

Significance Testing for INEX 2008-09 Ad Hoc Track

A thesis
submitted to the faculty of the graduate school
of the University of Minnesota
by

Ramakrishna Cherukuri

In partial fulfillment of requirements
for the degree of
Master of Science
August, 2010.

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

Acknowledgements

I would like to take this opportunity to acknowledge all the people who helped me during this thesis work. Firstly I would like to thank my advisor, Dr. Carolyn J. Crouch for her invaluable guidance, timely advice and support.

I would like to thank my committee members Dr. Donald B. Crouch and Dr. Todd W. Loushine for their suggestions and feedback during the important stages of my thesis.

I would like to thank all the faculty and staff (especially Jim Luttinen and Lori Lucia) at the Department of Computer Science at the University of Minnesota Duluth for their assistance during my master's course-work. I would thank my fellow graduate students who offered great help during my study and stay in Duluth.

I would like to thank Varun Sudhakar, Pavan Poluri, Chaitanya Polumetla, and Dinesh Bhirudh for sharing their experience and helping me on various aspects of thesis. I would like to thank my co-workers Sridhar Uppala, Sandeep Vadlamudi and Abhijet Mahule for their valuable time spent helping me in this thesis. I thank my friends Sathavahana Bhogapathi, Sunil Vejandla, Vivek Kasireddy, Bharat Siginam, Venkat Kovelamudi and others for supporting me and helping me in times of need.

Finally I would like to thank all my friends and family members for supporting me all throughout my life.

ABSTRACT

INEX (*The INitiative for the Evaluation of XML retrieval*) sponsors a competition that promotes the development/evaluation of XML-based retrieval systems. INEX provides a document collection, a query set (topics) and evaluation measures for use by the XML-based retrieval systems of the participants. We have developed a methodology for the retrieval of elements, at the appropriate level of granularity, within the XML document. This methodology is applied to the tasks of the INEX Ad Hoc Track.

In this thesis, the focus lies on significance tests that are performed to see how our methods compare with those used by the top-ranked INEX participants. Our approach to significance testing is identical to that used by INEX for evaluating its participants. We use the results of each individual query; find the variance and standard deviation between the scores, and the value of t , which maps to probability. We use a confidence interval of 95% in one-tailed t -test (so the probability must be less than 0.05 to assure significance) to see if one method is significantly better than another. We evaluate results of the 2008 and 2009 Ad Hoc tasks using this approach. The results of these significance tests for the basic INEX Ad Hoc tasks are given in this paper along with observations on these results.

TABLE OF CONTENTS

LIST OF TABLESiv

LIST OF FIGURES.....vi

1. Introduction1

2. Overview3

2.1 INEX3

2.2 2009 Retrieval Tasks8

2.3 2009 Evaluation Measures10

2.4 Retrieval Engine14

3. Background to Retrieval15

3.1 Pre-Processing15

3.2 Base Retrieval16

3.3 Flex Retrieval17

3.4 Post-Processing.....18

4. Significance Tests, Results and Analysis19

4.1 Evaluating the Results of the Participants.19

4.2 Testing Process22

4.3 Significance Test Results26

4.4 Observations.....32

5. Conclusion36

References38

Appendix A.....40

LIST OF TABLES

Table 1: Sample Document from the 2009 Document Collection from INEX.	5
Table 2: Example of a Query Submission for INEX Ad Hoc Task.....	7
Table 3: Sample qrel file from the INEX 2009 Ad Hoc Task.	8
Table 4: Results of UMD Focused task with Child.	20
Table 5: Sample Output from <code>inex_eval</code>	21
Table 6: Sample of the t-test-file input.	25
Table 7: The Comparison of Method 1 to Method 2.	26
Table 8: Top 10-ranked Participants in Focused Task – 2008.	27
Table 9: Top 10-ranked Participants in Focused Task – 2009.	28
Table 10: Top 10-ranked Participants in RIC Task – 2008.	29
Table 11: Top 10-ranked Participants in RIC Task – 2009.	30
Table 12: Top 10-ranked Participants in Thorough Task – 2009.	31
Table A1: Focused Task Significance Test Results for 2008 using Section Strategy.	40
Table A2: Focused Task Significance Test Results for 2008 using Child Strategy.	40
Table A3: Focused Task Significance Test Results for 2008 using Correlation Strategy.	41
Table A4: Focused Task Significance Test Results for 2009 using Section Strategy.....	41
Table A5: Focused Task Significance Test Results for 2009 using Correlation Strategy.	42
Table A6: Focused Task Significance Test Results for 2009 using Child Strategy.....	42

Table A7: Relevant in Context Task Significance Test Results for 2008 using Section Strategy.....	43
Table A8: Relevant in Context Task Significance Test Results for 2008 using Child Strategy.....	43
Table A9: Relevant in Context Task Significance Test Results for 2008 using Correlation Strategy.	44
Table A10: Relevant in Context Task Significance Test Results for 2009 using Correlation Strategy.	44
Table A11: Relevant in Context Task Significance Test Results for 2009 using Child Strategy.	45
Table A12: Relevant in Context Task Significance Test Results for 2009 using Section Strategy.	45
Table A13: Thorough Task Significance Test Results for 2009 using All Element.	46
Table A14: Thorough Task Significance Test Results for 2009.	46
Table A15: Relevant in Context Task Significance Test - II Results for 2008.....	47
Table A16: Relevant in Context Task Significance Test - II Results for 2009.....	47
Table A17: Thorough Task Significance Test - II Results for 2009.	47

