

UNIVERSITY OF MINNESOTA

This is to certify that I have examined this copy of a master's thesis by

Chaitanya Polumetla

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Name of Faculty Adviser

Signature of Faculty Adviser

Date

GRADUATE SCHOOL

Improving Results for the Relevant in Context Task

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA

BY

Chaitanya Polumetla

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Carolyn J. Crouch

August, 2009

Department of Computer Science

University of Minnesota, Duluth

Duluth, MN 55812

USA

Acknowledgements

I would like to take this opportunity to thank all the people who have stood beside me and supported me during the course of my thesis.

I would like to thank my advisor Dr. Carolyn Crouch, for providing me the opportunity to work in the field of Information Retrieval. She has been a source of tremendous support and guidance right from the day I joined this university.

I would like to thank Dr. Donald Crouch, for providing timely suggestions and valuable feedback during important stages of my thesis.

I would like to thank Varun Sudhakar, Dinesh Bhirud and Pavan Poluri, for being supportive co-workers as well as good friends throughout the completion of this thesis. Special thanks to Salil Bapat, Sarika Mehta and Darshan Paranjape, for sharing their experience and helping me on various aspects of this thesis.

I would like to thank the faculty and staff of the Department of Computer Science at UMD, especially Jim Luttinen for keeping the systems running, Lori Lucia and Linda Meek for their invaluable help.

I would like thank my brother Aditya Polumetla, for encouraging me to join UMD and guiding me throughout the course of this Masters program, I would like to especially thank my parents, family and friends for their support and wishes.

Abstract

Information retrieval (IR) is finding information (usually documents) that satisfies an information need from within large collections stored on computers [2]. Though the concept of IR did not begin with the World Wide Web, the introduction of the web has popularized IR, and it has become one of the major fields of study in Computer Science. With the introduction of XML to the World Wide Web and its widespread popularity, the application of IR techniques to XML data has become important and is now being widely studied. In XML search, document structure is used to determine which document fragments are more meaningful (i.e., relevant) to a query. *The INitiative for the Evaluation of XML retrieval (INEX)* was started to provide benchmarks to evaluate the effectiveness of various retrieval strategies on XML data.

The main goal of this thesis is to improve performance of our retrieval strategy in the INEX 2008 Ad hoc Relevant in Context task. The Relevant in Context task requires systems to find the focused results that correspond to the relevant elements in each relevant article. In this thesis, we discuss and evaluate various strategies to improve the performance of the Relevant in Context task.

Table of Contents

List of Figures	iv
List of Tables	v
1. Introduction	1
2. Overview	3
2.1 INEX	3
2.2 The INEX Environment	3
2.3 INEX tasks	7
2.4 Smart Retrieval Engine	9
3. The Relevant in Context Task	10
3.1 Flexible Retrieval	10
3.2 The RIC Task	15
3.3 RIC Experiments	18
4. Conclusions and Future Research	26
References	27
Appendix A	28

List of Figures

Fig 2.1: Typical Structure of a Document from the Wikipedia Collection	4
Fig 2.2: INEX 2008 Sample Query [Topic id = 566]	5
Fig 3.1(a): Pre Flex Operations	10
Fig 3.1(b): Flex Operations	11
Fig 3.1(c): Post Flex Operations	11
Figure 3.2: Sample RIC Evaluation Run	17

List of Tables

Table 3.1: Slope and Pivot values for 2008 Collection	13
Table 3.2(a): Tagset 1	19
Table 3.2(b): Tagset 2	19
Table 3.3: 2008 RIC Child Strategy Exact, Post Rearranged (Tagset 1)	20
Table 3.4: 2008 RIC Child Strategy Exact, Rearranged (Tagset 1)	20
Table 3.5: 2008 RIC Correlation Strategy Exact, Post Rearranged (Tagset 1)	21
Table 3.6: 2008 RIC Correlation Strategy Exact, Rearranged (Tagset 1)	21
Table 3.7: 2008 RIC Section Strategy Exact, Post Rearranged (Tagset 1)	21
Table 3.8: 2008 RIC Section Strategy Exact, Rearranged (Tagset 1)	22
Table 3.9: 2008 RIC Child Strategy UB, Rearranged (Tagset 1)	22
Table 3.10: 2008 RIC Correlation Strategy UB, Rearranged (Tagset 1)	22
Table 3.11: 2008 RIC Section Strategy UB, Rearranged (Tagset 1)	23
Table 3.12: 2008 RIC Child Strategy Exact, Rearranged (Tagset 2)	23
Table 3.13: 2008 RIC Correlation Strategy Exact, Rearranged (Tagset 2)	23
Table 3.14: 2008 RIC Section Strategy Exact, Rearranged (Tagset 2)	24
Table 3.15: 2008 RIC Child Strategy UB, Rearranged (Tagset 2)	24
Table 3.16: 2008 RIC Correlation Strategy UB, Rearranged (Tagset 2)	24
Table 3.17: 2008 RIC Section Strategy UB, Rearranged (Tagset 2)	25
Table A.2: 2007 RIC Child Strategy Exact, Post Rearranged (Tagset 1)	28
Table A.3: 2007 RIC Child Strategy Exact, Rearranged (Tagset 1)	28

Table A.4: 2007 RIC Correlation Strategy Exact, Post Rearranged (Tagset 1)	29
Table A.5: 2007 RIC Correlation Strategy Exact, Rearranged (Tagset 1)	29
Table A.6: 2007 RIC Section Strategy Exact, Post Rearranged (Tagset 1)	29
Table A.7: 2007 RIC Section Strategy Exact, Rearranged (Tagset 1)	30
Table A.8: 2007 RIC Child Strategy UB, Rearranged (Tagset 1)	30
Table A.9: 2007 RIC Correlation Strategy UB, Rearranged (Tagset 1)	30
Table A.10: 2007 RIC Section Strategy UB, Rearranged (Tagset 1)	31
Table A.11: 2007 RIC Child Strategy Exact, Rearranged (Tagset 2)	31
Table A.12: 2007 RIC Correlation Strategy Exact, Rearranged (Tagset 2)	31
Table A.13: 2007 RIC Section Strategy Exact, Rearranged (Tagset 2)	32
Table A.14: 2007 RIC Child Strategy UB,Rearranged (Tagset 2)	32
Table A.15: 2007 RIC Correlation Strategy UB, Rearranged (Tagset 2)	32
Table A.16: 2007 RIC Section Strategy UB, Rearranged (Tagset 2)	33

1. Introduction

Since its introduction in 1989, the World Wide Web (WWW) or 'web' has seen a tremendous increase in the information it holds. The indexed WWW contained about 25.75 billion web pages as of January 8th 2009 [14]. The problem is that the information on the WWW is stored in many different formats and for the most part is not organized at all, so extracting the exact information that we need becomes quite challenging. This is where Information Retrieval (IR) is used in dealing with many of these problems.

Information retrieval refers to finding information (usually documents) that satisfy an information need from within large collections stored on computers [2]. Though the concept of IR did not begin with the World Wide Web, the introduction of the web has popularized IR and it has become one of the major fields of study in Computer Science. Our frequent usage of search engines such as Google, Yahoo, and MSN Search is the proof that IR has become a part of our daily life (whether knowingly or not) due largely to the Internet.

XML or extensible markup language was introduced in 1998, and since then it has become one of the most widely used formats to store information on the web. XML was designed and is recommended by the World Wide Web Consortium (W3C) for creating custom markup languages [1].

With the Introduction of XML and its widespread popularity, using IR techniques on XML data has become important and is now being widely studied. In XML search,

document structure is used to determine which document fragments are more meaningful (i.e., relevant) to a query. It is also used to specify query conditions on structure that limit the search context to specific XML elements, as opposed to whole documents [1].

Though XML data was supposed to be structured, the relaxation of rules to increase the popularity of XML has resulted in the use of untagged text within XML documents, which poses a challenge to researchers. Flex (our system for flexible retrieval) was initially designed for structured retrieval but was later modified to handle semi-structured documents as well.

Researchers need to evaluate the effectiveness of their search techniques on XML data, and in 2002 the *INitiative for the Evaluation of XML retrieval* (INEX) was started to address this issue. The aim of the INEX initiative is to establish an infrastructure and to provide means, in the form of large test collections and appropriate scoring methods, for evaluating the effectiveness of content-oriented XML search systems [1].

The University of Minnesota Duluth (UMD) is an INEX participant. The main objective of this work is to build on our previous research and to improve our results in future competitions by experimentation within the various INEX tasks.

