

Improving Results for the 2009 and 2010 INEX Focused Tasks

A thesis

submitted to the faculty of the graduate school

of the University of Minnesota

By

Natasha Deepak Acquilla

In partial fulfillment of the requirements

for the degree of

Master of Science

Dr. Carolyn J. Crouch

August, 2011

Acknowledgments

I would like to take this opportunity to thank God and everyone else who has contributed to the successful completion of this thesis.

I extend my thanks to Dr. Carolyn Crouch for giving me the opportunity to work in the field of Information Retrieval. I would also like to thank her for the constant guidance, support and valuable feedback over the past two years.

I appreciate the knowledge in Computer Science imparted to me over the past two years by Dr. Pete Willemsen, Dr. Hudson Turner, Dr. Ted Pedersen, Dr. Chris Prince. I would also like to thank my Math Professors Dr. Joseph Gallian and Dr. Steve Trogdon for their help and support to gain knowledge in the field of mathematics.

It was indeed a pleasure to work with my team mates Bhagyashri Mahule, Radhika Banhatti, Reena Rachel for all their work, support and feedback during this period. I also thank Sai Subramanyam and Supraja Nagalla for their contribution to the research.

I would like to thank Lori Lucia, Clare Ford and Jim Luttinen of the Computer Science department for all their help and support extended to me.

None of this would have been possible without the continuous support and love of my parents Deepak and Justila and my brother Nikhil.

Abstract

Information retrieval systems aim to retrieve precise and relevant information in response to a user's query. In past years entire documents which were considered to be relevant or highly correlating were returned to users. However with growth of the web and large numbers of XML documents, smaller elements or passages can be returned to the user for more precise results.

This thesis explains Flex, our system for dynamic element retrieval, where in XML elements rather than entire documents are retrieved and returned to the user. It also gives an overview of the process of generating highly correlating elements (from a large document collection) for a set of queries. The aim of this thesis is to improve the results for the INEX 2009 and 2010 Ad Hoc Focused Tasks.

The Focused Tasks require that each query return a result set of non-overlapping elements. This thesis describes the techniques involved in producing such elements and compares the results produced.

Table of Contents

List of Tables.....	iv
List of Figures.....	v
1. Introduction.....	1
2. Overview.....	3
2.1 INEX.....	3
2.2 2009 Retrieval Tasks.....	6
2.3 2010 Retrieval Tasks.....	7
2.4 Evaluation Metrics.....	9
3. Flexible Element Retrieval.....	10
3.1 Operations Before Flex.....	12
3.2 Operations for Flex.....	14
3.3 Operations After Flex.....	19
4. Experiments, Results and Analysis.....	23
4.1 Focusing Strategies.....	23
4.2 2009 Ad Hoc Focused Results.....	24
4.3 Analysis of 2009 Ad Hoc Focused Results.....	26
4.4 2010 Ad Hoc Restricted Focused Results.....	27
4.5 Analysis of 2010 Ad Hoc Restricted Focused Results	29
5. Conclusions and Suggestions for Future Work.....	31
References.....	32

List of Tables

Table 1. Details of the Fields in a Given INEX Topic.....	6
Table 2. Tags Identified as Terminal Nodes.....	12
Table 3. Tags Identified as Non-terminal Nodes.....	12
Table 4. Levels of Parsing.....	13
Table 5. iP[0.01]- Child Strategy for 2009 Ad Hoc Focused Task.....	25
Table 6. iP[0.01]- Section Strategy for 2009 Ad Hoc Focused Task.....	26
Table 7. iP[0.01]- Correlation Strategy for 2009 Ad Hoc Focused Task.....	26
Table 8. 2009 Ad Hoc Focused Task Top 10 Results.....	27
Table 9. char_prec - Child Strategy for 2010 Ad Hoc Restricted Focused Task.....	28
Table 10. char_prec - Section Strategy for 2010 Ad Hoc Restricted Focused Task.....	28
Table 11. char_prec - Correlation Strategy for 2010 Ad Hoc Restricted Focused Task....	29
Table 12. 2010 Ad Hoc Restricted Focused Task Top 10 Results.....	30

List of Figures

Figure 1. Excerpt of INEX Document (ID 52502.xml).....	5
Figure 2. A Sample Topic, INEX Topic ID 2010006 (2010 Topics).....	6
Figure 3. Calculation of iP	9
Figure 4. Calculating Character Precision-Recall.....	9
Figure 5. Operations Before Flex.....	10
Figure 6. Operations for Flex.....	11
Figure 7. Operations After Flex.....	11
Figure 8. Doctree for Article 52502.xml.....	14
Figure 9. Docid-DocPath File for Article 52502.xml.....	15
Figure 10. The Seeded Doctree for Article 52502.xml.....	16
Figure 11. Flex Configuration File.....	17
Figure 12. Formula for ltu -weighting.....	17
Figure 13. Sample Output of Flex Ranked by Document.....	18
Figure 14. Sample Trec Format File.....	19
Figure 15. Expanded Xpaths.....	20
Figure 16. Sample FOL File.....	20
Figure 17. Sample 2009 Focused INEX Evaluation.....	21
Figure 18. Sample 2010 Restricted Focused INEX Evaluation.....	22
Figure 19. Example of Overlapping Elements.....	23

1. Introduction

Information retrieval consists of retrieving relevant information from large document collections, based on queries given by a user. With the growth of the World Wide Web, large amounts of information are represented as XML (Extensible Markup Language) documents. These XML documents provide information about the underlying structure of the documents. Such structure enables us to retrieve XML *elements* (e.g., paragraphs, subsections and sections), as opposed to retrieving larger pieces of a document or the entire document itself. Thus precise information is returned in response to the query. (Retrieval at this level of granularity is made possible here with flexible retrieval [4], our approach to element retrieval, which is dynamic as the XML elements are generated and retrieved at run time. See Chapter 3 for a complete description of this process). Most XML documents do not strictly follow a DTD (Document Type Definition) and hence are called *semi-structured* documents.

The retrieval engine we use is Smart 13.0 [11]. Smart is based on the Vector Space Model [12]. In the Vector Space Model, both documents and queries are represented as n -dimensional vectors. Each component in the vector represents a unique term in the vector. The similarity of a document to a query is determined by calculating the similarity between the document and query vectors, using, for example, the cosine of the angle between them. The Vector Space Model thus allows us to perform ranked retrieval of documents (or elements) from a large document collection.

This thesis describes the work done in our research group at the University of Minnesota Duluth. Our research group is one of the participants in the INEX (Initiative

for the Evaluation of XML Retrieval) competition. INEX focuses on the development and evaluation of XML-based retrieval systems; it provides its participants with the XML document collection (i.e., 50GB subset of Wikipedia), a set of user queries, relevance assessments and tools and metrics to evaluate the retrieved results.

Our research group at the University of Minnesota Duluth has been a participant in the Ad Hoc Retrieval Track for all INEX competitions to date (2001 to 2010). This thesis describes the improvements in our current Ad Hoc retrieval methods compared to previous years. The current INEX Ad Hoc Track along with its document collection, a summary of our retrieval system, and the metrics and tools used for the evaluation of results are described in Chapter 2. Chapter 3 describes our retrieval methods in detail. Chapter 4 describes the evaluations and results of our experiments. Conclusions and suggestions for future research are contained in Chapter 5.