LIST OF FIGURES

Figure 1: Formula for Average Interpolated Precision.	11
Figure 2: Formula for Mean Average Interpolated Precision.	11
Figure 3: Formula for Finding Interpolated Precision (iP).	12
Figure 4: Formula to Find Mean Average Generalized Precision (MAgP).	13
Figure 5: Finding the Value of t (for Significance Testing).	23
Figure 6: Formula to Find the t from the Differences in the Scores for Each Query....	23
Figure 7: Usage of out_sig_test script.	24
Figure 8: Significant Testing for the Focused Task-2008.	32
Figure 9: Significant Testing for the Focused Task-2009.	32
Figure 10: Significant Testing for the Relevant in Context Task-2008.	33
Figure 11: Significant Testing –II for the Relevant in Context Task-2008.	33
Figure 12: Significant Testing for the Relevant in Context Task-2009.	34
Figure 13: Significant Testing –II for the Relevant in Context Task-2009.	34
Figure 14: Significant Testing on Thorough Task-2009.	35
Figure 15: Significant Testing –II for the Thorough Task-2009.	35

1. Introduction

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)[8]. As defined in this way, information retrieval used to be an activity that only a few people engaged in: reference librarians, paralegals, and similar professional searchers [8]. Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email.

In *web search*, the system has to search over a large number (potentially billions) of documents and give the results to the user. To provide the user with relevant and precise information in response to a query, one must consider the large amount of textual data represented as XML (Extensible Markup Language)[8]. The focus of this research is to extend traditional information retrieval, which concentrates on retrieving entire documents, to the retrieval of textual elements represented as XML. 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. Flexible retrieval is also dynamic, as the elements are retrieved at run time. The XML documents referred to here are *semi-structured* documents [3], meaning that not all text is enclosed within XML tags.

The retrieval environment used for dynamic element retrieval makes use of the Vector Space Model [11]. 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. 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. 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, is more efficient with respect to file space, and is cost effective [9].

INEX 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 performance of the participant systems is compared based on these evaluation measures [3].

2. Overview

This chapter presents a brief overview of INEX and the INEX 2009 Ad Hoc track in terms of document collection (database of the documents provided by INEX), topic development (queries by users), tasks and runs, relevance assessments, and evaluation measures used for assessing system performance. The Vector Space Model, the basis for our basic retrieval system (Smart), is also described in this chapter.

2.1 INEX

INEX, the *Initiative for the Evaluation of XML Retrieval*, is an initiative in the field of Information Retrieval (IR) with a goal of promoting the evaluation of XML retrieval systems. It provides test collections (e.g., a collection of IEEE articles in 2002 and 2005, Wikipedia documents in 2006 - 2010, and a collection of scanned books licensed from Microsoft from 2007), uniform evaluation measures, and a forum for all the organizations participating in it (to compare their results and improve their strategies through discussion) [3]. INEX provides Ad Hoc, Book, Efficiency, Entity Ranking, Interactive (iTrack), Question Answering (QA@INEX), Link-the-Wiki, and XML-Mining tracks for its participants/organizations [3]. The Ad Hoc XML task contains four sub-tasks, namely: Thorough Task, Focused Task, Relevant In Context Task, and Best In Context Task [3]. The University of Minnesota Duluth (UMD) participates in the Ad Hoc track using the basic retrieval engine, Smart, and our own system for the dynamic retrieval of elements, Flex. We use the collection and topics provided by INEX and our

software to produce the relevant XML elements. These results are evaluated using the relevance assessments and software provided by INEX.