2. Overview.

This chapter provides an overview of INEX, the INEX environment and tasks, and the Smart search engine.

2.1 INEX

To evaluate XML-based retrieval systems, it is necessary to build test collections where the evaluation paradigms are provided according to criteria that take into account the imposed structural aspects. INEX establishes an infrastructure and provided means, in the form of large test collections and appropriate scoring methods, for evaluating the effectiveness of content-oriented XML search systems [7]. Many universities and research groups across the globe participate in the INEX competition and evaluate their results using the evaluation measures provided by INEX. These participants compare results, publish papers and share their knowledge with others groups; they also present and discuss their work during the yearly INEX workshops [7].

2.2 The INEX Environment

Document Collection

The INEX 2007 document collection is a set of Wikipedia documents in XML format. The collection is based on a 2006 version of the English Wikipedia collection that has been converted to the XML format. The collection contains 659,338 Wikipedia articles [6]. The typical structure of a document from the Wikipedia collection is presented in Figure 2.1.

```
<article>
<name> text </name>
<body>
  text
  <section>
    <title> text </title>
    text
    <p> text </p>
    ...
  </section>
  <p> text </p>
  ...
</body>
</article>
```

Figure 2.1: Typical Structure of a Document from the Wikipedia Collection

Queries

Based on INEX guidelines, the participants create the queries that are used with the collection. These queries are referred to as CO + S (Content only + Structured) queries. A sample query from the INEX 2008 Ad hoc track is presented in Figure 2.2.

```
<topic id="566" ct_no="38">

<title>+alignment sequence DNA</title>

<castitle>/*! [about(., +alignment sequence DNA)]</castitle>

<description>I'd like to have information about all kinds of sequence alignment
techniques for DNA.</description>

<narrative>

I am a researcher in information technology, and my speciality is not biology nor
genetics. However, I know that sequence alignment is a very commonly used
technique for discovering aligned sequences of DNA in genes, for example. I also
know that these algorithms can be adapted to natural language texts, and that's why
I'd like to know more about sequence alignment.

</narrative>

</topic>
```

Figure 2.2: INEX 2008 Sample Query [Topic id = 566]

Evaluation Measures

The INEX 2008 measures are solely based on the retrieval of highlighted text (similar to INEX 2007). To simplify the process, INEX evaluates all the Ad hoc Track tasks based on highlighted text retrieval and assumes that systems return all, and only, highlighted text. INEX then compares the characters of text retrieved by a search engine to the number and location of characters of text identified as relevant by the assessor. For the Best in Context task, INEX uses the distance between the best entry point in the run to the one that has been identified by an assessor [8].

In the **Focused Task**, Recall is measured as the fraction of all highlighted text that has been retrieved. Precision is measured as the fraction of retrieved text that was highlighted. The INEX 2008 official measure is interpolated precision at 1% recall (iP[0.01]), as the interest is directed at the elements retrieved at highest ranks [8].

The evaluation of the **Relevant in Context Task** is based on the measures of generalized precision and recall, where the per document score reflects how well the text which was retrieved matches the relevant text in the document. Since the focus is on the overall performances, the main measure used is the mean average generalized precision (MAgP) [8].

The evaluation of the **Best in Context Task** is based on the measures of generalized precision and recall where the per document score reflects how well the retrieved entry point matches the best entry point in the document. If the distance between the retrieved entry point is within 500 characters of the assessed entry point, then it gets a score; otherwise it gets no score (at INEX 2007 this number was 1000). Since the focus is on the overall performance, the main measure used is mean average generalized precision (MAgP) [8].

Relevance Assessments

INEX provides relevance assessments against which the results generated by each participant are compared and evaluated [7]. All the participants have a hand in contributing to the relevance assessments, as they manually select the relevant documents for the queries they submitted from the pool of documents given by

INEX. Participants determine the relevance of articles and elements using the INEX TopX tool.

Assessment is a crucial task. Negligence on the part of the assessor could result in poor results even though the search engine did a good job. This has been evident, as our group has done many studies on the pool by randomly selecting some queries and checking the relevant assessments against the query. In many cases irrelevant documents have been marked relevant and vice-versa. But documents are reviewed as relevant or non-relevant according to the assessor's perception, and there is always the possibility of error when humans are concerned.

2.3 INEX Tasks

Our IR group at UMD participates in the Ad hoc track that consists of three tasks, namely, the Focused, Best in Context, and the Relevance in Context tasks.

Focused Task

This task requires retrieval systems to return the most focused results that satisfy the conditions specified by the query; the main goal of the task is to return a ranked list of elements without any overlap with other elements. (Overlapping elements are those which contain both the parent and child node XPaths. For example, 1034168/article[1]/body[1]/figure[1] and 1034168/article[1]/body[1]/figure[1]/image[1] are said to overlap since 'image[1]' is a child node of 'figure[1]'. In a file with non-overlapping elements, only one of the

above two paths can exist.) The result is, for each document in question, a ranked list of focused elements [3].

Relevance in Context Task

The Relevance in Context (RIC) task requires, for each query, the retrieval of focused elements from articles that correlate highly with the query. Our system first identifies the articles that correlate with the query, and then it uses Flex to find the elements present in those documents. Details of the RIC experiments can be found in Chapter 3.

Best in Context Task

The purpose of the Best in Context task is first to identify the articles that correlate strongly with the query and then to determine the Best Entry Point (BEP) associated with each document. The BEP is the point in the text where the user should begin reading to find the information sought in the query. The BEP is frequently found near the beginning of the document but it could be elsewhere in the text. There is one BEP per document; INEX evaluates the BEP by calculating the distance between the actual BEP (from the relevance assessments) and the BEP returned by the participants. The smaller the distance, the better the result.

We are currently evaluating different strategies to find the best BEP. Using name always as the BEP has provided quite decent results [13]. We keep in mind that the document set returned has a large impact on the BEP results as only the relevant documents are counted.

2.4 Smart Retrieval Engine

We use Smart 13.0 [12] as our retrieval engine. Smart uses the Vector Space Model [11], in which the query and document are represented as a weighted term vectors. The distance between the document vector and the query vector in the vector space determines the correlation of the document with respect to the query. The basic functionalities of Smart include indexing of the document collection, weighting of the documents and query vectors and retrieving rank-ordered (document) elements [3].

3. The Relevant in Context Task

This chapter describes the overview of the flexible retrieval, the Relevant in Context (RIC) task, and different strategies used to improve the RIC results.

3.1 Flexible Retrieval

Flexible Retrieval refers to the task of retrieving specific elements in response to a query. Unlike traditional IR systems that retrieve entire documents, our system of flexible retrieval seeks to recognize those elements that are specific to (i.e., focused on) the query, thereby reducing the time the user has to expend to find what he is looking for [9]. The procedure for flexible retrieval is given in Figure 3.1.

