

UNIVERSITY OF MINNESOTA

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

Pavan K. Poluri

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

Donald B. Crouch

Name of Faculty Adviser

Signature of Faculty Adviser

August 14, 2009

Date

GRADUATE SCHOOL

Focused Retrieval using Exact Methodology

A thesis

submitted to the faculty of the graduate school
of the University of Minnesota

By

Pavan K. Poluri

In partial fulfillment of the requirements
for the degree of
Master of Science

August, 2009

Department of Computer Science,
University of Minnesota, Duluth
Duluth, MN 55812
USA

© Pavan K. Poluri 2009

Acknowledgements

I would like to thank Dr. Donald Crouch and Dr. Carolyn Crouch, for giving me this opportunity to work with them and for their consistent guidance during the research work.

I would like to thank my colleagues Dinesh Bhirud, Chaitanya Polumetla, and Varun Sudhakar for their support throughout this work.

I would like to thank my seniors Salil Bapat, Sarika Mehta, and Darshan Paranjape for sharing their valuable knowledge and for their consistent support to this work.

I would like to thank Dr. Pete Willemsen, Jim Luttenin, and Lori Lucia of Computer Science Department for their help in this research work.

Abstract

The goal of an information retrieval system is to retrieve information relevant to the user's query. In its earlier stages, the goal was to retrieve documents. But with the development of XML (Extensible Markup Language), identifying structures in a document became possible. The XML markup of a document gives the underlying structure in terms of its components or *elements*. The goal of XML-based retrieval systems is to retrieve relevant *elements* at the appropriate level of granularity. Our method of retrieving *elements* instead of entire documents is called *flexible retrieval* [3].

The objective of this research is to improve the results of the INEX [4] Ad Hoc track *Focused Task*. The purpose of the *Focused Task* is to return a list of non-overlapping elements in response to the user's query. In this thesis, we discuss various approaches used to perform *Focused* retrieval based on our *Exact* methodology and analyze the results.

Table of Contents

List of Figures.....	iv
List of Tables	vi
1. Introduction.....	1
2. Overview	3
2.1. INEX	3
2.2. Tasks	8
3. Background for Focused Retrieval	10
4. Experiments, Results and Analysis.....	23
4.1. Tag Sets and Term Weighting Parameters.....	23
4.2. Methodology for Focused Retrieval	27
4.3. Overlap Removal Strategies.....	28
4.4. Rearranging the Focused Output.....	32
4.5. Focused Task Experiments.....	32
4.6. Analysis of Results.....	48
5. Conclusion	54
6. Future Work.....	56
References.....	57
Appendix A	59
Appendix B.....	63

List of Figures

Figure 1: Sample Document, Document Id: 1516.xml (2008) [4].....	4
Figure 2: Sample Query, Query Id: 422 (2007) [4].....	6
Figure 3: Sample Relevance Assessment 2007 [4].....	7
Figure 4: Entry from Qrels File 2008 [4].....	7
Figure 5: Steps in Element Retrieval Process.....	10
Figure 6: Sample Document.....	11
Figure 7: Article Parsing of the Sample Document.....	12
Figure 8: Paragraph + mt Parsing of the Sample Document.....	12
Figure 9: Configuration File for Article Parsing.....	13
Figure 10: Doctree of a Document (Doc Id : 480380 [4]).....	14
Figure 11: Flex Configuration File.....	17
Figure 12: Sample Flex Output.....	18
Figure 13: Sample XML File Created.....	20
Figure 14: Sample File Produced after XML File Evaluation.....	21
Figure 15: Sample FOL File Produced from XML File.....	22
Figure 16: Sample Output (1) Produced by Section Strategy.....	28
Figure 17: Sample Output (2) Produced by Section Strategy.....	29
Figure 18: Sample Output (3) Produced by Section Strategy.....	29
Figure 19: Sample Output Produced by Child Strategy.....	30
Figure 20: Sample Output Produced by Correlation Strategy.....	30
Figure 21: General steps for Focused Retrieval using Exact Strategy.....	31
Figure 22: Comparison of Three Strategies on Basic version.....	48
Figure 23: Comparison of Three Strategies on RAC.....	49

Figure 24: Comparison of Three Strategies on RBC.....	50
Figure 25: Comparison of Section Strategy on Basic and RAC.....	51
Figure 26: Comparison of Child Strategy on Basic and RAC.....	52
Figure 27: Comparison of Correlation Strategy on Basic and RAC.....	52
Figure B.1: Query of Form 1.....	63
Figure B.2: Query of Form 2.....	63

List of Tables

Table 1: Details of Query sub fields [4]	5
Table 2: Tags-to-Index (Tag set 1)	23
Table 3: Tags-to-Keep (Tag Set 1)	24
Table 4: Tags-to-Index (Tag Set 2)	25
Table 5: Tags-to-Keep (Tag Set 2)	26
Table 6: Slope and Pivot Values	26
Table 7: iP[0.01] Section Strategy 2007 (Tag Set 1)	33
Table 8: iP[0.01] Section Strategy 2007 (Tag Set 2)	33
Table 9: iP[0.01] Section Strategy 2008 (Tag Set 1)	34
Table 10: iP[0.01] Section Strategy 2008 (Tag Set 2)	34
Table 11: iP[0.01] Child Strategy 2007 (Tag Set 1)	35
Table 12: iP[0.01] Child Strategy 2007 (Tag Set 2)	35
Table 13: iP[0.01] Child Strategy 2008 (Tag Set 1)	35
Table 14: iP[0.01] Child Strategy 2008 (Tag Set 2)	36
Table 15: iP[0.01] Correlation Strategy 2007 (Tag Set 1)	36
Table 16: iP[0.01] Correlation Strategy 2007 (Tag Set 2)	37
Table 17: iP[0.01] Correlation Strategy 2008 (Tag Set 1)	37
Table 18: iP[0.01] Correlation Strategy 2008 (Tag Set 2)	37
Table 19: iP[0.01] Section Strategy RAC 2007 (Tag Set 1)	38
Table 20: iP[0.01] Section Strategy RAC 2007 (Tag Set 2)	38
Table 21: iP[0.01] Section Strategy RAC 2008 (Tag Set 1)	39
Table 22: iP[0.01] Section Strategy RAC 2008 (Tag Set 2)	39
Table 23: iP[0.01] Child Strategy RAC 2007 (Tag Set 1)	40

Table 24: iP[0.01] Child Strategy RAC 2007 (Tag Set 1)	40
Table 25: iP[0.01] Child Strategy RAC 2008 (Tag Set 1)	40
Table 26: iP[0.01] Child Strategy RAC 2008 (Tag Set 2)	41
Table 27: iP[0.01] Correlation Strategy RAC 2007 (Tag Set 1).....	41
Table 28: iP[0.01] Correlation Strategy RAC 2007 (Tag Set 2).....	42
Table 29: iP[0.01] Correlation Strategy RAC 2008 (Tag Set 1).....	42
Table 30: iP[0.01] Correlation Strategy RAC 2008 (Tag Set 2).....	42
Table 31: iP[0.01] Section Strategy RBC 2007 (Tag Set 1).....	43
Table 32: iP[0.01] Section Strategy RBC2007 (Tag Set 2).....	43
Table 33: iP[0.01] Section Strategy RBC 2008 (Tag Set 1).....	44
Table 34: iP[0.01] Section Strategy RBC 2008 (Tag Set 2).....	44
Table 35: iP[0.01] Child Strategy RBC 2007 (Tag Set 1)	45
Table 36: iP[0.01] Child Strategy Exact RBC 2007 (Tag Set 2)	45
Table 37: iP[0.01] Child Strategy RBC 2008 (Tag Set 1)	45
Table 38: iP[0.01] Child Strategy RBC 2008 (Tag Set 2)	46
Table 39: iP[0.01] Correlation Strategy RBC 2007 (Tag Set 1).....	46
Table 40: iP[0.01] Correlation Strategy RBC 2007 (Tag Set 2).....	47
Table 41: iP[0.01] Correlation Strategy RBC 2008 (Tag Set 1).....	47
Table 42: iP[0.01] Correlation Strategy RBC 2008 (Tag 2)	47
Table A. 1: iP[0.01] Section Strategy.....	59
Table A. 2: iP[0.01] Correlation Strategy.....	59
Table A. 3: iP[0.01] Child Strategy	60
Table A. 4: iP[0.01] Section Strategy RAC	60
Table A. 5: iP[0.01] Child Strategy RAC.....	60
Table A. 6: iP[0.01] Correlation Strategy RAC	61

Table A. 7: iP[0.01] Section Strategy RBC.....	61
Table A. 8: iP[0.01] Child Strategy RBC.....	61
Table A. 9: iP[0.01] Correlation Strategy RBC	62
Table B. 1: iP[0.01] Focused Child Upper Bound 2007	64

1. Introduction

An information retrieval system is by definition a system that retrieves information relevant to the query of a user. In its initial stages the focus was document retrieval. But, with the development of XML (Extensible Markup Language), we have a mechanism to identify structures in a document. The XML markup of a document gives the underlying structure in terms of its components or *elements* (e.g., paragraphs, etc.). The focus of XML retrieval is the retrieval of relevant elements at the appropriate level of granularity, rather than the retrieval of entire documents. Our method for facilitating this type of retrieval, wherein elements of the document at various level of granularity are retrieved, is called *flexible* retrieval [3]. Flexible retrieval is also dynamic, as the elements are retrieved at run time [3].

The retrieval environment used for dynamic element retrieval makes use of the Vector Space Model [12]. In the Vector Space Model, both documents and queries are represented as weighted term vectors. Weights assigned to terms indicate the importance of the terms in the document. In flexible retrieval, we consider *paragraph* the basic indexing unit. (So the paragraph takes the position of the leaf node when the document is represented in the form of a tree.) Dynamic element retrieval produces a rank-ordered list of retrieved elements that is identical to the result produced by the same retrieval against an all-element index of the collection [9]. Normal element retrieval requires storing either an all-element index or multiple indices of the collection. Dynamic element retrieval has been proved to produce an identical result

to all-element retrieval in content only (CO) contexts and is more efficient with respect to file space and cost effective [9].

INEX [4] sponsors a competition that promotes the development of XML-based retrieval systems. INEX provides a document collection (set of documents), a query set (topics) and evaluation measures for the XML-based retrieval systems of the participants.

The objective of this research is to improve the results of the Focused Task of the INEX Ad Hoc track. The purpose of the Focused Task is to produce a ranked list of non-overlapping elements in response to the query. A detailed description of the Focused Task and other tasks of Ad Hoc track, the query set, relevance assessments and evaluation measures is given in Chapter 2.

2. Overview

This chapter gives a brief overview of INEX in terms of the document collection, topics (queries), tasks, relevance assessments, and the evaluation measures used.

2.1. INEX

INEX [4] stands for *Initiative for the Evaluation of XML Retrieval*. It is a competition in the field of Information Retrieval that attracts participants from all over the world. There are several tracks, namely, Ad Hoc, Book, Efficiency, Entity Ranking, Interactive (iTrack), Question Answering (QA@INEX), Link-the-Wiki, and XML-Mining. The University of Minnesota Duluth participates in the Ad Hoc track. The Ad Hoc track itself contains three tasks: Focused, Relevant In Context, and Best In Context. Any general purpose search engine would give a (ranked) list of documents in response to the user's query (document retrieval). The goal of the INEX Focused task is to produce focused (i.e., non-overlapping) elements that are relevant to the query. INEX provides the document collection to be used, set of topics, relevance assessments and the software to evaluate the results.

Document Collection

The 2007 and 2008 document collection is approximately 5.8 GB in size distributed over 22 parts. Each document in this collection is an XML document containing text enclosed in XML tags. This collection has approximately 5000 unique tags in it [8]. See [4] for more details regarding the document collection. Some text in the

documents is untagged, making the collection semi-structured. The 2009 document collection is approximately 50 GB (approximately 8.6 times the size of the previous collection) distributed over 1000 parts. This collection is also semi-structured and has over 30000 unique tags in it. A sample document from one of these collections is shown in Figure 1.

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
<name id="1516">
Albert the Degenerate
</name>
<conversionwarning>
0
</conversionwarning>
<body>
<template name="Merge">
Albrecht II, Markgraf of Meißen
</template>
<emph3>
Albert, surnamed "The Degenerate"
</emph3>
  (c. 1240 â□“
<unknownlink src="13 November">
13 November
...
</article>
```

Figure 1: Sample Document, Document Id: 1516.xml (2008) [4]

Topics

Each participant group in the competition is asked to create a set of queries and submit them to INEX. The final set of queries is decided by INEX and is supplied to the participants. The queries are in the form of CO + S (Content Only + Structure). A query in CO + S form would have the following fields in it.

Table 1: Details of Query sub fields [4]

Field	Description
<title>	In which Content Only (CO) queries are given
<castitle>	In which Content and Structure (CAS) queries are given
<description>	A one or two sentence natural language definition of the information need
<narrative>	In which the definitive definition of relevance and irrelevance are given

There are 285 queries in total given for the 2008 collection. Before participants their submit the queries to INEX they need to run them on Max Planck provided search engine called TopX [5] to ensure that the number of relevant documents retrieved for that query lies between 2 and 20. For 2007 collection, there are a total of 127 topics [1]. A sample query appears in Figure 2.

Relevance Assessments

INEX provides relevance assessments that are used to evaluate the participant's results. Relevance assessments are produced through manual assessment using a tool called GPXRai [4]. Assessments from all the participants are then merged and that becomes the final relevance assessment pool. In 2007, INEX released assessments containing the relevant text from each of the documents in the pool for each query. A sample relevance assessment for a query is seen in Figure 3.