Document Collection

The 2009 collection is approximately 50.7 GB in size distributed over 995 parts. This collection is 8.6 times the size of collection provided in 2007 and 2008 [9]. These Wikipedia collections contain some degree of uniform structure as XML documents but do not strictly follow DTD (Document Type Definition) convention. Hence the collection is treated as semi-structured. The present collection has over 30,000 tags (most of which are removed during the parsing) and 2,666,190 articles in it [9]. A sample document from the collection provided by INEX in 2009 is given in Table 1.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- generated by CLiX/Wiki2XML [MPI-Inf, MMCI@UdS] $LastChangedRevision: 92
$ on 17.04.2009 03:27:08[mciao0828] -->
<!DOCTYPE article SYSTEM "../article.dtd">
<article xmlns:xlink="http://www.w3.org/1999/xlink">
<header>
<title>Portal:English football/Selected picture/21</title>
<id>16183995</id>
<revision>
<id>196856544</id>
<timestamp>2008-03-08T21:32:02Z</timestamp>
<contributor>
<username>Garden</username>
<id>5019622</id>
</contributor>
</revision>
<categories>
<category>English football portal selected pictures</category>
</categories>
</header>
</article>
```

Table 1: Sample Document from the 2009 Document Collection from INEX.

Topic Collection

A set of topics/queries is framed by the participants in INEX under the guidelines of the INEX to be used with the collection provided. These topics are submitted to the INEX. Each topic has some mandatory nodes like title, description, phrase-title and so on. The Title node contains the CO (Content-Only) query used in the retrieval process [9]. This query does not provide structural information or hints about the relevant text. The other optional nodes like castitle and narrative provide the information for checking relevance. Castitle provides a CAS (Content-And- Structure) query. The 2009 topic set contains 115 topics. A sample query is given in Table2.

Relevance Assessments:

Once the participants submit the topics to the INEX, INEX chooses a set of topics to be evaluated manually. Each selected topic is paired with a subset of the document collection. This data is given to the participants for manual assessment. The relevance assessments produced are used for evaluating the results produced by the participants using their mechanisms/systems. In 2009, INEX provided the GPXRai assessment system, which serves as an interface enabling the user to mark the text and Best Entry Point (BEP) found relevant to a particular query. These assessments are converted into qrel files, which give the file offset and length format for each query for every document relevant to it. These qrel files are used for evaluating the results [1].

A sample of INEX 2009 relevance assessment (inex2009qrel file) is given below (in Table 3).

<p>OpenGL Shading Language GLSL (id=229, INEX 2009) <u>45/100 Relevant</u></p> <p><title>OpenGL Shading Language GLSL</title></p> <p><castitle>//article [about(.,GLSL) or about(., "OpenGL Shading Language")]//sec[about(., GLSL OpenGL Shading)]</castitle></p> <p><phrasetitle> "OpenGL Shading Language"</phrasetitle></p> <p><description></p> <p>Find information about GLSL-OpenGL Shading Language.</p> <p></description></p> <p><narrative></p> <p>GLSL (OpenGL Shading Language), also known as GLSLang, is a high level shading language based on the C programming language..... From my point of view I would like to get the collection of articles with description about GLSL or the History of GLSL and the details of the functions and coding methodologies used in it and any constructive information about the shading language. If I find these elements then it would be relevant from my point of view.</p> <p></narrative></p>
--

Table 2: Example of a Query Submission for INEX Ad Hoc Task.

2009001 Q0 3260094 4213 4436 144 144:4213
2009001 Q0 21201 24903 33106 137 137:542 704:1871 2578:2506 5089:19984
2009001 Q0 52502 23232 39116 197 197:2085 2287:21147
2009001 Q0 19653466 16901 30413 288 288:2150 2458:14751
2009001 Q0 141921 22899 39637 124 124:11553 11685:11346
2009001 Q0 3260076 0 5071
2009001 Q0 80144 0 22137

Table 3: Sample qrel file from the INEX 2009 Ad Hoc Task.

2.2 2009 Retrieval Tasks

The 2009 Ad hoc retrieval tasks include these sub-tasks:

Thorough Task

“The aim of this task is to find all relevant elements or passages ranked in relevance order. It will be therefore the case that, due to the nature of relevance in XML retrieval (e.g. if a child element is relevant, so will be its parent, although to a greater or lesser extent), an XML retrieval system that has estimated an element to be relevant may decide to return all its ancestor elements. This means that runs for this task may contain a large number of overlapping elements. It is however a challenge to rank these elements appropriately. The evaluation will consider only the unseen text retrieved.” [6, p. 5].

Focused Task

In this task we find the most focused element that is retrieved for a query. The aim of the **Focused Task** is to return a ranked-list of elements or passages, where no result may be overlapping with any other result [1]. In general when we try to find the elements relevant to a query, the parents of that element share some relevance. For example, we find relevance in the paragraph, the sub-section that contains the paragraph, the section containing that subsection, and others in the upper level of the tree share some relevance. To remove the overlap in elements, we use the specific method (e.g., correlation score) to choose the most relevant element. This process is called overlap removal [1].

Relevant In Context Task

In this task we return all the non-overlapping or focused elements associated with the query. These elements are returned as a ranked list of focused elements from the top ranked documents in document order [3].

Best In Context Task

In this task we return the Best Entry Point along with the ranked list of articles. The Best Entry Point is that point in the article where the user begins reading to get the relevant information from the document or article [9].

2.3 2009 Evaluation Measures

INEX specifies the evaluation measures for the Thorough, Focused, Relevant in Context, and Best in Context tasks. In 2008 there were slight modifications to the evaluation process [6]. The Xpaths are to be converted into FOL format before evaluation [1]. Evaluation measures for the tasks are explained below.

Thorough Task

In this task we measure the performance across a set of topics. We use Mean Average Interpolated Precision (MAiP) as the metric for evaluating Thorough task. The overall performance of focused retrieval [see below] over the entire list is assessed using the average interpolated precision (AiP) measure [1]. The average interpolated precision measure is calculated by averaging interpolated precisions at 101 standard recall levels. We get MAiP by calculating the mean of the AiP values obtained for each individual topic. [2]. The formula for AiP and MAiP are given below (in Figure 1).

$$AiP = \frac{1}{101} \sum_{x=0.0,0.01,\dots,1.0} iP[x]$$

where $iP[x]$ is interpolated precision at recall x .

Figure 1: Formula for Average Interpolated Precision

Assuming there are n topics

$$MAiP = \frac{1}{n} \sum_t AiP(t)$$

Figure 2: Formula for Mean Average Interpolated Precision

Focused Task

Interpolated precision at 1% recall, also known as $iP[0.01]$, is the metric used for evaluating the Focused task. *Recall* is the term used to specify the fraction of highlighted text that is retrieved. *Precision* is the term used to specify the fraction of retrieved text that is highlighted. For details see [9].

$$iP[x] = \begin{cases} \max_{1 \leq r \leq |L_q|} (P[r] \wedge R[r]) & \text{if } x \leq R[|L_q|] \\ 0 & \text{otherwise} \end{cases}$$

where L_q is the ranked list of elements

and $|L_q|$ is the length of this ranked list, which is 1500 for INEX competition

$P[r]$ 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: Formula for Finding Interpolated Precision (iP).

Relevant in Context Task

Generalized precision and recall are the basis for the Relevant in Context task evaluation. Here per document score evaluates the match between retrieved text and relevant text in the document [5]. The focus in Relevant in Context is on overall performance so the main evaluation measure used is MAgP (Mean Average generalized Precision) as given below [3]. The formula for MAgP is given below in Figure 4.

Generalized Precision: $gP[r]$: sum of document scores up to (and including) document-rank r , divided by the rank r .

$$gP[r] = \frac{\sum_{i=1}^r S(d_i)}{r}$$