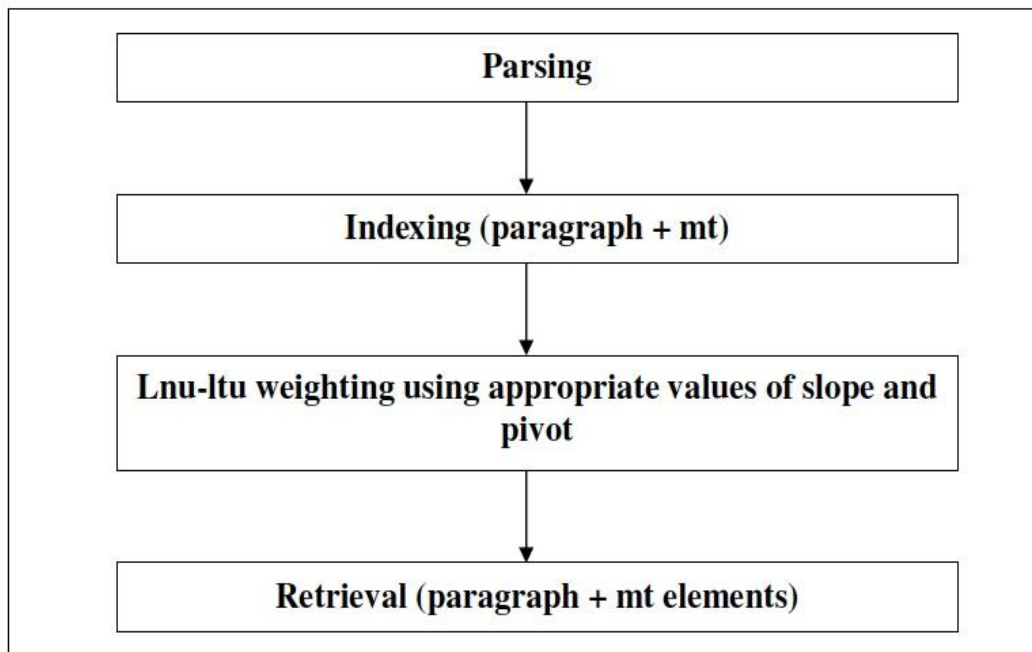


Figure 3.1 (a): Pre Flex Operations

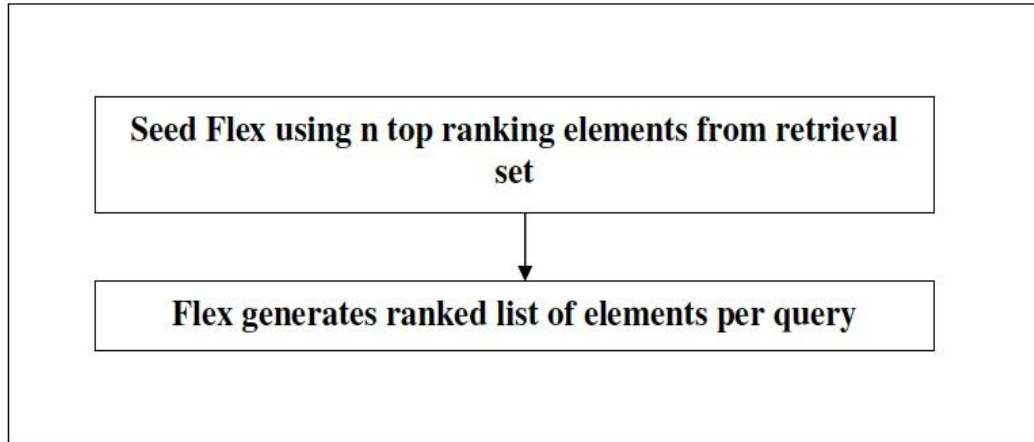


Figure 3.1(b): Flex Operations

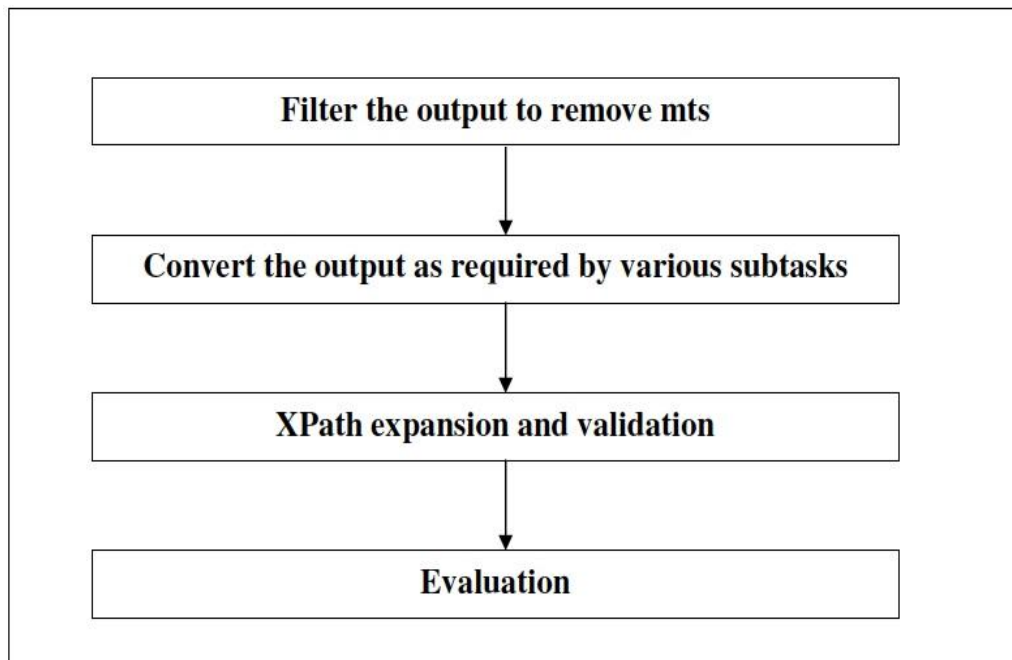


Figure 3.1(c): Post Flex Operations [3]

Pre Flex

Pre Flex operations begin by parsing the document collection and indexing the terminal node elements. After indexing, we perform *Lnu-ltu* term weighting and retrieval against the terminal node index. A set of top-ranked elements (used as an input to Flex) are retrieved for every query.

Parsing

The first step is to parse the documents. We generate a *paragraph-and-mt* parse to create the *paragraph-and-mt* index required as input to flexible retrieval. (This parse generates all terminal nodes along with the magic text elements.) Once a terminal node tag is recognized in the XML hierarchy, any tags contained in it are removed and all the text is aggregated to form the leaf node [3].

Indexing

Parsing is followed by the creation of indices using Smart. We generate a *paragraph-and-mt* index by taking the *paragraph-and-mt* parse as input. Queries are also indexed along with the elements (the title field of every query is used as the text for query indexing).

Term Weighting and Pivoted Normalization

Our Smart index of the leaf nodes uses the *nnn* weighting scheme [9] to weight terms. In this weighting scheme, the weight assigned to a term is equal to its frequency in the element. (This weighting scheme takes only term frequency into

account.) Larger elements have more terms with higher term frequency. Thus the probability of retrieval for larger elements is higher than that of the smaller elements irrespective of their relevance. In order to avoid this bias towards larger elements, we need to use a weighting scheme that takes element length into consideration and normalizes term weights based on it. Thus we convert the *nnn* weights on element vectors to *Lnu* weights [3].

The conversion of *nnn* weighting scheme to *Lnu* weighting scheme requires additional parameters, which are pivot and slope. (More details on the conversion of *nnn* to *Lnu* can be found in [9]). Pivot represents the average element length; the normalization is pivoted at pivot and tilted so that the elements on one side of pivot get larger normalization factors and the others get smaller normalization factors. The amount of tilting required is represented by slope. To reduce the difference between the probability of relevance and the probability of retrieval for all elements of different lengths, we also calculate slope. [9] Slope and pivot are empirically determined constants. The pivot value varies from collection to collection and slope needs to be calculated accordingly. Table 3.1 presents the values of slope and pivot used for INEX 2008 document collection.

Table 3.1: Slope and pivot values for 2008 collection

Element Level	Slope	Pivot
Article	0.04	120
Paragraph	0.12	15
All-element	0.12	38

Retrieval

In this step, the *Lnu*-weighted element vectors are correlated with the *ltu*-weighted query vector using inner product, and a correlation score is assigned to each element. The elements are then sorted according to the correlation score which produces the final list of ranked elements. We specify the number of elements to be retrieved for each query [3].

Flex

In this stage, we take highly correlating terminal node elements (from Pre Flex processing) as input and then construct the vectors for their parent elements. The output of this stage is a ranked list of elements.

Seeding

Retrieval against a set of terminal nodes (i.e., *paragraph-and-mt*) is required to seed Flex. The set of *n* top-ranked elements from this retrieval is used as the seed. In this context, seeding is the basis of document tree generation by Flex [3].

Flex Execution

Flex reads the schema information of each seeded document. It then constructs the vectors for the non-terminal elements and computes the correlation of every element with the query vector. It outputs a ranked list of elements (with correlation scores) for every query [3].

Post Flex

In Post Flex operations, we first remove the *mt* elements from Flex results. The results are then converted according to the various subtasks. For each subtask, the results are converted to INEX-specified XML format and evaluated.

Filtering Flex Output

The output of Flex is a ranked list of elements. We first remove from this list all the nodes tagged with the *mt* tag, which is not a valid tag for the Wikipedia collection. After removing the *mt* elements, the output is further filtered based on the subtask. Outputs for all the subtasks are converted to INEX-specified XML format before evaluation. Of particular interest is the output of the RIC task.

3.2 The RIC task

The RIC task can be divided into two major tasks: one is to identify the relevant articles (the fetching phase), and the other is to identify the relevant elements from those articles (the browsing phase). In the fetching phase, articles are ranked according to their topical relevance. Browsing should then provide a set of non-overlapping elements that cover the relevant information in the article [3].

