# From Focused Elements to Snippets

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Supraja Nagalla

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

Dr. Carolyn J. Crouch

April, 2013

Acknowledgements

This thesis is part of my two years of research at the University of Minnesota,Duluth. I have had the opportunity to meet some wonderful people who have inspired me from start to finish of my graduate work. I would like to thank them all for their help in making this thesis possible.

I owe my deepest gratitude to my advisor, Dr. Crouch, C. J. for guiding me through out my Masters and thesis. Her doors were always open whenever I needed advice and support. I will forever be grateful to her for encouraging me to attend the MinneWIC conference, from which I benefited greatly.

I would also like to thank my committee members Dr. Crouch, D. B. and Dr. Stech, H. for their suggestions and feedback during various milestones of my thesis.

It has been a pleasure working with my colleague, Sai. Many of the ideas in my work originated in discussions with him and Dr. Crouch. I am deeply grateful to them for investing their time and energy in discussing various ideas with me. I would also like to thank Deepak Acquilla, N., Banhatti, R., Mahule, B. A. and Narendravarapu, R. for imparting the knowledge necessary to start my thesis.

To Jim Luttinen, Lori Lucia and Clare Ford, thank you so much for assisting me with the administrative tasks necessary for completing my Master's program.

I would like to thank my mom and dad for their unflagging belief in me. I am indebted to them for encouraging me to pursue my dreams and for inculcating in me the dedication and discipline to do whatever I undertake well. I would also like to express my gratitude to my Uncle, Chandra, C. and family, our family friend, Sekhar.M and family for their constant support and care. Madhu and Vasu, thank you for all the fun in between work. Finally, I thank Siva, Amar and Gangesh for their support all through my Masters.

## Dedication

I dedicate this book to my grandmother who sacrificed everything for our family and without whom I wouldn't be where I am today. Thank you for telling me those bed time stories of brave women and for instilling in me the desire to be who I am . Thank you for staying with me and telling me to be strong and brave when I had nightmares. Thank you, for all that u have done for us.

Abstract

Information Retrieval is a field of computing which traditionally deals with searching a large collection of documents and retrieving documents based on their similarity to the query. INEX [10] provides a platform (e.g., document collection, queries and uniform evaluation metrics) for the development and evaluation of retrieval algorithms for XML documents. The focus of INEX is to reduce the granularity of search results from the entire document to the element level.

In 2011, INEX introduced a new track, called the Snippet Retrieval Track. In 2012, INEX improved this track to make the task of assessment easier. Its goal is to determine how best to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document without viewing the document itself. The Snippet Retrieval track uses the 50.7GB INEX Wikipedia collection of about 2.7 million articles. We use the Smart [15] experimental retrieval system, based on the Vector Space Model [16], for indexing and retrieval.

This thesis describes the approaches taken by UMD to generate runs to participate in the INEX 2011 and 2012 Snippet Retrieval track. We use our method of dynamic element retrieval [7] to generate the element vectors of the XML document tree at run time, thus producing a rank-ordered list of elements from each highly correlated document. These elements are further processed using our methods to generate snippets. The methods used, experimental results, and conclusions are described herein.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

Information Retrieval (IR) is the area of study concerned with searching for documents, for information within documents, and for meta-data about documents. Sources include structured storage, relational databases, and the World Wide Web. An information retrieval process begins when a user enters a query into the system. Queries are statements of information need. Queries are matched against a group of objects within a database, an object being an entity consisting of information. The objects of concern here are text documents, represented on the web as XML documents. Most IR systems compute a numerical score to determine how well the object correlates with the query and to rank the objects according to this value. The top-ranked objects are then displayed to the user. Salton's Vector Space Model (VSM) [16] is perhaps the best known model for information retrieval. The VSM allows humans to form search requests in natural language, with the objective of producing a list of documents ranked in the order of their correlation with the query. The Vector Space Model produces ranked output; it does not require document structure. The retrieval units are usually entire documents. Smart [15] is an experimental retrieval system that is based on the Vector Space Model. With the growth of the World Wide Web, large amounts of information are represented as XML (Extensible Markup Language) documents. Hence the focus has changed to XML retrieval. XML retrieval focuses on the content-based retrieval of information structured in XML. Whereas in traditional IR systems queries are based only on content (i.e., query terms), in XML retrieval queries may also include structural hints. The focus of XML retrieval is to reduce the granularity of search results from the entire document to the element level. This thesis describes the work done in our research group at the University of Minnesota Duluth (UMD). Our research group is one of the participants in the annual international INEX (Initiative for the Evaluation of XML Retrieval) competition [10]. INEX focuses on the development and evaluation of XML-based retrieval systems; it provides its participants with an XML document collection, a set of user queries, relevance assessments, and tools and metrics to evaluate the retrieved results. Our research group at UMD has participated in INEX since its inception (2001). In 2011, INEX introduced a new track, called Snippet Retrieval [18]. Its goal is to determine how best to generate informative snippets as search results. A snippet is defined as an element that provides sufficient information to allow the user to determine the relevance of each document to the query without needing to view the document itself[[18]. The Snippet Retrieval track uses the 50.7GB 2009 INEX Wikipedia collection of about 2.7 million articles. We use the Smart experimental retrieval system [15], based on the Vector Space Model [16], for indexing and retrieval. This thesis describes various algorithmic approaches for snippet retrieval. Some of these algorithms use our method of dynamic element retrieval [7] to generate the element vectors of the XML document tree at run time, thus producing a rank-ordered list of elements from each top-ranked document. Details of the algorithms, experimental results, and conclusions are found in subsequent chapters.