```

<inex_topic topic_id="422" ct_no="19">
<title>birds or passerine songs</title>
<castitle>//article[(about(., bird) or
about(.,passerine))]/p[about(.,song)]</castitle>
<description>Find information in paragraphs about the
different kinds of songs of birds, especially
passerines.</description>
<narrative>I have always been interested in finding out
information on the different kinds of songs that birds
sing, especially passerines.

For my own personal interest, I would like to learn more
information on the complexity of the songs in term of
rhythm and melody. I am also interested on finding
information on the periods when different birds sing and
how birds learn to sing.

To be relevant an paragraph element should describe one
or several songs of birds. To be relevant, the document
should explain the songs.</narrative>

</inex_topic>

```

Figure 2: Sample Query, Query Id: 422 (2007) [4]

In 2008, INEX changed the way results are evaluated. Rather than using the text in the path to the specific element, they evaluated with respect to what is called File-Offset-Length (FOL). File Offset refers to the starting point in the document and length refers to the size (in number of characters) of the relevant text from the offset. To achieve this, they produced a qrels file containing the information needed for evaluation. A typical entry line in a qrels file is seen in Figure 4. This file contains relevance assessments for only 70 queries.

```
<file collection="wikipedia" name="233116">
  <best-entry-point
path="/article[1]/body[1]/p[4]/text()[4].108"/>
  <passage
start="/article[1]/body[1]/p[4]/text()[4].108"
end="/article[1]/body[1]/p[4]/text()[4].134"
size="26"/>
  <element path="/article[1]/body[1]/p[4]"
exhaustivity="2" size="339" rsize="26"/>
  <element path="/article[1]/body[1]"
exhaustivity="2" size="1238" rsize="26"/>
  <element path="/article[1]" exhaustivity="2"
size="1256" rsize="26"/>
```

Figure 3: Sample Relevance Assessment 2007 [4]

```
544 Q0 177316 572 4798 1915 1915:299 3711:273
544 - Query Number
Q0 - Query type
177316 - Document ID
1915 - Best entry Point
1915:299 - A relevant passage starting at an offset of
1915 and continuing for 299 characters
3711:273 - Another relevant passage starting at an
offset of 3711 and continuing for 273 characters
```

Figure 4: Entry from Qrels File 2008 [4]

Evaluation Measures

Different tasks are evaluated based on different metrics [4].

Focused Task

The results of the Focused task are evaluated using the metric interpolated precision, or more precisely interpolated precision at 1% recall, i.e., $iP[0.01]$. Recall is defined as fraction of highlighted text that is retrieved and Precision is defined as fraction of retrieved text that is highlighted. See [4] for details.

Relevant In Context Task

The results of this task are evaluated using generalized precision and recall. Mean average generalized precision (MaGP) is used to get an overall performance estimate. See [4] for details.

Best In Context Task

The results of this task are evaluated using generalized precision and recall. Mean average generalized precision is used to get an overall performance estimate. See [4] for details.

2.2. Tasks

The three INEX Ad Hoc tasks are: Focused, Relevant In Context, and Best In Context.

Focused Task

[4] For each query, the output of the Focused task is a ranked list of the most specific (focused) elements from the relevant documents. The size of this list depends upon the number of relevant articles per query and number of focused elements per document. The rank of an element is based on its correlation score with respect to the query. Depending upon the number of articles and elements in question, the R top-ranked elements may come from a few, or many, documents. The list of elements being reported must not contain overlap, which means the same text should not be reported more than once. For example, if the text enclosed in a section tag is being returned, then the text enclosed in a paragraph tag, which itself is enclosed in the section tag (the section in this scenario is the parent of paragraph), should not be returned in the list because it is already being returned as a part of the section.

Relevant In Context Task

[4] As a response to the user's query, the Relevant In Context task returns a ranked list of focused elements from each of the top-ranked documents presented in document order. See [11] for details.

Best In Context Task

[4] As a response to the user's query, the Best In Context task returns the best entry point associated with each relevant document. The best entry point of a document is defined as that point in the document where you start reading to meet the informational need expressed by the query. See [8, 14] for details.

3. Background for Focused Retrieval

This chapter gives an overview of retrieval, from the processing of the document collection and the query set to the retrieval of the elements associated with each query. See Figure 5 below for details.

1. Cleaning the document collection (2009 collection)
2. Parsing the document collection
3. Indexing the document collection and the queries
4. Smart retrieval (para + mt parse)
5. Seeding
6. Flex element retrieval
7. Converting Flex output into the required task format
8. Evaluating the result

Figure 5: Steps in the Element Retrieval Process

Cleaning the Document Collection

The cleaning stage ensures that the collection is in a good form for parsing. For each document in the entire collection, we guarantee that each XML tag in the document also has a corresponding closing XML tag in the document – essential if the parsing scripts are to do their job. (The 2009 collection has many tags along with the text in the documents.) All unwanted tags are removed from the collection in this phase while the text enclosed between those tags is kept. (Thus the text enclosed within tags that have been removed is retained as a part of the containing element.)

Parsing the Document Collection

Once the data has been cleaned, it is parsed. The data is parsed with respect to articles (article parsing) and paragraphs (paragraphs + mt parsing). Parsing is the process of recognizing as an entity all the text enclosed within a matching set of open and closed XML tags. Article parsing results in all text associated within an article being enclosed with article tags (all other tags removed). The word *paragraph* here actually refers to a set of tags we have chosen as terminal node tags, which include the paragraph tag. Magic text (*mt*), refers to untagged text in the document. Magic text is important because Flex builds the documents from the terminal nodes. In order to build a parent node correctly, Flex needs to have all its children. Untagged text is enclosed between artificial `<mt> ... </mt>` tags. See [6] for detailed discussion of the importance of untagged text. Figure 6 shows an example of a paragraph element and untagged text in the article.

```
<article>
<body>
Roger Federer
<p>
Roger Federer equals Pete Sampras's 14 Grand Slams Record
</p>
</body>
</article>
```

Figure 6: Sample Document

An article parse for this document would result in the article element as seen in Figure 7.

```
Roger Federer Roger Federer equals Pete Sampras's 14  
Grand Slams Record
```

Figure 7: Article Parsing of the Sample Document

A *paragraph + mt* parse for the document would produce the following elements as seen in Figure 8.

```
<p>  
Roger Federer equals Pete Sampras's 14 Grand Slams Record  
</p>  
<mt>  
Roger Federer  
</mt>
```

Figure 8: Paragraph + mt Parsing of the Sample Document

We need to specify a configuration file for parsing. A configuration file for article parsing is seen in Figure 9. *tags-to-keep* are all the tags we identify as recognizable elements in the document; i.e., these tags identify elements that can be returned. *tags-to-index* are the tags that are identified as terminal nodes. *tags-to-index* is always a subset of *tags-to-keep*. Chapter 4 discusses the configuration files used for the experiments.

```
tags_to_keep
article,body
tags_to_index
body
```

Figure 9: Configuration File for Article Parsing

Indexing the Document Collection and Queries

Indexing of the document collection is done with respect to the *article*, *section* and *paragraph + mt* parses. Article indexing uses the *article* parse as input, paragraph + mt indexing uses the *paragraph + mt* parse as input. The result is a set of nnn weighted element vectors (term frequency vectors), the dictionary (dict.words all the unique words in the collection), the inverted file (inv.words) and the textloc file. Indexing is performed using Smart.

Smart Retrieval

Lnu-ltu [1] term weighting is used for all vectors. The similarity measure used is inner product. Retrieval produces a rank-ordered list of elements (e.g., articles, paragraph + mt elements). See [13] for details regarding the *Lnu-ltu* weighting scheme. The collection-dependent parameters, slope and pivot, must be determined for each vector set.

Seeding

Seeding takes place in two steps:

Generation of Doctrees

As a precursor to seeding, we generate document schemas, which represent a preorder traversal of the document tree. The schema or doctree contains the Xpath of every recognizable element. An example of a schema or doctree can be seen in Figure 10.

/article[1]	2	0
/article[1]/name[1]	0	1
/article[1]/body[1]	10	0
/article[1]/body[1]/mt[1]	0	1
/article[1]/body[1]/p[1]	0	1
/article[1]/body[1]/p[2]	0	1
/article[1]/body[1]/p[3]	0	1
/article[1]/body[1]/p[4]	0	1
/article[1]/body[1]/p[5]	0	1
/article[1]/body[1]/p[6]	0	1
/article[1]/body[1]/p[7]	0	1
/article[1]/body[1]/p[8]	0	1
/article[1]/body[1]/p[9]	0	0

Figure 10: Doctree of a Document (Doc Id : 480380 [4])

The number in the square braces (e.g., [3]) refers to the number of times the preceding tag has occurred in the document. The general case is that a document

would have only one set of <article> ... </article> tags and one set of <body> ... </body> tags. A document can have multiple occurrences of any other tag. (Doctree generation takes as input the configuration file used by *paragraph + mt* parsing and the unparsed document collection. The configuration file of the *paragraph + mt* parse contains all the tags considered terminal tags).

Seeding

Once the doctrees are generated, a large *paragraph +mt* element retrieval is performed (we use 125,000 elements) to ensure that every element of each top-ranked document is retrieved. Then the output of this retrieval, along with the doctrees already created and the *docid_docpath_mappings* file (generated during indexing with the mapping of the Smart document identifier to the original identifier of the document) are input to the seeding script. The process of seeding populates the trees; i.e., fills in the content of all the terminal elements of the doctree. When the seeding process is done, all the terminal nodes of the doctree are populated. If we fail to populate all terminal nodes of the doctrees, Flex will generate incomplete trees—hence the very large retrieval.

Focused retrieval requires us to combine the article retrieval [1] of step 4 with the *paragraph + mt* retrieval (used for seeding) to generate the trees of the documents in interest. Thus, given a query, for every document retrieved in the corresponding article retrieval, we extract the tree that corresponds to that

document from set of all seeded documents. This set of selected trees (of all the documents retrieved from article retrieval) is called seed subsets. With the generation of seed subsets the seeding process comes to an end.

Flex

Flexible retrieval is performed after seeding. The seed subsets produced are taken as input by Flex. Flex uses the content of all terminal nodes to generate the parent vector and calculate the score of the parent vector with the query. This process of calculating scores starts at the leaf nodes of the document tree and proceeds towards the root of the document tree. So by the end of this process we have correlations for each element in the document tree with the query. Flex cannot calculate the score of a parent correctly unless it has available all the child nodes of the parent (otherwise producing incomplete trees). Having calculated the correlation of each element with the query, Flex then outputs a ranked list of elements for each document retrieved. The correlation score of each element along with the element itself is present in the output of Flex. Flex takes as input a configuration file such as seen in Figure 11.

Once the Flex output of elements in rank order is generated, the *mt* elements inserted during parsing must be removed. *mt* tags have been introduced to capture the untagged text in a document, enabling Flex to generate the doctrees properly. In the *mt* removal stage, each element having an *mt* tag in its Xpath is removed.

1. DOC_INDEX_PATH: path to the doc.nnn file of the paragraph + mt indexing
 2. QUERY_INDEX_PATH: path to the query.nnn file of the paragraph + mt indexing
 3. OUTPUT_PATH: path to the output produced by Flex
 4. TREES_PATH: path to the output of seeding
 5. NUM_OUTPUT_ELEMS: Number of elements in the ranked list per query
 6. SLOPE_PARA: slope value used for paragraph + mt**
 7. SLOPE_SEC: slope value used for sections**
 8. SLOPE_BDY: slope value used for body**
 9. SLOPE_ALLELEMS: slope value of all elements*
 10. PIVOT_PARA: pivot value used for paragraph + mt**
 11. PIVOT_SEC: pivot value used for sections**
 12. PIVOT_BDY: pivot value used for body**
 13. PIVOT_ALLELEMS: pivot value of all elements*
 14. N_PARA_VALUES: Number of paragraphs. Can get this value from the textloc.txt file of the paragraph + mt indexing
 15. N_SEC_VALUES: Number of sections. Can get this value from the textloc.txt file of the section indexing
 16. N_BDY_VALUES: Number of articles. Can get this value from the textloc.txt file of the article indexing
 17. N_STATS_FOLDER: path to the paragraph + mt indexing
 18. N_STATS_FILES: ctype0
- * Used by Flex
- **Not used by Flex

Figure11: Flex Configuration File

(For the 2007 collection, after removal of the *mt*, we needed to do expand the Xpaths [1]). This step is eliminated in 2008 collection because evaluation of results has changed and Xpath expansion no longer improves results [2]. A sample Flex output (after removing *mt* elements) is shown in Figure 12. For each query (1 in this example), we see the Xpath of the element and the correlation score of the element with the query.

