficiently good-enough scoring system which is also capable of retrieving a *comprehensive* set of items on the query topic(s) (the structural fusion, in this study), blended (fused) with a selection scheme (INEX 2009 submitted run – itself a semi-structured retrieval result, in this study), improves the retrieval effectiveness of the selection scheme, independent of the selection scheme.

A selection criteria or scheme, for example, which craves to retrieve only items at the ⟨`paragraph`⟩ granularity levels (deep and thorough elements), based on our propositions in this study, should get improved retrieval after the selection fusion methodology. Similarly, another example selection criteria, requiring only ⟨`article`⟩ level elements (larger results), should as well be improved as a result of selection fusion methodology. The selection criteria could be a mix of different granularity levels; it could be taken from the users personalization settings; from system's pre- or post-defined conventions; or it could be formed as a result of a particular domain or system specific constraints.

The proposed selection fusion methodology, in this study, is experimentally applied to the semantically annotated Wiki-pedia XML collection using the INEX[2] [28] evaluation measures and test-bed. The multiple and diverse set of submitted runs at INEX 2009 are fused together using the known fusion methods [30, 8] and then structurally contextualized [25]. The results obtained because of selection fusion methodology, from the ad-hoc track (INEX) on all the four tasks (i) focused, (ii) thorough, (iii) relevant in context (RiC) and (iv) best in context (BiC), measured as focused / semi-structured retrieval, exhibit clear improvements over the submitted runs at INEX 2009, and over a strong and competitive baseline system – data fusion over all INEX 2009 submitted runs (Section G.3).

Summarizing, the contributions of this study include:

[2] INEX (Initiative for the Evaluation of XML retrieval) 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 at INEX.

1. Creating decent scoring system using structural fusion – based on the primitive, simplistic data fusion methods and structural contextualization (Section G.2) and applying that system for selection fusion.
2. Developing selection fusion methodology - simple and effective yet flexible fusion approach (Section G.3).
3. Construction of a test setting for evaluating the retrieval of focused items (Section G.4).
4. Experimentally evaluating the competitive scoring system (the structural fusion) using varying selection scenarios (INEX 2009 submitted runs), the selection fusion, on the large semantically annotated Wikipedia XML corpora at INEX [28] (Section G.4.3).

Section G.5 concludes and highlights the future work.

G.2 Semi-structured Retrieval

```
<A>
        Text A1
        <B> Text B
                <D> Text D  </D>
                <E> Text E  </E>
        </B>
         Text A2
         <C> Text C
                <F> Text F  </F>
                <G> Text G  </G>
        </C>
        Text A3
</A>
```

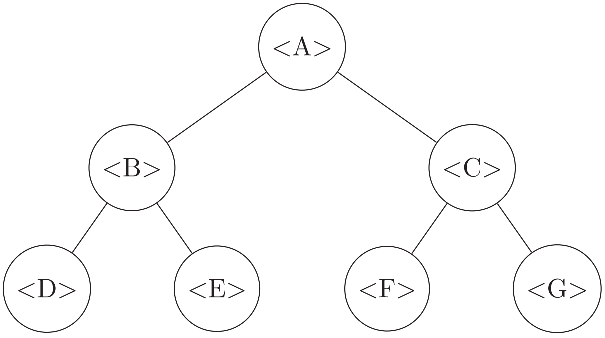


Figure G.1: *Sample semi-structured data and its tree representation.*

Information Retrieval (IR) is about finding relevant material of an *unstructured nature* (typically text). The notion of "unstructured-ness" in the retrieved material / item refers to the distinction between structured data in the databases and unstructured text in the documents – considering the latter to be the focus of IR. However, many text documents, such as newspaper articles or books, have a well-defined structure consisting of coherent blocks or parts such as, titles, paragraphs, sections, and so on. These coherent chunks of document parts form the retrievable units (or the logical documents – a document would contain many logical documents) of semi structured retrieval. For digital data representation, storage, manipulation, and also for data semantics, the structure of a document is often presented using the mark-up language, typically XML or likes. Thus, XML retrieval is semi-structured in nature, and hence in an XML document, the coherent parts are referred to as *elements*.

G.2.1 Ad-hoc Retrieval tasks

Measuring semi-structured retrieval has a relatively long history in IR research. An essential part of the research carried out in the evaluation and standardization of XML IR has been done at the yearly INEX (Initiative for the Evaluation of Xml retrieval) workshop since 2002 [10]. The initiative also offers standard collections of documents (mostly or entirely in XML), with a set of query topics (per collection per track) and corresponding relevance assessments, as well as various evaluation metrics and toolkits for the semi-structured retrieval [3].

Within the INEX ad-hoc track, several tasks are proposed, tested and discussed over the time for being representative use cases of semi-structured retrieval [19, 32]. One of the common perspectives from most of the state of the art use cases in XML retrieval is; “how are the results organized?”. The ability to describe the relevant retrievable content within the document with the metadata (structure) which is a part of the markup (in the form of tags around the content). The results organization defines the interpretation (may be semantically) of the enclosed content, to be presented to the users. The results organization at INEX 2009 has been done in the following ways:

1. element (fine grained) retrieval tasks: thorough and focused, and
2. in-Context retrieval tasks: Relevant-in-Context (RiC) and Best-in-Context (BiC).

Thorough and Focused tasks

The *element retrieval* tasks differ in the way they treat the overlap in the result list [18]. The feature of elements being part of and containing other elements is called overlapping, or nested elements. In other words, in overlapping and nested results, the same text fragment may appear several times in the results list. For

[3] It is worth noting that INEX result lists may contain arbitrary passages in results. However, vast majority of the runs retrieve elements only, and in this paper we focus only on the semi-structured retrieval task (i.e., elements retrieval).

example, in Figure G.1, element A is overlapping with all other elements. In addition, B is overlapping with D and E, and C with F and G. In the *thorough task*, elements in the result list may be overlapping, whereas in *focused task* the elements in the result list should not overlap. The thorough task is considered system oriented, while there are no user interface nor user-related assumptions underlying the task [7]. The task is to rank elements based on their relevance with overlap. The comprehensive result list with overlapping elements does not remove the challenge, apart from relevance, deciding which (kind / size) of elements should be preferred in the results.