2. Overview

This chapter provides a description of the INEX 2009 and 2010 Ad Hoc tracks and the evaluation measures used to measure system performance.

2.1 INEX

INEX is an organization that supports the development of effective information retrieval systems for XML documents. INEX conducts competitions which attract participants from around the world. It has developed tracks (such as the Ad Hoc, Book, Data-Centric, Interactive, Link-the-Wiki, Question Answering, Relevance Feedback, Web service discovery and XML-mining) to aid participants in developing and evaluating effective XML retrieval systems. In 2009, the Ad Hoc track included the (1) Thorough, (2) Focused and (3) Relevant in Context Tasks. In 2010, these tasks changed slightly to include (1) Efficiency (similar to Thorough), (2) Restricted Focused, (3) Relevant In Context (RiC) and (4) Restricted Relevance In Context. These tasks emphasized focused retrieval of information in terms of *snippets*. A snippet is defined as a text segment that enables a user to assess document relevance and decide on that basis if the document is of interest.

In 2009, our research centered on the Thorough, Focused and Relevance In Context Tasks. In the Thorough Task, the set of all (i.e., overlapping) elements along a path is returned. The Focused Task aims to produce non-overlapping elements which are relevant to the query (i.e., a single element along the path). The Relevance in Context Task requires output in the form of a ranked list of all the focused elements per document, sorted in document order. The document collection, the topics, and an

evaluation package is provided by INEX for these tasks.

Document Collection

The document collection for the 2009 and 2010 tasks is the INEX 2009 Wikipedia collection. It is approximately 50GB in size and contains 2,666,190 XML documents. Each article or document is semi-structured, as it may contain some untagged text. The collection has 32,311 unique tags in it. A sample document from the collection is shown in Figure 1.

Topics

INEX requests each participant group to submit a set of queries at the beginning of the year; INEX then finalizes the set of queries to be included in the topic set and makes them available to the participants. Query fields are described in Table 1. A sample topic is shown in Figure 2. Our experiments use CO (i.e., title only) topics.

Relevance Assessments

Relevance assessments are provided each year by participants, using the GPXRai tool provided by INEX [1]. Using this tool each participant is requested to highlight the pieces of text s/he considers relevant to the query. INEX then gathers all the highlighted text and creates an assessment pool, which is then converted into File Offset Length (FOL) [6] format and is used to evaluate the results.

```

- <!--
  generated by CLiX/Wiki2XML [MPI-InF, MMCI@UdS] $LastChangedRevision: 92 $ on 16.04.2009 15:45:25[mcia0827]
-->
- <article>
- <symbol confidence="0.8" wordnetid="106806469">
- <award confidence="0.8" wordnetid="106696483">
- <signal confidence="0.8" wordnetid="106791372">
- <header>
  <title>Nobel Prize in Physiology or Medicine</title>
  <id>52502</id>
- <revision>
  <id>244569061</id>
  <timestamp>2008-10-11T13:55:36Z</timestamp>
- <contributor>
  <username>VolkovBot</username>
  <id>3035831</id>
</contributor>
</revision>
- <categories>
  <category>Medicine awards</category>
  <category>Nobel Prize</category>
</categories>
</header>
- <bdy>
- <image width="150px" src="1950Nobel.JPG" type="thumb">
- <caption>
  Front side of an award medal in physiology or medicine.
</caption>
</image>
  The
- <award wordnetid="106696483" confidence="0.9508927676800064">
  <link xlink:type="simple" xlink:href="..201/21201.xml"> Nobel Prize</link>
</award>
  in
  <link xlink:type="simple" xlink:href="..597/23597.xml"> Physiology</link>
  or
  <link xlink:type="simple" xlink:href="..957/18957.xml"> Medicine</link>
  (
- <language wordnetid="106282651" confidence="0.9508927676800064">
  <link xlink:type="simple" xlink:href="..689/26689.xml"> Swedish</link>
</language>
  :
  <it>Nobelpriset i fysiologi eller medicin</it>
  ) is awarded once a year by the Swedish
+ <university wordnetid="108286163" confidence="0.9508927676800064"></university>
  . It is one of the five Nobel Prizes established by the will of
+ <person wordnetid="100007846" confidence="0.9508927676800064"></person>
  in 1895, awarded for outstanding contributions in
+ <physical_entity wordnetid="100001930" confidence="0.8"></physical_entity>
  ,
+ <symbol wordnetid="106806469" confidence="0.8"></symbol>
  ,
+ <symbol wordnetid="106806469" confidence="0.8"></symbol>

```

Figure 1. Excerpt of INEX Document (ID 52502.xml)

Table 1. Details of the Fields in a Given INEX Topic [7]

Field	Description
Title	Contains Content Only (CO) queries
Castitle	Contains Content and Structure (CAS) queries
Phrasetitle	Contains phrase titles
Description	A brief description of the information needed is given here
Narrative	Information about what is relevant and irrelevant to the topic is given here

```

- <topic id="2010006" ct_no="320">
  <title>Evolution of Storage devices</title>
  - <castitle>
    //article[about(., Evolution or History of Storage devices) or about(., "Primary or Secondary Storage devices")]//sec[about(., RAM
    ROM)]
  </castitle>
  <phrasetitle>"Evolution / History of storage devices"</phrasetitle>
  - <description>
    Find information about the development of storage devices over time.
  </description>
  - <narrative>
    A Storage device is a major component a computer system. Data has to be stored for various processing needs. With faster technology
    the need for better and faster storage devices is pertinent.I thought it would be interesting to know about how the storage devices changed
    over time to see how it catered to the needs with the evolution of technology.
  </narrative>
</topic>

```

Figure 2. A Sample Topic, INEX Topic ID 2010006 (2010 Topics)

2.2 2009 Retrieval Tasks

The 2009 Ad Hoc tasks are described below. Full descriptions can be found at [6].

Thorough Task

This task returns a ranked list of elements for each query. This list may contain overlapping elements. For the evaluation of the Thorough Task, Mean Average interpolated Precision (MAiP) is used [6].

Focused Task

For this task, a set of non-overlapping elements, ranked in terms of perceived relevance to the query, is returned to the user. For example, if a positively correlated paragraph is found, the subsection or subsection containing that paragraph (i.e., its parent) can also be considered to be relevant to the query. In such cases, either the child or the parent is returned to prevent overlap in the elements returned. A metric called interpolated Precision at 1% recall (iP[0.01]) [6] is used to evaluate focused results. See Section 2.4 for details.

Relevant In Context (RiC) Task

This task is similar to the Focused Task, except that the ranked elements are grouped by document before being returned to the user. Here, for each query we first retrieve the most highly correlated articles and then retrieve focused elements from those articles. Results are evaluated based on an overall performance estimate [6]. The metric used is Mean Average generalized Precision (MAgP) [6].

2.3 2010 Retrieval Tasks

The 2010 Ad Hoc tasks are described below. Full descriptions can be found at [1].