```
1 985414/article[1]/body[1]/section[1] 28.8246
1 20347/article[1]/body[1]/section[3]/section[6] 28.4672
1 261763/article[1]/body[1] 28.1136
1 20347/article[1]/body[1]/section[3]/section[7] 27.8816
1 20347/article[1]/body[1]/section[6] 27.2642
1 682628/article[1]/body[1] 27.1834
1 20347/article[1]/body[1]/p[1] 26.1209
```

Figure 12: Sample Flex output

Converting Flex Output into Required Rask Format

With the completion of Flexible retrieval, we have the specified number *m* of elements from the *n* retrieved documents along with their correlation scores. The Flex output must be converted to the corresponding task, namely, Focused, Relevant in Context and Best in Context. See Chapter 4 for details.

Evaluation of the Results

After the output of Flex is converted into the required task format, we evaluate the results. Evaluation takes place in three steps. We first convert our Focused (or Relevant in Context or Best in Context) files into XML format. This conversion is achieved with the help of a script provided by INEX [4]. A sample XML file produced by this script is given in Figure 13.

In the second step, we validate the XML files created in the first step. This step checks for any invalid Xpaths or for overlap in the XML files (the final result should not contain overlap). INEX provides a package for this purpose [4], which reports an error in case of invalid Xpaths or overlap. In the final step, we use scripts provided by INEX (`foc_dom_eval2.pl` for Focused, `ric_dom_eval2.pl` for Relevant in Context, `bic_dom_eval2.pl` for Best in Context) to evaluate the XML files. These files take in as input the DB_File (database of all elements with global offsets), relevance assessments [7] (which can be downloaded from INEX website), and the XML files (produced after validation). A sample file obtained through evaluation of an XML file can be seen in Figure 14.

In 2008, INEX changed the method of evaluating the results (although evaluation metrics are the same). After the XML files are produced, they are converted into File Offset Length format using the package provided by INEX. No relevance assessments are provided, but a `qrels` file with manual relevance assessments converted into File Offset Length format is made available. The evaluation program (in java) [4] takes as

```

<inex-submission participant-id="51" run-id="focused"
task="Focused" query="automatic" result-type="element">
  <topic-fields title="yes" mmtitle="no" castitle="no"
description="no" narrative="no"/>
  <description></description>
  <collections>
    <collection>wikipedia</collection>
  </collections>

  <topic topic-id="544">

    <result>
      <file>985414</file>
      <path>/article[1]/body[1]/section[1]</path>
      <rank>1</rank>
    </result>
    <result>
      <file>20347</file>

      <path>/article[1]/body[1]/section[3]/section[6]</path>
      <rank>2</rank>
    </result>

    <result>
      <file>20347</file>

      <path>/article[1]/body[1]/section[3]/section[7]</path>
      <rank>3</rank>
    </result>
  </topic>
</inex-submission>

```

Figure 13: Sample XML File Created

input the qrels file provided by INEX, the File Offset Length file (obtained from converting xml files) and a file that contains File Offset Length of every element in the collection (provided by INEX) and gives the result. A sample FOL (File-Offset-Length) file produced from XML file can be seen in Figure 15. The file obtained through evaluating the FOL file would resemble the file shown in Figure 14.

```
<eval run-id="p51_focused"
file="/smart/2007_coll_08_parsing/flex_removingminusword/foc
used_section_exact/xml/25/focused_25_50.xml">

num_q          all          106
num_ret        all          6374145
num_rel        all          10614635
num_rel_ret    all          1210084
iP[0.00]       all          0.516475240613331
iP[0.01]       all          0.470123208937047
iP[0.05]       all          0.359490393028904
iP[0.10]       all          0.289673978169276
MAiP           all          0.0923127000252057
ircl_prn.0.00  all          0.516475240613331
ircl_prn.0.01  all          0.470123208937047
```

Figure 14: Sample File Produced after XML File Evaluation

```
544 Q0 985414 1 -1 p51-focused 393 785
544 Q0 20347 2 -1 p51-focused 13174 2365
544 Q0 20347 3 -1 p51-focused 15539 1658
544 Q0 20347 4 -1 p51-focused 23262 1670
544 Q0 20347 5 -1 p51-focused 74 637
544 Q0 20347 6 -1 p51-focused 2344 3475
544 Q0 233013 7 -1 p51-focused 721 2076
544 Q0 1654632 8 -1 p51-focused 622 1400
544 Q0 20347 9 -1 p51-focused 21806 1456
544 Q0 261763 10 -1 p51-focused 1405 727
544 O0 20347 11 -1 p51-focused 17197 1395
```

Figure 15: Sample FOL File Produced from XML File

4. Experiments, Results and Analysis

This chapter discusses the experiments performed on the INEX 2007 and 2008 collections.

4.1. Tag Sets and Term Weighting Parameters

We need to specify two tag sets: *tags-to-keep* (tags that are kept in parsing) and *tags-to-index* (tags we have chosen to represent terminal nodes). The tag set used in the results submitted in 2007 can be found in [1]. Table 2 lists the *tag-to-index* and Table 3 lists *tags-to-keep*. We identify this tag set as tag set 1.

Table 2: Tags-to-Index (Tag set 1)

Tag Name	Tag Representation
Name	<name> ... </name>
Magic Text	<mt> ... </mt>
Paragraph	<p> ... </p>
Figure	<figure> ... </figure>
Emphasis	<emph3> ... </emph3>

In an attempt to improve the results, we produced a revised tag set called tag set 2. Tables 4 and 5 identify these tags. The same sets of experiments are done with both tag sets.

Table 3: Tags-to-Keep (Tag Set 1)

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

Slope and pivot values must also be specified for retrieval and in the configuration file for Flex. The purpose of using slope and pivot values is to moderate the effect of the length differences in the vectors. This is done in order to avoid the dominance of longer vectors over shorter vectors. The *Lnu-ltu* [13] term weighting scheme uses the values of slope and pivot in its weighting formulae. Table 6 shows the slope and pivot values used in our experiments.

Table 4: Tags-to-Index (Tag Set 2)

Tag Name	Tag Representation
Paragraph	<p> ... </p>
Figure	<figure> ... </figure>
Name	<name> ... </name>
Emphasis	<emph3> ... </emph3>
Magic Text	<mt> ... </mt>
Table	<table> ... </table>
Normal-list	<normallist> ... </normallist>
Ordered-list	 ...
Unordered-list	 ...
Number-list	<numberlist> ... </numberlist>
Definition-list	<definitionlist> ... </definitionlist>

Table 5: Tags-to-Keep (Tag Set 2)

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

Table 6: Slope and Pivot Values

Retrieval Type	Slope	Pivot
Article	0.04	120
Paragraph	0.12	18

4.2. Methodology for Focused Retrieval

We have developed two strategies for Focused Retrieval. One is called the *Upper Bound Strategy* and the second is called the *Exact Strategy*. This research describes the *Exact Strategy* and the experiments based on it. A brief overview of both strategies follows.

Upper Bound Strategy

In this strategy, Flex gives for each query a ranked list of m elements from n documents. The value of m (e.g., 50, 100, etc.) is specified in the configuration file for Flex. After Flex produces its output, we must eliminate the *magic text* (mt) elements from it. In this stage, some of the m elements may be eliminated (i.e., the mts). Now the Flex output contains up to m elements from n documents where m is the upper bound on the number of elements produced (per query)—hence the name, *Upper Bound Strategy*. See [2] for details.

Exact Strategy

In this strategy, we guarantee that for each query exactly m elements are retrieved from the n documents of interest even after the removal of all mt elements. Here, instead of specifying m as the number of elements required from Flex, we specify a larger number (in our case 30,000 in the configuration file) to ensure that we get all the elements from each retrieved article. From the rank-ordered list of elements provided by Flex for each query, we remove the elements with mt in their Xpath.

Even though some elements are eliminated, we still have a long ranked list of elements. We return the top ranked m elements from the list.

4.3. Overlap Removal Strategies

In order to convert Flex output into focused format, we need to remove overlap from Flex output (overlap is not allowed in the final result). We have three strategies for overlap removal, namely: section, child and correlation strategy.

Section Strategy

This strategy selects the highest correlating non-body element. If the parent has a higher correlation score than the child, and if the parent is not a body, parent is preferred to child. If the child has a higher correlation score, then child is preferred to parent. Figures 16, 17 and 18 illustrate this strategy. In Figure 16, child has a higher correlation than parent and hence finds a place in the Focused output. In Figure 17, parent has a higher correlation score than child, but parent is a body. Hence, child element finds a place in the Focused output. In Figure 18, parent has a higher correlation score than child and parent is not a body. Hence, the parent element occurs in the Focused output.

```
1 20347/article[1]/body[1]/section[4]/p[3] 32.04
...
1 20347/article[1]/body[1]/section[4] 23.98

Focused output:

1 20347/article[1]/body[1]/section[4]/p[3]
```

Figure 16: Sample Output (1) Produced by Section Strategy

```
1 1516/article[1]/body[1] 56.04
.
.
.
1 1516/article[1]/body[1]/section[2]/p[4] 32.98

Focused output:

1 1516/article[1]/body[1]/section[2]/p[4]
```

Figure 17: Sample Output (2) Produced by Section Strategy

```
1 1516/article[1]/body[1]/section[2] 56.04
.
.
.
1 1516/article[1]/body[1]/section[2]/p[4] 32.98

Focused output:

1 1516/article[1]/body[1]/section[2]
```

Figure 18: Sample Output (3) Produced by Section Strategy

Child Strategy

In this strategy, preference is given to child over parent. So if there are two overlapping elements, where one is a child of the other, the child element appears in the Focused output whereas the parent element is discarded. Figure 19 shows an example of the output produced by this strategy. From the figure it can be observed that the second element (child) is placed in the Focused output and the first element (parent) is discarded (even though parent had a higher correlation score than the child). Only the element itself is present in the Focused output (i.e., correlation scores are not present in the output file). See [1, 10] for details.

```
1 20347/article[1]/body[1]/section[4] 46.04
.
.
.
1 20347/article[1]/body[1]/section[4]/p[3] 32.98

Focused output:

1 20347/article[1]/body[1]/section[4]/p[3]
```

Figure 19: Sample Output Produced by Child Strategy

Correlation Strategy

In this strategy, preference is given to the element that has the higher correlation score. So if there are two overlapping elements, the element with the higher correlation score appears in Focused output and the other element is discarded. Figure 20 shows an example of the output produced by this strategy. Figure 21 shows the general steps for Focused Retrieval using *Exact Strategy*.

```
1 20347/article[1]/body[1]/section[4] 46.04
.
.
.
1 20347/article[1]/body[1]/section[4]/p[3] 32.98

Focused output:

1 20347/article[1]/body[1]/section[4]
```

Figure 20: Sample Output Produced by Correlation Strategy

- Parse the document collection to produce *article* and *paragraph + mt* parses.
- Index the parses to produce *article* and *paragraph + mt* indexes.
- Index the query set.
- Apply *Lnu-ltu* [12] term weighting to the vectors.
- Generate the doctrees.
- For each query
 - Using Smart, retrieve 25, 50, 100, 150, 200, 250, and 500 documents.
 - Perform a large 125000 *paragraph + mt* Smart retrieval.
 - Seed the doctrees.
 - Generate the seed subsets for the articles retrieved in step 5.
 - Flex populates the doctrees, producing a ranked list of elements (the maximum size of this list is 30,000).
- Using one of the three overlap removal strategies (*Child, Correlation, and Section*), convert the Flex output into *Focused* format.
- For each document set retrieved n (= 25, 50 etc.), the list of *Focused* elements is restricted to m (= 50, 100, 150, 200, 250, 500, 1000, 1500, 2000, 3000, and 4000) elements.
- Convert the Focused output into XML format.
- Convert the XML files into File-Offset-Length (FOL) format.
- Evaluate the XML files.

Figure 21: General steps for Focused Retrieval using Exact Strategy

4.4. Rearranging the Focused Output

The Focused task requires that its output be a ranked list of focused elements. We conducted experiments to determine if arranging this output by document would produce an improved result for the Focused task.

Given the ranked list of focused elements, and an upper bound (30,000) on the length of this list, the *Exact Strategy* has two possibilities of rearrangement.

1. *Rearrangement after chopping the big Focused output*

Take the top m elements from the long list of Focused output and then sort the list by document. This methodology is referred to as *RAC*.

2. *Rearrangement before chopping the big Focused output*

Sort the long list of Focused output by document, and then take top m elements from the sorted list. This methodology is referred to as *RBC*.

4.5. Focused Task Experiments

The experiments described in this section, are performed on 2007 data using both 2007 and 2008 evaluation packages. All results reported within this section are performed using the 2008 evaluation package. For corresponding results evaluated using 2007 package, see Appendix A. All experiments are performed on the 2007 and 2008 collections using tag set 1 and tag set 2. In all these tables, the column headings

refer to the number of *elements* and the row headings refer to number of *articles* retrieved.

Experiment 1

In this experiment, *Section Strategy* is used as the overlap removal strategy. The retrieval is performed according to the steps shown in Figure 21. Tables 7 - 10 give the results of this experiment.

Table 7: iP[0.01] Section Strategy 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4769	0.4778	0.4780	0.4780	0.4780	0.4780	0.4780	0.4780	0.4780	0.4780	0.4780
50	0.4661	0.4677	0.4681	0.4683	0.4683	0.4683	0.4683	0.4683	0.4683	0.4683	0.4683
100	0.4549	0.4573	0.4576	0.4581	0.4581	0.4584	0.4584	0.4584	0.4584	0.4584	0.4584
150	0.4515	0.4541	0.4544	0.4547	0.4549	0.4551	0.4552	0.4552	0.4552	0.4552	0.4552
200	0.4508	0.4527	0.4533	0.4536	0.4536	0.4540	0.4541	0.4541	0.4541	0.4541	0.4541
250	0.4490	0.4511	0.4516	0.4517	0.4518	0.4519	0.4520	0.4520	0.4520	0.4520	0.4520
500	0.4460	0.4481	0.4485	0.4486	0.4486	0.4487	0.4488	0.4488	0.4488	0.4488	0.4488

Table 8: iP[0.01] Section Strategy 2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4660	0.4680	0.4683	0.4684	0.4688	0.4688	0.4688	0.4688	0.4688	0.4688	0.4688
50	0.4456	0.4489	0.4493	0.4496	0.4496	0.4498	0.4498	0.4498	0.4498	0.4498	0.4498
100	0.4411	0.4436	0.4444	0.4450	0.4452	0.4457	0.4458	0.4458	0.4458	0.4458	0.4458
150	0.4412	0.4433	0.4439	0.4445	0.4447	0.4452	0.4452	0.4452	0.4452	0.4452	0.4452
200	0.4398	0.4418	0.4423	0.4427	0.4429	0.4432	0.4432	0.4432	0.4432	0.4432	0.4432
250	0.4389	0.4411	0.4413	0.4417	0.4418	0.4420	0.4420	0.4420	0.4420	0.4420	0.4420
500	0.4364	0.4374	0.4380	0.4381	0.4382	0.4382	0.4383	0.4383	0.4383	0.4383	0.4383

Table 9: iP[0.01] Section Strategy 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6428	0.6455	0.6455	0.6455	0.6455	0.6455	0.6455	0.6455	0.6455	0.6455	0.6455
50	0.6496	0.6517	0.6521	0.6521	0.6521	0.6521	0.6521	0.6521	0.6521	0.6521	0.6521
100	0.6415	0.6429	0.6429	0.6429	0.6434	0.6443	0.6443	0.6443	0.6443	0.6443	0.6443
150	0.6383	0.6388	0.6397	0.6397	0.6397	0.6410	0.6410	0.6410	0.6410	0.6410	0.6410
200	0.6299	0.6310	0.6314	0.6314	0.6314	0.6315	0.6327	0.6327	0.6327	0.6327	0.6327
250	0.6324	0.6338	0.6342	0.6342	0.6342	0.6343	0.6355	0.6355	0.6355	0.6355	0.6355
500	0.6289	0.6316	0.6322	0.6322	0.6322	0.6322	0.6323	0.6336	0.6336	0.6336	0.6336

Table 10: iP[0.01] Section Strategy 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6377	0.6415	0.6420	0.6426	0.6426	0.6426	0.6426	0.6426	0.6426	0.6426	0.6426
50	0.6208	0.6214	0.6232	0.6232	0.6232	0.6233	0.6233	0.6233	0.6233	0.6233	0.6233
100	0.6155	0.6166	0.6170	0.6170	0.6170	0.6170	0.6170	0.6170	0.6170	0.6170	0.6170
150	0.6118	0.6123	0.6129	0.6130	0.6130	0.6130	0.6131	0.6131	0.6131	0.6131	0.6131
200	0.6092	0.6100	0.6106	0.6107	0.6107	0.6108	0.6108	0.6108	0.6108	0.6108	0.6108
250	0.6118	0.6123	0.6124	0.6126	0.6126	0.6126	0.6127	0.6127	0.6127	0.6127	0.6127
500	0.6101	0.6104	0.6104	0.6104	0.6105	0.6105	0.6106	0.6106	0.6106	0.6106	0.6106

Experiment 2

In this experiment, *Child Strategy* is used as the overlap removal strategy. The retrieval is performed according to the steps shown in Figure 21. Tables 11 - 14 give the results of this experiment.

Table 11: iP[0.01] Child Strategy 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.5271	0.5295	0.5311	0.5311	0.5311	0.5311	0.5311	0.5311	0.5311	0.5311	0.5311
50	0.5083	0.5122	0.5137	0.5138	0.5141	0.5143	0.5143	0.5143	0.5143	0.5143	0.5143
100	0.4967	0.5002	0.5016	0.5024	0.5027	0.5032	0.5035	0.5035	0.5035	0.5035	0.5035
150	0.4935	0.4971	0.4978	0.4985	0.4992	0.4998	0.5000	0.5000	0.5000	0.5000	0.5000
200	0.4849	0.4887	0.4895	0.4907	0.4908	0.4913	0.4916	0.4916	0.4916	0.4916	0.4916
250	0.4804	0.4839	0.4852	0.4859	0.4862	0.4866	0.4869	0.4869	0.4869	0.4869	0.4869
500	0.4712	0.4757	0.4761	0.4769	0.4770	0.4774	0.4774	0.4775	0.4775	0.4775	0.4775

Table 12: iP[0.01] Child Strategy 2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4771	0.4821	0.4828	0.4834	0.4839	0.4846	0.4846	0.4846	0.4846	0.4846	0.4846
50	0.4637	0.4679	0.4700	0.4708	0.4713	0.4717	0.4718	0.4718	0.4718	0.4718	0.4718
100	0.4520	0.4557	0.4562	0.4569	0.4577	0.4581	0.4588	0.4588	0.4588	0.4588	0.4588
150	0.4500	0.4542	0.4546	0.4552	0.4553	0.4564	0.4567	0.4569	0.4569	0.4569	0.4569
200	0.4457	0.4494	0.4506	0.4510	0.4513	0.4520	0.4521	0.4523	0.4523	0.4523	0.4523
250	0.4409	0.4465	0.4483	0.4484	0.4485	0.4493	0.4494	0.4495	0.4495	0.4495	0.4495
500	0.4344	0.4419	0.4430	0.4435	0.4435	0.4440	0.4442	0.4442	0.4442	0.4442	0.4442

Table 13: iP[0.01] Child Strategy 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6089	0.6130	0.6145	0.6150	0.6150	0.6150	0.6150	0.6150	0.6150	0.6150	0.6150
50	0.6023	0.6045	0.6050	0.6053	0.6054	0.6057	0.6057	0.6057	0.6057	0.6057	0.6057
100	0.5965	0.5983	0.5984	0.5984	0.5986	0.6003	0.6003	0.6003	0.6003	0.6003	0.6003
150	0.5812	0.5840	0.5846	0.5846	0.5846	0.5855	0.5868	0.5868	0.5868	0.5868	0.5868
200	0.5728	0.5754	0.5760	0.5771	0.5771	0.5770	0.5783	0.5783	0.5783	0.5783	0.5783
250	0.5733	0.5757	0.5763	0.5763	0.5774	0.5774	0.5784	0.5786	0.5786	0.5786	0.5786
500	0.5666	0.5686	0.5688	0.5690	0.5705	0.5705	0.5705	0.5705	0.5705	0.5705	0.5705

Table 14: iP[0.01] Child Strategy 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6076	0.6120	0.6137	0.6146	0.6151	0.6156	0.6156	0.6156	0.6156	0.6156	0.6156
50	0.5889	0.5917	0.5919	0.5919	0.5931	0.5933	0.5933	0.5933	0.5933	0.5933	0.5933
100	0.5810	0.5814	0.5815	0.5829	0.5829	0.5832	0.5835	0.5835	0.5835	0.5835	0.5835
150	0.5699	0.5709	0.5725	0.5726	0.5726	0.5728	0.5731	0.5731	0.5731	0.5731	0.5731
200	0.5702	0.5715	0.5737	0.5738	0.5738	0.5738	0.5740	0.5741	0.5741	0.5741	0.5741
250	0.5649	0.5667	0.5690	0.5691	0.5691	0.5691	0.5691	0.5693	0.5693	0.5693	0.5693
500	0.5593	0.5641	0.5645	0.5645	0.5645	0.5645	0.5645	0.5645	0.5645	0.5645	0.5645

Experiment 3

In this experiment, *Correlation Strategy* is used as the overlap removal strategy. The retrieval is performed according to the steps shown in Figure 21. Tables 15 - 18 give the results of this experiment.

Table 15: iP[0.01] Correlation Strategy 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4962	0.4964	0.4964	0.4964	0.4964	0.4964	0.4964	0.4964	0.4964	0.4964	0.4964
50	0.4837	0.4845	0.4846	0.4846	0.4846	0.4847	0.4847	0.4847	0.4847	0.4847	0.4847
100	0.4743	0.4749	0.4752	0.4753	0.4754	0.4755	0.4755	0.4755	0.4755	0.4755	0.4755
150	0.4739	0.4750	0.4750	0.4750	0.4750	0.4751	0.4752	0.4752	0.4752	0.4752	0.4752
200	0.4778	0.4790	0.4792	0.4792	0.4792	0.4792	0.4792	0.4792	0.4792	0.4792	0.4792
250	0.4757	0.4768	0.4768	0.4768	0.4768	0.4769	0.4769	0.4769	0.4769	0.4769	0.4769
500	0.4853	0.4864	0.4867	0.4867	0.4867	0.4867	0.4867	0.4867	0.4867	0.4867	0.4867

Table 16: iP[0.01] Correlation Strategy 2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4745	0.4757	0.4758	0.4758	0.4758	0.4758	0.4758	0.4758	0.4758	0.4758	0.4758
50	0.4590	0.4602	0.4602	0.4602	0.4602	0.4603	0.4603	0.4603	0.4603	0.4603	0.4603
100	0.4570	0.4587	0.4590	0.4591	0.4593	0.4593	0.4595	0.4595	0.4595	0.4595	0.4595
150	0.4562	0.4576	0.4578	0.4578	0.4578	0.4580	0.4580	0.4580	0.4580	0.4580	0.4580
200	0.4554	0.4570	0.4571	0.4572	0.4572	0.4573	0.4573	0.4573	0.4573	0.4573	0.4573
250	0.4549	0.4567	0.4569	0.4569	0.4570	0.4570	0.4570	0.4570	0.4570	0.4570	0.4570
500	0.4521	0.4547	0.4548	0.4549	0.4549	0.4549	0.4549	0.4549	0.4549	0.4549	0.4549

Table 17: iP[0.01]Correlation Strategy 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6603	0.6621	0.6621	0.6621	0.6621	0.6621	0.6621	0.6621	0.6621	0.6621	0.6621
50	0.6596	0.6607	0.6611	0.6611	0.6611	0.6611	0.6611	0.6611	0.6611	0.6611	0.6611
100	0.6575	0.6595	0.6595	0.6595	0.6599	0.6599	0.6599	0.6599	0.6599	0.6599	0.6599
150	0.6575	0.6583	0.6586	0.6586	0.6586	0.6593	0.6593	0.6593	0.6593	0.6593	0.6593
200	0.6539	0.6556	0.6558	0.6558	0.6558	0.6565	0.6565	0.6565	0.6565	0.6565	0.6565
250	0.6553	0.6565	0.6568	0.6570	0.6570	0.6571	0.6576	0.6576	0.6576	0.6576	0.6576
500	0.6533	0.6545	0.6550	0.6551	0.6551	0.6554	0.6557	0.6561	0.6561	0.6561	0.6561

Table 18: iP[0.01] Correlation Strategy 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6855	0.6887	0.6897	0.6898	0.6898	0.6898	0.6898	0.6898	0.6898	0.6898	0.6898
50	0.6700	0.6718	0.6721	0.6723	0.6726	0.6726	0.6726	0.6726	0.6726	0.6726	0.6726
100	0.6686	0.6692	0.6692	0.6692	0.6692	0.6692	0.6692	0.6692	0.6692	0.6692	0.6692
150	0.6677	0.6682	0.6682	0.6682	0.6682	0.6682	0.6682	0.6682	0.6682	0.6682	0.6682
200	0.6678	0.6684	0.6684	0.6684	0.6684	0.6685	0.6685	0.6685	0.6685	0.6685	0.6685
250	0.6689	0.6692	0.6692	0.6692	0.6692	0.6693	0.6693	0.6693	0.6693	0.6693	0.6693