The motive of the Relevant in Context Task is the return of a ranked list of articles and to find relevant information within those articles. A relevant article will most likely contain relevant information that could be spread across different (non-overlapping) elements. The task requires systems to find a set of results that correlate well to all relevant information in each relevant article [3].

The task is evaluated by mean average generalized precision (MAGP) where the generalized score per article is based on the retrieved highlighted text. Specifically, the per document score is the harmonic mean of precision and recall in terms of the fractions of retrieved and highlighted text in the document. INEX uses an $F\beta$ score with $\beta = 1/4$ making precision four times as important as recall (at INEX 2007, F1 was used) [8]. A sample RIC evaluation is given in Figure 3.2.

In our implementation of the RIC task, we first perform article retrieval to identify highly correlating articles. These articles are used to identify the seed subset for Flex. All overlapping elements are then removed from Flex output (using one of several strategies) leaving only positively correlating terminal nodes. The results are then converted to INEX format. The overlap removal strategies are covered in detail in [4]. A brief description follows.

Section Strategy

In the section strategy, we search for the focused elements by highest correlating non-body element.

Child Strategy

In the child strategy, we search for the focused elements by always giving preference to the child node.

```

<eval run-id="p51-ric" file="/smart/2008_analysis/ric_sec_UB/fol/25/ric_25_50.txt">
num_q      all      70
num_ret    all      1256
num_rel    all      4887
num_rel_ret all    538
ret_size all 3305377
rel_size all 11471649
rel_ret_size all 1081420
MAgP      all      0.10044556038542417
gP[1]     all      0.5344643767950747
gR[1]     all      0.03534910982997681
gP[2]     all      0.48628470725322603
gR[2]     all      0.05993101592460779
gP[3]     all      0.4446092782192075
gR[3]     all      0.07662966804685174
gP[5]     all      0.4040346417456151
gR[5]     all      0.10464880390291778
gP[10]    all      0.34713157413237866
gR[10]    all      0.15369225960267197
gP[25]    all      0.28397679858004693
gR[25]    all      0.21326693536200034
gP[50]    all      0.28397679858004693
gR[50]    all      0.21326693536200034
ircl_prn.0.00 all 0.6795235362089384
ircl_prn.0.10 all 0.32878829980015695
ircl_prn.0.20 all 0.16969252042732183
ircl_prn.0.30 all 0.13357594646370755
ircl_prn.0.40 all 0.06644685472268735
ircl_prn.0.50 all 0.03759858866063658
ircl_prn.0.60 all 0.021517425341811548
ircl_prn.0.70 all 0.015139850711852742
ircl_prn.0.80 all 0.00846012416031593
ircl_prn.0.90 all 0.001150504035450319
ircl_prn.1.00 all 0.001150504035450319
</eval>

```

Figure 3.2: Sample RIC Evaluation Run

Correlation Strategy

In the correlation strategy, we search for the focused elements by selecting the element purely based on the highest correlation score.

Exact and Upper Bound (UB) Methods for Element Retrieval

The *Exact* method is designed to return the exact number of elements from the documents if possible. The *Upper Bound* (UB) method specifies an upper bound of elements for both number of elements produced in the output and the number of articles from which they may come [4].

3.3 RIC Experiments

The results are organized into two subsets:

- 2008 queries with 2008 evaluation
- 2007 queries with 2008 evaluation

and

- Old tagset (Tagset 1)
- New tagset (Tagset 2)

Tag set here refers to the tags that are indexed and the tags that are kept when the collection is indexed. The tagsets used are presented in Tables 3.2.

Table 3.2(a): Tagset 1 [3]

Tag Name	Tag Representation
Tags to Index	
Paragraph	<p> ... </p>
Magic Text	<mt> ... </mt>
Emphasis	<emph3> ... </emph3>
Name	<name> ... </name>
Figure	<figure> ... </figure>
Tags to Keep	
Article	<article> </article>
Emphasis	<emph3> ... </emph3>
Title	<title> ... </title>
Magic Text	<mt> ... </mt>
Name	<name> ... </name>
Figure	<figure> ... </figure>
Definition-list	<definitionlist> ... </definitionlist>
Number - list	<numberlist> </numberlist>
Unordered-list	
Ordered-list	
Normal-list	<normallist> </normallist>
Paragraph	<p> ... </p>
Body	<body> </body>
Section	<section> </section>

Table 3.2(b): Tagset 2 [4].

Tag Name	Tag Representation
Tags to Index	
Paragraph	<p> ... </p>
Figure	<figure> ... </figure>
Name	<name> ... </name>
Emphasis	<emph3> ... </emph3>
Magic Text	<mt> ... </mt>
Table	<table> ... </table>
Ordered-list	
Normal-list	<normallist> </normallist>
Unordered-list	...
Number - list	<numberlist>...</numberlist>
Definition-list	<definitionlist> ... </definitionlist>
Tags to Keep	
Article	<article> </article>
Paragraph	<p> ... </p>
Body	<body> </body>
Section	<section> </section>
Unordered-list	
Ordered-list	
Number - list	<numberlist> </numberlist>
Unordered-list	
Name	<name> ... </name>
Figure	<figure> ... </figure>
Definition-list	<definitionlist> ... </definitionlist>
Emphasis	<emph3> ... </emph3>
Title	<title> ... </title>
Magic Text	<mt> ... </mt>
Table	<table> ... </table>

RIC Results

The best results in each table are marked in bold. Columns represent the *number of articles* and the rows represent the *number of para's*. The 2008 results are given in Tables 3.3 – 3.17.

Table 3.3: 2008 RIC Child Strategy Exact, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0838	0.1031	0.0936	0.1031	0.0991	0.1031	0.1018	0.1031	0.1031	0.0572	0.1031
50	0.0867	0.1265	0.0987	0.1265	0.1065	0.1265	0.1138	0.1265	0.1265	0.0573	0.1262
100	0.0887	0.1465	0.1011	0.1465	0.1102	0.1465	0.1180	0.1465	0.1465	0.0573	0.1393
150	0.0887	0.1540	0.1011	0.1557	0.1102	0.1557	0.1185	0.1557	0.1557	0.0573	0.1406
200	0.0887	0.1567	0.1011	0.1606	0.1102	0.1606	0.1185	0.1606	0.1606	0.0573	0.1410
250	0.0887	0.1574	0.1011	0.1632	0.1102	0.1632	0.1185	0.1632	0.1632	0.0573	0.1412
500	0.0887	0.1584	0.1011	0.1659	0.1102	0.1659	0.1185	0.1659	0.1659	0.0573	0.1414

Table 3.4: 2008 RIC Child Strategy Exact, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1021	0.1031	0.1031	0.1031	0.1031	0.1031	0.1033	0.1031	0.1031	0.0928	0.1031
50	0.1202	0.1265	0.1242	0.1265	0.1253	0.1265	0.1273	0.1265	0.1265	0.1034	0.1265
100	0.1255	0.1465	0.1357	0.1465	0.1403	0.1465	0.1440	0.1465	0.1465	0.1026	0.1476
150	0.1265	0.1561	0.1372	0.1557	0.1443	0.1557	0.1478	0.1557	0.1557	0.1035	0.1556
200	0.1234	0.1622	0.1372	0.1617	0.1448	0.1607	0.1497	0.1606	0.1606	0.1026	0.1601
250	0.1229	0.1657	0.1369	0.1648	0.1443	0.1640	0.1505	0.1632	0.1632	0.1026	0.1616
500	0.1252	0.1726	0.1357	0.1740	0.1447	0.1705	0.1499	0.1670	0.1667	0.1036	0.1620

Table 3.5: 2008 RIC Correlation Strategy Exact, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1005	0.1044	0.1044	0.1044	0.1044	0.1044	0.1044	0.1044	0.1044	0.0829	0.1044
50	0.1075	0.1277	0.1198	0.1277	0.1259	0.1277	0.1273	0.1277	0.1277	0.0847	0.1277
100	0.1097	0.1478	0.1240	0.1478	0.1332	0.1478	0.1393	0.1478	0.1478	0.0847	0.1478
150	0.1097	0.1572	0.1242	0.1572	0.1338	0.1572	0.1404	0.1572	0.1572	0.0847	0.1554
200	0.1097	0.1625	0.1242	0.1625	0.1339	0.1625	0.1406	0.1625	0.1625	0.0847	0.1575
250	0.1097	0.1659	0.1242	0.1661	0.1339	0.1661	0.1408	0.1661	0.1661	0.0847	0.1579
500	0.1097	0.1700	0.1242	0.1738	0.1339	0.1738	0.1408	0.1738	0.1738	0.0847	0.1585