In focused task, however, only one of the overlapping elements should be selected in the final result list, i.e., one element from each branch. For example, in Figure G.1, from elements A and B, either A or B should be selected in the result list, not both (means overlap). The most straightforward way to perform this selection is to calculate a score for individual elements as if they were documents (logical documents) and filter the resulting comprehensive ranked list to remove overlap (see [13]). A typical strategy is, to select the elements on the basis of their scores, so that among the overlapping elements, the element with the best relevance score is selected. We call this score based selection as the *first-come first-choice* (FCFC) selecting principle. The FCFC principle is a widely used, but a rather straightforward technique, having no pre-defined and / or intellectual considerations, for example, one based on screen-size requirements, user preferences or any other constraints.

In-Context tasks

In addition to handling the overlap, two different approaches are also modelled at INEX; characterizing how the results ought to be grouped. In the focused and thorough tasks, the users prefer a single element that is relevant to the query, while in the so-called in-Context retrieval tasks, the users are more interested in elements within the relevant articles – they want to see parts of the (highly) relevant document that satisfy their information need most effectively [7, 16]. With in-Context retrieval tasks, the ideal user is interested only in elements within the highly relevant articles – they want to see which parts of the highly relevant document will best satisfy their information need. In RiC task, all the relevant, non-overlapping elements are sought together, whereas in BiC task, the focus is only on the best entry point. In RiC the challenge is to find appropriate amount of elements strictly with no overlaps. For the selectivity (as described earlier), FCFC or any other selection method could be used. Further details are found in [7, 16].

G.2.2 Elements Size and Type

Traditionally, the aim of the semi-structured retrieval system primarily is, to retrieve (1) the relevant elements, which are (2) at the right level of granularity. In other words, only those elements should be answer to the user query which are; as focused as possible, while still covering the users' query topic(s). In a nutshell, the evaluation of XML / semi-structured retrieval system, addresses the combination of (a) the scoring quality, i.e., how well a system scores the relevant elements, and (b) the selection quality, which means the selection of the right granularity level of elements, for the required task and of appropriate size (as hinted earlier as well). The appropriate size or type is dependent on relevance as well as many application and user specific attributes.

Thus, the right granularity level, i.e., element size / type is very much subject to different use cases and user interfaces [17, 25]. For example, the need for a shorter excerpt of information as a result-set, is felt when a user is querying the system using a device with screen-size constraints, e.g., from a smart-phone (which is quite common, nowadays). In a laptop device the relevant content is more easily obtained without automatic search by skim reading or using some simple interface gadget such as, find on page. Also, many other different use cases require retrieval of elements with various granularities. For a snippet retrieval (e.g., for result list presentation) or fragment search, a small element is more suitable than a whole section or full article, or for example a user interested in abstracts of search results, to skim through them before opening the whole document itself. One could imagine a number of other such use cases and user interface scenarios. However, apart from the tasks, measuring the performance follows rather a one-size-fits-all principle, where system developers have to guess, what is the correct granularity level appropriate for the metrics involved in the evaluation [17].

G.2.3 Scoring and Selection Systems

Based on the above considerations, we argue that *selection* and the *scoring* can be considered as two different but related tasks and hence two distinct systems can be employed separately for these two tasks. The selection result list may contain only the (unranked or poorly ranked) elements, that are required, for example, because of certain type / size constraint or for a particular use case scenario. While the scoring result list is a comprehensive and properly ranked result list containing diverse, overlapping list of elements. Intuitively, one way of combining the scoring and the selection is by fusing the selection scheme of one system with the scoring (or ranking) ability of another. In other words, the elements provided by the selection result list are (re) ranked by the ranking or scoring result list. In the Section G.3 and G.3.2, we have defined our scoring and selection systems, for this study, and subsequently applied the selection fusion methodology on them. The results of selection fusion on all the 98 submitted runs at INEX 2009 are discussed in Section G.4.3 and the improvements are graphically shown in Figures G.2 and G.2, for all the 4 ad-hoc track tasks described in this section.

In the next section we have outlined the dichotomy of the fusion in the semi-structured information retrieval settings.

G.3 Dichotomy of Fusion in Semi-structured retrieval

Data fusion combines two or more retrieval results and has been shown to be effective in full document retrieval – better than the individual component systems [30, 8, 24]. In this study, we experiment with the fusion of semi-structured retrieval results from a set of individual and diverse component systems, where the retrievable units are treated as if they are documents. The outcome of such a fusion is a comprehensive list of elements treated as the ranking system (Section G.3.1). In other words, the fusion methods from document retrieval are applied directly to the semi-structured retrieval results without taking into account the features, such as the selectivity strategy (the granularity level). As this study purposefully separates the ranking and the selection systems – fusion in semi-structured retrieval is also *dichotomized* in a similar fashion. The fusion methods responsible for the ranking system, and the *selection* fusion methods accountable for the selection system. More specifically we have:

1. Fusion for the ranking system.
2. Fusion for the selection system (Selection fusion).

G.3.1 Fusion for the Ranking System

Document retrieval fusion methods (rank-based and score-based) can be applied to semi-structured retrieval results as well, when considering the results as a flat list. For the scoring or ranking system in semi-structured retrieval settings, there is yet another family of fusion models – fusion methods utilizing the structural or hierarchical relationships of elements in the re-scoring process.

Rank based methods

There is an analogy between the fusion method based on rank and the social voting system [24, 20]. The rank of a particular retrievable item is decided based on how many different systems vote (in one or another form, e.g., in *majoritarian* view) for this particular retrievable item to be ranked to a particular rank (position in the result list). We have tested, in several intuitive settings, earlier [25, 26], the following rank-based fusion methods:

- Borda Count
- Condorcet
- Reciprocal Ranking

Reciprocal Rank fusion [8] is found to be well suited with our current theoretical and experimental settings (see Section G.4.1). Hence, in the rank based fusion methods, we have reported only the results from reciprocal rank fusion (Section G.4).