Much of this research was performed within a team environment. The team consisted of the author and and a fellow graduate student, S. Chittilla, whose work is reported in [6].

# 2 Background

This chapter presents the background upon which our experiments are based. It describes Salton's VSM, the Smart retrieval system, and our method of dynamic element retrieval (which is called Flex). It also gives an overview of INEX, the INEX Snippet Retrieval tasks of 2011 and 2012, and the tools provided by INEX used in conducting this research.

## 2.1 INEX

INEX [10] is a competition started in 2001 to investigate (facilitate) XML retrieval. It's main goal is to facilitate experimentation, retrieval and evaluation of XML documents by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results. The focus of INEX is to reduce the granularity of search results from the entire document to the element level.

### The INEX 2011 Snippet Retrieval Track

The goal of the Snippet Retrieval track [18] is to return, for each query, a ranked list of documents and for each document, a corresponding text snippet describing or representing that document. This snippet should attempt to convey the relevance of the underlying document, without the user's needing to view the document itself.

Each run is allowed to return up to 500 documents per topic, with a maximum of 300 characters per snippet.

### The INEX 2011/2012 Document Collection

The Snippet Retrieval Track uses the INEX Wikipedia collection [1] introduced in 2009 (an XML version of the English Wikipedia, based on a dump taken on 8 October 2008). This corpus contains 2,666,190 documents and more than 30,000 unique tags. This collection of XML documents does not strictly follow a DTD (Document Type Definition). The documents are semi-structured; i.e., untagged text is present in them. A sample Wikipedia document is shown in Fig. 2.1.

### The 2011 INEX Queries

The 2011 queries [2] are topics that have been reused from the INEX 2009 Ad Hoc Track. Each topic contains a short, content only (CO) query, a content and structure (CAS) query, a phrase title, a one line description of the search request, and a narrative with a detailed explanation of the information need, the context and motivation of the information need, and a description of what makes a document relevant or not [18]. The 2009 topics are ranked in order of the number of relevant documents found in the corresponding relevance judgments, and the 50 queries with the lowest number were selected for the 2011 query set. This method ensures that queries with fewer rather than many relevant documents make up the query set. Part of a sample topic file is shown in Fig. 2.2.

```
-<!--
   generated by CLiX/Wiki2XML [MPI-Inf, MMCI@UdS] $LastChangedRevision: 92 $ on 16.04.2009 17:40:30[m
 -->
-<article>
 -<laureate confidence="0.9511911446218017" wordnetid="110249011">
  -<alumnus confidence="0.9511911446218017" wordnetid="109786338">
   -<scientist confidence="0.9508927676800064" wordnetid="110560637">
    -<physiologist confidence="0.9511911446218017" wordnetid="110429965">
     -<header>
        <title>Herbert Spencer Gasser</title>
        <id>782023</id>
      -<revision>
         <id>242006978</id>
         <timestamp>2008-09-30T13:36:00Z</timestamp>
       -<contributor>
          <username>Thijs!bot</username>
          <id>1392310</id>
         </contributor>
        </revision>
      +<categories></categories>
      </header>
     -<bdy>
      +<template></template>
      +<p></p>
      -<sec>
         <st> References</st>
       +<p></p>
       +<p></p>
        </sec>
      </bdy>
     </physiologist>
    </scientist>
   </alumnus>
  </laureate>
 </article>
```

Fig. 2.1: Wikipedia XML Document

## The 2011 INEX Reference Run

For those participants who only wish to generate snippets and not use their own search engine, a reference run was generated using BM25 [3]. The reference run consists of the 500 top-ranked documents for each of the 50 queries. A part of the reference run is shown in Fig. 2.3.

```xml
<inex-topic-file>
<topic id="2011011" ct_no="11">
<title>Nobel prize</title>
<castitle>//article[about(., Nobel prize)]</castitle>
<phrasetitle>"Nobel prize"</phrasetitle>
<description>information about Nobel prize</description>
<narrative>
I need to prepare a presentation about the Nobel prize. Therefore, I want to collect
information about it as much as possible. Information, the history of the Nobel prize or the
stories of the award-winners for example, is in demand.
</narrative>
</topic>
<topic id="2011012" ct_no="12">
<title>best movie</title>
<castitle>//article[about(., best movie)]</castitle>
<phrasetitle>best movie</phrasetitle>
<description>information of classical movies</description>
<narrative>
I spend most of my free time seeing movies. Recently, I want to retrospect some classical
movies. Therefore, I need information about the awarded movies or movies with good
reputation. Any information, such as the description or comments of the awarded movies
on famous filmfests or movies with good fame, is in demand.
</narrative>
</topic>
```

Fig. 2.2: 2011 Sample Topic File

## INEX 2011 Assessment

Manual assessment is used to determine whether a document is relevant to a query. For each topic, each document was assessed for relevance based on the snippet alone, as the goal is to determine the snippet's ability to represent the document.