$S(d)$ – Score for individual document

Generalized Recall: $gR[r]$: Number of relevant documents retrieved up to (and including) document-rank r , divided by the total number of relevant documents.

$$gR[r] = \frac{\sum_{i=1}^r IsRel(d_i)}{N_{rel}}$$

Average Generalized Precision: AgP : Average of the generalized precision scores obtained for each natural recall points, where generalized recall increases.

$$AgP = \frac{\sum_{r=1}^{|\mathcal{L}|} IsRel(d_r) \cdot gP[r]}{N_{rel}}$$

The mean average generalized precision (MAgP) is simply the mean of the average generalized precision scores over all topics.

Assuming there are n topics:

$$MAgP = (1/n) * \sum_t AgP(t)$$

Figure 4: Formula to find Mean Average Generalized Precision

Best In Context Task:

Similar to the Relevant in Context Task, generalized precision and recall are the basis for the Best in Context Task. However the per document score is the match between user-specified entry point and the best entry point in the document [5]. The Best in Context Task uses MAgP as the evaluation measure, since the focus is on overall performance even in Best in Context Task [2].

2.4 Retrieval Engine

Vector Space Model

In the Vector Space Model, both the query and document are represented by weighted term vectors using the n-dimensional vector representation. Components of the vector are all the word types (unique terms) in the document (query) vector. We use a similarity measure to get the correlation scores for each document with the query vector. This model is the basis for all the retrieval in our work [11].

Smart Retrieval Engine

The Smart Retrieval Engine uses the Vector Space Model to perform its functions of indexing, term weighting, and retrieval of rank ordered elements. We use the Smart 13.0 as our retrieval engine for this research work [10].

3. Background to Retrieval

This chapter gives an insight into the process of retrieval. INEX provides all its participants with the Wikipedia document collection and a query set. The job of the participants is to retrieve relevant information from the document collection with respect to the queries in the query set, as declared by the particular INEX task. The steps in the retrieval are discussed below.

3.1 Pre-Processing

Cleaning the Document Collection

Once we download the documents, we screen them to see if all the text in the collection is clean. If not, we remove the non-meaningful text (e.g., `?"<!@#$$%^&**>:"`) and the unwanted tags (e.g., `<u>` `<i>` `` `<place>` `<Country>`). There are more than 30,000 different tags used in the Wikipedia collection; some of them are formatting tags and user-defined tags that are not useful in retrieval.

Parsing

“Parsing is the process of recognizing as an entity all the text enclosed within a matching set of open and closed XML tags.” [9]. We parse with respect to articles, sections, sub-sections, paragraphs and paragraphs including magic text. Magic text is the

untagged text in the document. Flex needs all the terminal nodes in the document to build the parent nodes correctly so we must take magic text (enclosed within `<mt>` and `</mt>` tags) into consideration. For details, see [9] and [1].

3.2 Base Retrieval

Indexing

Indexing is the process of creating the element and query vectors. Input for the indexing step is the element set generated by parsing. We use Smart to produce the nnn vectors (also called as term frequency vectors) for all the parses. Each vector represents an element in the parse. Indexing needs *tags-to-keep* and *tags-to-index* based on which the indexes are created. For details, see [9].

Weighting

Many different weighing schemes are used to find the correlation score between the element vectors and query vectors. For details see [12] and [13]. The *Lnu-ltu* weighing scheme is used here to minimize the advantage given to longer vectors by earlier weighting schemes.

Retrieval

The elements are *Lnu-weighted* and query is *ltu-weighted*. We use inner product as the similarity measure between the two vectors. Based on this correlation score, a ranked list of elements is produced for each and every query. For more details, see [1].

3.3 Flex Retrieval

Seeding

This is the Flex part of our work. Flex produces the doc trees (document trees) including the correlation scores for the elements. We normally generate doc-trees of interest based on the ranked list produced in the article retrieval. Seeding is the process of filling in the content of all the terminal elements of the doc-tree. So after seeding, all the terminal nodes of the doc-tree are populated. For details see [1] and [9].

Flex

This step is done after seeding. The input for this step is the seed subsets produced using Flex in the seeding process. Having correlated the query with each terminal node, Flex generates each parent node (from its children) and then correlates that node with the query using a bottom up approach. Having generated all the correlation

scores for all the nodes in the tree, the *mt* tags created during parsing are removed. The result is a rank-ordered list of elements from the tree. This process continues until all trees are generated.

3.4 Post-Processing

Conversion

Depending on the task we are going to perform, we filter or rearrange the rank-ordered list generated by Flex retrieval. The Focused Task needs a ranked list without any overlapping elements. The Relevance in Context needs the rank list with elements grouped by document. To remove overlapping elements, we used *child*, *section* and *correlation* strategies (for more details see, [1] chapter 4).

Evaluation

Once Flex output is converted into the required format, we use the evaluation tool provided by INEX to get the results. We use *iP[0.01]* for Focused task, *MAgP* for Relevance in Context and *MAgP* for Best in Context task. We specify the task we are interested in by giving the arguments for the evaluation tool. This evaluation tool uses specific arguments to provide results for each individual query. We use this functionality in our significance testing discussed in Chapter 4.

4. Significance Tests

Once the participants submit the baseline runs, INEX evaluates them and publishes the results of all the participants. Our aim here is to improve our results. So we compare our results with those of the top 10 participants. As mentioned in Chapter 3, the INEX evaluation tool can produce the evaluation scores for each individual query (along with the score for the entire query set as a whole). We performed significance testing to compare our results (Tables 7-9 [2] and results from [7, 14]) with those of the top ranked participants.

We downloaded the baseline results of the top ten INEX participants from the website. We need to compare the results produced by our retrieval method with other participants to decide if our method produces statistically better results. Our approach to this problem is to apply the paired t-test to the differences between scores for each query. The comparison is between our results and those from the top ten participant groups.

4.1 Evaluating the results

A sample of results downloaded from INEX is given in Table 4. The first column is the query number, second column is a dummy tag, third column is the document id, fourth and fifth columns are the rank and inverse ranks, sixth column is the participant run id, and the last column is the Xpath of the element found relevant to the query.

2009001	Q0	3260094	1	4000	umd_focused_2	/article[1]/bdy[1]/p[2]
2009001	Q0	52497	2	3999	umd_focused_2	/article[1]/bdy[1]/sec[3]
2009001	Q0	52497	3	3998	umd_focused_2	/article[1]/bdy[1]/sec[7]/p[1]
2009001	Q0	52497	4	3997	umd_focused_2	/article[1]/bdy[1]/sec[2]/p[1]
2009001	Q0	52497	5	3996	umd_focused_2	/article[1]/bdy[1]/sec[1]/p[1]
2009001	Q0	52497	6	3995	umd_focused_2	/article[1]/bdy[1]/sec[6]/p[1]
2009001	Q0	52497	7	3994	umd_focused_2	/article[1]/bdy[1]/sec[2]/p[2]
2009001	Q0	52497	8	3993	umd_focused_2	/article[1]/bdy[1]/sec[8]
2009001	Q0	52497	9	3992	umd_focused_2	/article[1]/bdy[1]/sec[1]/p[6]
2009001	Q0	52497	10	3991	umd_focused_2	/article[1]/bdy[1]/sec[1]

Table 4: Results of UMD Focused Task with Child Strategy (top 10 rank list for query 001)

We then use the INEX evaluation tool (`inex_eval`) on these results (i.e., complete results of Table 4) and specify the task (Focused or Relevant in Context) along with the argument specifying a query-by-query result.

The command used to run evaluation tool is:

“Java -jar `inex_eval.jar`(evaluation tool) -(argument to specify the task f-focused r-relevant in context) -q(to get metrics for every query) <qrels file> <result file downloaded from INEX>”

Example:

```
java -jar inex_eval.jar -f -q inex2009.qrels 664.txt >> 664results.txt
```

```
<eval run-id="umd_focused_2" file="664.txt">
```

num_ret	2009001	129
num_rel	2009001	22
num_rel_ret	2009001	19
ret_size	2009001	75493
rel_size	2009001	260694
rel_ret_size	2009001	15223
iP[0.00]	2009001	0.9990198480764518
iP[0.01]	2009001	0.9990198480764518
iP[0.05]	2009001	0.22093377646836856
iP[0.10]	2009001	0.0
AiP	2009001	0.028532423782439392

Table 5: Sample Output from `inex_eval` for Query 2009001 on UMD_Focused Child Strategy

Once we get these results for all the queries we repeat the step for all the top-ranked participant groups. We do this for each task.

Metrics Used

We use the interpolated precision at 1% recall ($iP[0.01]$) as the metric for the Focused task, Mean Average interpolated Precision (MAiP) as metric for the Thorough task, and Mean Average generalized Precision (MAgP) as the metric for the Relevant in Context task.

4.2 Testing Process

We used David Hull's Statistical Testing for Evaluating Retrieval Experiments [4] as the basis for our significance testing. We want to know if the differences in the scores are significant. We used the one tailed t-test to do this. We need to calculate the value of t to establish whether the results produced by one method differ significantly over that produced by a second method [4].

T-test

The one-tailed t-test gives an indication of difference between the two sets of results and is thus used to check whether the two sets of results are essentially different. We did this t-test with the null hypothesis (two sets of results are equal). "The t-test assumes that the error follows the normal distribution, but it often performs well even when this assumption is violated" [4]. The value of t is calculated as:

$t = \text{experimental effect} / \text{variability}$

$= \text{difference between group means} / \text{standard error of difference between group means}$

Figure 5: Finding the Value of t (for Significance Testing)

To find the value of t, we need the values below:

N: number of queries

df: degrees of freedom = n-1

Paired t-test

$$t = \frac{\bar{D}}{s(D_i) / \sqrt{n}}$$

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i, \quad s(D_i) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2}$$

assumption: errors are normally distributed
distribution under H_0 : Student's t with n-1 degrees
of freedom.

Here X_i and Y_i are the scores of retrieval methods X and Y for each query I where

$i=1 \dots n$ and $D_i=Y_i-X_i$

Figure 6: Formula to Find the t from the Differences in the Scores for Each Query

We took the confidence interval as 95%. This approach is identical to that used by Kamps [6] in the significance testing on participant scores for the Ad Hoc Track.

Interpretation

We need to find the probability using the value of t . The resultant t -value is looked up in a t -table to determine the probability that a significant difference between the results exists and so that we can claim the efficacy of one method over the other. Since our confidence interval is 95%, the probability should be less than 0.05 for one method to differ significantly from the other. After looking at the t -table and finding t -inverse, the t -value should be more than 1.67 for the first method to vary significantly from the other.

We used the `t-test-file` script (provided by Kamps [6]) to find the value of t . To run this script we need to provide an input file in a required format. The details for generating the input file using the output of `inex_eval` follow. Place the runs that you want to compare in a folder and provide it as input to the `out-sig-test.pl` as seen in Figure 7. (This script converts the results into the required format.)