Score based methods

A fusion method which is concentrated on combining search results based on the *similarity values* of individual retrievable items, in the semi-structured retrieval, for each query topic and from a set of varied runs (employing presumably different retrieval models). In the fusion based on scores (similarity values), we have explored some of the state of the art fusion methods, for combining the similarity scores from the set of runs used in this study:

- CombANZ
- CombMNX
- CombSUM

We have found (in conjunction to what has also been reported by Shaw et. al [30]) that simply combining the similarity values in a linear fashion, summing the similarity values, works better than trying to select a given similarity value. This fusion method is named as the CombSUM fusion [30]. CombSUM fusion method is therefore chosen, in the experimental evaluations, to fuse only the set of runs containing similarity scores in their result lists.

Structure based methods

Here, we propose a type of fusion, which uses the *structural features* in the semi-structured documents, for *structurally* fusing together a set of different semi-structured retrieval systems, using the hierarchical structure (relationships) of the elements in the semi-structured documents. By doing that, the aim is to produce an extensive (covering the query topics noticeably) and effective (good ranking outcomes) results list. The structural features in semi-structured documents could possibly originate from the *structural context* (elements in the structural kinship [26]) of the relevant item (Section G.3.1) or from the selection methodology described in the Section G.3.2.

Structural context [26] is at the core of this type of fusion. Structural fusion is done in two steps:

1. The results from different semi-structured retrieval systems are first fused together based on their ranks or scores (if scores are available). Described in detail in Section G.4.
2. The fused results are then structurally contextualized [25] – which could also be considered as the *internal fusion* of the elements within a result list with their respective structural contexts. Here, the aim is to rank higher an element in a good structural context (strong evidence in the structural context) than an identical element in a not so good context (less or no evidence in the structural context), within the document. In the hierarchical and tree structure of the semi-structured documents (e.g., Figure G.1), the structural context comes from either (i) the ancestor-descendant relationships [3, 27, 31],

(ii) the sibling relationships [6] or (iii) the elements which are in the hierarchically kinship relationships [26]. In this study, we take the idea (i) for the structural context (ancestor-descendant relationships).

This way we would be able to get results which are both comprehensive (containing thoroughly almost all the relevant elements on the query topics) and at the same time highly relevant as well [25].

With structural fusion, we are able to retrieve elements independent of their sizes (or independent of granularity levels). A small element, in term of size, can be viewed and hence scored in relation to its structural context, and its smaller size (which means having less textual evidence) doesn't stop it from being selected as one of the top ranked results. The weight or score of a relevant and / or small element is adjusted (re-enforced) by the basic weights (content based weight) of the elements in its structural context (its contextualizing elements in hierarchy). In addition to basic weights, each element in the structural context of the contextualized element, should possess an *impact* factor [25]. An higher impact factor shows the importance of the contextualizing element and vice versa.

G.3.2 Selection Fusion

Selection fusion is also a fusion method based on structure, as the name suggests, but intended for the selection process instead ranking or scoring the semi-structured result sets. The structurally fused and structurally contextualized results; the scoring system, is *selection* fused with one particular semi-structured retrieval system. The reason for selection fusing the scoring and the selection system is to choose the selectivity of the overlapping elements of the scoring system from the selection system. Hence, the scoring of the *focused* elements come from the scoring system and the selectivity comes from the selection scheme. In a nutshell, we would like to retrieve highly relevant set of elements with controlled granularity levels (selectivity scheme).

A user querying the semi-structured retrieval system from a handheld device, the selection criteria in this case (should) take into account the limitations of the results presentation – they should preferably fit-in the limited screen-size – which in this example means that, they should be small and focused enough to satisfy the device requirements and users' needs. Therefore we name this type of fusion as the *Selection fusion.* A toy example for the use of selection fusion is shown in the Table G.1, with illustration in the caption. The selection fusion results in the Table G.1 are based on the ranking order $<C, A, G, F, B, D, E>$, which is assumed to be the result of the scoring system.

Selection fusion approach formulated in this way, is flexible and independent, any selection scheme and any retrieval or scoring systems could be applied. We have applied the selection methodology to a set of semi-structured retrieval systems, the INEX 2009 officially submitted runs by the participants, and got an overall steady and statistically significant results over most of the retrieval systems and ad-hoc retrieval tasks (as described in Section G.2.1) at INEX 2009, measured as focused retrieval (see Figures G.2 and G.2).

Table G.1: *Selection fusion ranks above are based on the ranking order (for this example, is assumed to be): <C, A, G, F, B, D, E>. The selection scenarios cover all combinations of the example in Figure G.1, for the focused task. FCFC method delivers C, B.*

Selection schemes	Selection fusion results
{A}	<A>
{B, C}	<C, B>
{B, F, G}	<G, F, B>
{D, E, C}	<C, D, E>
{D, E, F, G}	<G, F, D, E>

In addition to the selection fusion methodology described above, there is another type of selection fusion as well, which is based on the selection fusion of ⟨`article`⟩ and element runs. In the in-Context tasks (RiC and BiC, Section G.2.1), the articles are selected first and thereafter the elements within the article are selected. Thus, the two systems can be used for another type of fusion – document retrieval system fused with element retrieval system [14]. In this study, we are not considering this type of selection fusion.

G.4 Experimental Evaluations and Test Settings

In this section, the proposed ideas presented in this study are empirically tested and the results are analysed in light of the posed hypothesis and the theoretical foundations established. Rest of the section is organized as follows; in Section G.4.1, we define the ranking system for experimentations; Section G.4.2, we lay down the test settings in the semi-structured environment and Section G.4.3, we interpret and assess the experimental outcomes.

G.4.1 The Ranking System

Given the English Wikipedia test collection [12, 28], containing 2.66 million semantically annotated XML documents (50.7 Gb), 68 related topics, and 98 submitted runs [4], provided by the INEX 2009 ad-hoc track [9]; we performed a data fusion

[4] A total of 173 runs were submitted by participants to INEX 2009. 13 runs were not element runs, i.e. they contained ranges of fragments or file-offset-lengths (FOL) as retrievable units and

based on sum of normalized scores (CombSUM) [30] and reciprocal rank fusion [8]. The reason for using the INEX 2009 test topics (instead of 2010) is the larger variety of elements in the participants' results which was primarily due to the existence of the thorough task.