Efficiency Task

In this task retrieved elements or passages may overlap. Either 15, 150 or 1500 elements per topic may be returned to the user. The evaluation measure is mean average interpolated precision (MAiP), calculated over 101 standard recall points (0.00, 0.01, 0.02, ..., 1.00).

Restricted Focused Task

This task is similar to the Focused Task of the 2009 track, the difference being that only 1000 characters per topic may be returned to the user. Evaluation is in terms of set-based precision over the retrieved characters (`char_prec`) [1]. See Section 2.4 for details.

Relevant in Context (RiC) Task

This task is similar to the 2009 Relevant in Context Task, but it is viewed in the form of snippet retrieval (i.e., snippets must be returned). The measure to evaluate snippets is the T2I(300) measure [2], which penalizes the retrieval of irrelevant text.

Restricted Relevant in Context Task

This task is similar to the 2010 RiC Task, but only 500 characters per element may be returned. This is to simulate information retrieval on small screen mobile devices or a document summary on a hit-list, and the best elements or passages that convey relevant information are to be selected [1]. The T2I(300) [2] measure is used to evaluate the restricted RiC snippets.

The research described in this thesis is based on producing a methodology for the solution of the generic Ad Hoc Focused Tasks (2009 Focused and 2010 Restricted Focused). The metrics used for evaluation of these tasks are described in Section 2.4. See [3] and [8] for detailed descriptions of the other Ad Hoc tasks for 2009 and 2010.

2.4 Evaluation Metrics

Interpolated Precision

For the 2009 Focused Task, interpolated Precision is calculated as shown in

Figure 3.

$$iP[x] = \begin{cases} \max(P[r] \wedge R[r]), & \text{if } x \leq R[|L_q|] \\ 0, & \text{otherwise} \end{cases}$$

Here L_q is the ranked list of elements
 $|L_q|$ is the length of the ranked list. In INEX, this length is 1500
 $P[r]$ is the precision at rank r
 $R[r]$ is the recall at rank r
 $R[|L_q|]$ is the recall over all the documents retrieved

Figure 3. Calculation of iP

Character Precision-Recall

For the 2010 Restricted Focused Task, the character average precision-recall is calculated as shown in Figure 4. See [2] for more details. Average character precision is calculated only for a relevant document. For non-relevant documents, the character precision-recall for the document is 0.

$$\text{Average character precision } (d) = \frac{\sum_{p=1}^{|d|} (P_d(p) \times RL_d(p))}{NRC_d}$$

Here p is the character position from the point the reading starts
 RL is a binary valued function on the relevance of a given position
 NRC is the number of relevant characters in document d
 P precision at a given position in d .

Figure 4. Calculating Character Precision-Recall [2]

3. Flexible Element Retrieval

This chapter explains the procedures involved in our approach to Flexible Element Retrieval (Flex). Flex enables us to retrieve specific elements in response to a query rather than entire documents. We attempt to minimize the time a user has to spend to obtain specific information in response to his/her query.

The procedures preceding Flex operation are illustrated in Figure 5. Flex actions are seen in Figure 6. The procedures performed after Flex are described in Figure 7.