## INEX 2011 Evaluation Measures

Submissions are evaluated by comparing the snippet-based relevance judgments with the existing document-based relevance judgments, the latter of which are treated as ground truth. This section gives a brief summary of the specific metrics used. In all cases, the metrics are averaged over all topics. All these metrics and related

**Reference Run(given by INEX)**
**(query-1 → "Nobel Prize")**

| query_no | Doc_no | Doc_Score |
|---|---|---|
| query_1 | 13218231 | 1411.00 |
| query_1 | 900383 | 1410.00 |
| query_1 | 236582 | 1410.00 |
| query_1 | 42909 | 1410.00 |
| query_1 | 92418 | 1410.00 |
| query_1 | 24509 | 1410.00 |
| query_1 | 13032 | 1409.00 |
| query_1 | 3712 | 1409.00 |
| query_1 | 1312121 | 1409.00 |
| query_1 | 11462560 | 1409.00 |
| query_1 | 323750 | 1409.00 |
| query_1 | 1322983 | 1408.00 |
| query_1 | 708064 | 1408.00 |
| query_1 | 630821 | 1408.00 |
| query_1 | 34675 | 1408.00 |
| query_1 | 782023 | 1408.00 |

Fig. 2.3: 2011 Sample Reference run

information can be found in [18]. The focus is to determine how effective the snippets are at providing the user with sufficient information to determine the relevance of the underlying document. The simplest metric is the *Mean Precision Accuracy* (MPA), defined as the percentage of results that the assessor correctly assessed, averaged over all topics [18, p. 286].

$$MPA = \frac{TP + TN}{TP + FP + FN + TN}. \tag{2.1}$$

where $TP$ is the *number of true positives*, $TN$ is the *number of true negatives*, $FP$ is the *number of false positives* and $FN$ is the *number of false negatives*.

Since most topics have a much higher percentage of irrelevant than relevant documents, MPA weights relevant results much higher than irrelevant results. MPA can be considered the raw agreement between two assessors, one of whom assessed the actual documents (i.e., the ground truth relevance judgments) and one who assessed the snippets. Because the relative size of the two groups (relevant and irrelevant documents) can skew this result, it is also useful to look at positive agreement and negative agreement to see the effects of these two groups. *Positive agreement* (PA) is the conditional probability that, given one of the assessors judges a document as relevant, the other will also do so [18, p. 286].

$$PA = \frac{2TP}{2TP + FP + FN}. \tag{2.2}$$

Likewise, *Negative Agreement* (NA) is the conditional probability that, given one of the assessors judges a document as irrelevant, the other will also do so [18, p. 286].

$$NA = \frac{2TN}{2TN + FP + FN}. \tag{2.3}$$

*Mean Normalized Prediction Accuracy* (MNPA) calculates the rates for relevant and irrelevant documents separately and averages the results (to avoid relevant results being weighted higher than irrelevant results) [18, p. 286].

$$MNPA = 0.5\frac{TP}{TP + FN} + 0.5\frac{TN}{TN + FP}. \tag{2.4}$$

MNPA can also be thought of as the arithmetic mean of recall and negative recall. These two metrics are interesting in themselves and so are also reported separately. *Recall* is the percentage of relevant documents that are correctly assessed [18, p. 286].

$$Recall = \frac{TP}{TP + FN}. \tag{2.5}$$

*Negative recall* (NR) is the percentage of irrelevant documents that are correctly assessed [18, p. 286].

$$NR = \frac{TN}{TN + FP} \tag{2.6}$$

The primary evaluation metric, which is used to rank the submissions, is the *Geometric Mean of recall and negative recall* (GM) [18, p. 287]. A high value of GM requires a high value in both recall and negative recall; i.e., the snippets must help the user to accurately predict both relevant and irrelevant documents. If a submission has high recall but zero negative recall (in which case everything is judged relevant), GM will be zero. Likewise, if a submission has high negative recall but zero recall (e.g., in the case that everything is judged irrelevant), GM is zero.

$$GM = \sqrt{\frac{TP}{TP + FN} \frac{TN}{TN + FP}}. \tag{2.7}$$

## 2.2 Vector Space Model

The Vector Space Model (VSM) [16] is a model wherein both documents and queries are represented as vectors. E.g., both document D and query Q, seen below, are composed of sets of weighted terms.

$$d_j = (w_{1,j} \quad w_{2,j} \quad \ldots \quad w_{t,j}). \tag{2.8}$$

$$q = (w_{1,q} \quad w_{2,q} \quad \ldots \quad w_{t,q}). \tag{2.9}$$

Many different methods of computing these values, also known as term weights, have been developed; all are based on functions of term frequency (the number of times a term occurs in the document/element) and document frequency (the number of documents/elements the term occurs in). Similarity between vectors is computed based on a similarity measure such as cosine or inner product.