The element scores (for each runs, per topic) were normalized for the fusion based on scores (CombSUM, Section G.3.1) as follows:

$$score_x = \frac{score_x - min(scores)}{max(scores) - min(scores)} \tag{G.1}$$

where, $max(scores)$ and $min(scores)$ denote the maximal and minimal scores respectively.

Although, the INEX 2009 runs have the largest variety in the results for the fusion – in comparison to other years of the initiative, most of the participants (56 of them), unfortunately, did not report any real element scores, just the ranking orders (without relevance scores), because the INEX evaluation toolkit did not require that information. For those systems (without scores), an artificial score was computed for each element, based on their reciprocal rank, before applying the normalization function, Equation G.1. For the reciprocal rank fusion, a score for an element e is calculated as follows:

$$RRScore(e, q) = \sum_{r \in R} \frac{1}{k + rank(r, e, q)} \tag{G.2}$$

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. Value of the parameter k is considered to be in the range [0, 5], based on our earlier findings [25].

In other words, as a result of Equation G.2, when $k = 0$, the top ranked element in the result list is given a score of 1, the second ranked gets 1/2, third 1/3, fourth 1/4 and so on.

First, we apply the reciprocal rank fusion on the 56 runs (without scores), and then we apply the CombSUM fusion on the overall 98 runs, the resultant fusion run is named as the CombSUM_Reciprocal fusion. Finally, the ComSUM_ Reciprocal fusion is structurally contextualized using the combination function, from our earlier work [25]:

were omitted from the fusion. In addition, in order to avoid noise, we made a deliberate 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 and thorough) of the ad-hoc track were used in fusion.

$$CR(x, f, C_x, g^k) \quad = \quad BS(x) + f \cdot \frac{\sum\limits_{y \in C_x} BS(y) \cdot g^k(y)}{\sum\limits_{y \in C_x} g^k(y)} \tag{G.3}$$

where

- $BS(x)$ is the basic score of contextualized element x (text-based score, e.g., $tf_e \cdot ief$[5])
- f is a parameter which determines the weight of the context in the overall scoring. The optimal values of contextualization force f lies in the range: ($f \in \{1.0,\ 1.25,\ 1.50,\ 2.75,\ 3.00,\ 3.25,\ 3.50\}$) (using 68-fold cross-validation technique) [25, 26].
- C_x is the context surrounding the contextualizing element x, i.e., $C_x \subseteq$ *structural_ context*(x), $\subseteq$, because only the structural context containing the query terms are considered.
- $g^k(y)$ is the generalized contextualization vector based on random walk principle, which gives the authority weight of y, the contextualizing elements (ancestors) of x in XML graph (for details see [25]).

The final result-set constitute the ranking system – the structural fusion results (or CR – Contextualization Re-rank). For each of the 98 submitted runs (the selection run); we take the selectivity scheme from the selection run, while scoring is taken from the structural fusion run (CR), the ranking system. The selection fusion, for the rest of experiments, is the fusion of the ranking system formulated here with a particular selection run or system.

Figures G.2 draw pictorial overview of the effects of selection fusion (fusion of the ranking system above, and the selection system) on each of the 98 INEX 2009 submitted runs (the selection runs). Each of the points (blue and orange) on the Figures G.2 represent the gain / loss effects of selection fusion methodology (gain if above and loss if below the red diagonal line) on each of the participating (task-wise) semi-structured retrieval systems.

G.4.2 Test Settings

The effectiveness of selection fusion methodology is experimentally analysed using the INEX 2009 evaluation test-bed. We have conducted retrieval experiments within the ad-hoc track at INEX, featured by four (4) tasks (focused, thorough, RiC and BiC) and measured them as focused (semi-structured) retrieval. The results are pairwise (the selection fusion run and the selection run) compared, evaluated and reported, see Figures G.2.

[5] tf_e term frequency and ief inverse element frequency at focused granularities.

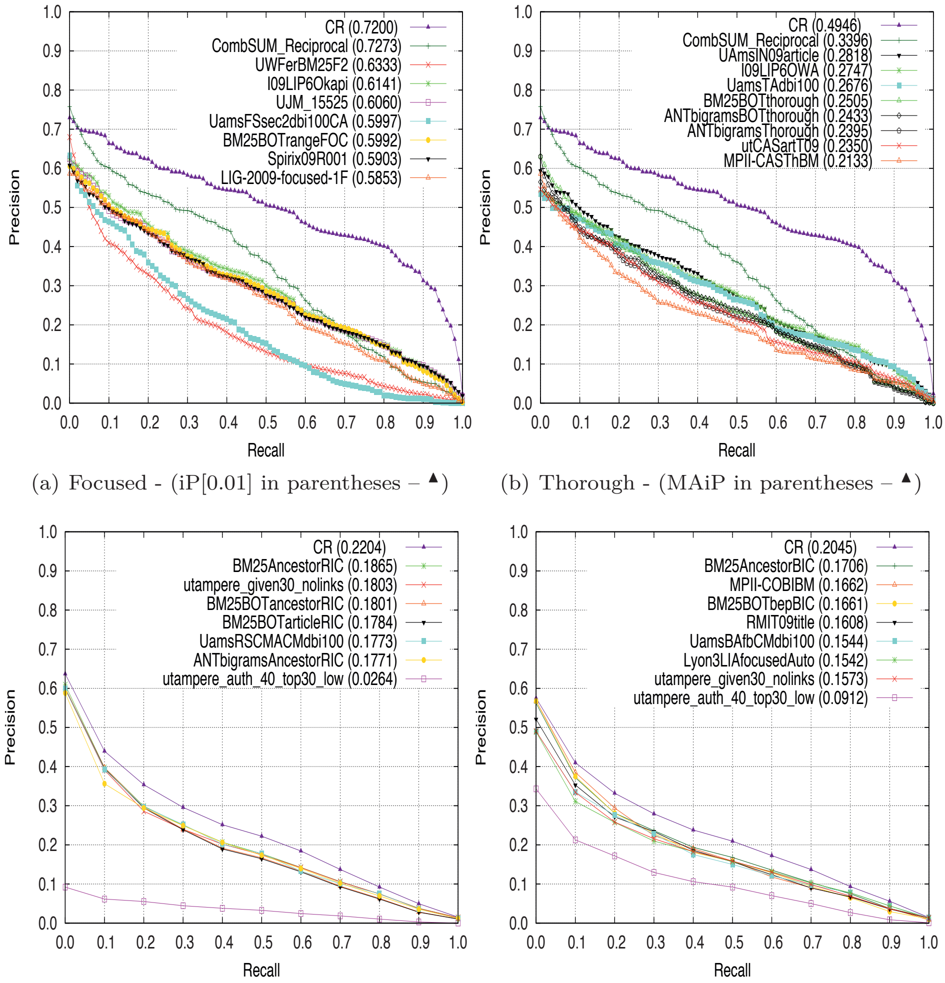
We report the improvements for Mean Average interpolated Precision (MAiP), interpolated Precision at interpolated Recalls (iP[@iRecall]), Mean Average generalized Precision (MAgP) provided by the INEX evaluation toolkit, and the precision-recall curves. In addition, we calculate the specificity of the results in terms of how deep in the XML hierarchy the retrieved elements are on average, calculated over topics. We call this feature of the result set as Mean Average element Depth (MAeD). The depth of an element can be determined, for example, by calculating the number of slashes in its path expression.