500	0.6682	0.6685	0.6685	0.6685	0.6685	0.6685	0.6685	0.6686	0.6686	0.6686	0.6686

Experiment 4

In this experiment, *Section Strategy* is used as the overlap removal strategy. Focused output is rearranged using RAC. The retrieval is performed according to the steps shown in Figure 21. Tables 19 - 22 give the results of this experiment.

Table 19: iP[0.01] Section Strategy RAC 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4912	0.4933	0.4906	0.4919	0.4919	0.4919	0.4919	0.4919	0.4919	0.4919	0.4919
50	0.4742	0.4875	0.4903	0.4903	0.4874	0.4885	0.4885	0.4885	0.4885	0.4885	0.4885
100	0.4690	0.4782	0.4885	0.4864	0.4909	0.4889	0.4891	0.4891	0.4891	0.4891	0.4891
150	0.4690	0.4748	0.4801	0.4864	0.4868	0.4899	0.4892	0.4892	0.4892	0.4892	0.4892
200	0.4698	0.4740	0.4760	0.4802	0.4859	0.4915	0.4891	0.4892	0.4892	0.4892	0.4892
250	0.4692	0.4775	0.4774	0.4794	0.4880	0.4856	0.4892	0.4894	0.4894	0.4894	0.4894
500	0.4651	0.4728	0.4757	0.4780	0.4791	0.4871	0.4917	0.4890	0.4890	0.4892	0.4892

Table 20: iP[0.01] Section Strategy RAC 2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4824	0.4811	0.4806	0.4788	0.4804	0.4804	0.4804	0.4804	0.4804	0.4804	0.4804
50	0.4655	0.4798	0.4801	0.4763	0.4780	0.4773	0.4773	0.4773	0.4773	0.4773	0.4773
100	0.4590	0.4706	0.4713	0.4791	0.4777	0.4771	0.4776	0.4776	0.4776	0.4776	0.4776
150	0.4579	0.4701	0.4705	0.4713	0.4801	0.4766	0.4773	0.4776	0.4776	0.4776	0.4776
200	0.4490	0.4645	0.4702	0.4725	0.4728	0.4773	0.4769	0.4773	0.4777	0.4777	0.4777
250	0.4499	0.4635	0.4717	0.4717	0.4719	0.4764	0.4779	0.4775	0.4777	0.4776	0.4776
500	0.4483	0.4590	0.4681	0.4691	0.4712	0.4703	0.4760	0.4779	0.4781	0.4774	0.4776

Table 21: iP[0.01] Section Strategy RAC 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.7016	0.7097	0.7090	0.7093	0.7093	0.7093	0.7093	0.7093	0.7093	0.7093	0.7093
50	0.6912	0.6972	0.6976	0.7091	0.7091	0.7090	0.7090	0.7090	0.7090	0.7090	0.7090
100	0.6826	0.6989	0.6965	0.6923	0.6949	0.7104	0.7104	0.7104	0.7104	0.7104	0.7104
150	0.6838	0.7010	0.6992	0.6980	0.6930	0.7130	0.7114	0.7114	0.7114	0.7114	0.7114
200	0.6838	0.7012	0.7010	0.6978	0.6976	0.7100	0.7112	0.7112	0.7112	0.7112	0.7112
250	0.6834	0.6892	0.6974	0.6992	0.6973	0.6961	0.7112	0.7116	0.7113	0.7113	0.7113
500	0.6749	0.6884	0.6965	0.6974	0.6999	0.6941	0.7103	0.7120	0.7117	0.7118	0.7118

Table 22: iP[0.01] Section Strategy RAC 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.7046	0.7181	0.7182	0.7198	0.7193	0.7195	0.7195	0.7195	0.7195	0.7195	0.7195
50	0.6968	0.7039	0.7200	0.7188	0.7207	0.7204	0.7204	0.7204	0.7204	0.7204	0.7204
100	0.7000	0.7042	0.7098	0.7075	0.7081	0.7214	0.7211	0.7211	0.7211	0.7211	0.7211
150	0.6901	0.6982	0.7109	0.7097	0.7081	0.7201	0.7209	0.7214	0.7214	0.7214	0.7214
200	0.6914	0.6984	0.7064	0.7145	0.7088	0.7195	0.7210	0.7220	0.7217	0.7217	0.7217
250	0.6916	0.6980	0.6997	0.7150	0.7128	0.7200	0.7223	0.7213	0.7217	0.7217	0.7217
500	0.6858	0.6914	0.6997	0.7072	0.7128	0.7096	0.7200	0.7217	0.7211	0.7216	0.7218

Experiment 5

In this experiment, *Child Strategy* is used as the overlap removal strategy. Focused output is rearranged using RAC. The retrieval is performed according to the steps shown in Figure 21. Tables 23 - 26 give the results of this experiment.

Table 23: iP[0.01] Child Strategy RAC 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4950	0.4967	0.4968	0.5024	0.5009	0.5015	0.5015	0.5015	0.5015	0.5015	0.5015
50	0.4939	0.4964	0.4927	0.4907	0.4919	0.4985	0.4985	0.4985	0.4985	0.4985	0.4985
100	0.4975	0.4923	0.4991	0.4944	0.4929	0.4983	0.4994	0.4993	0.4993	0.4993	0.4993
150	0.5079	0.4918	0.4933	0.4968	0.4951	0.5025	0.4992	0.4994	0.4994	0.4994	0.4994
200	0.5070	0.4961	0.4918	0.4908	0.4951	0.4916	0.4982	0.4995	0.4994	0.4994	0.4994
250	0.5081	0.4947	0.4915	0.4898	0.4913	0.4928	0.4984	0.4995	0.4995	0.4994	0.4994
500	0.5034	0.5041	0.4942	0.4929	0.4913	0.4916	0.5015	0.5002	0.4992	0.4993	0.4994

Table 24: iP[0.01] Child Strategy RAC 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4813	0.4864	0.4922	0.4919	0.4896	0.4884	0.4884	0.4884	0.4884	0.4884	0.4884
50	0.4758	0.4793	0.4834	0.4857	0.4872	0.4851	0.4855	0.4855	0.4855	0.4855	0.4855
100	0.4742	0.4693	0.4706	0.4851	0.4885	0.4904	0.4857	0.4857	0.4857	0.4857	0.4857
150	0.4762	0.4680	0.4714	0.4712	0.4845	0.4863	0.4869	0.4856	0.4858	0.4858	0.4858
200	0.4740	0.4723	0.4715	0.4707	0.4705	0.4861	0.4879	0.4863	0.4855	0.4858	0.4858
250	0.4703	0.4717	0.4765	0.4729	0.4693	0.4867	0.4901	0.4870	0.4861	0.4858	0.4858
500	0.4682	0.4718	0.4712	0.4751	0.4737	0.4761	0.4881	0.4886	0.4881	0.4864	0.4856

Table 25: iP[0.01] Child Strategy RAC 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6777	0.6941	0.7088	0.7139	0.7159	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158
50	0.6841	0.6858	0.6948	0.7104	0.7115	0.7161	0.7160	0.7160	0.7160	0.7160	0.7160
100	0.6831	0.6798	0.6836	0.6902	0.6932	0.7185	0.7179	0.7176	0.7176	0.7176	0.7176
150	0.6750	0.6784	0.6807	0.6842	0.6913	0.7108	0.7173	0.7183	0.7181	0.7181	0.7181
200	0.6732	0.6801	0.6799	0.6836	0.6825	0.7113	0.7182	0.7185	0.7182	0.7183	0.7183
250	0.6671	0.6811	0.6849	0.6798	0.6887	0.6974	0.7178	0.7179	0.7187	0.7184	0.7184
500	0.6674	0.6888	0.6831	0.6852	0.6798	0.6914	0.7110	0.7175	0.7182	0.7184	0.7189

Table 26: iP[0.01] Child Strategy RAC 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6684	0.6925	0.7162	0.7175	0.7151	0.7179	0.7179	0.7179	0.7179	0.7179	0.7179
50	0.6681	0.6799	0.6909	0.7068	0.7173	0.7184	0.7192	0.7192	0.7192	0.7192	0.7192
100	0.6705	0.6738	0.6830	0.6926	0.6929	0.7196	0.7195	0.7198	0.7198	0.7198	0.7198
150	0.6695	0.6751	0.6824	0.6857	0.6902	0.7093	0.7188	0.7200	0.7201	0.7201	0.7201
200	0.6684	0.6760	0.6761	0.6854	0.6894	0.6984	0.7196	0.7196	0.7204	0.7203	0.7203
250	0.6586	0.6789	0.6716	0.6843	0.6856	0.6952	0.7196	0.7189	0.7199	0.7204	0.7204
500	0.6609	0.6768	0.6763	0.6751	0.6807	0.6889	0.7105	0.7134	0.7201	0.7195	0.7202

Experiment 6

In this experiment, *Correlation Strategy* is used as the overlap removal strategy. Focused output is rearranged using RAC. The retrieval is performed according to the steps shown in Figure 21. Tables 27 – 30 give the results of this experiment.

Table 27: iP[0.01] Correlation Strategy RAC 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4927	0.4887	0.4888	0.4886	0.4886	0.4886	0.4886	0.4886	0.4886	0.4886	0.4886
50	0.4900	0.4885	0.4867	0.4847	0.4850	0.4850	0.4850	0.4850	0.4850	0.4850	0.4850
100	0.4907	0.4815	0.4899	0.4869	0.4869	0.4850	0.4852	0.4852	0.4852	0.4852	0.4852
150	0.4817	0.4856	0.4801	0.4896	0.4871	0.4849	0.4853	0.4853	0.4853	0.4853	0.4853
200	0.4806	0.4857	0.4823	0.4885	0.4896	0.4851	0.4853	0.4853	0.4853	0.4853	0.4853
250	0.4802	0.4881	0.4834	0.4815	0.4883	0.4858	0.4851	0.4852	0.4852	0.4852	0.4852
500	0.4840	0.4879	0.4863	0.4820	0.4831	0.4873	0.4854	0.4851	0.4852	0.4852	0.4852

Table 28: iP[0.01] Correlation Strategy RAC 2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4868	0.4828	0.4802	0.4796	0.4796	0.4796	0.4796	0.4796	0.4796	0.4796	0.4796
50	0.4726	0.4829	0.4771	0.4770	0.4760	0.4759	0.4759	0.4759	0.4759	0.4759	0.4759
100	0.4757	0.4745	0.4741	0.4808	0.4784	0.4758	0.4761	0.4761	0.4761	0.4761	0.4761
150	0.4762	0.4738	0.4752	0.4742	0.4809	0.4773	0.4761	0.4761	0.4761	0.4761	0.4761
200	0.4755	0.4741	0.4743	0.4750	0.4737	0.4767	0.4758	0.4761	0.4761	0.4761	0.4761
250	0.4758	0.4769	0.4730	0.4747	0.4737	0.4771	0.4760	0.4761	0.4761	0.4761	0.4761
500	0.4766	0.4752	0.4752	0.4743	0.4740	0.4724	0.4772	0.4768	0.4759	0.4761	0.4761

Table 29: iP[0.01] Correlation Strategy RAC 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.7010	0.6995	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
50	0.6998	0.7010	0.6996	0.7004	0.7005	0.7004	0.7004	0.7004	0.7004	0.7004	0.7004
100	0.6925	0.6972	0.7015	0.7024	0.7058	0.7010	0.7010	0.7010	0.7010	0.7010	0.7010
150	0.6959	0.7005	0.7007	0.7041	0.7015	0.7020	0.7018	0.7018	0.7018	0.7018	0.7018
200	0.6846	0.6955	0.6981	0.7005	0.7031	0.7060	0.7015	0.7015	0.7015	0.7015	0.7015
250	0.6866	0.6956	0.6980	0.6986	0.7046	0.7034	0.7017	0.7016	0.7016	0.7016	0.7016
500	0.6857	0.7021	0.6992	0.7004	0.6996	0.7043	0.7057	0.7021	0.7019	0.7019	0.7019

Table 30: iP[0.01] Correlation Strategy RAC 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.7185	0.7151	0.7167	0.7159	0.7166	0.7166	0.7166	0.7166	0.7166	0.7166	0.7166
50	0.7108	0.7185	0.7168	0.7180	0.7175	0.7175	0.7175	0.7175	0.7175	0.7175	0.7175
100	0.7115	0.7134	0.7208	0.7183	0.7171	0.7175	0.7180	0.7180	0.7180	0.7180	0.7180
150	0.7105	0.7101	0.7191	0.7197	0.7189	0.7182	0.7181	0.7182	0.7182	0.7182	0.7182
200	0.7104	0.7104	0.7133	0.7237	0.7205	0.7171	0.7184	0.7184	0.7184	0.7184	0.7184
250	0.7100	0.7098	0.7129	0.7194	0.7241	0.7174	0.7175	0.7179	0.7184	0.7184	0.7184
500	0.7091	0.7048	0.7105	0.7140	0.7182	0.7211	0.7186	0.7187	0.7184	0.7187	0.7187

Experiment 7

In this experiment, *Section Strategy* is used as the overlap removal strategy. Focused output is rearranged using RBC. The retrieval is performed according to the steps shown in Figure 21. Tables 31 - 34 give the results of this experiment.

Table 31: iP[0.01] Section Strategy RBC 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4783	0.4910	0.4917	0.4919	0.4919	0.4919	0.4919	0.4919	0.4919	0.4919	0.4919
50	0.4739	0.4864	0.4878	0.4883	0.4885	0.4885	0.4885	0.4885	0.4885	0.4885	0.4885
100	0.4739	0.4864	0.4878	0.4883	0.4886	0.4891	0.4891	0.4891	0.4891	0.4891	0.4891
150	0.4739	0.4864	0.4878	0.4883	0.4886	0.4892	0.4892	0.4892	0.4892	0.4892	0.4892
200	0.4739	0.4864	0.4878	0.4883	0.4886	0.4892	0.4892	0.4892	0.4892	0.4892	0.4892
250	0.4738	0.4864	0.4878	0.4883	0.4886	0.4894	0.4894	0.4894	0.4894	0.4894	0.4894
500	0.4738	0.4864	0.4878	0.4883	0.4886	0.4892	0.4892	0.4892	0.4892	0.4892	0.4892

Table 32: iP[0.01] Section Strategy RBC2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4666	0.4756	0.4782	0.4803	0.4804	0.4804	0.4804	0.4804	0.4804	0.4804	0.4804
50	0.4622	0.4712	0.4739	0.4764	0.4767	0.4773	0.4773	0.4773	0.4773	0.4773	0.4773
100	0.4622	0.4712	0.4739	0.4764	0.4767	0.4775	0.4776	0.4776	0.4776	0.4776	0.4776
150	0.4622	0.4712	0.4739	0.4764	0.4767	0.4775	0.4776	0.4776	0.4776	0.4776	0.4776
200	0.4622	0.4712	0.4739	0.4764	0.4767	0.4775	0.4777	0.4777	0.4777	0.4777	0.4777
250	0.4622	0.4712	0.4739	0.4764	0.4767	0.4775	0.4776	0.4776	0.4776	0.4776	0.4776
500	0.4622	0.4712	0.4739	0.4764	0.4767	0.4775	0.4776	0.4776	0.4776	0.4776	0.4776

Table 33: iP[0.01] Section Strategy RBC 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6971	0.7083	0.7093	0.7093	0.7093	0.7093	0.7093	0.7093	0.7093	0.7093	0.7093
50	0.6967	0.7079	0.7090	0.7090	0.7090	0.7090	0.7090	0.7090	0.7090	0.7090	0.7090
100	0.6966	0.7079	0.7089	0.7089	0.7097	0.7104	0.7104	0.7104	0.7104	0.7104	0.7104
150	0.6971	0.7084	0.7094	0.7094	0.7102	0.7114	0.7114	0.7114	0.7114	0.7114	0.7114
200	0.6966	0.7079	0.7089	0.7089	0.7097	0.7110	0.7112	0.7112	0.7112	0.7112	0.7112
250	0.6967	0.7079	0.7090	0.7090	0.7098	0.7111	0.7113	0.7113	0.7113	0.7113	0.7113