## 2.3 Smart

The Smart Retrieval System [15] is an IR system conceived at Cornell University in the 1960s and developed continually for the next 25 years under Salton and his students. Many important concepts in information retrieval were developed using Smart. We use Smart in our research for all the basic functions of retrieval. This engine is based on the Vector Space Model [16]. A Smart retrieval produces a ranked list of articles or elements for each query.

## 2.4 Flex

For element retrieval, we use our method of dynamic element or flexible retrieval called Flex [7, 8]. Given the structure of the document in the form of a parse tree and all the terminal nodes of the tree in vector form, Flex generates the correlation between child nodes and the query first. In the next step, it uses the content of each child to build the parent node. The elements are generated dynamically and correlated with the query. It is time and memory efficient. We use *Lnu-Ltu* term weighting [17] in this process. Complete descriptions of Flex and it's methodology are found in [7].

# 3 2011 Snippet Retrieval Track

## 3.1 Focused Task

The INEX 2009/2010 Focused Task aims to return a ranked list of non-overlapping (or focused) elements for each document for each topic. Since ancestor elements almost always contain more information (compared to their children), it is a challenge to choose the correct level of granularity. Ranking is based on the correlation score of the element with the query. This discussion about the Focused Task is important because we realize that the best focused element may serve as the basis for generating the snippet that best represents the document. Based on this view, we conduct our experiments by applying different snippet generation strategies to focused elements and documents. This chapter gives a brief overview of the Focused Task and the snippet generation methods we adopted in 2011. Chapter 4 describes our 2012 snippet experiments, results and observations.

## 3.2 Methodology

Two important issues arise with respect to retrieving good focused elements in response to a query: (1) how the documents of interest are identified, and (2) the method by which the focused elements are selected from those documents. Fig. 3.1 shows a sample query.

```
<topic id="2011011" ct_no="11">
<title>Nobel prize</title>
<castitle>//article[about(., Nobel prize)]</castitle>
<phrasetitle>"Nobel prize"</phrasetitle>
<description>information about Nobel prize</description>
<narrative>
I need to prepare a presentation about the Nobel prize. Therefore, I want
to collect information about it as much as possible. Information, the
history of the Nobel prize or the stories of the award-winners for
example, is in demand.
</narrative>
</topic>
```

Fig. 3.1: Sample Query 2011011

For the Snippet Retrieval track, we use the reference run provided by INEX instead of our own article retrieval. The reference run consists of 50 queries and 500 documents retrieved for each query (See Chapter 2 for details). We then use dynamic element retrieval (i.e., Flex) to produce the elements themselves. Fig. 3.2 gives a sample Flex output for query 1 and document 782023. Dynamic element retrieval allows us to build the document tree at execution time, based on a stored schema of the document and a terminal-node (para+mt) index of the collection [7, 8]. *Lnu-ltu* term weighting [17], designed to deal with differences in the lengths of vectors, is utilized to produce a rank-ordered list of elements from each document for a given query. One of two focusing strategies [14] (described below) is then used to remove overlap. A set of focused elements is then reported for each document for a given query.

Three focusing or overlap removal strategies [14] were investigated in earlier experiments. We reduced this set to two. The *correlation strategy* [14] chooses the highest correlating element along a path as the focused element, without restriction on ele-

**Flex output**

| Query | Doc_id/Xpath | Similarity_Score |
|---|---|---|
| 1 | 782023/article[1]/bdy[1]/ | 22.7654 |
| 1 | 782023/article[1]/bdy[1]/sec[1]/ | 18.8078 |
| 1 | 782023/article[1]/bdy[1]/sec[1]/p[1]/ | 17.7904 |
| 1 | 782023/article[1]/bdy[1]/p[1]/ | 13.038 |
| 1 | 782023/article[1]/bdy[1]/template[1]/ | 12.9778 |
| 1 | 782023/article[1]/header[1]/categories[1]/ | 7.12426 |
| 1 | 782023/article[1]/header[1]/ | 6.5986 |
| 1 | 782023/article[1]/bdy[1]/sec[1]/p[2]/ | 5.82003 |

Fig. 3.2: Sample Flex Output

ment type. The *child strategy* [14] always chooses the terminal element along a path as the focused element. Fig. 3.3 and Fig. 3.4 show the output of the child and correlation strategies respectively. The 2011 snippet generation experiments were then conducted on these elements to obtain the "best" snippet to represent the document.