```
perl out_sig_test.pl folder >> 705_705.txt
```

Here “folder” is the folder name that contains the results of the two methods from INEX evaluation tool.

Figure 7: Usage of `out_sig_test` script

A sample t-test-file input file is given in Table 6.

2009001	0.9990198480764518	0.9990198480764518
2009002	0.0	0.0
2009003	0.0	0.5120564388754344
2009004	0.34397119086054884	0.5815299446401611
2009005	0.12936590721932473	0.12024428931579069
2009006	0.0	0.0
2009010	0.06138073065902579	0.04249008051157595

Table 6: Sample of the t-test-file input.

We provide this as input to the t-test file and it calculates the sum of differences, mean, standard variance, and t-value and gives the results. Running the script: “perl t-test-file.pl 663_664.txt >> 663_664_stats” produces results as seen Table 7.

In this example, the value of t produced is 0.34 and since it is less than 1.67, the results produced by method of run 663 do not vary significantly from those produced by the method of run 664.

n	:	68
df	:	67
sum	:	1.39276029
mean	:	0.02048177
s ²	:	15.6978612
t	:	0.3489269

Table 7: The Comparison of Method 1 (run 663) to Method 2 (run 664)

4.3 Significance Test Results

We compared our experimental results for Focused, Thorough and Relevant in context with the baseline runs of the top-ranked participants. We did these experiments on 2008 and 2009 results. For more details about 2008 runs, see [A methodology for producing improved focused elements – table 7-9]. For more details about 2009 Focused Task and Relevant in Context Task runs, see [14]. For more details about Thorough Task runs, see [7]. The top 10 ranked participants results along with the UMD results using different strategies are given below in Tables 8-12.

Participant ID	ip[0.01]	Institute
P72 (Section Strategy)*	0.7236	UMD
P72 (Child Strategy)*	0.7230	UMD
P72 (Correlation Strategy)*	0.7225	UMD
P78	0.6897	University of Waterloo
P10	0.6799	Max-Planck-Institute
P48	0.6678	LIG
P92	0.6664	University of Lyon
P9	0.6648	University of Helsinki
P60	0.6640	Saint Etienne University
P14	0.6427	University of California
P29	0.6365	Indian Statistical Inst.
P25	0.6346	Renmin Univ. of China
P5	0.6344	Queensland Univ. of Tech.

*** UMD results compared to official runs.**

Table 8: Top 10-ranked Participants in Focused Task – 2008

Participant ID	ip[0.01]	Institute
P72 (Section Strategy)*	0.6594	UMD
P72 (Correlation Strategy)*	0.6488	UMD
P72 (Child Strategy)*	0.6482	UMD
P78	0.6333	University of Waterloo
P68	0.6141	Univ. Pierre et Marie Curie
P10	0.6134	Max-Planck-Institute
P60	0.6060	Saint Etienne University
P6	0.5997	Univ. of Amsterdam
P5	0.5592	Queensland Univ. of Tech.
P16	0.5903	Univ. of Applied Science
P48	0.5853	LIG
P22	0.5844	ENSM - SE
P25	0.4973	Renmin Univ. of China

*** UMD results compared to official runs.**

Table 9: Top 10-ranked Participants in Focused Task - 2009

Participant ID	MAGP	Institute
P78	0.2278	University of Waterloo
P92	0.2106	Univ. of Lyon3
P5	0.2106	Queensland Univ. of Tech.
P10	0.1947	Max-Plank-Institute
P4	0.1929	Univ. of Otago
P72 (Correlation Strategy)*	0.1765	UMD
P72 (Child Strategy)*	0.1761	UMD
P36	0.1758	Univ. of Amsterdam
P72 (Section Strategy)*	0.1743	UMD
P72	0.1724	UMD
P12	0.1582	Univ. of Granada
P56	0.1500	Just systems Corp.
P48	0.1497	LIG

* UMD results compared to official runs.

Table 10: Top 10-ranked Participants in Relevant in Context Task - 2008

Participant ID	MAGP	Institute
P5	0.1885	Queensland Univ. of Tech.
P4	0.1847	Univ. of Otago
P6	0.1773	Univ. of Amsterdam
P48	0.1760	LIG
P72 (Correlation Strategy)*	0.1731	UMD
P36	0.1720	Univ. of Tampere
P72 (Child Strategy)*	0.1689	UMD
P72 (Section Strategy)*	0.1636	UMD
P346	0.1188	Univ. of Twente
P60	0.1075	Saint Etienne University
P167	0.1045	School of Ele. Engg. & CS
P25	0.1028	Renmin Univ. of China
P72	0.0424	UMD

*** UMD results compared to official runs.**

Table 11: Top 10-ranked Participants in Relevant in Context Task - 2009

Participant ID	MAiP	Institute
P48	0.2855	LIG
P6	0.2818	Univ. of Amsterdam
P5	0.2585	Queensland Univ. of Tech.
P92	0.2496	Univ. of Lyon
P60	0.2435	Saint Etienne University
P346	0.2350	Univ. of Twente
P10	0.2133	Max-Planck-Institute
P72 (Thorough)*	0.2120	UMD
P72 (All Element)*	0.1920	UMD
P167	0.1390	School of Ele. Engg. & CS
P68	0.0630	Univ. of Pierre
P25	0.0577	Renmin Univ. of China

*** UMD results compared to official runs.**

Table 12: Top 10-ranked Participants in Thorough Task - 2009

The significance tests performed on the results given in the table 8-12 are displayed in Appendix A (Tables A1-A17).

4.4 Observations

The results of the significance testing are presented in this section of the paper. Dash (-) indicates no significant difference between methods, whereas asterisk (*) indicates that a statistically significant difference is present. Significant testing for the INEX Ad Hoc Tasks is given below in Figures 8-15.

Participant #	1	2	3	4	5	6	7	8	9	10
UMD Section Strategy	-	-	-	-	-	-	-	-	-	-
UMD Child Strategy	-	-	-	-	-	-	-	-	-	-
UMD Correlation Strategy	-	-	-	-	-	-	-	-	-	-

Figure 8: Significant Testing for the Focused Task-2008.

Participant #	1	2	3	4	5	6	7	8	9	10
UMD Section Strategy	-	-	-	-	-	-	-	-	-	-
UMD Child Strategy	-	-	-	-	-	-	-	-	-	-
UMD Correlation Strategy	-	-	-	-	-	-	-	-	-	-

Figure 9: Significant Testing for the Focused Task-2009.

Participant ID	1	2	3	4	5	6	7	8	9	10
UMD Section Strategy	-	-	-	-	-	-	*	*	*	*