Table 3.6: 2008 RIC Correlation Strategy Exact, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1047	0.1044	0.1044	0.1044	0.1044	0.1044	0.1044	0.1044	0.1044	0.1064	0.1044
50	0.1307	0.1277	0.1290	0.1277	0.1282	0.1277	0.1277	0.1277	0.1277	0.1253	0.1277
100	0.1438	0.1478	0.1506	0.1478	0.1519	0.1478	0.1501	0.1478	0.1478	0.1281	0.1478
150	0.1459	0.1572	0.1547	0.1572	0.1590	0.1572	0.1614	0.1572	0.1572	0.1272	0.1584
200	0.1460	0.1625	0.1556	0.1625	0.1606	0.1625	0.1644	0.1625	0.1625	0.1267	0.1644
250	0.1460	0.1669	0.1551	0.1661	0.1610	0.1661	0.1656	0.1661	0.1661	0.1274	0.1693
500	0.1464	0.1765	0.1562	0.1755	0.1617	0.1747	0.1661	0.1738	0.1738	0.1276	0.1760

Table 3.7: 2008 RIC Section Strategy Exact, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0951	0.1038	0.1024	0.1038	0.1038	0.1038	0.1038	0.1038	0.1038	0.0729	0.1038
50	0.0999	0.1271	0.1115	0.1271	0.1195	0.1271	0.1250	0.1271	0.1271	0.0736	0.1271
100	0.1017	0.1471	0.1145	0.1471	0.1238	0.1471	0.1313	0.1471	0.1471	0.0735	0.1463
150	0.1018	0.1565	0.1145	0.1565	0.1241	0.1565	0.1318	0.1565	0.1565	0.0736	0.1499
200	0.1018	0.1615	0.1145	0.1617	0.1241	0.1617	0.1320	0.1617	0.1617	0.0736	0.1506
250	0.1017	0.1634	0.1145	0.1652	0.1240	0.1652	0.1319	0.1652	0.1652	0.0735	0.1509
500	0.1018	0.1653	0.1145	0.1707	0.1241	0.1707	0.1320	0.1707	0.1707	0.0736	0.1511

Table 3.8: 2008 RIC Section Strategy Exact, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1044	0.1038	0.1039	0.1038	0.1038	0.1038	0.1038	0.1038	0.1038	0.0985	0.1038
50	0.1254	0.1271	0.1261	0.1271	0.1275	0.1271	0.1275	0.1271	0.1271	0.1084	0.1271
100	0.1302	0.1471	0.1391	0.1471	0.1449	0.1471	0.1465	0.1471	0.1471	0.1102	0.1473
150	0.1253	0.1565	0.1404	0.1565	0.1471	0.1565	0.1512	0.1565	0.1565	0.1079	0.1571
200	0.1272	0.1626	0.1401	0.1617	0.1481	0.1617	0.1521	0.1617	0.1617	0.1067	0.1614
250	0.1272	0.1654	0.1396	0.1652	0.1456	0.1652	0.1524	0.1652	0.1652	0.1072	0.1642
500	0.1285	0.1736	0.1394	0.1743	0.1399	0.1725	0.1519	0.1707	0.1707	0.1080	0.1645

Table 3.9: 2008 RIC Child Strategy UB, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1030	0.1031	0.1048	0.1031	0.1042	0.1031	0.1039	0.1031	0.1031	0.0980	0.1031
50	0.1206	0.1265	0.1245	0.1265	0.1283	0.1265	0.1284	0.1265	0.1265	0.1113	0.1276
100	0.1317	0.1472	0.1372	0.1465	0.1404	0.1465	0.1434	0.1465	0.1465	0.1133	0.1496
150	0.1339	0.1580	0.1417	0.1562	0.1461	0.1563	0.1482	0.1557	0.1557	0.1126	0.1561
200	0.1346	0.1640	0.1427	0.1627	0.1472	0.1619	0.1509	0.1607	0.1606	0.1129	0.1583
250	0.1333	0.1676	0.1447	0.1671	0.1481	0.1659	0.1516	0.1641	0.1632	0.1132	0.1615
500	0.1333	0.1716	0.1450	0.1759	0.1536	0.1761	0.1548	0.1735	0.1685	0.1139	0.1651

Table 3.10: 2008 RIC Correlation Strategy UB, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1072	0.1044	0.1058	0.1044	0.1057	0.1044	0.1050	0.1044	0.1044	0.0996	0.1044
50	0.1251	0.1277	0.1287	0.1277	0.1308	0.1277	0.1310	0.1277	0.1277	0.1123	0.1282
100	0.1332	0.1484	0.1410	0.1479	0.1442	0.1478	0.1471	0.1478	0.1478	0.1148	0.1517
150	0.1339	0.1593	0.1432	0.1579	0.1495	0.1578	0.1526	0.1572	0.1572	0.1147	0.1599
200	0.1345	0.1659	0.1443	0.1640	0.1506	0.1636	0.1543	0.1625	0.1625	0.1150	0.1637
250	0.1341	0.1698	0.1454	0.1679	0.1514	0.1673	0.1550	0.1669	0.1661	0.1153	0.1663
500	0.1353	0.1761	0.1459	0.1783	0.1534	0.1773	0.1575	0.1760	0.1750	0.1161	0.1688

Table 3.11: 2008 RIC Section Strategy UB, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1047	0.1038	0.1050	0.1038	0.1048	0.1038	0.1048	0.1038	0.1038	0.1004	0.1038
50	0.1231	0.1271	0.1260	0.1271	0.1294	0.1271	0.1296	0.1271	0.1271	0.1141	0.1279
100	0.1343	0.1478	0.1385	0.1471	0.1418	0.1471	0.1445	0.1471	0.1471	0.1153	0.1506
150	0.1362	0.1586	0.1437	0.1571	0.1478	0.1570	0.1499	0.1565	0.1565	0.1139	0.1580
200	0.1367	0.1650	0.1444	0.1633	0.1488	0.1627	0.1524	0.1617	0.1617	0.1143	0.1607
250	0.1353	0.1689	0.1464	0.1674	0.1499	0.1665	0.1533	0.1659	0.1652	0.1146	0.1635
500	0.1350	0.1734	0.1468	0.1771	0.1551	0.1768	0.1561	0.1752	0.1728	0.1153	0.1662

Table 3.12: 2008 RIC Child Strategy Exact, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1040	0.1056	0.1035	0.1056	0.1054	0.1056	0.1056	0.1056	0.1056	0.0945	0.1056
50	0.1188	0.1292	0.1242	0.1292	0.1264	0.1292	0.1275	0.1292	0.1292	0.0999	0.1292
100	0.1256	0.1496	0.1348	0.1496	0.1403	0.1496	0.1446	0.1496	0.1496	0.1025	0.1493
150	0.1253	0.1591	0.1369	0.1591	0.1440	0.1586	0.1477	0.1585	0.1585	0.1023	0.1567
200	0.1252	0.1649	0.1368	0.1643	0.1455	0.1629	0.1491	0.1626	0.1626	0.1027	0.1604
250	0.1259	0.1683	0.1387	0.1688	0.1458	0.1653	0.1501	0.1641	0.1640	0.0988	0.1621
500	0.1260	0.1741	0.1368	0.1769	0.1472	0.1740	0.1510	0.1674	0.1660	0.0986	0.1652

Table 3.13: 2008 RIC Correlation Strategy Exact, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1058	0.1052	0.1053	0.1052	0.1052	0.1052	0.1052	0.1052	0.1052	0.1065	0.1052
50	0.1311	0.1289	0.1303	0.1289	0.1296	0.1289	0.1294	0.1289	0.1289	0.1248	0.1289
100	0.1464	0.1496	0.1509	0.1496	0.1516	0.1496	0.1511	0.1496	0.1496	0.1246	0.1497
150	0.1475	0.1593	0.1567	0.1593	0.1596	0.1593	0.1612	0.1593	0.1593	0.1248	0.1604
200	0.1482	0.1651	0.1579	0.1651	0.1637	0.1651	0.1658	0.1651	0.1651	0.1243	0.1669
250	0.1484	0.1692	0.1592	0.1690	0.1653	0.1690	0.1680	0.1690	0.1690	0.1242	0.1716
500	0.1488	0.1797	0.1604	0.1786	0.1671	0.1759	0.1708	0.1749	0.1749	0.1247	0.1789