Figure 5. Operations Before Flex

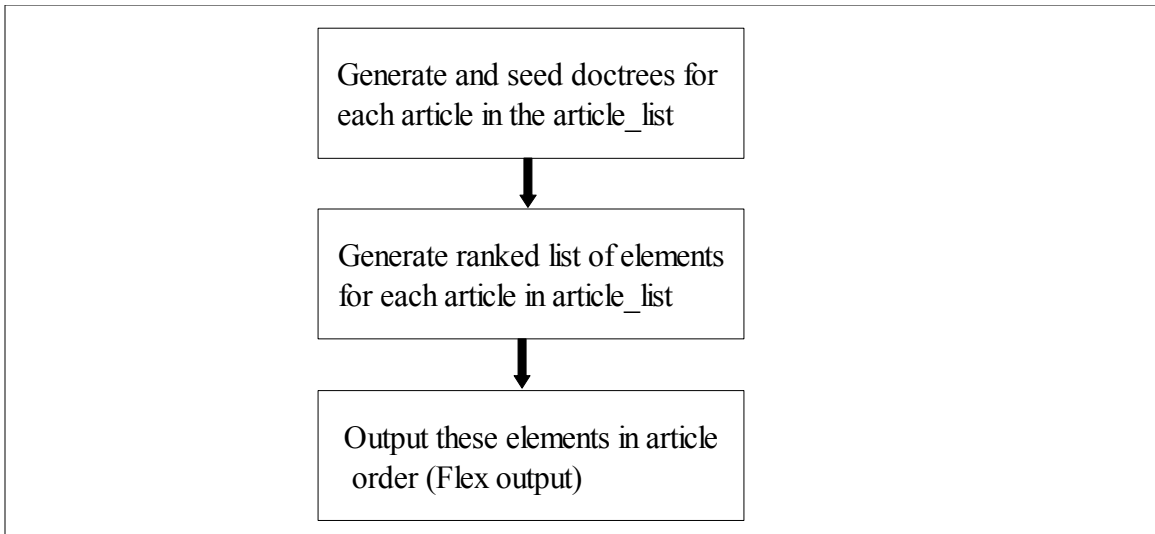


Figure 6. Operations for Flex

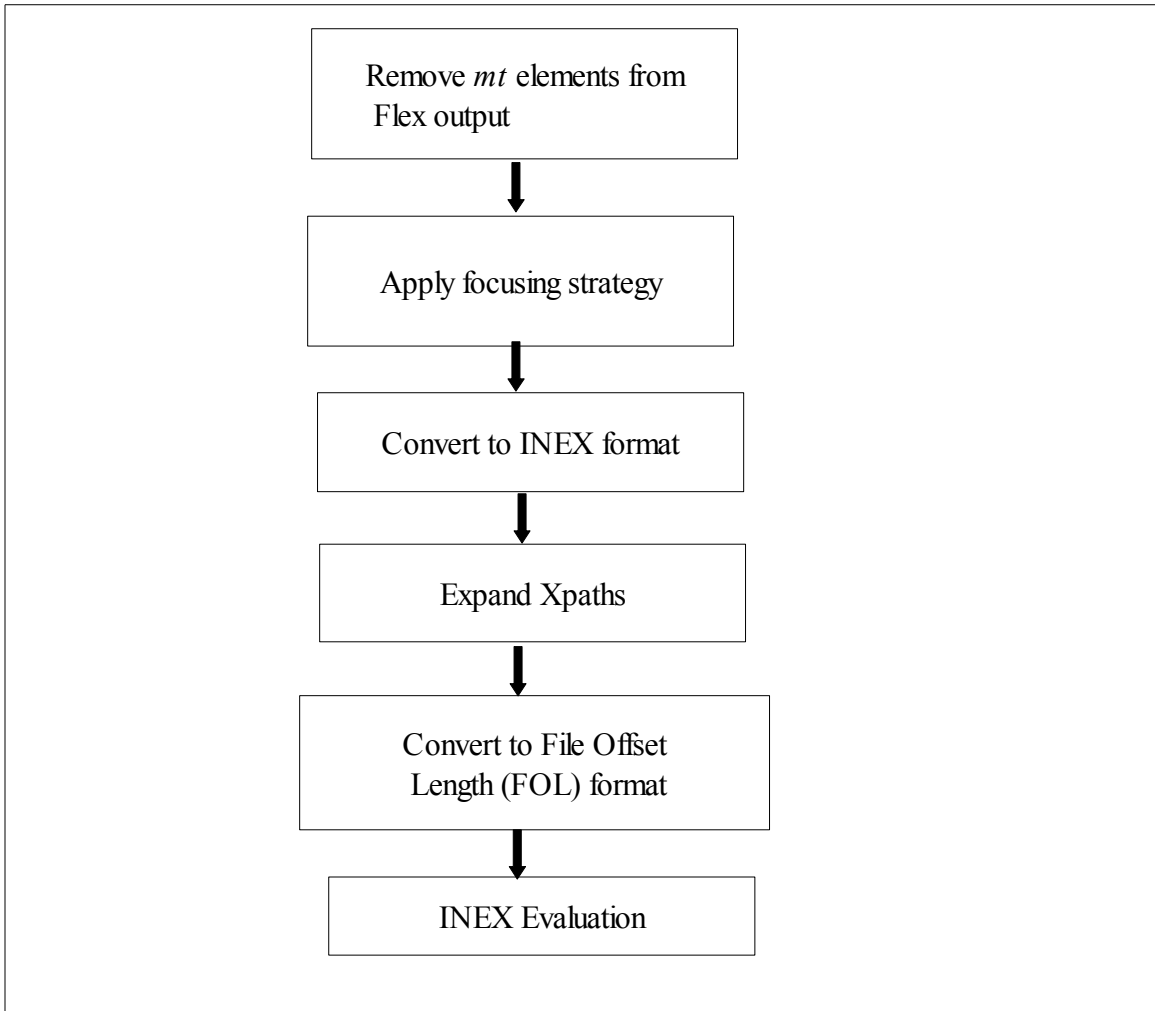


Figure 7. Operations After Flex

3.1 Operations Before Flex

The document collection and query set are parsed using the libxml2 parser (version libxml2-2.7.8). The XML elements in each document are considered to be either terminal or non-terminal nodes. The tags identified as terminal nodes and non-terminal nodes are listed in Table 2 and Table 3, respectively.

Table 2. Tags Identified as Terminal Nodes

Terminal-node	Tag
Title	<title> ... </title>
Template	<template> ... </template>
Table	<table> ... </table>
Section title	<st> ... </st>
Image	<image> ... </image>
Ordered list	 ...
Unordered list	 ...
Normal list	<normallist> ... </ normallist>
Number list	<numberlist> ... </ numberlist>
Definition list	<definitionlist> ... </definitionlist>
Paragraph	<p> ... </p>
List	<list> ... </list>

Table 3. Tags Identified as Non-terminal Nodes

Non-terminal node	Tag
Article	<article> ... </article>
Body	<bdy> ... </bdy>
Section	<sec> ... </sec>
Subsection	<ss1> ... </ss1>, <ss2> ... </ss2> ...
Header	<header> ... </header>

The untagged text at every level of the document is enclosed within *magic text* (*mt*) tags at that level. The *mt* nodes are present in order for Flex to build all parent vectors correctly. See [9] for details. The collection is first scrubbed or cleaned to remove unwanted tags and non-alphanumeric characters; this text is then input to the parsing phase. Each scrubbed XML document is parsed in terms of levels; for example, level 1 contains all the nodes at level 1 in the documents. As seen in Table 4, parsing of the document collection includes 8 parses at the interior node levels (0-7), para parse which contains the terminal nodes, and the para+*mt* parse for all the terminal nodes including *mts*. See [3] for details.

Table 4. Levels of Parsing

article	Level 0
bdy/header	Level 1
section/subsection	Level 2 - Level 7
Terminal nodes	Level 8

The parsed documents and queries are indexed. The document vectors are *Lnu* weighted and the query vectors are *ltu* weighted [13]. See [3] for details. A Smart retrieval is performed to obtain a list of the top n most highly correlated articles for each topic. For focused retrieval, we first identify the highly correlated articles corresponding to a topic and then identify the focused elements within them. In our experiments we use the Reference Runs provided by INEX as the list of articles.

3.2 Operations for Flex

Generation of Doctrees

A doctree provides the structure or schema of a document. Doctrees are generated by performing a preorder traversal of the parsed document collection. A sample doctree is shown in Figure 8. In this figure, the first column represents the Xpath of the element, the second represents its number of children, and the third column indicates if it has a sibling to its right (1 if it has a sibling to the right and 0 otherwise). A node is identified as a terminal node if it has 0 children. The doctree for each article is stored in a file named with its Wiki-id.

/article[1]/	2	0
/article[1]/header[1]/	2	1
/article[1]/header[1]/title[1]/	0	1
/article[1]/header[1]/categories[1]/	0	0
/article[1]/bdy[1]/	6	0
/article[1]/bdy[1]/image[1]/	0	1
/article[1]/bdy[1]/p[1]/	0	1
/article[1]/bdy[1]/sec[1]/	3	1
/article[1]/bdy[1]/sec[1]/st[1]/	0	1
/article[1]/bdy[1]/sec[1]/p[1]/	0	1
/article[1]/bdy[1]/sec[1]/p[2]/	0	0
/article[1]/bdy[1]/sec[2]/	2	1
/article[1]/bdy[1]/sec[2]/st[1]/	0	1
/article[1]/bdy[1]/sec[2]/p[1]/	0	0
/article[1]/bdy[1]/sec[3]/	4	1
/article[1]/bdy[1]/sec[3]/st[1]/	0	1
/article[1]/bdy[1]/sec[3]/p[1]/	0	1
/article[1]/bdy[1]/sec[3]/p[2]/	0	1
/article[1]/bdy[1]/sec[3]/p[3]/	0	0
/article[1]/bdy[1]/mt[1]/	0	0

Figure 8. Doctree for Article 52502.xml

Docid-DocPath Mapping

The docid-docPath mapping contains a mapping between the Smart id of the

element and its Xpath. The docid-docPath file is produced by the script *generate_docid_docpath_mapping.pl* which takes as input the *textloc* file produced during document indexing. A sample of the docid-docPath file is shown in Figure 9. We use the docid-docPath file of the paras+mt vectors for seeding since it contains all the terminal and mt nodes. The docid-docPath mapping enables Flex to build the doctrees by mapping the Smart identifiers of the terminal nodes to their corresponding Xpaths. The first column in the docid-docPath file represents the Smart identifier of the element and the second column represents the Xpath of the element, including the Wiki-id of the document it is contained in.