UMD Child Strategy	-	-	-	-	-	-	*	*	*	*
UMD Correlation Strategy	-	-	-	-	-	-	*	*	*	*

Figure 10: Significant Testing for the Relevant in Context Task-2008.

Figure 11 compares the methods of higher-ranking participants with our method to see if those methods produce statistically better results than our method.

Participant Rank	P72 (Correlation Strategy)
1	*
2	*
3	*
4	-
5	-

Figure 11: Significant Testing –II for the Relevant in Context Task-2008.

Participant ID	1	2	3	4	5	6	7	8	9	10
UMD Section Strategy	-	-	-	-	-	*	*	*	*	*
UMD Child Strategy	-	-	-	-	-	*	*	*	*	*
UMD Correlation Strategy	-	-	-	-	-	*	*	*	*	*

Figure 12: Significant Testing for the Relevant in Context Task-2009.

Figure 13 compares the methods of higher-ranking participants with our method to see if those methods produce statistically better results than our method.

Participant Rank	P72 (Correlation Strategy)
1	*
2	-
3	-
4	-

Figure 13: Significant Testing –II for the Relevant in Context Task-2009.

Participant ID	1	2	3	4	5	6	7	8	9	10
UMD All Element	-	-	-	-	-	-	-	-	*	*
UMD Thorough	-	-	-	-	-	-	-	-	*	*

Figure 14: Significant Testing for the Thorough Task-2009.

Figure 15 compares the methods of higher-ranking participants with our method to see if those methods produce statistically better results than our method.

Participant Rank	P72
1	*
2	-
3	-
4	-
5	-
6	-
7	-

Figure 15: Significant Testing –II for the Thorough Task-2009.

5. Conclusion

For the Focused Tasks in 2008-09, our scores are comparable and better than all top-ranked INEX Ad Hoc participant scores (baseline), but the results produced by our method are not statistically significant compared to the results of the top-ranked participants. For the Relevant in Context Task in 2008-09, our results are better than some of the top ranked participants. For the Relevant in Context Task, our method produces statistically significant results with respect to those produced by participants with rankings 6-10 in 2009 and 7-10 in 2008.

For the Thorough Task in 2009, our results are better than those of participants ranked 8-10, and we produce significantly better results than participants in ranks 9-10. For the 2008-09 Focused Tasks, there is no statistically significant difference between top-ranked participant results and our results. For the 2008 Focused Task, participants ranked 1-3 (University of Waterloo, University of Lyon³, Queensland University of Technology) produced statistically better results when compared with our results; the results of participants ranked 4-6 were comparable with our results. For the 2009 Focused Task, the first ranked participant (Queensland University of Technology) produced statistically better results than ours and the results of other participants (2-5 ranked) were comparable with our results. For the 2009 Thorough Task, only the participant ranked 1 (LIG) produced statistically better results than ours and other participant (2-7 ranked) results were comparable with our results.

The only participant group which produced significantly better results for the 2009 Relevant in Context Task, Queensland University of Technology, used a GPX run

using a `/**[about(.,keywords)]` query, serving non-overlapping elements grouped per article, with the articles ordered by their best scoring element [6]. The only participant group, which produced significantly better results for the 2009 Thorough Task, LIG Grenoble, used a language Model using a Dirichlet smoothing, and equally weighting element score and its context score, where the context score are based on the collection-links in Wikipedia [6].

References:

- [1] Bhirud, D., Focused Retrieval Using Upper Bound Methodology, MS Thesis, Department of Computer Science, University of Minnesota Duluth (2009). <http://www.d.umn.edu/cs/thesis/Bhirud.pdf>
- [2] Crouch C., et al., A Methodology for Producing Improved Focused Elements, *Proceedings of INEX 2009*, LNCS 6203, 70-80 (2010). (to appear).
- [3] Geva, S., Kamps, J., Trotman, A., INEX 2008 Workshop Pre-proceedings. <http://www.inex.otago.ac.nz/tracks/adhoc/gtd.asp>
- [4] Hull, D., Using Statistical Testing in the Evaluation of Retrieval Experiments, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 329-338, (1993).
- [5] Kamps, J., et al., INEX 2007 Evaluation Measures *INEX 2007*, LNCS 4862, pp. 24–33 (2008).
- [6] Kamps, J., Geva, S., Trotman, A. et al., Overview of the INEX 2008 Ad Hoc Track, *INEX 2008*, LNCS 5631, 1–28 (2009).
- [7] Mahule, A. Improving Results for the INEX Thorough Tasks, MS Thesis, Department of Computer Science, University of Minnesota Duluth (2010). <http://www.d.umn.edu/cs/thesis/Mahule.pdf>
- [8] Manning, C., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press (2008).
- [9] Poluri, P., Focused Retrieval Using Exact Methodology, MS Thesis, Department of Computer Science, University of Minnesota Duluth (2009). <http://www.d.umn.edu/cs/thesis/Poluri.pdf>
- [10] Salton, G., (ed.) *The Smart Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs (1971)
- [11] Salton, G., Wong, A., Yang, C.S., A vector space model for automatic indexing. *Comm. ACM* 18(11), 613–620 (1975)
- [12] Singhal, A. AT&T at TREC-6, *The Sixth Text REtrieval Conf (TREC-6)*, 215 – 225 (1998).
- [13] Singhal, A., Buckley, C., Mitra, M., Pivoted Document Length Normalization. *Proc. of the 19th Annual International ACM SIGIR Conference*, 19-21 (1996).