Table 3.14: 2008 RIC Section Strategy Exact, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1035	0.1050	0.1041	0.1050	0.1048	0.1050	0.1050	0.1050	0.1050	0.1018	0.1050
50	0.1230	0.1284	0.1252	0.1284	0.1270	0.1284	0.1276	0.1284	0.1284	0.1048	0.1284
100	0.1278	0.1486	0.1380	0.1486	0.1423	0.1486	0.1443	0.1486	0.1486	0.1027	0.1481
150	0.1275	0.1581	0.1391	0.1582	0.1457	0.1582	0.1500	0.1582	0.1582	0.0994	0.1573
200	0.1280	0.1636	0.1390	0.1638	0.1454	0.1638	0.1515	0.1638	0.1638	0.0993	0.1611
250	0.1272	0.1678	0.1390	0.1674	0.1452	0.1669	0.1516	0.1669	0.1669	0.0987	0.1629
500	0.1247	0.1748	0.1387	0.1762	0.1446	0.1732	0.1505	0.1698	0.1698	0.0984	0.1659

Table 3.15: 2008 RIC Child Strategy UB, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1040	0.1056	0.1055	0.1056	0.1058	0.1056	0.1055	0.1056	0.1056	0.0986	0.1057
50	0.1232	0.1293	0.1269	0.1292	0.1288	0.1292	0.1297	0.1292	0.1292	0.1068	0.1297
100	0.1316	0.1506	0.1400	0.1496	0.1439	0.1496	0.1458	0.1496	0.1496	0.1088	0.1495
150	0.1318	0.1601	0.1421	0.1598	0.1482	0.1590	0.1520	0.1586	0.1585	0.1070	0.1594
200	0.1319	0.1659	0.1428	0.1661	0.1505	0.1657	0.1548	0.1637	0.1627	0.1051	0.1636
250	0.1323	0.1705	0.1435	0.1701	0.1513	0.1696	0.1563	0.1659	0.1643	0.1050	0.1650
500	0.1336	0.1766	0.1452	0.1787	0.1529	0.1791	0.1580	0.1780	0.1727	0.1049	0.1696

Table 3.16: 2008 RIC Correlation Strategy UB, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1067	0.1052	0.1063	0.1052	0.1056	0.1052	0.1056	0.1052	0.1052	0.0999	0.1052
50	0.1265	0.1290	0.1303	0.1289	0.1306	0.1289	0.1310	0.1289	0.1289	0.1085	0.1299
100	0.1341	0.1506	0.1428	0.1497	0.1479	0.1496	0.1495	0.1496	0.1496	0.1105	0.1515
150	0.1348	0.1610	0.1453	0.1601	0.1510	0.1594	0.1556	0.1593	0.1593	0.1102	0.1609
200	0.1347	0.1669	0.1461	0.1665	0.1530	0.1658	0.1574	0.1651	0.1651	0.1084	0.1659
250	0.1350	0.1716	0.1462	0.1710	0.1549	0.1701	0.1588	0.1687	0.1690	0.1082	0.1679
500	0.1366	0.1790	0.1481	0.1804	0.1562	0.1800	0.1620	0.1791	0.1786	0.1083	0.1735

Table 3.17: 2008 RIC Section Strategy UB, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.1057	0.1050	0.1058	0.1050	0.1052	0.1050	0.1053	0.1050	0.1050	0.0997	0.1050
50	0.1253	0.1284	0.1284	0.1284	0.1297	0.1284	0.1297	0.1284	0.1284	0.1075	0.1293
100	0.1330	0.1498	0.1418	0.1487	0.1456	0.1486	0.1467	0.1486	0.1486	0.1104	0.1498
150	0.1337	0.1594	0.1439	0.1588	0.1498	0.1582	0.1534	0.1582	0.1582	0.1086	0.1596
200	0.1337	0.1654	0.1447	0.1651	0.1521	0.1645	0.1562	0.1639	0.1638	0.1067	0.1639
250	0.1341	0.1697	0.1450	0.1696	0.1527	0.1686	0.1575	0.1673	0.1670	0.1067	0.1652
500	0.1357	0.1774	0.1468	0.1780	0.1542	0.1786	0.1591	0.1778	0.1749	0.1068	0.1703

The following observations can be made from the 2008 RIC results with the 2008 evaluations:

- Correlation Strategy (UB) with Tagset 2 (new tagset) gives us the best result of 0.1800 (MAGP)
- The best result places UMD at position 15 in the INEX rankings (up 3 positions from the previous position of 18)
- The best results are concentrated at 500 articles and 150-1500 paras for most of the experiments

2007 results can be found in the Appendix Tables A.2 – A.16.

4. Conclusions and Future Research

Since the input to the RIC is the output of the Focused method, the Focused results have a direct impact to the RIC results. It is evident from [4] and [10] that UMD's rankings for the focused task are at the very top. The improvement has been tremendous but the same is not the case for the RIC results. UMD's rankings improved (to the 15th position from the 18th position) but the improvement is not in par with the Focused results, further analysis needs to be done to uncover the reasons behind this phenomenon.

Future Research

To improve our results for the RIC task we could make sure we have good retrieval all over the window and not just over the task. We can also try and increase the number of articles (>500) in the window since all our best results are given when the number of articles are equal to 500.

References

- [1] Amer-Yahia, S., and Lalmas, M. 2006. XML search: languages, INEX and scoring. *SIGMOD*, 16-23, 2006.
- [2] Manning, C., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [3] Paranjape, D. Improving Focused Retrieval. MS Thesis. University of Minnesota Duluth, 2008. <http://www.d.umn.edu/cs/thesis>
- [4] Bhirud, D. Focused Retrieval using Upper Bound Methodology. MS Thesis, University of Minnesota Duluth, 2009. <http://www.d.umn.edu/cs/thesis>
- [5] EvalJ – Evaluation Software <http://evalj.sourceforge.net/>
- [6] INEX Document Collection
<http://www.inex.otago.ac.nz/data/documentcollection.asp>
- [7] INitiative for the Evaluation of XML retrieval (INEX 2007)
<http://inex.is.informatik.uni-duisburg.de/2007/>
- [8] Kamps, J., Geva, S., Trotman, A., Woodley, A., and Koolen, M. Overview of the INEX 2008 Ad hoc track. *INEX 2008 Pre-proceedings*, 12-41, 2008.
- [9] Khanna, S. Design and Implementation of a Flexible Retrieval System. MS Thesis, University of Minnesota Duluth, 2005. <http://www.d.umn.edu/cs/thesis>
- [10] Poluri, P. Focused Retrieval using Exact methodology. MS Thesis, University of Minnesota Duluth, 2009. <http://www.d.umn.edu/cs/thesis>
- [11] Salton, G., Wong A., and Yang C. A Vector Space Model for Information Retrieval. *Journal of American Society for Information Science*, 18(11):613-620, 1975.
- [12] Salton, G. *The SMART Retrieval System – Experiments in Automatic Documents Retrieval*. Prentice-Hall, Eaglewood Cliffs, NJ, 1971.
- [13] Sudhakar, V. Improving the Best in Context Task. MS Project, University of Minnesota Duluth, 2009.
- [14] WorldWideWebSize www.worldwidewebsize.com

APPENDIX A

A.1 2007 RIC Results with the 2007 Evaluations

The results for the 2007 RIC task (removing minus from query) with the 2007 evaluations is presented in Table A.1 Columns represent the *number of articles* and the rows represent the *number of para's*.