Table G.2: *Comparison between Mean Average interpolated Precision and Mean Average element Depth (>1) of INEX 2009 submitted runs, with correlation coefficient -0.4046, over all the runs including the ones with MAeD=1. Bold-face entries are top ranked runs at INEX 2009 thorough task.*

Run ID	MAiP	MAeD	**Run ID**	MAiP	MAeD
UAmsIN09article	**0.2818**	**1.0**	utampere_given30_nolinks_low	0.0503	3.4
I09LIP6OWA	**0.2747**	**1.0**	UamsRSCWartCWdocbi100	0.2132	4.3
UamsTAbi100	**0.2676**	**1.0**	ANTbigramsThorough	0.2395	4.7
utCASartT09	**0.2350**	**1.0**	**ANTbigramsBOTthorough**	**0.2433**	**4.7**
ANTbigramsFOC	0.2721	2.2	**BM25BOTthorough**	**0.2505**	**5.3**
ANTbigramsRIC	0.2721	2.2	BM25thorough	0.2585	5.3
ANTbigramsBOTFOC	0.2740	2.2	UWFERBase2	0.2489	5.4
ANTbigramsBOTRIC	0.2740	2.2	doshisha09f	0.0093	5.4
BM25FOC	0.2920	2.4	MPII-COThRF	0.1445	5.7
BM25RIC	0.2920	2.4	UAmsIN09section	0.1429	6.2
BM25BOTFOC	0.2822	2.4	UamsTSbi100	0.1712	6.3
BM25BOTRIC	0.2822	2.4	UamsFSdocbi100	0.1727	6.3
emse2009-151	0.1114	2.6	UamsFSsec2docbi100	0.1928	6.3
emse2009-150	0.1470	2.7	**MPII-COThBM**	**0.2079**	**6.7**
emse2009-152	0.0968	2.7	MPII-COFoBM	0.1973	7.1
emse2009-153	0.1389	2.8	UWFERBM25F2	0.1854	7.2
MPII-CASFoBM	0.2128	3.3	utampere_auth_40_top10_low	0.0057	10.5
MPII-CASThBM	**0.2133**	**3.3**	utampere_auth_40_top30_low	0.0057	10.5

In the focused retrieval tasks in the ad-hoc track of INEX, the aim is to retrieve the most focused elements satisfying an information need without overlapping elements. An overlapping result list means that the elements in the result list may have an ancestor-descendant relationship with each other and therefore share the same text content. For instance, the ⟨`sec`⟩ element within the ⟨`article`⟩ element, is overlapping with the parent ⟨`article`⟩ element (because they are nested). In this study, we are following the focused approach, considering a result list where only one of the overlapping elements from each branch is selected. This means that including the ⟨`article`⟩ element would mean excluding the ⟨`sec`⟩ element in the result-set or vice versa, depending on which selectivity scheme is used.

The fused result list contains all the elements delivered by the 98 component systems. This comprehensive result list obviously (by definition) contains overlapping elements. In order to remove the overlap, one intuitive solution is to select the



(a) Focused - (iP[0.01] in parentheses – $^{\blacktriangle}$)

(b) Thorough - (MAiP in parentheses – $^{\blacktriangle}$)

(c) Relevant in Context - (MAgP in parentheses – $^{\triangle}$)

(d) Best in Context - (MAgP in parentheses – $^{\triangle}$)

Figure G.2: *Precision recall curves for all four INEX 2009 focused retrieval tasks. On CR and CombSUM runs, FCFC is applied after pruning the* ⟨`article`⟩ *elements. Where* ($^{\blacktriangle}$) *denotes statistical significance at* $p < 0.01$ *and* ($^{\triangle}$) *stat. significance at* $p < 0.05$ *(1-tailed, t-test).*

elements having highest score from each branch, the FCFC approach, as described in Section G.2. In Figures G.2(a-d), the FCFC approach is used to prune the overlapping elements from the CR and CombSUM_Reciprocal runs. The purpose of sketching Figures G.2(a-d), is to demonstrate the effectiveness of the baseline scoring systems CR and CombSUM_Reciprocal against the best INEX 2009 submitted runs. The scoring systems are clearly and visibly better than the best reported runs at INEX 2009, in all the four tasks. The improvements are always statistically significant at $p < 0.01$ and $p < 0.05$ in pairwise comparisons, as indicated in the Figure G.2 caption. As a result of the findings from Figure G.2, now the CR run is capable to be used in the selection fusion as the scoring system.

Many participants, however, returned runs containing only full-article results, which therefore led to a full-article bias in the fusion results, as can be visually seen in the Figures G.2, orange points with MAeD=1, are in majority.