- [14] Vadlamudi, S., Producing Improved Focused Results for INEX Focused and Relevant in Context Tasks, MS Thesis, Department of Computer Science, University of Minnesota Duluth (2010). <http://www.d.umn.edu/cs/thesis/Vadlamudi.pdf>

Appendix A:

Focused task Significance test results for 2008 using Section strategy in overlap

removal: $iP[0.01]= .7236$

Participant ID	t-value
P78	0.697
P10	0.732
P48	0.773
P92	0.777
P9	0.820
P60	0.857
P14	0.899
P29	0.939
P25	0.996
P5	1.225

Table A1: Focused Task Significance Test Results for 2008 using Section Strategy.

Focused task Significance test results for 2008 using Child strategy in overlap

removal: $iP[0.01]= .7230$

Participant ID	t-value
P78	0.547
P10	0.639
P48	0.773
P92	0.781
P9	0.828
P60	0.857
P14	0.899
P29	0.910
P25	0.964
P5	1.067

Table A2: Focused Task Significance Test Results for 2008 using Child Strategy.

Focused task Significance test results for 2008 using Correlation strategy in overlap

removal: $iP[0.01]= .7225$

Participant ID	t-value
P78	0.526
P10	0.568
P48	0.617
P92	0.704
P9	0.721
P60	0.833
P14	0.867
P29	0.899
P25	0.906
P5	1.037

Table A3: Focused Task Significance Test Results for 2008 using Correlation Strategy.

Focused task Significance test results for 2009 using Section strategy in overlap

removal: $iP[0.01]= 0.6594$

Participant ID	t-value
P78	0.122
P68	0.171
P10	0.348
P60	0.598
P6	0.857
P5	0.996
P16	1.253
P48	1.274
P22	1.314
P25	1.381

Table A4: Focused Task Significance Test Results for 2009 using Section Strategy.

Focused task Significance test results for 2009 using Correlation strategy in overlap

removal: $iP[0.01]= 0.6488$

Participant ID	t-value
P78	0.108
P68	0.160
P10	0.291
P60	0.516
P6	0.573
P5	0.778
P16	0.935
P48	1.212
P22	1.278
P25	1.336

Table A5: Focused Task Significance Test Results for 2009 using Correlation Strategy.

Focused task Significance test results for 2009 using Child strategy in overlap

removal: $iP[0.01]= 0.6482$

Participant ID	t-value
P78	0.108
P68	0.153
P10	0.278
P60	0.487
P6	0.547
P5	0.732
P16	0.899
P48	1.160
P22	1.255
P25	1.325

Table A6: Focused Task Significance Test Results for 2009 using Child Strategy.

Relevant in Context task Significance test results for 2008 using Section strategy in overlap removal: MAgP= 0.1743

Participant ID	t-value
P78	0.381
P92	0.550
P5	0.669
P10	0.696
P4	0.932
P36	1.377
P72	1.820
P12	1.828
P56	1.883
P48	1.857

Table A7: Relevant in Context Task Significance Test Results for 2008 using Section Strategy.

Relevant in Context task Significance test results for 2008 using Child strategy in overlap removal: MAgP= 0.1761

Participant ID	t-value
P78	0.291
P92	0.406
P5	0.668
P10	0.704
P4	0.732
P36	1.456
P72	1.777
P12	1.778
P56	1.791
P48	1.810

Table A8: Relevant in Context Task Significance Test Results for 2008 using Child Strategy.

Relevant in Context task Significance test results for 2008 using Correlation strategy in overlap removal: MAgP= 0.1765

Participant ID	t-value
P78	0.244
P92	0.542
P5	0.668
P10	0.696
P4	0.732
P36	1.556
P72	1.762
P12	1.781
P56	1.792
P48	1.799

Table A9: Relevant in Context Task Significance Test Results for 2008 using Correlation Strategy.

Relevant in Context task Significance test results for 2009 using Correlation strategy in overlap removal: MAgP= 0.1731

Participant ID	t-value
P5	0.381
P4	0.579
P6	0.906
P48	1.262
P36	1.644
P346	1.704
P60	1.773
P167	1.781
P25	1.820
P72	1.828

Table A10: Relevant in Context Task Significance Test Results for 2009 using Correlation Strategy.

Relevant in Context task Significance test results for 2009 using Child strategy in overlap removal: MAgP= 0.1689

Participant ID	t-value
P5	0.348
P4	0.558
P6	0.878
P48	1.240
P36	1.623
P346	1.696
P60	1.756
P167	1.777
P25	1.781
P72	1.790

Table A11: Relevant in Context Task Significance Test Results for 2009 using Child Strategy.

Relevant in Context task Significance test results for 2009 using Section strategy in overlap removal: MAgP= 0.1636

Participant ID	t-value
P5	0.336
P4	0.548
P6	0.867
P48	1.236
P36	1.579
P346	1.668
P60	1.732
P167	1.756
P25	1.777
P72	1.778

Table A12: Relevant in Context Task Significance Test Results for 2009 using Section Strategy.

Thorough task Significance test results for 2009 using All Element: MAiP= 0.192

Participant ID	t-value
P48	0.024
P6	0.050
P5	0.067
P92	0.122
P60	0.269
P346	0.385
P10	0.612
P167	1.293
P68	1.704
P25	1.781

Table A13: Thorough Task Significance Test Results for 2009 using All Element.

Thorough task Significance test results for 2009: MAiP= 0.212

Participant ID	t-value
P48	0.037
P6	0.057
P5	0.108
P92	0.188
P60	0.291
P346	0.406
P10	0.657
P167	1.325
P68	1.732
P25	1.820

Table A14: Thorough Task Significance Test Results for 2009.

Relevant in Context Task Significance Test – II results for 2008: MAgP= 0.1765

Participants ID	t-value
P78	1.7219
P92	1.6918
P5	1.6703
P10	1.326
P4	1.101

Table A15: Relevant in Context Task Significance Test - II Results for 2008.

Relevant in Context Task Significance Test – II results for 2009: MAgP= 0.1731

Participants ID	t-value
P5	1.6762
P4	1.2591
P6	0.7216
P48	0.3927

Table A16: Relevant in Context Task Significance Test - II Results for 2009.

Thorough Task Significance Test – II results for 2009: MAiP= 0.2120

Participants ID	t-value
P48	1.6890
P6	1.5638
P5	1.0274
P92	0.8059
P60	0.4927
P346	0.3173
P10	0.1952

Table A17: Thorough Task Significance Test - II Results for 2009.