Table A.1: 2007 RIC, Removing minus from query

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0767	0.0821	0.0865	0.0915	0.0951	0.0978	0.0977	0.0977	0.0977	0.0977	0.0977
50	0.0772	0.0889	0.0946	0.0980	0.1034	0.1129	0.1155	0.1154	0.1154	0.1154	0.1154
100	0.0772	0.0908	0.0989	0.1021	0.1098	0.1190	0.1305	0.1310	0.1315	0.1313	0.1313
150	0.0741	0.0913	0.0991	0.1048	0.1084	0.1197	0.1288	0.1345	0.1343	0.1346	0.1343
200	0.0750	0.0918	0.0982	0.1041	0.1088	0.1203	0.1299	0.1350	0.1375	0.1371	0.1374
250	0.0771	0.0921	0.0987	0.1040	0.1113	0.1243	0.1330	0.1385	0.1408	0.1408	0.1415
500	0.0772	0.0911	0.0985	0.1032	0.1112	0.1230	0.1326	0.1373	0.1393	0.1391	0.1405

The results for the 2007 RIC task with the 2008 evaluations are presented in Table

A.2 – Table A.16

Table A.2: 2007 RIC Child Strategy Exact, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0652	0.0786	0.0727	0.0786	0.0763	0.0786	0.0774	0.0786	0.0786	0.0458	0.0786
50	0.0652	0.0914	0.0743	0.0914	0.0817	0.0914	0.0852	0.0914	0.0914	0.0455	0.0908
100	0.0655	0.1036	0.0748	0.1036	0.0824	0.1036	0.0871	0.1036	0.1036	0.0455	0.0990
150	0.0655	0.1085	0.0751	0.1092	0.0827	0.1092	0.0875	0.1092	0.1092	0.0455	0.1003
200	0.0655	0.1111	0.0751	0.1128	0.0829	0.1128	0.0878	0.1128	0.1128	0.0455	0.1010
250	0.0655	0.1121	0.0752	0.1148	0.0831	0.1148	0.0881	0.1148	0.1148	0.0455	0.1013
500	0.0655	0.1133	0.0752	0.1174	0.0831	0.1174	0.0883	0.1174	0.1174	0.0455	0.1017

Table A.3: 2007 RIC Child Strategy Exact, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0762	0.0786	0.0764	0.0786	0.0785	0.0786	0.0787	0.0786	0.0786	0.0724	0.0786
50	0.0871	0.0914	0.0889	0.0914	0.0893	0.0914	0.0889	0.0914	0.0914	0.0758	0.0915
100	0.0894	0.1036	0.0964	0.1036	0.0998	0.1036	0.1006	0.1036	0.1036	0.0799	0.1015
150	0.0906	0.1095	0.0969	0.1093	0.1017	0.1092	0.1034	0.1092	0.1092	0.0802	0.1073
200	0.0912	0.1111	0.0974	0.1132	0.1009	0.1128	0.1057	0.1128	0.1128	0.0802	0.1097
250	0.0912	0.1131	0.0980	0.1157	0.1009	0.1150	0.1052	0.1148	0.1148	0.0801	0.1125
500	0.0927	0.1182	0.0976	0.1179	0.1013	0.1178	0.1029	0.1178	0.1174	0.0788	0.1125

Table A.4: 2007 RIC Correlation Strategy Exact, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0756	0.0778	0.0775	0.0778	0.0778	0.0778	0.0778	0.0778	0.0778	0.0660	0.0778
50	0.0854	0.0912	0.0834	0.0913	0.0876	0.0912	0.0887	0.0913	0.0913	0.0812	0.0901
100	0.0874	0.0918	0.0865	0.0998	0.0879	0.0978	0.0943	0.1032	0.0932	0.0819	0.0987
150	0.0891	0.0932	0.0873	0.1032	0.0892	0.0965	0.1031	0.1108	0.1032	0.0821	0.1032
200	0.0903	0.1003	0.0891	0.1064	0.0913	0.0978	0.1054	0.1132	0.1123	0.0831	0.1102
250	0.0905	0.1011	0.0912	0.1135	0.0942	0.1078	0.1032	0.1134	0.1132	0.0832	0.1123
500	0.0912	0.1132	0.0923	0.1172	0.1003	0.1132	0.1043	0.1165	0.1146	0.0843	0.1132

Table A.5: 2007 RIC Correlation Strategy Exact, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0776	0.0778	0.0779	0.0778	0.0778	0.0778	0.0778	0.0778	0.0778	0.0786	0.0778
50	0.0915	0.0906	0.0903	0.0906	0.0907	0.0906	0.0906	0.0906	0.0906	0.0857	0.0906
100	0.0987	0.1026	0.1031	0.1026	0.1025	0.1026	0.1025	0.1026	0.1026	0.0862	0.1027
150	0.0987	0.1084	0.1034	0.1084	0.1081	0.1084	0.1082	0.1084	0.1084	0.0844	0.1084
200	0.0988	0.1123	0.1058	0.1122	0.1094	0.1122	0.1115	0.1122	0.1122	0.0847	0.1124
250	0.0989	0.1147	0.1062	0.1146	0.1093	0.1146	0.1116	0.1146	0.1146	0.0847	0.1148
500	0.0965	0.1198	0.1045	0.1196	0.1091	0.1194	0.1120	0.1193	0.1193	0.0846	0.1176

Table A.6: 2007 RIC Section Strategy Exact, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0743	0.0790	0.0783	0.0790	0.0790	0.0790	0.0790	0.0790	0.0790	0.0572	0.0790
50	0.0757	0.0917	0.0844	0.0917	0.0885	0.0917	0.0905	0.0917	0.0917	0.0569	0.0917
100	0.0760	0.1039	0.0852	0.1039	0.0907	0.1039	0.0945	0.1039	0.1039	0.0569	0.1035
150	0.0760	0.1096	0.0855	0.1096	0.0911	0.1096	0.0950	0.1096	0.1096	0.0569	0.1064
200	0.0760	0.1132	0.0855	0.1134	0.0913	0.1134	0.0952	0.1134	0.1134	0.0569	0.1073
250	0.0760	0.1151	0.0856	0.1157	0.0914	0.1157	0.0955	0.1157	0.1157	0.0569	0.1078
500	0.0760	0.1171	0.0856	0.1196	0.0915	0.1196	0.0957	0.1196	0.1196	0.0569	0.1082

Table A.7: 2007 RIC Section Strategy Exact, Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0777	0.0790	0.0788	0.0790	0.0790	0.0790	0.0790	0.0790	0.0790	0.0742	0.0790
50	0.0869	0.0917	0.0898	0.0917	0.0910	0.0917	0.0915	0.0917	0.0917	0.0762	0.0917
100	0.0919	0.1039	0.0969	0.1039	0.0990	0.1039	0.1016	0.1039	0.1039	0.0770	0.1040
150	0.0907	0.1096	0.0974	0.1096	0.1013	0.1096	0.1034	0.1096	0.1096	0.0745	0.1085
200	0.0902	0.1135	0.0979	0.1134	0.1005	0.1134	0.1041	0.1134	0.1134	0.0742	0.1115
250	0.0897	0.1157	0.0971	0.1157	0.1024	0.1157	0.1043	0.1157	0.1157	0.0743	0.1117
500	0.0882	0.1188	0.0960	0.1202	0.1001	0.1197	0.1028	0.1196	0.1196	0.0739	0.1115

Table A.8: 2007 RIC Child Strategy UB,Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0655	0.1036	0.0748	0.1036	0.0824	0.1036	0.0871	0.1036	0.1036	0.0455	0.0990
50	0.0655	0.1085	0.0751	0.1092	0.0827	0.1092	0.0875	0.1092	0.1092	0.0455	0.1003
100	0.0655	0.1111	0.0751	0.1128	0.0829	0.1128	0.0878	0.1128	0.1128	0.0455	0.1010
150	0.0652	0.0786	0.0727	0.0786	0.0763	0.0786	0.0774	0.0786	0.0786	0.0458	0.0786
200	0.0655	0.1121	0.0752	0.1148	0.0831	0.1148	0.0881	0.1148	0.1148	0.0455	0.1013
250	0.0652	0.0914	0.0743	0.0914	0.0817	0.0914	0.0852	0.0914	0.0914	0.0455	0.0908
500	0.0655	0.1133	0.0752	0.1174	0.0831	0.1174	0.0883	0.1174	0.1174	0.0455	0.1017

Table A.9: 2007 RIC Correlation Strategy UB, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0769	0.0784	0.0784	0.0784	0.0780	0.0784	0.0782	0.0784	0.0784	0.0740	0.0784
50	0.0853	0.0924	0.0894	0.0924	0.0898	0.0924	0.0926	0.0924	0.0924	0.0769	0.0925
100	0.0906	0.1052	0.0953	0.1049	0.0973	0.1048	0.1021	0.1048	0.1048	0.0774	0.1046
150	0.0908	0.1086	0.0977	0.1087	0.1010	0.1085	0.1030	0.1084	0.1084	0.0753	0.1078
200	0.0910	0.1124	0.0980	0.1123	0.1018	0.1125	0.1042	0.1123	0.1123	0.0766	0.1103
250	0.0913	0.1170	0.0982	0.1175	0.1021	0.1172	0.1077	0.1172	0.1171	0.0786	0.1144
500	0.0906	0.1200	0.0969	0.1217	0.1023	0.1223	0.1084	0.1223	0.1224	0.0789	0.1158