The selection fusion approach do not take any stand on which elements should be selected from each branch. It provides a comprehensive set of highly relevant items (CR run) and a selection possibility. In the selection fusion process, we therefore have to perform a structural fusion, where we take the element-level selections from each of the 98 submitted runs at INEX 2009, one by one, and subsequently re-rank them. For instance, if an INEX submitted run suggests the ⟨`body`⟩ element of document d as the top (1^{st}) ranked result for query q, while the structural fusion (CR) rank it at 5^{th} position, the ⟨`body`⟩ element would eventually be selected to the 5^{th} rank position and so on.

G.4.3 Experimental Findings

For the *focused* task a ranked-list of non-overlapping results (elements or passages) must be returned [9]. It is evaluated at the early precision, interpolated Precision at 0.01 interpolated recall (iP[0.01]) measure. As it is visible from the Figures G.2(b) and G.2(a-d), in the focused task, the early precision values are improved notably both when FCFC selection approach (Figure G.2(b)) is used and when selection fusion is applied (Figures G.2(a-d)), respectively. Especially in Figure G.2(b), for the runs with Mean Average element Depth (MAeD) > 1, the improvements in iP[0.01] values are more than the systems with MAeD $= 1$ (orange points without numbers).

For the *thorough* task a ranked-list of results (elements or passages) by estimated relevance must be returned [9]. It is evaluated by Mean Average interpolated Precision (MAiP) measure. The comparison of MAeD and MAiP, for this task, is shown in Table G.2. The correlation coefficient value of -0.4046, characterizes an inverse relationship between the two measures. A higher value of MAeD means more deep elements (result list with more hierarchies) or focused results, and a negative correlation value implies that, INEX evaluation metrics penalizes the MAiP scores of the more focused runs (with MAeD > 1). It has also been observed that the top ranked retrieval systems were mostly those runs retrieving only the ⟨`article`⟩ elements (having MAeD $= 1$, see also Table G.2, bold-faced entries) [4]. As the runs contain thorough results, the margin of improvements in those runs with MAeD > 1 is found to be larger, as it is depicted in the Figures G.2(f-h), runs with MAeD > 1 are in the top right corners (best improvements). Because of FCFC selection approach, the behaviour of CR run in Figure G.2(d) is no different than Figure G.2(b).

For the *RiC* task, non-overlapping results (elements or passages) must be returned, these are grouped by documents. It is evaluated by Mean Average generalized Precision (MAgP) measure, where the generalized score per article is based on the retrieved highlighted text [9]. The overall improvement in this task is observed to be not as significant as it was in the focused and thorough tasks (as it is also evident from the Figures G.2(a) and G.2(i)). The primary reason could be attributed to the complexity of the evaluation metrics [9]. The other reason could be, as the Figure G.2(j) also indicate, that most of the runs are ⟨`article`⟩ or document retrieval runs (MAeD=1), which is because of the definition of the task – the results are grouped by documents. Therefore, the room for improvements, in focused retrieval, was minimal in this task. The overall MAgP values of the top runs are also pretty low, which could again be attributed to the aforementioned challenges in this task.

For the *BiC* task, a single starting point (element's starting tag or passage offset) per article must be returned. It is also evaluated by Mean Average generalized Precision (MAgP) measure, but with the generalized score (per article) based on the distance to the assessor's best-entry point [9]. A similar reasoning as that of RiC task is applicable here as well, the runs are mostly document retrieval runs, the MAgP overall are very low, and the metric is complex [9]. The margin of improvements was low because almost all of the runs having MAeD=1, i.e., document retrieval.

Selection fusion is directly applicable for focused, thorough, RiC, BiC tasks as well as Content-and-Structure tasks (i.e., CAS) queries [21], where the required elements are explicitly expressed by path expressions, or the path expressions are used as mere structural hints of the possible location of the information needed [15]. However, in the RiC task, one has to decide the order of the documents first. According to Kamps et. al [14], the article run determines the article ranking best. This means that, in the comprehensive ranking list the existence of the root element (⟨`article`⟩ node) determines the article ranking.

The overall improvements in all the four tasks are found to be extraordinary. What makes the overall methodology work, in most of the cases, could be accredited to the comprehensiveness and the flexibility of the selection fusion method (the ability to cover the query topic exhaustively). It is comprehensive both on the documents and elements levels, which is essentially due to the large variety of documents and deep elements in the participants runs, which were then structurally fused (using structural contextualization). On elements level, the structural fusion help to provide a comprehensive set of focused answers, using the structural context. These comprehensively focused and highly relevant set (with good relevance scores) [25] of answer-set help the selection methodology to improve almost all of the runs irrespective of their selection scheme. Thus we can conclude with theoretical, statistical and experimental confidence that a good comprehensive semi-structured scoring system can deliver improved focused retrieval experience, flexible enough to serve a diverse set of selection schemes.

G.5 Concluding remarks

In this study, the semi-structured retrieval is executed in two phases; (i) ranking/scoring and (ii) selection. Accordingly, the dichotomy of fusion is presented, where the *ranking* system was based on the fusion of ranks and scores of the official submitted runs of INEX, together with utilizing the context of the relevant retrievable unit. The *selection*, in turn, was based on individual selection schemes, taken again from the INEX submitted runs, one by one. The retrieval experience was notably improved after reordering the elements from the selection system, by fusing them with the ranking system. Extensive and favourable empirical results have validated the hypothesis, the selection fusion methodology enhances the retrieval of semi-structured retrieval item specified in the selection criteria. The selection fusion methodology is independent of selection scheme as well as generalized enough that any method can be applied for fusing and scoring, based on the user, system or domain preferences.

Measuring the effectiveness of IR systems should be well-defined and intuitive, yet simple and convenient enough for the system developers to comprehend and subsequently improve their retrieval and representational approaches. This study argues that developing good ranking system without taking the selection criteria into account, would lead to an overall good performance in a number of different selection scenarios and retrieval tasks. The scoring or ranking of elements should be considered an independent task in the course of system development. In contrast, developing scoring methods for a particular or set of selection scenarios and corresponding metrics, might end up in favouring certain type or size of elements, and therefore, lead to over-fitting the systems according to the metrics [5, 25].