**Result of applying Child strategy on Flex output**

| Query | Doc_id/Xpath |
|---|---|
| 1 | 782023/article[1]/bdy[1]/sec[1]/p[1]/ |
| 1 | 782023/article[1]/bdy[1]/p[1]/ |
| 1 | 782023/article[1]/bdy[1]/template[1]/ |
| 1 | 782023/article[1]/header[1]/categories[1]/ |
| 1 | 782023/article[1]/bdy[1]/sec[1]/p[2]/ |

Fig. 3.3: Child Strategy as Applied to Flex Output in Fig. 3.2

**Result of applying Correlation strategy to flex output**

| Query | Doc_id/Xpath |
|---|---|
| 1 | 782023/article[1]/bdy[1]/ |
| 1 | 782023/article[1]/header[1]/categories[1]/ |

Fig. 3.4: Correlation Strategy as Applied to Flex Output in Fig. 3.2

## 3.3 Snippet Generation

By definition, a snippet is designed to provide the user with information sufficient to determine the relevance of a document without viewing the document itself, thus quickly allowing the user to identify what s/he is looking for [18]. By definition of the 2011 INEX Snippet Retrieval task [18], a snippet must be less than 300 characters in length. Initial experiments were based on our intuition that the best focused element may also be the basis for generating the snippets. Dynamic element retrieval produces for each query, a rank-ordered list of elements for each document. Hence the highest ranked focused element in a document forms the initial basis for snippet generation. Given a query, dynamic element retrieval produces an overlapping list of elements for each document. To remove overlap, a focusing strategy is applied to the list. We then extract the highest ranked focused element from each document. This forms the basis for the snippet of that document and is called the raw snippet. The raw snippet must be further processed to generate the snippet. The following sections discuss these methods.

### 3.3.1 Method 1 (Extracting the first 300 characters)

A snippet representing a document must be less than 300 characters in length. The raw snippet can be of any length. In this experiment, we extract the first 300 characters of the raw snippet to form the snippet. This method is applied to elements generated by both the child strategy and the correlation strategy [14]. The results are discussed below.

#### Generating the Snippet Based on the Child Strategy

Dynamic element retrieval using Flex produces an overlapping list of elements. To remove overlap, the child strategy is applied to this list. The result is a non-overlapping list of elements. The highest correlating element is then selected from this list and this is the raw snippet. The first 300 characters of the element are then extracted and returned. These 300 characters form the snippet. Fig. 3.5 shows a sample snippet generated by this method. (The Snippets generated by Method 1 using the child strategy were submitted to INEX as p65-UMD_SNIPPET_RETRIEVAL_RUN_2.)

<snippet rsv="1564.00" doc-id="21201">
"Nobel Prize" (2007), in Encyclopædia Britannica, accessed 14 November 2007, from Encyclopædia Britannica Online: . An additional award, the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel, was established in 1968 by the Bank of Sweden and was first awarded in 1969. not technically a Nobel Prize, it is identified with the award its
</snippet>

**xpath:** /article[1]/bdy[1]/sec[8]/p[1]/

Fig. 3.5: Method 1: Snippet generated by Child Focusing Strategy for query 2011011 (Fig. 3.1)

### Generating the Snippet Based on the Correlation Strategy

Dynamic element retrieval using Flex gives an overlapping list of elements. To remove overlap, the correlation focusing strategy is applied to this list. The result is a non-overlapping list of elements. The highest correlating element is then selected from this list and this is the raw snippet. The first 300 characters of the raw snippet form the snippet. Fig. 3.5 shows a sample snippet generated by this method. (The snippets generated by Method 1 using the correlation strategy are submitted to INEX as p65-UMD_SNIPPET_RETRIEVAL_RUN_4.)

```
<snippet rsv="1564.00" doc-id="21201">
Infobox award Outstanding contributions in Physics , Chemistry , Literature , Peace, and
Physiology or Medicine . The Sveriges Riksbank Prize in Economic Sciences in Memory of
Alfred Nobel, commonly identified with the Nobel Prize, is awarded for outstanding
contributions in Economics. The Nobel Prize 1901 Swedish Academy Royal Swedish
Academy of Sciences
</snippet>
xpath: /article[1]/bdy[1]/
```

Fig. 3.6: Method 2: Snippet Generated by Correlation Focusing Strategy for query 2011011 (Fig. 3.1)

## 3.3.2 Method 2 (Ranking the Sentences by Relevance

In this experiment, in an attempt to improve the readability of the snippet, we employed a new approach to process the raw snippet. In this method, the raw snippet is split into sentences. These sentences are then ranked based on a scoring method. The score is the ratio of the total number of query terms present in the sentence to the length of the sentence. (Shorter sentences of length less than 6 words are discarded). The sentences are then sorted in the descending order and then merged together in that order to form a single paragraph. The first 300 characters of the paragraph form the snippet.