Table A.10: 2007 RIC Section Strategy UB, Post Rearranged (Tagset 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0774	0.0792	0.0776	0.0792	0.0782	0.0792	0.0788	0.0792	0.0792	0.0740	0.0792
50	0.0851	0.0934	0.0904	0.0934	0.0903	0.0934	0.0929	0.0934	0.0934	0.0774	0.0934
100	0.0914	0.1062	0.0963	0.1059	0.0984	0.1058	0.1038	0.1058	0.1058	0.0776	0.1032
150	0.0914	0.1095	0.0979	0.1098	0.1021	0.1096	0.1041	0.1095	0.1095	0.0752	0.1086
200	0.0917	0.1114	0.0988	0.1128	0.1032	0.1136	0.1055	0.1134	0.1134	0.0756	0.1116
250	0.0919	0.1173	0.0990	0.1179	0.1034	0.1176	0.1094	0.1182	0.1180	0.0778	0.1163
500	0.0911	0.1210	0.0977	0.1216	0.1036	0.1225	0.1106	0.1223	0.1222	0.0780	0.1179

Table A.11: 2007 RIC Child Strategy Exact, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0763	0.0777	0.0769	0.0777	0.0774	0.0777	0.0777	0.0777	0.0777	0.0666	0.0777
50	0.0839	0.0906	0.0876	0.0906	0.0892	0.0906	0.0898	0.0906	0.0906	0.0689	0.0906
100	0.0839	0.1030	0.0937	0.1028	0.0977	0.1028	0.0995	0.1028	0.1028	0.0721	0.1029
150	0.0855	0.1084	0.0935	0.1086	0.0994	0.1082	0.1022	0.1083	0.1083	0.0727	0.1075
200	0.0876	0.1124	0.0932	0.1123	0.0990	0.1114	0.1025	0.1112	0.1112	0.0728	0.1103
250	0.0870	0.1144	0.0943	0.1147	0.0991	0.1131	0.1032	0.1124	0.1124	0.0722	0.1110
500	0.0869	0.1185	0.0943	0.1192	0.0993	0.1172	0.1019	0.1149	0.1139	0.0707	0.1120

Table A.12: 2007 RIC Correlation Strategy Exact, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0842	0.0829	0.0830	0.0829	0.0829	0.0829	0.0829	0.0829	0.0829	0.0838	0.0829
50	0.0978	0.0961	0.0972	0.0961	0.0967	0.0961	0.0962	0.0961	0.0961	0.0917	0.0961
100	0.1066	0.1085	0.1092	0.1085	0.1104	0.1085	0.1097	0.1085	0.1085	0.0947	0.1087
150	0.1077	0.1143	0.1135	0.1143	0.1151	0.1143	0.1153	0.1143	0.1143	0.0956	0.1153
200	0.1078	0.1185	0.1148	0.1183	0.1179	0.1183	0.1186	0.1183	0.1183	0.0960	0.1193
250	0.1088	0.1209	0.1148	0.1206	0.1185	0.1206	0.1201	0.1206	0.1206	0.0956	0.1219
500	0.1088	0.1269	0.1153	0.1266	0.1189	0.1253	0.1217	0.1245	0.1245	0.0948	0.1267

Table A.13: 2007 RIC Section Strategy Exact, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0767	0.0778	0.0781	0.0778	0.0778	0.0778	0.0779	0.0778	0.0778	0.0709	0.0778
50	0.0876	0.0908	0.0890	0.0908	0.0892	0.0908	0.0913	0.0908	0.0908	0.0724	0.0908
100	0.0869	0.1030	0.0943	0.1030	0.0987	0.1030	0.1002	0.1030	0.1030	0.0713	0.1031
150	0.0877	0.1088	0.0956	0.1086	0.0993	0.1086	0.1033	0.1086	0.1086	0.0717	0.1077
200	0.0869	0.1128	0.0955	0.1126	0.1003	0.1126	0.1033	0.1126	0.1126	0.0712	0.1114
250	0.0860	0.1152	0.0953	0.1150	0.1004	0.1144	0.1030	0.1144	0.1144	0.0710	0.1112
500	0.0838	0.1176	0.0938	0.1198	0.0991	0.1180	0.1033	0.1167	0.1165	0.0697	0.1111

Table A.14: 2007 RIC Child Strategy UB,Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0757	0.0777	0.0789	0.0777	0.0776	0.0777	0.0780	0.0777	0.0777	0.0742	0.0777
50	0.0847	0.0908	0.0882	0.0906	0.0913	0.0906	0.0922	0.0906	0.0906	0.0779	0.0922
100	0.0897	0.1042	0.0964	0.1029	0.1008	0.1029	0.1022	0.1028	0.1028	0.0809	0.1032
150	0.0917	0.1095	0.0990	0.1095	0.1037	0.1088	0.1059	0.1082	0.1083	0.0810	0.1088
200	0.0923	0.1127	0.0993	0.1135	0.1046	0.1131	0.1080	0.1119	0.1112	0.0815	0.1114
250	0.0927	0.1154	0.0995	0.1159	0.1049	0.1167	0.1088	0.1144	0.1126	0.0819	0.1139
500	0.0921	0.1185	0.0999	0.1206	0.1050	0.1209	0.1093	0.1215	0.1176	0.0815	0.1167

Table A.15: 2007 RIC Correlation Strategy UB, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0830	0.0829	0.0846	0.0829	0.0838	0.0829	0.0837	0.0829	0.0829	0.0797	0.0829
50	0.0919	0.0961	0.0951	0.0961	0.0976	0.0961	0.0979	0.0961	0.0961	0.0850	0.0968
100	0.0980	0.1094	0.1033	0.1087	0.1075	0.1085	0.1084	0.1085	0.1085	0.0877	0.1098
150	0.1010	0.1152	0.1056	0.1151	0.1105	0.1146	0.1124	0.1143	0.1143	0.0879	0.1155
200	0.1018	0.1197	0.1068	0.1190	0.1117	0.1189	0.1140	0.1184	0.1183	0.0882	0.1184
250	0.1019	0.1218	0.1071	0.1216	0.1121	0.1219	0.1151	0.1210	0.1206	0.0886	0.1203
500	0.1017	0.1264	0.1084	0.1270	0.1127	0.1271	0.1158	0.1274	0.1266	0.0875	0.1234

Table A.16: 2007 RIC Section Strategy UB, Rearranged (Tagset 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.0757	0.0778	0.0790	0.0778	0.0782	0.0778	0.0779	0.0778	0.0778	0.0742	0.0779
50	0.0846	0.0909	0.0882	0.0908	0.0912	0.0908	0.0923	0.0908	0.0908	0.0774	0.0916
100	0.0900	0.1039	0.0965	0.1033	0.1005	0.1030	0.1022	0.1030	0.1030	0.0799	0.1035
150	0.0923	0.1093	0.0990	0.1092	0.1037	0.1090	0.1055	0.1086	0.1086	0.0800	0.1089
200	0.0928	0.1130	0.0993	0.1130	0.1047	0.1128	0.1076	0.1128	0.1126	0.0803	0.1117
250	0.0932	0.1155	0.0995	0.1156	0.1049	0.1163	0.1085	0.1153	0.1147	0.0807	0.1140
500	0.0928	0.1193	0.1000	0.1206	0.1047	0.1208	0.1091	0.1212	0.1192	0.0802	0.1168

The following observations can be made from the 2007 RIC results with the 2008 evaluations:

- Correlation Strategy (UB) with Tagset 2 gives us the best result of 0.1274 (MAgP)
- For both 2008 and 2007 INEX Results the Correlation Strategy (UB) rearranged with Tagset 2 gives us the best RIC results.
- The best results are concentrated at 500 articles and 150-1500 para's.

A.2 Evaluation of the RIC results

The command to evaluate the RIC results from the fol (file offset and length) is given below:

```
./java -jar /smart/software_2008-assessment-tool/INEX_2008_results/inex_eval.jar -r  
-q /smart/software_2008-assessment-tool/INEX_2008_results/inex2007.qrels  
/smart/2007_coll_08_parsing/flex_removingminusword/focused_rearranged/fol/25/f  
ocused_25_50.txt > /smart/polum002/ric/2007_old/exact_child_post/evalj/25/50.txt
```

The parameter ‘-r’ is given to the `inex_eval.jar` while evaluating the results for the RIC task. The parameter ‘-q’ is an optional parameter to the `inex_eval.jar` for the per query results.