500	0.6967	0.7079	0.7089	0.7089	0.7098	0.7110	0.7114	0.7118	0.7118	0.7118	0.7118

Table 34: iP[0.01] Section Strategy RBC 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6921	0.7139	0.7186	0.7195	0.7195	0.7195	0.7195	0.7195	0.7195	0.7195	0.7195
50	0.6921	0.7139	0.7186	0.7202	0.7204	0.7204	0.7204	0.7204	0.7204	0.7204	0.7204
100	0.6921	0.7139	0.7186	0.7202	0.7204	0.7210	0.7211	0.7211	0.7211	0.7211	0.7211
150	0.6921	0.7139	0.7186	0.7202	0.7204	0.7210	0.7214	0.7214	0.7214	0.7214	0.7214
200	0.6921	0.7139	0.7186	0.7202	0.7204	0.7210	0.7217	0.7217	0.7217	0.7217	0.7217
250	0.6921	0.7139	0.7186	0.7202	0.7204	0.7210	0.7217	0.7217	0.7217	0.7217	0.7217
500	0.6921	0.7139	0.7186	0.7202	0.7204	0.7210	0.7217	0.7218	0.7218	0.7218	0.7218

Experiment 8

In this experiment, *Child Strategy* is used as the overlap removal strategy. Focused output is rearranged using RBC. The retrieval is performed according to the steps shown in Figure 21. Tables 35 – 38 give the results of this experiment.

Table 35: iP[0.01] Child Strategy RBC 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4912	0.5048	0.5112	0.5118	0.5120	0.5120	0.5120	0.5120	0.5120	0.5120	0.5120
50	0.4866	0.5001	0.5064	0.5078	0.5084	0.5089	0.5089	0.5089	0.5089	0.5089	0.5089
100	0.4872	0.5008	0.5072	0.5085	0.5092	0.5100	0.5101	0.5101	0.5101	0.5101	0.5101
150	0.4872	0.5008	0.5072	0.5085	0.5092	0.5100	0.5102	0.5102	0.5102	0.5102	0.5102
200	0.4872	0.5008	0.5072	0.5085	0.5092	0.5100	0.5102	0.5102	0.5102	0.5102	0.5102
250	0.4872	0.5008	0.5072	0.5085	0.5092	0.5100	0.5102	0.5102	0.5102	0.5102	0.5102
500	0.4872	0.5008	0.5072	0.5085	0.5092	0.5100	0.5102	0.5102	0.5102	0.5102	0.5102

Table 36: iP[0.01] Child Strategy Exact RBC 2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4639	0.4753	0.4831	0.4860	0.4869	0.4884	0.4884	0.4884	0.4884	0.4884	0.4884
50	0.4595	0.4709	0.4786	0.4817	0.4828	0.4850	0.4855	0.4855	0.4855	0.4855	0.4855
100	0.4595	0.4709	0.4786	0.4817	0.4828	0.4850	0.4857	0.4857	0.4857	0.4857	0.4857
150	0.4595	0.4709	0.4786	0.4817	0.4828	0.4850	0.4857	0.4858	0.4858	0.4858	0.4858
200	0.4595	0.4709	0.4786	0.4817	0.4828	0.4850	0.4857	0.4858	0.4858	0.4858	0.4858
250	0.4595	0.4709	0.4786	0.4817	0.4828	0.4850	0.4857	0.4858	0.4858	0.4858	0.4858
500	0.4595	0.4709	0.4786	0.4817	0.4828	0.4850	0.4857	0.4858	0.4858	0.4858	0.4858

Table 37: iP[0.01] Child Strategy RBC 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6895	0.7108	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158
50	0.6895	0.7108	0.7158	0.7158	0.7158	0.7160	0.7160	0.7160	0.7160	0.7160	0.7160
100	0.6895	0.7108	0.7158	0.7158	0.7158	0.7176	0.7176	0.7176	0.7176	0.7176	0.7176
150	0.6895	0.7108	0.7158	0.7158	0.7158	0.7177	0.7181	0.7181	0.7181	0.7181	0.7181
200	0.6895	0.7108	0.7158	0.7158	0.7158	0.7177	0.7183	0.7183	0.7183	0.7183	0.7183
250	0.6895	0.7108	0.7158	0.7158	0.7158	0.7177	0.7184	0.7184	0.7184	0.7184	0.7184
500	0.6895	0.7108	0.7158	0.7158	0.7158	0.7170	0.7184	0.7187	0.7187	0.7187	0.7187

Table 38: iP[0.01] Child Strategy RBC 2008 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6693	0.7009	0.7107	0.7156	0.7168	0.7179	0.7179	0.7179	0.7179	0.7179	0.7179
50	0.6693	0.7009	0.7107	0.7157	0.7168	0.7192	0.7192	0.7192	0.7192	0.7192	0.7192
100	0.6693	0.7009	0.7107	0.7157	0.7168	0.7194	0.7198	0.7198	0.7198	0.7198	0.7198
150	0.6693	0.7009	0.7107	0.7157	0.7168	0.7194	0.7201	0.7201	0.7201	0.7201	0.7201
200	0.6693	0.7009	0.7107	0.7157	0.7168	0.7194	0.7202	0.7203	0.7203	0.7203	0.7203
250	0.6693	0.7009	0.7107	0.7157	0.7168	0.7194	0.7202	0.7204	0.7204	0.7204	0.7204
500	0.6693	0.7009	0.7107	0.7157	0.7168	0.7194	0.7202	0.7204	0.7204	0.7204	0.7204

Experiment 9

In this experiment, *Correlation Strategy* is used as the overlap removal strategy. Focused output is rearranged using RBC. The retrieval is performed according to the steps shown in Figure 21. Tables 39 – 42 give the results of this experiment.

Table 39: iP[0.01] Correlation Strategy RBC 2007 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4864	0.4882	0.4886	0.4886	0.4886	0.4886	0.4886	0.4886	0.4886	0.4886	0.4886
50	0.4818	0.4841	0.4846	0.4850	0.4850	0.4850	0.4850	0.4850	0.4850	0.4850	0.4850
100	0.4818	0.4841	0.4846	0.4851	0.4851	0.4852	0.4852	0.4852	0.4852	0.4852	0.4852
150	0.4818	0.4841	0.4845	0.4850	0.4851	0.4852	0.4853	0.4853	0.4853	0.4853	0.4853
200	0.4818	0.4841	0.4845	0.4850	0.4851	0.4852	0.4853	0.4853	0.4853	0.4853	0.4853
250	0.4818	0.4841	0.4845	0.4850	0.4851	0.4852	0.4852	0.4852	0.4852	0.4852	0.4852
500	0.4818	0.4841	0.4845	0.4850	0.4851	0.4852	0.4852	0.4852	0.4852	0.4852	0.4852

Table 40: iP[0.01] Correlation Strategy RBC 2007 (Tag Set 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4678	0.4770	0.4782	0.4796	0.4796	0.4796	0.4796	0.4796	0.4796	0.4796	0.4796
50	0.4634	0.4725	0.4739	0.4755	0.4759	0.4759	0.4759	0.4759	0.4759	0.4759	0.4759
100	0.4634	0.4725	0.4739	0.4756	0.4760	0.4760	0.4761	0.4761	0.4761	0.4761	0.4761
150	0.4634	0.4725	0.4739	0.4756	0.4760	0.4760	0.4761	0.4761	0.4761	0.4761	0.4761
200	0.4634	0.4725	0.4739	0.4756	0.4760	0.4761	0.4761	0.4761	0.4761	0.4761	0.4761
250	0.4634	0.4725	0.4739	0.4756	0.4760	0.4760	0.4761	0.4761	0.4761	0.4761	0.4761
500	0.4634	0.4725	0.4739	0.4756	0.4760	0.4760	0.4761	0.4761	0.4761	0.4761	0.4761

Table 41: iP[0.01] Correlation Strategy RBC 2008 (Tag Set 1)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6919	0.6995	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
50	0.6915	0.6991	0.7004	0.7004	0.7004	0.7004	0.7004	0.7004	0.7004	0.7004	0.7004
100	0.6915	0.6990	0.7003	0.7006	0.7010	0.7010	0.7010	0.7010	0.7010	0.7010	0.7010
150	0.6919	0.6995	0.7008	0.7011	0.7015	0.7018	0.7018	0.7018	0.7018	0.7018	0.7018
200	0.6915	0.6990	0.7003	0.7006	0.7010	0.7015	0.7015	0.7015	0.7015	0.7015	0.7015
250	0.6915	0.6991	0.7004	0.7006	0.7010	0.7016	0.7016	0.7016	0.7016	0.7016	0.7016
500	0.6915	0.6990	0.7003	0.7006	0.7010	0.7015	0.7018	0.7019	0.7019	0.7019	0.7019

Table 42: iP[0.01] Correlation Strategy RBC 2008 (Tag 2)

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.6929	0.7146	0.7166	0.7166	0.7166	0.7166	0.7166	0.7166	0.7166	0.7166	0.7166
50	0.6930	0.7146	0.7175	0.7175	0.7175	0.7175	0.7175	0.7175	0.7175	0.7175	0.7175
100	0.6929	0.7146	0.7175	0.7175	0.7176	0.7180	0.7180	0.7180	0.7180	0.7180	0.7180
150	0.6929	0.7146	0.7175	0.7175	0.7176	0.7182	0.7182	0.7182	0.7182	0.7182	0.7182
200	0.6930	0.7146	0.7175	0.7175	0.7177	0.7183	0.7184	0.7184	0.7184	0.7184	0.7184
250	0.6930	0.7146	0.7175	0.7175	0.7177	0.7183	0.7184	0.7184	0.7184	0.7184	0.7184
500	0.6930	0.7146	0.7175	0.7175	0.7177	0.7183	0.7185	0.7187	0.7187	0.7187	0.7187

4.6. Analysis of Results

This section discusses the observations that can be made from the results.

Comparison of the Three Strategies

Basic Version:

The below figure (Figure 22) shows a comparison of the best iP[0.01] score obtained by all the three focusing strategies, namely, *Section*, *Child* and *Correlation*, on the basic version (no rearrangement performed) of the Focused output using tag set 1 and tag set 2 on 2008 collection.

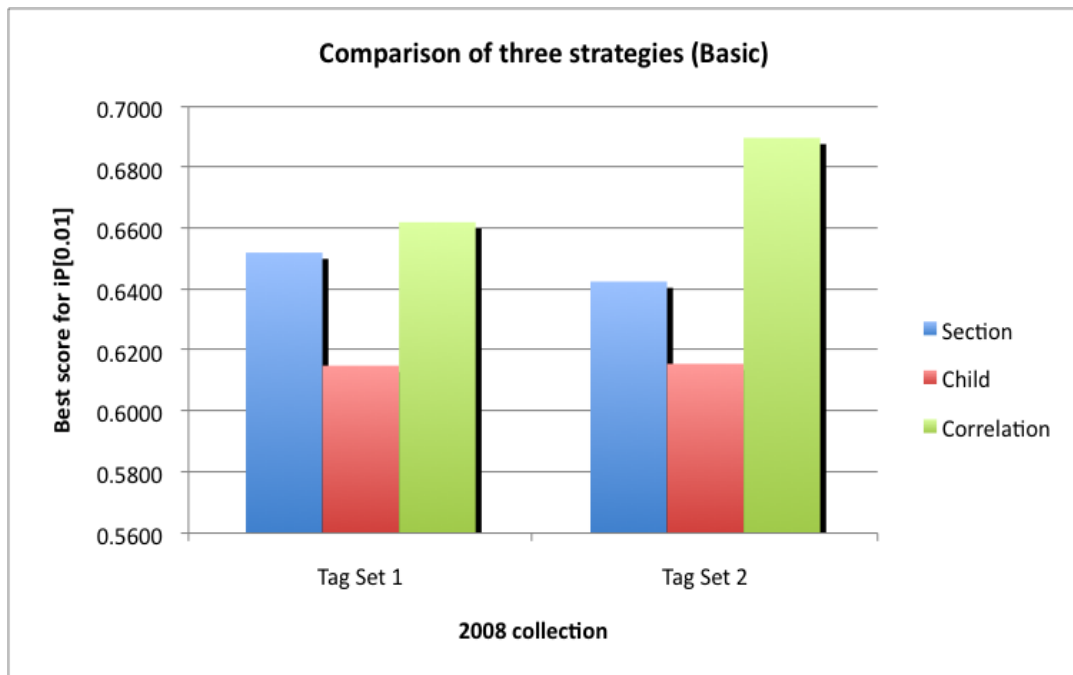


Figure 22: Comparison of Three Strategies on Basic Version

The *Correlation Strategy* has the highest value of **0.6898** on tag set 2. From the figure it can be inferred that *Section* and *Correlation Strategy* did well when compared to *Child Strategy* using both tag sets.

Rearrangement after Chopping (RAC)

The below figure (Figure 23) shows a comparison of the best iP[0.01] score obtained by all the three focusing strategies namely, *Section*, *Child* and *Correlation* on the RAC version of the Focused output using tag set 1 and tag set 2 on 2008 collection.

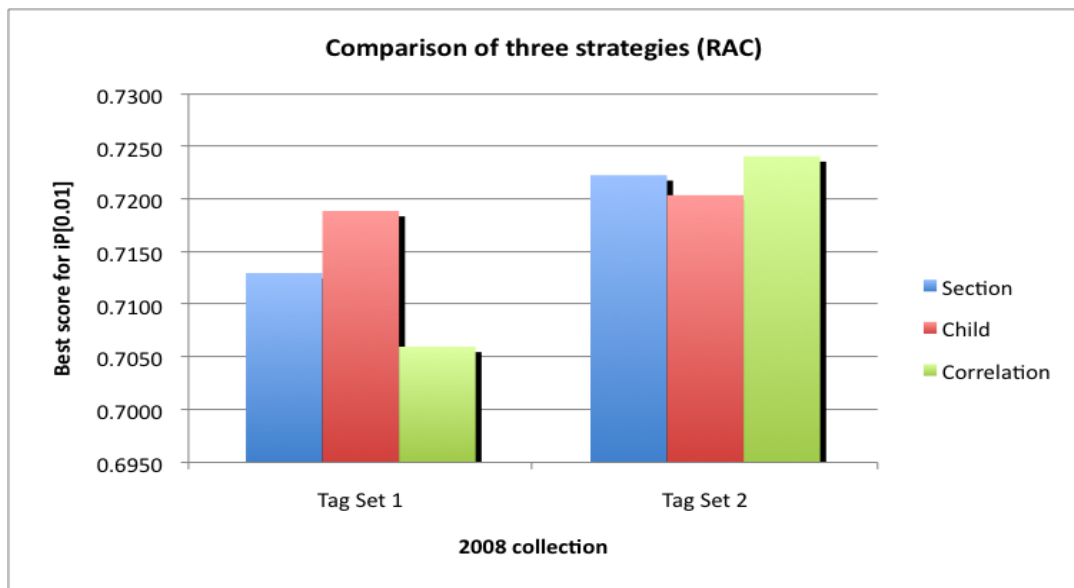


Figure 23: Comparison of Three Strategies on RAC

The *Correlation Strategy* has the highest value of **0.7241** on tag set 2. *Child Strategy* did consistently well on both tag sets.

Rearrangement before Chopping (RBC)

The below figure (Figure 24) shows a comparison of the best $iP[0.01]$ score obtained by all the three focusing strategies namely, *Section*, *Child* and *Correlation*, on the RBC version of the Focused output using tag set 1 and tag set 2 on 2008 collection. The *Section Strategy* has the highest value of **0.7218** on tag set 2. Again *Child Strategy* did consistently well on both the tag sets.

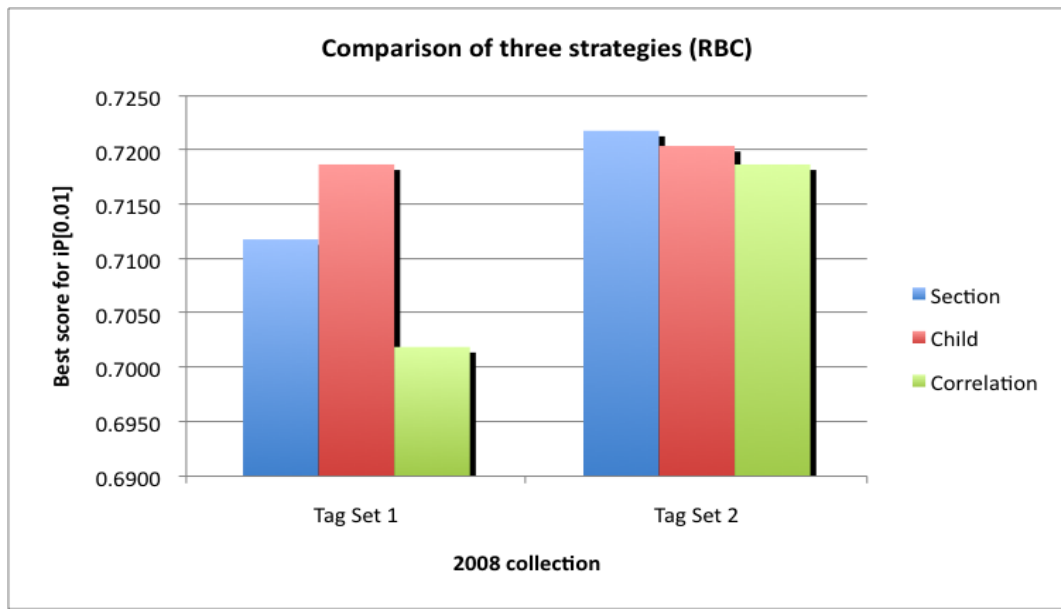


Figure 24: Comparison of Three Strategies on RBC

Section Strategy (Basic vs. RAC)

The below figure (Figure 25) shows a comparison of the *Section Strategy* on the basic version and RAC of the Focused output using tag set 1 and tag set 2 on 2008 collection. The values used in the graph are averaged over 77 runs. It can be inferred

from the figure that on both the tag sets, RAC showed better results when compared to basic.

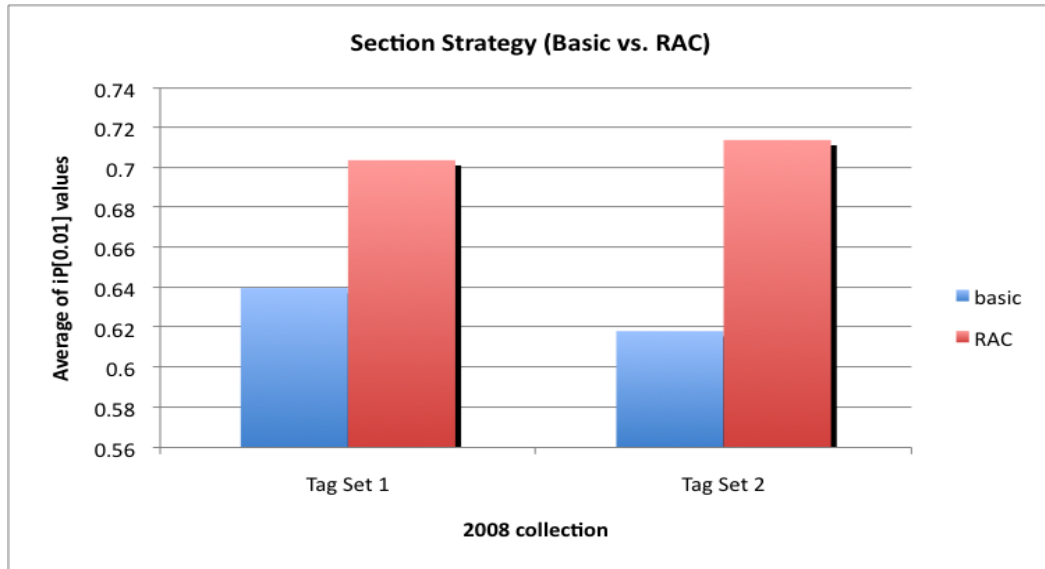


Figure 25: Comparison of Section Strategy on Basic and RAC

Child Strategy (Basic vs. RAC)

The below figure (Figure 26) shows a comparison of the *Child Strategy* on the basic version and RAC of the Focused output using tag set 1 and tag set 2 on 2008 collection. The values used in the graph are averaged over 77 runs. It can be inferred from the figure that on both the tag sets, RAC showed better results when compared to basic.

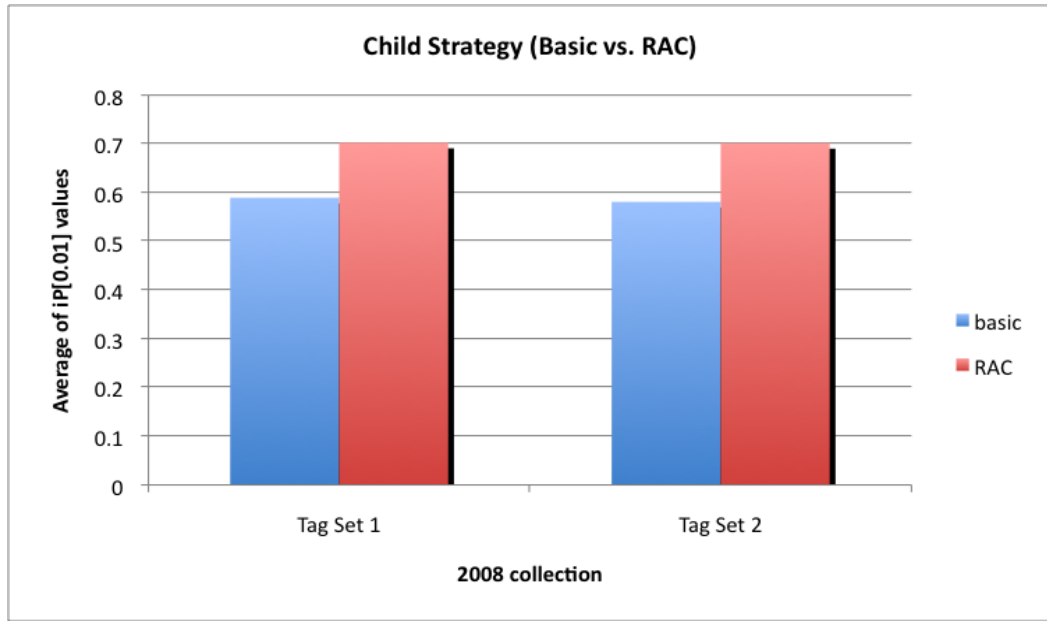


Figure 26: Comparison of Child Strategy on Basic and RAC

Correlation Strategy (Basic vs. RAC)

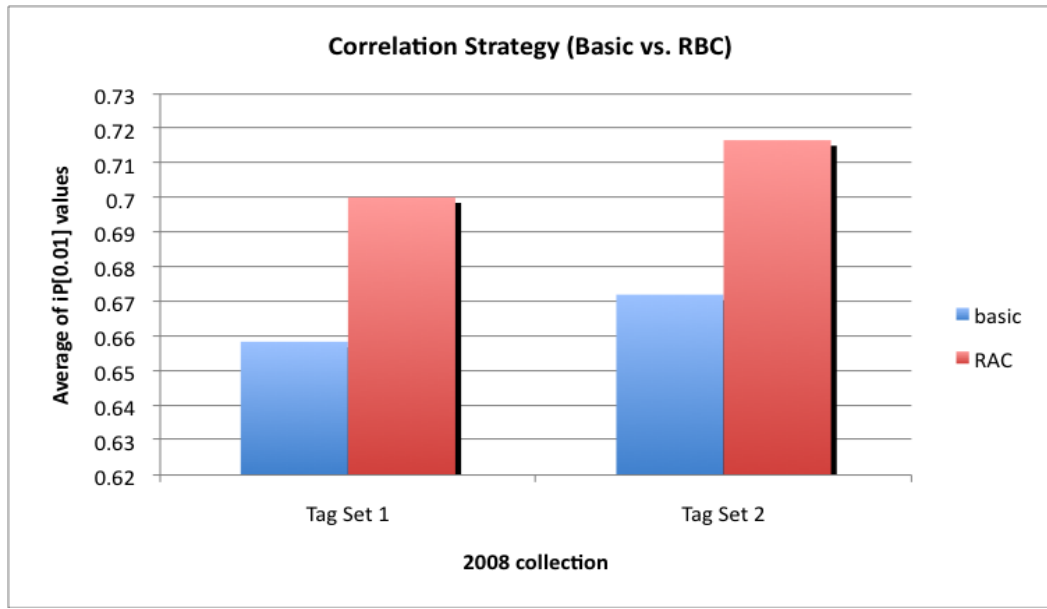


Figure 27: Comparison of Correlation Strategy on Basic and RAC

The above figure (Figure 27) shows a comparison of the *Correlation Strategy* on the basic version and RAC of the Focused output using tag set 1 and tag set 2 on 2008 collection. The values used in the graph are averaged over 77 runs. It can be inferred from the figure that on both the tag sets, RAC showed better results when compared to basic.

Observations

From all the experiments performed it is clear that the results produced using tag set 2 are better than those produced by tag set 1. The results show the rearranged version of the Focused output always did well when compared to the basic version.

5. Conclusion

Results are very much dependent on the *tag set*. The 2007 and 2008 collections are identical. The experiments show that all the three overlap removal strategies, *Section*, *Child* and *Correlation*, did well with tag set 1 on 2007 collection and tag set 2 on 2008 collection. In all the three versions of the Focused output, namely, basic, rearrangement after chopping, and rearrangement before chopping, for 2007 tag set 1 produced better results and for 2008, tag set 2 produced better results. Among all the three versions of the Focused output, for 2007, using tag set 1 *Child Strategy* produced the best score and for 2008, using tag set 2 *Correlation Strategy* has the best score. The strategy of rearranging the Focused output always worked well for 2008, but did not show the same consistency for 2007. The highest score of **0.7241** for $iP[0.01]$ achieved using *Exact Strategy* is produced by rearranging the 2008 Focused output (produced by *Correlation Strategy*) after chopping, when 250 articles and 250 elements from those articles are retrieved.