19146225	52502/article[1]/header[1]/title[1]/
19146226	52502/article[1]/header[1]/categories[1]/
19146227	52502/article[1]/bdy[1]/image[1]/
19146228	52502/article[1]/bdy[1]/p[1]/
19146229	52502/article[1]/bdy[1]/sec[1]/st[1]/
19146230	52502/article[1]/bdy[1]/sec[1]/p[1]/
19146231	52502/article[1]/bdy[1]/sec[1]/p[2]/
19146232	52502/article[1]/bdy[1]/sec[2]/st[1]/
19146233	52502/article[1]/bdy[1]/sec[2]/p[1]/
19166103	52502/article[1]/bdy[1]/sec[3]/st[1]/
19166106	52502/article[1]/bdy[1]/sec[3]/p[1]/
19166107	52502/article[1]/bdy[1]/sec[3]/p[2]/
19166109	52502/article[1]/bdy[1]/sec[3]/p[3]/
19166112	52502/article[1]/bdy[1]/mt[1]/

Figure 9. Docid-DocPath File for Article 52502.xml

Seeded Doctrees

Only terminal nodes of the doctree are seeded. The Smart id of the corresponding element is fetched from the docid-docPath file and is listed alongside the element. A sample seeded doctree is shown in Figure 10. The first column represents the Wiki-id of the document containing the element and its Xpath. The second column represents the

number of children of the element. The third column indicates if there a sibling to the right of the element (1 if there is a sibling to the right and 0 otherwise). The fourth, highlighted column represents the Smart id of the terminal element and the fifth column represents a dummy correlation value for the element with respect to the query (in our case the dummy correlation is 0).

52502/article[1]/	2	0		
52502/article[1]/header[1]/	2	1		
52502/article[1]/header[1]/title[1]/	0	1	19146225	0
52502/article[1]/header[1]/categories[1]/	0	0	19146226	0
52502/article[1]/bdy[1]/	6	0		
52502/article[1]/bdy[1]/image[1]/	0	1	19146227	0
52502/article[1]/bdy[1]/p[1]/	0	1	19146228	0
52502/article[1]/bdy[1]/sec[1]/	3	1		
52502/article[1]/bdy[1]/sec[1]/st[1]/	0	1	19146229	0
52502/article[1]/bdy[1]/sec[1]/p[1]/	0	1	19146230	0
52502/article[1]/bdy[1]/sec[1]/p[2]/	0	0	19146231	0
52502/article[1]/bdy[1]/sec[2]/	2	1		
52502/article[1]/bdy[1]/sec[2]/st[1]/	0	1	19146232	0
52502/article[1]/bdy[1]/sec[2]/p[1]/	0	0	19146233	0
52502/article[1]/bdy[1]/sec[3]/	4	1		
52502/article[1]/bdy[1]/sec[3]/st[1]/	0	1	19166103	0
52502/article[1]/bdy[1]/sec[3]/p[1]/	0	1	19166106	0
52502/article[1]/bdy[1]/sec[3]/p[2]/	0	1	19166107	0
52502/article[1]/bdy[1]/sec[3]/p[3]/	0	0	19166109	0
52502/article[1]/bdy[1]/mt[1]/	0	0	19166112	0

Figure 10. The Seeded Doctree for Article 52502.xml

Flex calculates the correlation for each element in the seeded doctree by taking the inner product of the Lnu-weighted element vector and the corresponding ltu-weighted query vector. The formula for calculating the ltu weight for each term in the query vector

is shown in Figure 12. The best all-element slope value is 0.11 and the best all-element pivot value is 38. See [3] for details.

Flex takes as input a configuration file as shown in Figure 11.

<p>DOC_INDEX_PATH (path to doc.nnn file from the <i>para+mt</i> indexing)</p> <p>ARTICLE_LIST (path to ranked list of articles retrieved for the topic set)</p> <p>QUERY_LTU_PATH (path to <i>ltu</i> weighted query.nnn vectors)* (* required when using a pre-existing <i>ltu</i>-weighted query file)</p> <p>QUERY_INDEX_PATH (path to query.nnn vectors)</p> <p>OUTPUT_PATH (path to output file generated by Flex)</p> <p>RESULT_TREES_PATH (path to seeded doc trees of documents in the <i>article_list</i>)</p> <p>SLOPE_ALLELEMS (all-element slope value, for use in <i>Lnu</i>-element term weighting)</p> <p>PIVOT_ALLELEMS (all-element pivot value, for use in <i>Lnu</i>-element term weighting)</p>

Figure 11. Flex Configuration File

$\frac{1 + \log(tf) * \log(N/n_k)}{(1 - slope) + slope * (number\ of\ unique\ terms) / pivot}$ <p>Here <i>tf</i> – term frequency <i>N</i> – collection size <i>n_k</i> - number of documents that contain this term <i>slope</i> and <i>pivot</i> – empirically determined constants number of unique terms is the number of distinct terms in the query vector</p>

Figure 12. Formula for *ltu*-weighting

The *ltu* weight for each query term can be calculated by Flex in two ways: (1) by determining *N* and *n_k* from the global statistics of the collection as in [5]. This method can be calculated at execution time. (2) The *ltu* weights for the query can be read from

the query.ltu file which is created by *ltu* weighting of the query.nnn vectors. This method can only be used (in lieu of the first method) if an all-element collection has been created. The global values of n_k required by *ltu*-weighting can easily be obtained at execution time (as illustrated in [5]) using the nstats file.

The output of Flex is a ranked list of elements for each topic, which are ordered by document (according to the `article_list`) and sorted by correlation within each document. Figure 13 shows a sample output of Flex. The first column represents the query number for which the element was retrieved. The second column represents the Wiki-id of the document and the Xpath of the element. The third column represents the correlation of the element with the query.

1	21201/article[1]/bdy[1]/sec[4]/p[3]/	6.5477
1	21201/article[1]/bdy[1]/sec[6]/ss1[1]/p[4]/	5.85175
1	21201/article[1]/bdy[1]/sec[2]/p[4]/	5.34255
1	21201/article[1]/bdy[1]/sec[3]/p[5]/	4.75056
1	52502/article[1]/bdy[1]/	37.4588
1	52502/article[1]/bdy[1]/sec[3]/	35.6705
1	52502/article[1]/bdy[1]/sec[3]/p[1]/	23.2623
1	52502/article[1]/bdy[1]/p[1]/	22.1915
1	52502/article[1]/bdy[1]/sec[3]/p[2]/	19.1709
1	52502/article[1]/bdy[1]/sec[2]/	18.0332
1	52502/article[1]/bdy[1]/sec[2]/p[1]/	17.6994
1	52502/article[1]/header[1]/	15.7204
1	52502/article[1]/header[1]/title[1]/	13.387
1	52502/article[1]/header[1]/categories[1]/	13.387
1	52502/article[1]/bdy[1]/sec[1]/p[1]/	7.53524
1	52502/article[1]/bdy[1]/sec[3]/p[3]/	7.39033
1	52502/article[1]/bdy[1]/sec[1]/	4.39528
1	141921/article[1]/bdy[1]/	33.8861
1	141921/article[1]/bdy[1]/sec[9]/	31.4253
1	141921/article[1]/bdy[1]/sec[9]/p[1]/	31.395
1	141921/article[1]/bdy[1]/sec[10]/	29.59
1	141921/article[1]/bdy[1]/sec[5]/	28.7504

Figure 13. Sample Output of Flex Ranked by Document

3.3 Operations After Flex

After Flex has generated its output, the *mts* are removed (since these artificial nodes were introduced to deal with semi-structured documents and do not exist in the original document collection). Since overlapping elements may not be returned for the 2009 Ad Hoc Focused and 2010 Ad Hoc Restricted Focused Tasks, we apply focusing strategies (child, section and correlation strategy) to remove the overlaps (See Chapter 4 for details).