### Generating the Snippet Based on the Child Strategy

The highest correlating element produced by applying the child focusing strategy to Flex output is the raw snippet. The text of the raw snippet is then split into sentences (Sentences of length less than 6 words are discarded). The sentences are then scored based on the scoring method described above, sorted in the decreasing order, and merged together to form a paragraph. The first 300 characters of this paragraph are returned as the snippet. Fig. 3.7 gives a sample snippet generated by this method. (The snippets generated by Method 2 using child strategy are submitted to INEX as p65-UMD_SNIPPET_RETRIEVAL_RUN_1.)

```
<snippet rsv="1564.00" doc-id="21201">
Birgitta Lemmel, "The Nobel Prize Medals and the Medal for the Prize in Economics", nobelprize .
not technically a Nobel Prize, it is identified with the award its winners are announced with the
Nobel Prize recipients, and the Prize in Economic Sciences is presented at the Nobel Prize Award
Cerem
</snippet>

xpath: /article[1]/bdy[1]/sec[8]/p[1]/
```

Fig. 3.7: Method 2: Snippet Generated by the Child Focusing Strategy for query 2011011 (Fig. 3.1)

### Generating the Snippet Based on the Correlation Strategy

The highest correlating element produced by applying the correlation focusing strategy to Flex output is the raw snippet; it's content is extracted and processed. The text is split into sentences (Sentences of length less than 6 words are discarded). The sentences are scored based on the scoring method described above, sorted in the decreasing order, and joined to form a paragraph. The first 300 characters of this paragraph are returned as the snippet. Fig. 3.8 gives a sample snippet generated by this method. (The snippets generated by Method 2 using correlation strategy are submitted to INEX as p65-UMD_SNIPPET_RETRIEVAL_RUN_3.)

<snippet rsv="1564.00" doc-id="21201">
Niels Bohr won the Nobel prize in Physics in 1922, and his son Aage Bohr won the Nobel prize in Physics in 1975 . Manne Siegbahn won the Nobel prize in Physics in 1924 he was the father of Kai Siegbahn who shared the Nobel prize in Physics in 1981 . Raman won the Nobel prize in Physics in 1930 h
</snippet>

**xpath:** /article[1]/bdy[1]/

Fig. 3.8: Method 2: Snippet Generated by the Correlation Focusing Strategy for query 2011011 (Fig. 3.1)

## 3.4 2011 INEX Snippet Retrieval Track Results

The runs generated by the methods described in Section 3.3 were submitted to INEX in 2011. Table 3.1 displays the values of the metrics used for evaluation of the runs submitted. Results are ranked by GM. Table 3.2 shows the top ten results of the INEX 2011 Snippet Retrieval Track.

Table 3.1 indicates that the snippets generated by applying the correlation focusing strategy (runs 3 and 4) produced better GM results than those generated by applying child focusing strategy (runs 1 and 2). It can also be deduced that processing the raw snippet further using the scoring method (as in run 3) improved that result significantly. Table 3.2, which displays the ten top-ranked runs, shows that processing the raw snippet generated by the correlation focusing strategy(run 3) gave us a rank 9 among the top ten submitted runs. The corresponding run produced by applying the child(run 1) strategy ranked 30. The runs generated based on the first 300 characters of the raw snippet ranked 26(run 4) and 35(run 2) respectively. It can be observed

| Run | MPA | MNPA | Recall | NR | GM | PA | NA |
|---|---|---|---|---|---|---|---|
| p65-UMD _SNIPPET_ RETRIEVAL _RUN_3 | 0.7850 | 0.6207 | 0.3811 | 0.8602 | 0.5264 | 0.3498 | 0.8542 |
| p65-UMD _SNIPPET_ RETRIEVAL _RUN_4 | 0.7976 | 0.5982 | 0.3027 | 0.8937 | 0.4680 | 0.3078 | 0.8645 |
| p65-UMD _SNIPPET_ RETRIEVAL _RUN_1 | 0.7652 | 0.5877 | 0.3229 | 0.8525 | 0.4470 | 0.2849 | 0.8365 |
| p65-UMD _SNIPPET_ RETRIEVAL _RUN_2 | 0.7724 | 0.5813 | 0.2904 | 0.8723 | 0.4270 | 0.2736 | 0.8500 |

Table 3.1: 2011 INEX Snippet Retrieval Evaluation Measures of UMD's Runs

that (1) our negative recall values are comparable to runs that ranked higher than ours but that our low recall values contributed to lower GM values, and (2) the negative recall values of all the runs are comparable. Hence higher recall values can contribute to better GM values and there by a higher rank.

| Rank | Run | GM |
|---|---|---|
| 1 | p72-LDKE-1111 | 0.5705 |
| 2 | p23-baseline | 0.5505 |
| 3 | p72-LDKE-0101 | 0.5472 |
| 4 | p20-QUTFirst300 | 0.5416 |
| 5 | p73-PKU_ICST_REF_11a | 0.5341 |
| 6 | p72-LDKE-1110 | 0.5317 |
| 7 | p23-expanded-40 | 0.5294 |
| 8 | p72-LDKE-0111 | 0.5270 |
| 9 | p65-UMD_SNIPPET_RETRIEVAL_RUN_3 | 0.5264 |
| 10 | p20-QUTFocused | 0.5242 |