At the end of this research, three issues regarding Flex came into light, namely: (1) ensuring that all terminal nodes are populated, (2) that all correlation scores are calculated at execution time, and (3) that the correct (all-element) values of slope and pivot are used for all correlations between the query and the elements by Flex.

Flex should get the content of all the terminal nodes with the help of *docid_docpath_mappings* file (contains pointers to all terminal nodes). This

guarantees that all terminal nodes are populated. Given a query, note that all the positively correlating leaf nodes of the document tree are available from the *para+mt* parse. The scores of the zero-correlating terminal nodes do not affect the similarity calculations themselves, but their content must be present in the generated trees. Thus Flex must ensure that all leaf nodes are populated (whether they are present in the set of *para+mt* elements retrieval or not) and all correlations are performed at execution time. (I.e., the correlations of terminal nodes which may be available through the *para+mt* retrieval cannot be used in Flex operations. Note that the slope and pivot values of the *para+mt* search are not the same as those used for all-element retrieval, which are required by Flex.)

The experiments reported herein utilized the *para+mt* parse to populate the terminal nodes. As a result, the values reported in may vary from those calculated using the *docid_docpath_mapping* file as a source. (This work is currently in progress.) Nevertheless, the *Exact* methodology remains the same. Thus the conclusions reached herein are expected to hold true; verification will be reported in our INEX 2008 paper.

6. Future Work

Fine-tuning the collection-dependent values of *slope* and *pivot* may further improve the results. Rearranging the Focused output did very well with respect to the 2008 collection. More research in the area of rearranging Focused output might improve the results. Identifying the perfect tag set for a document collection is important. Various experiments can be performed on different tag sets and may give insight into finding the “best” tag set across all experiments.

References

- [1] Bapat, S. Improving Results for Focused and Relevance-in-context tasks, MS Thesis, Department of Computer Science, University of Minnesota Duluth, 2007. http://www.d.umn.edu/cs/thesis/salil_bapat_ms.pdf
- [2] Bhirud, D. Focused Retrieval using Upper Bound Methodology, MS Thesis, Department of Computer Science, University of Minnesota Duluth, 2009.
- [3] Crouch, C. Dynamic element retrieval in structured environment. *ACM TOIS*, 24(4): 437-454, 2006.
- [4] Geva, S., Kamps, J., Trotman, A. INEX 2008 Workshop Pre-proceedings. <http://www.inex.otago.ac.nz/>
- [5] <http://topx.sourceforge.net/>
- [6] Kamat, N. Impact of Untagged Text in Dynamic Element Retrieval, MS Thesis, Department of Computer Science, University of Minnesota Duluth, 2007. http://www.d.umn.edu/cs/thesis/nachiket_kamat_ms.pdf
- [7] Lalmas, M., Piwowarski, B. INEX 2007 Relevance Assessment Guide, http://inex.is.informatik.uni-duisburg.de/2007/inex07/adhoc_protected/downloads/Relevance_Assessment2007.pdf
- [8] Mehta, S. Finding the Best Entry Point. MS Thesis, Department of Computer Science, University of Minnesota Duluth, 2008. http://www.d.umn.edu/cs/thesis/sarika_mehta_ms.pdf
- [9] Mone, A. Dynamic Element Retrieval for Semi-Structured Documents, MS Thesis, Department of Computer Science, University of Minnesota Duluth, 2007. http://www.d.umn.edu/cs/thesis/aditya_mone_ms.pdf
- [10] Paranjape, D. Improving Focused Results, MS Thesis, Department of Computer Science, University of Minnesota Duluth, 2008. http://www.d.umn.edu/cs/thesis/darshan_paranjape_ms.pdf
- [11] Polumetla, C. Improving the Results for the Relevant In Context Task, MS Thesis, Department of Computer Science, University of Minnesota Duluth, 2009.
- [12] Salton, G., Wong, A., Yang, C. A vector space model for information retrieval. *JASIS*, 18(11): 613-620, 1975.

- [13] Singhal, A., Buckley, C., Mitra M. Pivoted Document Length Normalization, *Proceedings of 19th Annual International Conference on Research and Development in Information Retrieval*, Zurich. 19-21, 1996.
- [14] Sudhakar, V. Improving Results for the Best In Context Task, MS Project, Department of Computer Science, University of Minnesota Duluth, 2009.

Appendix A

This Appendix contains the results of the experiments of three overlap removal strategies (*Section, Child, and Correlation*) on 2007 collection using tag set 1 and evaluated according to 2007 evaluation method.

Table A. 1: iP[0.01] Section Strategy

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4701	0.4710	0.4714	0.4715	0.4715	0.4715	0.4715	0.4715	0.4715	0.4715	0.4715
50	0.4595	0.4606	0.4609	0.4610	0.4610	0.4610	0.4610	0.4610	0.4610	0.4610	0.4610
100	0.4539	0.4564	0.4566	0.4569	0.4570	0.4573	0.4573	0.4573	0.4573	0.4573	0.4573
150	0.4504	0.4531	0.4535	0.4536	0.4537	0.4540	0.4541	0.4541	0.4541	0.4541	0.4541
200	0.4498	0.4518	0.4526	0.4529	0.4528	0.4531	0.4532	0.4532	0.4532	0.4532	0.4532
250	0.4481	0.4502	0.4507	0.4508	0.4508	0.4509	0.4509	0.4510	0.4510	0.4510	0.4510
500	0.4450	0.4470	0.4475	0.4475	0.4476	0.4475	0.4476	0.4476	0.4476	0.4476	0.4476

Table A. 2: iP[0.01] Correlation Strategy

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4932	0.4933	0.4934	0.4934	0.4934	0.4934	0.4934	0.4934	0.4934	0.4934	0.4934
50	0.4813	0.4820	0.4822	0.4822	0.4822	0.4823	0.4823	0.4823	0.4823	0.4823	0.4823
100	0.4731	0.4738	0.4740	0.4742	0.4742	0.4744	0.4744	0.4744	0.4744	0.4744	0.4744
150	0.4729	0.4740	0.4740	0.4740	0.4741	0.4742	0.4742	0.4742	0.4742	0.4742	0.4742
200	0.4751	0.4763	0.4765	0.4765	0.4765	0.4765	0.4765	0.4765	0.4765	0.4765	0.4765
250	0.4754	0.4765	0.4765	0.4765	0.4765	0.4766	0.4766	0.4766	0.4766	0.4766	0.4766
500	0.4850	0.4860	0.4863	0.4863	0.4863	0.4863	0.4863	0.4863	0.4863	0.4863	0.4863

Table A. 3: iP[0.01] Child Strategy

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.5211	0.5231	0.5247	0.5247	0.5245	0.5245	0.5245	0.5245	0.5245	0.5245	0.5245
50	0.5061	0.5099	0.5111	0.5112	0.5119	0.5120	0.5120	0.5120	0.5120	0.5120	0.5120
100	0.4972	0.5009	0.5023	0.5029	0.5033	0.5038	0.5041	0.5041	0.5041	0.5041	0.5041
150	0.4941	0.4975	0.4984	0.4992	0.4997	0.5002	0.5005	0.5005	0.5005	0.5005	0.5005
200	0.4862	0.4900	0.4908	0.4921	0.4922	0.4927	0.4930	0.4930	0.4930	0.4930	0.4930
250	0.4816	0.4851	0.4864	0.4872	0.4875	0.4879	0.4882	0.4883	0.4883	0.4883	0.4883
500	0.4723	0.4765	0.4769	0.4777	0.4780	0.4784	0.4785	0.4785	0.4785	0.4785	0.4785

Table A. 4: iP[0.01] Section Strategy RAC

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4716	0.4690	0.4672	0.4685	0.4685	0.4685	0.4685	0.4685	0.4685	0.4685	0.4685
50	0.4666	0.4682	0.4663	0.4666	0.4649	0.4658	0.4658	0.4658	0.4658	0.4658	0.4658
100	0.4611	0.4693	0.4697	0.4618	0.4669	0.4663	0.4663	0.4663	0.4663	0.4663	0.4663
150	0.4602	0.4670	0.4701	0.4675	0.4633	0.4665	0.4662	0.4662	0.4662	0.4662	0.4662
200	0.4608	0.4673	0.4686	0.4698	0.4671	0.4675	0.4662	0.4662	0.4662	0.4662	0.4662
250	0.4603	0.4710	0.4705	0.4692	0.4692	0.4615	0.4663	0.4663	0.4663	0.4663	0.4663
500	0.4559	0.4642	0.4683	0.4709	0.4686	0.4675	0.4676	0.4663	0.4662	0.4662	0.4662

Table A. 5: iP[0.01] Child Strategy RAC

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4950	0.4967	0.4968	0.5024	0.5009	0.5015	0.5015	0.5015	0.5015	0.5015	0.5015
50	0.4939	0.4964	0.4927	0.4907	0.4919	0.4985	0.4985	0.4985	0.4985	0.4985	0.4985
100	0.4975	0.4923	0.4991	0.4944	0.4929	0.4983	0.4994	0.4993	0.4993	0.4993	0.4993
150	0.5079	0.4918	0.4933	0.4968	0.4951	0.5025	0.4992	0.4994	0.4994	0.4994	0.4994
200	0.5070	0.4961	0.4918	0.4908	0.4951	0.4916	0.4982	0.4995	0.4994	0.4994	0.4994
250	0.5081	0.4947	0.4915	0.4898	0.4913	0.4928	0.4984	0.4995	0.4995	0.4994	0.4994
500	0.5034	0.5041	0.4942	0.4929	0.4913	0.4916	0.5015	0.5002	0.4992	0.4993	0.4994

Table A. 6: iP[0.01] Correlation Strategy RAC

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4826	0.4738	0.4739	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737
50	0.4802	0.4783	0.4720	0.4699	0.4703	0.4703	0.4703	0.4703	0.4703	0.4703	0.4703
100	0.4908	0.4718	0.4800	0.4716	0.4719	0.4703	0.4704	0.4704	0.4704	0.4704	0.4704
150	0.4821	0.4768	0.4713	0.4798	0.4719	0.4702	0.4705	0.4705	0.4705	0.4705	0.4705
200	0.4810	0.4769	0.4727	0.4786	0.4798	0.4701	0.4705	0.4704	0.4704	0.4704	0.4704
250	0.4805	0.4794	0.4747	0.4726	0.4784	0.4706	0.4703	0.4704	0.4704	0.4704	0.4704
500	0.4843	0.4799	0.4775	0.4722	0.4735	0.4763	0.4703	0.4703	0.4704	0.4704	0.4704

Table A. 7: iP[0.01] Section Strategy RBC

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4554	0.4681	0.4683	0.4685	0.4685	0.4685	0.4685	0.4685	0.4685	0.4685	0.4685
50	0.4516	0.4642	0.4648	0.4653	0.4658	0.4658	0.4658	0.4658	0.4658	0.4658	0.4658
100	0.4516	0.4642	0.4648	0.4653	0.4659	0.4663	0.4663	0.4663	0.4663	0.4663	0.4663
150	0.4516	0.4642	0.4648	0.4653	0.4659	0.4662	0.4662	0.4662	0.4662	0.4662	0.4662
200	0.4516	0.4642	0.4648	0.4653	0.4659	0.4662	0.4662	0.4662	0.4662	0.4662	0.4662
250	0.4516	0.4642	0.4648	0.4653	0.4659	0.4663	0.4663	0.4663	0.4663	0.4663	0.4663
500	0.4516	0.4642	0.4648	0.4653	0.4659	0.4662	0.4662	0.4662	0.4662	0.4662	0.4662

Table A. 8: iP[0.01] Child Strategy RBC

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4813	0.4963	0.5010	0.5013	0.5015	0.5015	0.5015	0.5015	0.5015	0.5015	0.5015
50	0.4768	0.4918	0.4965	0.4971	0.4978	0.4983	0.4985	0.4985	0.4985	0.4985	0.4985
100	0.4775	0.4926	0.4973	0.4979	0.4986	0.4991	0.4993	0.4993	0.4993	0.4993	0.4993
150	0.4775	0.4926	0.4973	0.4979	0.4986	0.4991	0.4994	0.4994	0.4994	0.4994	0.4994
200	0.4775	0.4926	0.4973	0.4979	0.4986	0.4991	0.4994	0.4994	0.4994	0.4994	0.4994
250	0.4775	0.4926	0.4973	0.4979	0.4986	0.4991	0.4994	0.4994	0.4994	0.4994	0.4994
500	0.4775	0.4926	0.4973	0.4979	0.4986	0.4991	0.4994	0.4994	0.4994	0.4994	0.4994

Table A. 9: iP[0.01] Correlation Strategy RBC

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4716	0.4733	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737
50	0.4672	0.4694	0.4698	0.4703	0.4703	0.4703	0.4703	0.4703	0.4703	0.4703	0.4703
100	0.4672	0.4694	0.4698	0.4703	0.4703	0.4704	0.4704	0.4704	0.4704	0.4704	0.4704
150	0.4671	0.4693	0.4698	0.4703	0.4704	0.4705	0.4705	0.4705	0.4705	0.4705	0.4705
200	0.4671	0.4693	0.4698	0.4703	0.4704	0.4704	0.4704	0.4704	0.4704	0.4704	0.4704
250	0.4671	0.4693	0.4698	0.4703	0.4704	0.4704	0.4704	0.4704	0.4704	0.4704	0.4704
500	0.4671	0.4693	0.4698	0.4703	0.4704	0.4704	0.4704	0.4704	0.4704	0.4704	0.4704

Appendix B

Modifications to the Query Set

The query set used for the above result had two different forms of queries in it. The first form of query looked like this:

```
milan italy football monuments
```

Figure B.1: Query of form 1

The second form of query looked like this:

```
machine learning theory algorithm -bayesian  
or  
Three Greatest Rivers +Japan
```

Figure B.2: Query of form 2

The difference between the two forms of queries is that one form consists of only words where as the second form also contains symbols like “+”and “-“. Reading through the query description (given by the author of the query) suggested that, in case of a “-“ sign the authors did not want the word after “-“ sign in the result and in case of a “+” sign the authors wanted the word after “+” sign in the result. For example, consider the queries in Figure B.2. In the first query, there is the word “bayesian” after “-“ sign. The author of this query wanted to know about any machine learning algorithm other than “bayesian”. In the second query, there is the word

“Japan” after the “+” sign. The author of this query wanted to know only about three greatest rivers in “Japan”.

Since the authors did not want the word after “-“, we developed a new query set such that, in all queries the words after the “-“ are removed. The queries that have a “+” sign are untouched. The queries that are in form 1 (no symbols) are also untouched. Table B.1 gives the effect of removing words after the “-“ sign.

Table B. 1: iP[0.01] Focused Child Upper Bound 2007

	50	100	150	200	250	500	1000	1500	2000	3000	4000
25	0.4972	0.4734	0.5089	0.5144	0.5383	0.5386	0.5381	0.5381	0.5381	0.5381	0.5381
50	0.4751	0.4736	0.4747	0.4943	0.5338	0.5367	0.5343	0.5343	0.5343	0.5343	0.5343
100	0.4510	0.4429	0.4315	0.4538	0.4962	0.5242	0.5273	0.5238	0.5238	0.5238	0.5238
150	0.4398	0.4435	0.4351	0.4369	0.4557	0.5016	0.5066	0.5002	0.4964	0.4964	0.4964
200	0.4379	0.4445	0.4477	0.4361	0.4480	0.4829	0.4894	0.4928	0.4932	0.4893	0.4893
250	0.4697	0.4376	0.4513	0.4361	0.4603	0.4920	0.5031	0.5148	0.5136	0.5098	0.5098
500	0.4664	0.4244	0.4508	0.4444	0.4589	0.4746	0.4990	0.4900	0.4995	0.5052	0.5058

Our best result for Focused 2007 is **0.5266** (with words after “-“ kept in the queries) [1] and the result with words after “-“ sign removed is **0.5386**. So removing the words after “-“ improved the result.