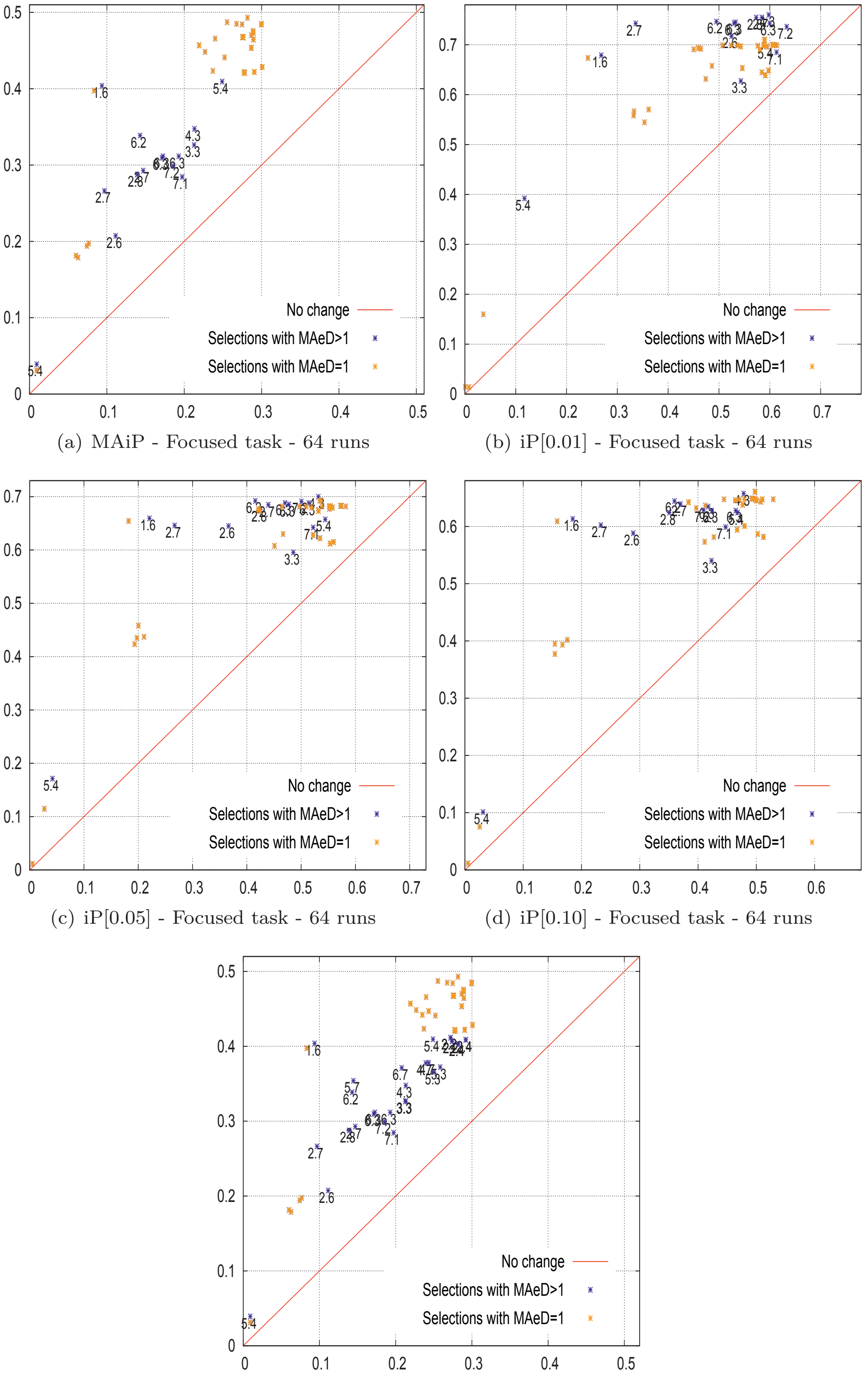
Fundamentally a number of issues need to be addressed in the further studies – (a) Application of the proposition (the selection fusion methodology) to a real world example; (b) Evaluating and finding the optimal number of retrieval systems (methods) needed for the selection fusion methodology to perform even better; (c) Using another collection, if available, with different representation of semi-structured data, e.g., overlapping structures; (d) Finding a different selection criteria, for example, a selection criteria based on the users' personalization.

G.6 References

[1] S. Abiteboul. *Querying semi-structured data.* Springer, 1997.

[2] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: from relations to semistructured data and XML.* Morgan Kaufmann Pub, 2000.

[3] P. Arvola, M. Junkkari, and J. Kekäläinen. Generalized Contextualization Method for XML Information Retrieval. In *Proc. of the 14th ACM CIKM*, pages 20–27. ACM, 2005.