Table 3.2: INEX 2011 Snippet Retrieval Track Run Ranking

Significance tests [18, 14] were performed to determine whether higher ranked systems were signicantly better than lower ranked systems. A one-tailed t-test at 95 percent was used. Table 3.3 shows,for each submission in the ten top ranked runs, whether it is significantly better than each lower ranked run (indicated by "*"). As can be seen from Table 3.3, although the GM of the runs differ, there is no significant difference between them.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | - | - | - | - | - | - | - | - | - |
| 2 | | | - | - | - | - | - | - | - | - |
| 3 | | | | - | - | - | - | - | - | - |
| 4 | | | | | - | - | - | - | - | - |
| 5 | | | | | | - | - | - | - | - |
| 6 | | | | | | | - | - | - | - |
| 7 | | | | | | | | - | - | - |
| 8 | | | | | | | | | - | - |
| 9 | | | | | | | | | | - |
| 10 | | | | | | | | | | |

Table 3.3: INEX 2011 Snippet Retrieval Track Significance Results [18]

Our methods certainly succeeded in achieving the a top ranked run in the 2011 INEX Snippet Retrieval track, but there was much room for improvement. Recall indicates the percentage of relevant documents correctly assessed. Our recall values and NR values indicate that our scoring methods did a better job of recognizing irrelevant documents than they did in recognizing relevant documents. Our methods need improvement so that we can produce better snippets for relevant documents. One approach would be to develop new scoring methods to produce snippets that better represent documents. Another approach is to research the use entire documents as the source of the snippet.

# 4 2012 Snippet Retrieval Track

## 4.1 INEX 2012 Snippet Retrieval Track

The 2012 Snippet Retrieval Track is similar to the 2011 track. It has the same description and uses same evaluation measures as the 2011 Snippet Retrieval Track. The major goal of 2012 is to improve results compared to the previous year. Changes were made to make the manual assessment easier and better. To increase the accuracy of the assessments, full document-based assessment is now used in addition to snippet-based assessment, so that both the snippet and the full document are assessed by the same assessor. To keep the assessment load manageable, the number of topics and snippets has been reduced. The query set contains 35 topics (down from 50 in 2011), with 20 snippets per topic (rather than 100 in 2011), and the snippets are now limited to 180 characters (300 in 2011)[19]. As in 2011, a reference run is provided. As discussed in Section 3.4, our 2012 experiments were conducted with major emphasis on producing snippets that better represent documents. The experiments aimed at increasing the probability of relevant documents being marked relevant. For the 2012 Snippet Retrieval track, we decided to use the reference run provided by INEX. It consists of 35 queries with 20 documents retrieved for each query. We then use dynamic element retrieval (i.e., Flex) to produce the focused elements. The following sections describe the experiments conducted for the 2012 Snippet Retrieval Track. Our major goal for 2012 was to produce snippets that could better determine the relevance of the document. We use focused elements as the source of the snippets in 2011. In 2012, we use both focused elements and the entire documents as the source of the snippet. (Further details on it can be found in chapter 3.)We conducted 17 different experiments, with two different scoring methods; eleven runs were submitted to INEX. These 17 experiments are discussed in the following sections. A sample Query for 2012 is shown in Fig. 4.1.

```
<topic id="2012001" ct_no="1">
    <title>Death of John Lennon</title>
    <phrasetitle>Death of "John Lennon"</phrasetitle>
    <description>Information about John Lennon's death</description>
    <narrative>
    I want to know how where and when (including time of day) when John Lennon
    died. Now I know he was shot, but what was the name of the guy who shot him?
    </narrative>
</topic>
```

Fig. 4.1: Sample Query 2012001

## 4.2 Scoring Methods

A scoring method is used to determine the similarity of a sentence to the query. For the 2012 experiments, we use two different scoring methods. Method 1 is based on the number of unique query terms in the sentence [4], and Method 2 is inspired by BLEU, used for evaluating the quality of text which has been machine-translated from one natural language to another [13].

### Scoring Method 1

In this method, the sentence is scored based on the number of query terms present in it. This is the method employed by RMIT in the 2011 Snippet Retrieval Track [4].

The formula is given below.

$$Score = \frac{(number\ of\ unique\ query\ terms\ in\ the\ sentence)^2}{total\ number\ of\ query\ terms\ in\ the\ sentence} \tag{4.1}$$

### Scoring Method 2

In this method, unigrams, bigrams, trigrams and 4-grams [11] in both the query and the sentence are identified. We calculate the ratio of the total number of N-grams to unique number of N-grams where, N = 1 to 4, and apply the formula given below.

$$Score = \frac{1}{4}\ \left[\ \sum_{N=1}^{4} \frac{total\ number\ of\ query\ N-grams\ in\ sentence}{total\ unique\ N-grams\ in\ sentence}\ \right] \tag{4.2}$$

## 4.3 2012 Experiments

### Experiment 1

For each document in the reference run, a ranked list of focused elements is identified using the child strategy. The text of each element is obtained using it's xpath. The gathered text is merged, then split into sentences which are ranked based on the scoring method of equation 4.1.