Conversion to Trec Format

The focused output is converted to INEX Trec format as shown in Figure 14. Columns 1-7 represent the INEX topic id, a dummy value, the Wiki-id of the document in which the element resides, the rank of the element, its inverse rank, run name and the Xpath of the element, respectively.

```
2009001 Q0 52502 78 1500 UMD_FOC /article[1]/bdy[1]/
2009001 Q0 52502 79 1499 UMD_FOC /article[1]/bdy[1]/sec[3]/
2009001 Q0 52502 80 1498 UMD_FOC /article[1]/bdy[1]/sec[3]/p[1]/
2009001 Q0 52502 81 1497 UMD_FOC /article[1]/bdy[1]/p[1]/
2009001 Q0 52502 82 1496 UMD_FOC /article[1]/bdy[1]/sec[3]/p[2]/
2009001 Q0 52502 83 1495 UMD_FOC /article[1]/bdy[1]/sec[2]/
2009001 Q0 52502 84 1494 UMD_FOC /article[1]/bdy[1]/sec[2]/p[1]/
2009001 Q0 52502 85 1493 UMD_FOC /article[1]/header[1]/
```

Figure 14. Sample Trec Format File

Patching the Xpaths

In order to be evaluated properly, the Xpaths in the Trec format file must be expanded. (The Xpaths generated from the scrubbed collection do not include all the tags preceding *bdy* in the document collection). To produce the correct Xpath of the element, we patch the Xpaths with the tags which were eliminated during scrubbing. A sample of

the expanded paths for the Trec file in Figure 14 is shown in Figure 15. The columns are similar to the Trec format mentioned above except that the last column contains the expanded Xpaths.

```
2009001 Q0 52502 35 1500 UMD_FOC /article[1]/symbol[1]/award[1]/signal[1]/bdy[1]/p[1]
2009001 Q0 52502 36 1499 UMD_FOC /article[1]/symbol[1]/award[1]/signal[1]/bdy[1]/sec[3]
2009001 Q0 52502 37 1498 UMD_FOC /article[1]/symbol[1]/award[1]/signal[1]/bdy[1]
2009001 Q0 52502 38 1497 UMD_FOC /article[1]/symbol[1]/award[1]/signal[1]/bdy[1]/sec[3]/p[3]
2009001 Q0 52502 39 1496 UMD_FOC /article[1]/symbol[1]/award[1]/signal[1]/bdy[1]/sec[1]
```

Figure 15. Expanded Xpaths

Conversion to FOL Format

The Trec format with expanded Xpaths is converted to a FOL format file as shown in Figure 16. The 8 columns represent the topic id, a dummy value, the Wiki-id of the document in which the element resides, the rank of the element, its inverse rank, the run name, the start of the element in terms of character offset from the start of the document, and the number of characters to be read starting from the offset.

```
2009001 Q0 52502 78 1500.0 UMD_FOC 133 38979
2009001 Q0 52502 79 1499.0 UMD_FOC 33991 5120
2009001 Q0 52502 80 1498.0 UMD_FOC 34011 521
2009001 Q0 52502 81 1497.0 UMD_FOC 1541 746
2009001 Q0 52502 82 1496.0 UMD_FOC 34533 230
2009001 Q0 52502 83 1495.0 UMD_FOC 23467 10523
2009001 Q0 52502 84 1494.0 UMD_FOC 23483 10505
2009001 Q0 52502 85 1493.0 UMD_FOC 4 128
2009001 Q0 52502 86 1492.0 UMD_FOC 5 37
```

Figure 16. Sample FOL File

Evaluation

The FOL format is evaluated for the 2009 Ad Hoc Focused Task using $iP[0.01]$. A snippet of the evaluation is shown in Figure 17. The highlighted value shows the evaluation measure for the 2009 Ad Hoc Focused Task.

```
<eval run-id="UMD_FOCUSED" file="subtoFOL_1500.txt">
num_q      all    68
num_ret    all    1687
num_rel    all    4858
num_rel_ret all    734
ret_size   all    12663115
rel_size   all    18838137
rel_ret_size all    3286326
iP[0.00]  all    0.6565658650092696
iP[0.01]  all    0.6129733540211658
iP[0.05]  all    0.5512473622362251
iP[0.10]  all    0.4649842493930969
MAiP      all    0.18701074980416488
ircl_prn.0.00 all    0.6565658650092696
ircl_prn.0.01 all    0.6129733540211658
ircl_prn.0.02 all    0.5886923760752297
ircl_prn.0.03 all    0.5714328913896185
ircl_prn.0.04 all    0.5633877311868141
ircl_prn.0.05 all    0.5512473622362251
ircl_prn.0.06 all    0.5494238155828602
ircl_prn.0.07 all    0.5317690593908461
ircl_prn.0.08 all    0.5063255578708352
ircl_prn.0.09 all    0.49278008864086215
ircl_prn.0.10 all    0.4649842493930969
.
.
.
```

Figure 17. Sample 2009 Focused INEX Evaluation

The FOL format is evaluated for the 2010 Ad Hoc Restricted Focused Task using char_prec. A snippet of the evaluation is shown in Figure 18. The highlighted value shows the evaluation measure for the 2010 Ad Hoc Restricted Focused Task.

```

<eval run-id="UMD_FOC" file="subtoFOL_1500.trec++">
num_q      all  52
num_ret    all  56
num_rel    all  5471
num_rel_ret all  38
ret_size   all  52000
rel_size   all  17641119
rel_ret_size all  18666
art_prec   all  0.6826923076923077
art_rec    all  0.02184447460080518
art_fl     all  0.040522719877816006
char_prec  all  0.35896153846153844
char_rec   all  0.006101545466889521
char_fl    all  0.011457118282828806
iP[0.00]  all  0.3897996818360134
iP[0.01]  all  0.11778846153846154
iP[0.05]  all  0.038461538461538464
iP[0.10]  all  0.0
MAiP      all  0.007413287025032882
ircl_prn.0.00 all  0.3897996818360134
ircl_prn.0.01 all  0.11778846153846154
ircl_prn.0.02 all  0.08038461538461537
.
.
.

```

Figure 18. Sample 2010 Restricted Focused INEX Evaluation

4. Experiments, Results and Analysis

This chapter describes the experiments performed for the 2009 Ad Hoc Focused Task and the 2010 Ad Hoc Restricted Focused Task. It also describes the focusing strategies applied to the Flex output in order to avoid overlap of elements in the results.

4.1 Focusing Strategies

Since the 2009 Focused Task and 2010 Restricted Focused Task results cannot contain overlapping elements, we must apply one of several previously generated strategies to remove overlaps. Consider the elements with their correlations as shown in Figure 19 below.