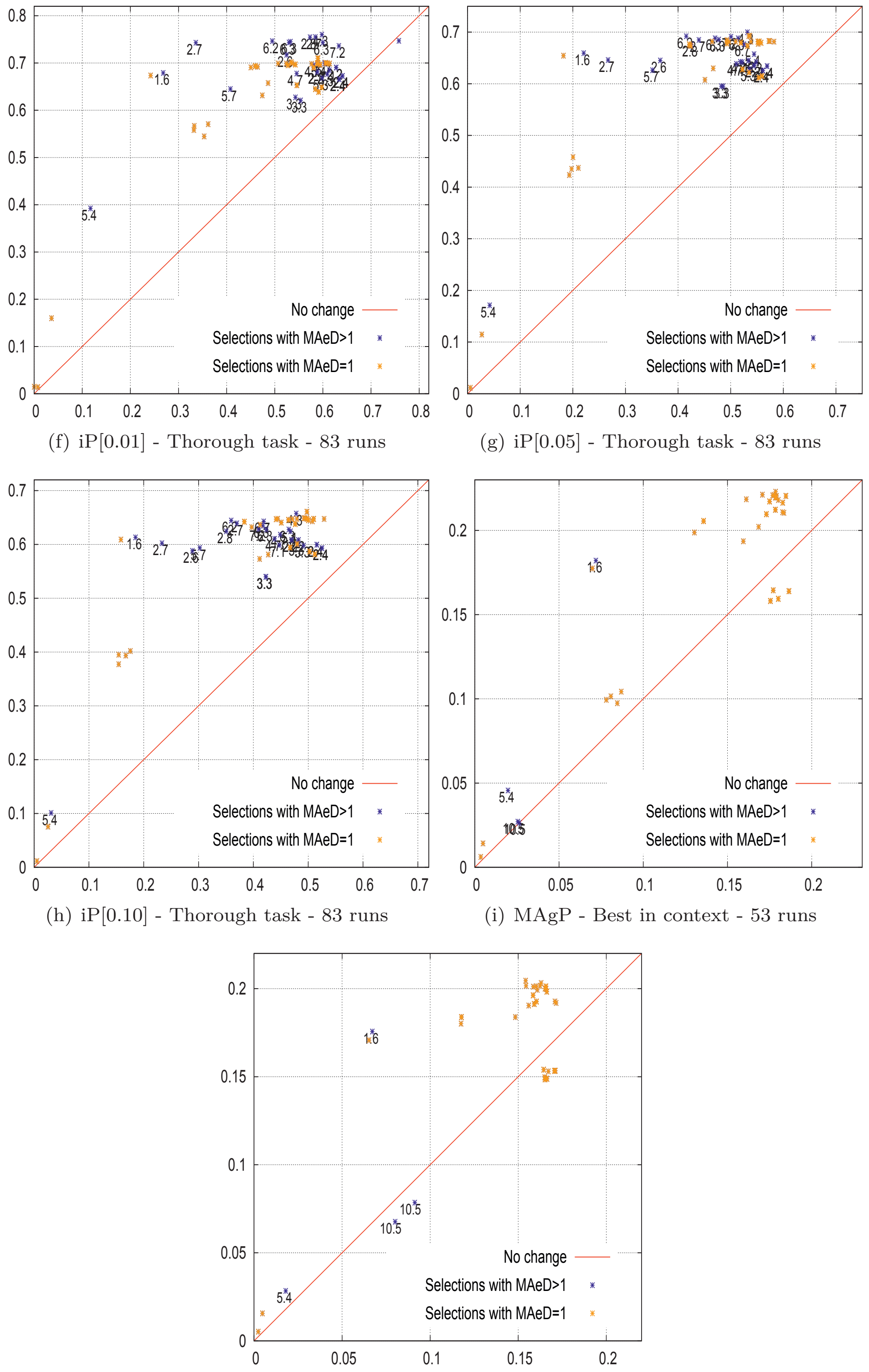
(a) MAiP - Focused task - 64 runs

(b) iP[0.01] - Focused task - 64 runs

(c) iP[0.05] - Focused task - 64 runs

(d) iP[0.10] - Focused task - 64 runs

(e) MAiP - Thorough task - 83 runs



(f) iP[0.01] - Thorough task - 83 runs

(g) iP[0.05] - Thorough task - 83 runs

(h) iP[0.10] - Thorough task - 83 runs

(i) MAgP - Best in context - 53 runs

(j) MAgP - Relevant in context - 53 runs

Figure G.2: *Effects of the selection fusion methodology on the 98 INEX 2009 submitted runs – ad-hoc track, four tasks. Numbers on each figure indicate the MAeD of that particular run (no numbers and orange marks, mean MAeD=1, i.e., article retrieval). Most of the improvements are statistically significant at* $p < 0.01$ *(1-tailed t-test).*

[4] P. Arvola, J. Kekäläinen, and M. Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13(5):460–484, 2010.

[5] 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, 2011a.

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

[7] C. Clarke, J. Kamps, and M. Lalmas. INEX 2006 retrieval task and result submission specification. In *INEX 2006 Workshop Pre-Proceedings*, pages 381–388. Citeseer, 2006.

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

[9] S. Geva, J. Kamps, M. Lehtonen, R. Schenkel, J. Thom, and A. Trotman. Overview of the INEX 2009 ad hoc track. *Focused Retrieval and Evaluation*, pages 4–25, 2010.

[10] N. Gövert and G. Kazai. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In *INEX Workshop*, pages 1–17. Citeseer, 2002.

[11] D. A. Grossman and O. Frieder. *Information retrieval: Algorithms and heuristics*, volume 15. Springer, 2004.

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

[13] K. Y. Itakura and C. L. Clarke. A framework for BM25F-based XML retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 843–844. ACM, 2010.

[14] J. Kamps and M. Koolen. The impact of document level ranking on focused retrieval. In *Advances in Focused Retrieval*, pages 140–151. Springer, 2009.

[15] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Structured queries in XML retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 4–11. ACM, 2005.

[16] J. Kamps, M. Lalmas, and J. Pehcevski. Evaluating relevant in context: document retrieval with a twist. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 749–750. ACM, 2007.

[17] J. Kamps, M. Koolen, and M. Lalmas. Locating relevant text within XML documents. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 847–848. ACM, 2008.

[18] G. Kazai, M. Lalmas, and A. P. de Vries. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 72–79. ACM, 2004.

[19] M. Lehtonen, N. Pharo, and A. Trotman. A Taxonomy for XML retrieval use cases. *Comparative Evaluation of XML IR Systems*, pages 413–422, 2007.

[20] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 387–396. ACM, 2006.

[21] S. Malik, M. Lalmas, and N. Fuhr. Overview of INEX 2004. In *Advances in XML Information Retrieval*, pages 1–15. Springer, 2005.

[22] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

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

[24] M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the 11th international conference on Information and knowledge management*, pages 538–548. ACM, 2002.

[25] M. A. Norozi, P. Arvola, and A. P. de Vries. Contextualization using hyperlinks and internal hierarchical structure of Wikipedia documents. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM)*, pages 734–743. ACM, 2012.

[26] M. A. Norozi and P. Arvola. Kinship Contextualization: Utilizing the Preceding and Following Structural Elements. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 837–840. ACM, 2013.

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

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

[29] T. Schlieder and H. Meuss. Querying and ranking XML documents. *Journal of the American Society for Information Science and Technology*, 53(6):489–503, 2002.

[30] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The 2nd TREC*. Citeseer, 1994.

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

[32] A. Trotman, N. Pharo, and M. Lehtonen. XML-IR users and use cases. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 400–412. Springer, 2007.