52502/article[1]/bdy[1]/ 35.5775 52502/article[1]/bdy[1]/sec[2]/ 17.1275 52502/article[1]/bdy[1]/sec[2]/p[1]/ 16.8104

Figure 19. Example of Overlapping Elements

These elements are overlapping since the content of paragraph 1 (*p[1]*) is a part of the content of its parent, section element (*sec[2]*), and the content of the section element is part of the content of the body element (*bdy[1]*). Hence there is an overlap of content in these elements. Using the focusing strategies, we remove such overlaps to obtain a set of non-overlapping focused elements.

Child Strategy [10]

This strategy selects the element at a deeper level of the tree, as compared to an element at a higher level, even if the parent has a greater correlation. Here the child node is always preferred to its parent. In the example, this strategy would select 52502/article[1]/bdy[1]/sec[2]/p[1] over the other two elements.

Section Strategy [10]

Using this strategy, preference is given to the element with the highest correlation along a path, provided it is not a body element (*bdy*). Most of the elements selected by this strategy were observed to be sections; hence this technique is called the section strategy. In the example, this strategy selects 52502/article[1]/bdy[1]/sec[2] over the other two elements.

Correlation Strategy [10]

Using this strategy, the element with the greatest correlation along a given path is chosen over all the other elements along the same path. Most of the elements selected using the correlation strategy were observed to be entire body nodes. In the example, this strategy selects 52502/article[1]/bdy[1] over the other two elements.

4.2 2009 Ad Hoc Focused Results

In the 2009 Ad Hoc Focused Task, a ranked list of 1500 non-overlapping elements per topic are returned to the user. Experiments were conducted by varying the number of highly correlating documents and the number of elements retrieved from them. In the following tables n indicates the number of documents from which the elements were retrieved and m denotes the number of focused elements retrieved from n documents. Experiments were conducted with Child, Section and Correlation strategies to remove the overlaps. The results of each of these experiments are listed below.

Child Strategy

In this experiment the Child Strategy is used for removing overlaps. The results for this experiment are shown below in Table 5. Best results are shown in bold.

Table 5. iP[0.01]- Child Strategy for 2009 Ad Hoc Focused Task

$\begin{matrix} m \\ n \end{matrix}$	# of Elements							
# of Docs	50	100	150	200	250	500	1000	1500
25	0.5524	0.5835	0.6028	0.6069	0.6081	0.6124	0.6129	0.6130
50	0.5524	0.5835	0.6028	0.6069	0.6081	0.6133	0.6145	0.6145
100	0.5524	0.5835	0.6028	0.6069	0.6081	0.6133	0.6146	0.6146
150	0.5524	0.5835	0.6028	0.6069	0.6081	0.6133	0.6146	0.6146
200	0.5524	0.5835	0.6028	0.6069	0.6081	0.6133	0.6146	0.6146
250	0.5524	0.5835	0.6028	0.6069	0.6081	0.6133	0.6146	0.6146
500	0.5524	0.5835	0.6028	0.6069	0.6081	0.6133	0.6146	0.6146
1000	0.5524	0.5835	0.6028	0.6069	0.6081	0.6133	0.6146	0.6146

Section Strategy

In this experiment the Section Strategy is used for removing overlaps. The results for this experiment are shown below in Table 6. Best results for this experiment are shown in bold.

Correlation Strategy

In this experiment the Correlation Strategy is used for removing overlaps. The results for this experiment are shown below in Table 7. Best results for this experiment are shown in bold.

Table 6. iP[0.01]- Section Strategy for 2009 Ad Hoc Focused Task

n \ m	# of Elements							
# of Docs	50	100	150	200	250	500	1000	1500
25	0.6324	0.6496	0.6508	0.6532	0.6537	0.6537	0.6537	0.6537
50	0.6324	0.6496	0.6508	0.6532	0.6552	0.6555	0.6555	0.6555
100	0.6324	0.6496	0.6508	0.6532	0.6552	0.6555	0.6557	0.6557
150	0.6324	0.6496	0.6508	0.6532	0.6552	0.6555	0.6557	0.6557
200	0.6324	0.6496	0.6508	0.6532	0.6552	0.6555	0.6557	0.6557
250	0.6324	0.6496	0.6508	0.6532	0.6552	0.6555	0.6557	0.6557
500	0.6324	0.6496	0.6508	0.6532	0.6552	0.6555	0.6557	0.6557
1000	0.6324	0.6496	0.6508	0.6532	0.6552	0.6555	0.6557	0.6557

Table 7. iP[0.01]- Correlation Strategy for 2009 Ad Hoc Focused Task

n \ m	# of Elements							
# of Docs	50	100	150	200	250	500	1000	1500
25	0.5912	0.5900	0.5921	0.5921	0.5921	0.5921	0.5921	0.5921
50	0.5912	0.5938	0.5938	0.5938	0.5938	0.5938	0.5938	0.5938
100	0.5912	0.5938	0.5940	0.5940	0.5940	0.5940	0.5940	0.5940
150	0.5912	0.5938	0.5940	0.5940	0.5940	0.5941	0.5941	0.5941
200	0.5912	0.5938	0.5940	0.5940	0.5940	0.5941	0.5941	0.5941
250	0.5912	0.5938	0.5940	0.5940	0.5940	0.5941	0.5941	0.5941
500	0.5912	0.5938	0.5940	0.5940	0.5940	0.5941	0.5941	0.5941
1000	0.5912	0.5938	0.5940	0.5940	0.5940	0.5941	0.5941	0.5941

4.3 Analysis of 2009 Ad Hoc Focused Results

For the 2009 Ad Hoc Focused Task the section strategy produced better results than the child and correlation strategies. The highest iP[0.01] value obtained is **0.6557**

using the section strategy. There were 19 organizations that submitted their results for the 2009 Ad Hoc Track. Our results compared with those of the top ten participants in INEX 2009 are shown in Table 8. The highest $iP[0.01]$ **0.6557** is better than that of the university ranked #1. We observe that the best results are obtained at 100 articles and 1000 elements. Using a confidence interval of 95% in a one-tailed t-test, we found that our best result was statistically significant compared to the results of the participants ranked 3, 5, 7, 9, 10 and not statistically significant from results of the other five teams.

Table 8. 2009 Ad Hoc Focused Task Top 10 Results

Participant	$iP[0.01]$	Rank
p72-UMD (section strategy)	0.6557	-
p78-University of Waterloo	0.6333	1
p72-UMD (child strategy)	0.6146	-
p68-Univ. Pierre et Marie Curie	0.6141	2
p10- Max-Planck-Institute	0.6134	3
p60-Saint Etienne University	0.6060	4
p6-Univ. of Amsterdam	0.5997	5
p72-UMD (correlation strategy)	0.5941	-
p5-Queensland Univ. of Tech	0.5592	6
p16-Univ. of Applied Science	0.5903	7
p48-LI	0.5853	8
p22-ENSM - S	0.5844	9
p25-Renmin Univ. of China	0.4973	10

4.4 2010 Ad Hoc Restricted Focused Results

In the 2010 Ad Hoc Restricted Focused Task, a ranked list of 1500 non-overlapping snippets per topic are returned to the user. Each snippet has a maximum of

1000 characters. Experiments were conducted by varying the number of documents and the number of elements retrieved from them. In the following tables n indicates the number of documents from which the elements were retrieved and m denotes the number of focused elements retrieved from the n documents. Experiments were conducted with Child, Section and Correlation strategies to remove overlaps. The results of each of these experiments are listed below. As the tables show, the result window is filled at $n=25$, $m=50$ and there is no further improvement to be gained for higher values of n and m .

Child Strategy

In this experiment the Child Strategy is used for removing overlaps; results are shown below in Table 9. The best results for this experiment are shown in bold.

Table 9. char_prec - Child Strategy for 2010 Ad Hoc Restricted Focused Task

$n \backslash m$	# of Elements							
# of Docs	50	100	150	200	250	500	1000	1500
25	0.3494	0.3494	0.3494	0.3494	0.3494	0.3494	0.3494	0.3494

Section Strategy

In this experiment the Section Strategy is used for removing overlaps; results are shown below in Table 10. The best results for this experiment are shown in bold.

Table 10. char_prec - Section Strategy for 2010 Ad Hoc Restricted Focused Task

$n \backslash m$	# of Elements							
# of Docs	50	100	150	200	250	500	1000	1500
25	0.3569	0.3569	0.3569	0.3569	0.3569	0.3569	0.3569	0.3569

Correlation Strategy

In this experiment the Correlation Strategy is used for removing overlaps; results are shown below in Table 11. The best results for this experiment are shown in bold.

Table 11. char_prec – CorrelationStrategy for 2010 Adhoc Restricted Focused Task

n \ m	# of Elements							
# of Docs	50	100	150	200	250	500	1000	1500
25	0.3347	0.3347	0.3347	0.3347	0.3347	0.3347	0.3347	0.3347

4.5 Analysis of 2010 Ad Hoc Restricted Focused Results

In the 2010 Ad Hoc Focused Task, since only 1000 characters per topic may be returned to the user, the window for each query is filled by a small number of elements. It was observed that retrieving the top 50 elements from the top 25 documents was sufficient to produce 1000 characters per topic. The best result for this task has a char_prec of **0.3569** using the section strategy. There were 18 organizations that submitted their results for the 2010 Ad Hoc Track. The comparison of our results with other participants of INEX are shown in Table 12. The best result for char_prec **0.3569** is better than that of the university ranked #3. We are awaiting data from INEX for significance testing of our 2010 results.

Table 12. 2010 Ad Hoc Restricted Focused Task Top 10 Results

Participant	char_prec	Rank
p68-University Pierre et Marie Curie	0.4125	1
p55-Doshisha University	0.3884	2
p72-UMD (section strategy)	0.3569	-
p72-UMD (child strategy)	0.3494	-
p9-University of Helsinki	0.3435	3
p98-LIA - University of Avignon	0.3434	4
p167-Peking University	0.3370	5
p65-Radboud University Nijmegen	0.3361	6
p72-UMD (correlation strategy)	0.3347	-
p5-Queensland University of Technology	0.3199	7
p557-Universitat Pompeu Fabra	0.3066	8
p4-University of Otago	0.3036	9
p29-Indian Statistical Institute	0.2451	10

5. Conclusions and Suggestions for Future Work

The 2010 INEX Ad Hoc track has been changed to the 2011 Snippet Retrieval Track. This track requires 500 snippets per topic; a snippet may not exceed 300 characters. Shorter snippets are to be returned for the 2011 track and research has to be done to select the best 300 characters out of each highly correlating element.

Since the 2011 Snippet Track requires shorter snippets, a *Sub-section* Strategy could be implemented (similar to Section Strategy) where subsections are given preference over other elements. This could produce smaller elements compared to *sec* but sufficient to fill the window of 300 characters of a snippet. A good foundation has been laid this year to retrieve snippets and we wish success to the next year's team participating in the Snippet Track.

During evaluation of results, we observed that some of the terminal nodes (*st*, *title*, *categories*) we had considered in our parsing and indexing were termed “small irrelevant nodes” (due to their small size) by the INEX evaluation tool. Excluding such nodes might improve results.

For the 2010 Ad Hoc Restricted Focused Task the first 1000 characters from the element are selected for return to INEX. However, there is no surety that the relevant data is present in those 1000 characters. Future work could be done to identify the best set of 1000 characters from each element to be returned to INEX; this may improve results of the 2010 Ad Hoc Restricted Focused Task.

References

- [1] Arvola, P., Geva, S., Kamps, J., Schenkel, R., Trotman, A., and Vainio, J. Overview of the INEX 2010 Ad Hoc Track.
<http://www.cs.otago.ac.nz/homepages/andrew/2010-13.pdf>
- [2] Arvola, P., Kekalainen, J. and Junkkari, M. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13:460–484, 2010.
- [3] Banhatti, R. Improving Results for the INEX 2009 Thorough and 2010 Efficiency Tasks, MS Thesis, Department of CS, UMD, August 2011.
<http://www.d.umn.edu/cs/thesis/banhatti.pdf>
- [4] Crouch, C. Dynamic element retrieval in structured environment. *ACM TOIS*, 24(4): 437-454, 2006.
- [5] Ganapathibhotla, S. Query Processing in a Flexible Retrieval Environment, MS Thesis, Department of CS, UMD, July 2006.
<http://www.d.umn.edu/cs/thesis/ganapathibhotla.pdf>
- [6] Geva, S., Kamps, J., Lethonen, M., Schenkel, R., Thom, J. and Trotman, A. Overview of the INEX 2009 Ad Hoc Track.
<http://www.cs.otago.ac.nz/homepages/andrew/2009-13.pdf>
- [7] INEX 2009 Guidelines for Topic Development.
<http://www.inex.otago.ac.nz/tracks/adhoc/gtd.asp>
- [8] Narendravarapu, R. Improving Results for the INEX 2009 and 2010 Relevant in Context Tasks, MS Thesis, Department of CS, UMD, August 2011.
<http://www.d.umn.edu/cs/thesis/narendravarapu.pdf>
- [9] Paranjape, D. Improving Focused Retrieval, MS Thesis, Department of CS, UMD, July 2008.
<http://www.d.umn.edu/cs/thesis/paranjape.pdf>
- [10] Poluri, P. Focused Retrieval using Exact Methodology, MS Thesis, Department of CS, UMD, August 2009.
<http://www.d.umn.edu/cs/thesis/poluri.pdf>
- [11] Salton, G. *The Smart Retrieval System – Experiments in Automatic Documents Retrieval*. Prentice-Hall, Eaglewood Cliffs, NJ, 1971.
- [12] Salton, G., Wong, A., Yang, C.S., A vector space model for automatic indexing. *Comm. ACM* 18(11), 613–620 (1975).

- [13] Singhal, A., Buckley, C., Mitra, M. Pivot document length normalization. In *Proceedings of the 19th Annual International ACM Special Interest Group in Information Retrieval (SIGIR) Conference, Zurich*. 19-21, 1996.