<snippet rsv="1890.00" doc-id="2023257">
Goldman implies that Mark David Chapman's murder of John Lennon may have been part of a conspiracy by Fundamentalist Christians. According to Goldman, on the day Lennon was murde
</snippet>

Fig. 4.2: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 1

The sentences are sorted in the decreasing order and merged in that order to form a paragraph. The first 180 characters of this paragraph form the snippet. Fig. 4.2 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

### Experiment 2

This experiment is identical to experiment 1 with one exception. It uses the first 60 characters from each of the top ranked three sentences and merges them to form the snippet. Fig. 4.3 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

```
<snippet rsv="1890.00" doc-id="2023257">
The Lives of John Lennon . Goldman implies that Mark David
Chapman's murder of John Le According to Goldman, on the day
Lennon was murdered he was
</snippet>
```

Fig. 4.3: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 2

### Experiment 3

This experiment is identical to experiment 1 but uses the correlation strategy to produce the focused elements. Fig. 4.4 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

```
<snippet rsv="1890.00" doc-id="2023257">
Rock Bottom, p. 605 et al The Lives of John Lennon by Ray Albert Goldman
(1988, William Morrow and Co.  The Lives of John Lennon by Ray Albert
Goldman (1988, William Morrow and Co
</snippet>
```

Fig. 4.4: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1)by Experiment 3

## Experiment 4

This experiment is identical to experiment 2; the only difference is that correlation strategy is used to produce the focused elements. Fig. 4.5 shows a sample snippet generated for document 2023257 and query 201001 (Fig. 4.1).

```
<snippet rsv="1890.00" doc-id="2023257">
The Lives of John Lennon by Ray Albert Goldman (1988, Willi
"Rock Bottom", p. 605 et al The Lives of John Lennon by Ray
"Making Magic", p. 580 The Lives of John Lennon by Ray Albe
</snippet>
```

Fig. 4.5: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 4

## Experiment 5

This experiment is identical to experiment 1 but uses the scoring method specified in equation 4.2. Fig. 4.6 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

```
<snippet rsv="1890.00" doc-id="2023257">
He also mocked what he called the stupidity of the magazine employees who were
assigned the task of smearing me and my book, and concluded by saying that Sante
was a young man of
</snippet>
```

Fig. 4.6: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 5

## Experiment 6

This experiment replicates experiment 2 but uses equation 4.2 for scoring the sentences. Fig. 4.7 shows a sample snippet generated for document 2023257 and the query 2012001 (Fig. 4.1).

```
<snippet rsv="1890.00" doc-id="2023257">
He also mocked what he called "the stupidity of the ["News" As such,
The Lives of John Lennon was extremely controvers He also enumerates
what he described as Yoko Ono's lavish s
</snippet>
```

Fig. 4.7: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 6

## Experiment 7

This experiment replicates experiment 3 but uses equation 4.2 for scoring the sentences. Fig. 4.8 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet rsv="1890.00" doc-id="2023257">
John Lennon (2006)Â· Chapter 27 (2008) Â· The Killing of John Lennon (2008) Family Julia Lennon Â· Alfred Lennon Â· Mimi Smith Â· Cynthia Lennon Â· Julian Lennon Â· Yoko Ono Â· Se
</snippet>

Fig. 4.8: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 7

## Experiment 8

This experiment replicates experiment 4 but uses equation 4.2 for scoring the sentences. Fig. 4.9 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet rsv="1890.00" doc-id="2023257">
John Lennon (2006)Â· Chapter 27 (2008) Â· The Killing of Jo The Lives of John Lennon is a 1988 biography of musician Jo John Lennon Studio albums John Lennon/Plastic Ono Band Â· I
</snippet>

Fig. 4.9: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 8

## Experiment 9

In this experiment, the first 180 characters of text from the entire article forms the snippet. Fig. 4.10 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet doc-id="2023257">
The Lives of John Lennon is a 1988 biography of musician John Lennon by American author Albert Goldman . It is best known for its criticism and generally negative representation o
</snippet>

Fig. 4.10: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 9

## Experiment 10

In this experiment, the text of the article is split in to sentences, which are ranked based on the scoring method specified in equation 4.1. The sentences are sorted and merged. The first 180 characters form the snippet. Fig. 4.11 shows a sample snippet generated for document 2023257 and the query 2012001 (Fig. 4.1).

<snippet doc-id="2023257">
Rock Bottom, p. 605 et al The Lives of John Lennon by Ray Albert Goldman (1988, William Morrow and Co. The Lives of John Lennon by Ray Albert Goldman (1988, William Morrow and Co
</snippet>

Fig. 4.11: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 10

## Experiment 11

This experiment is identical experiment 10 but merges the first 60 characters from each of the three top ranked sentences to form the snippet. Fig. 4.12 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet rsv="1890.00" doc-id="2023257">
The Lives of John Lennon by Ray Albert Goldman (1988, Willi
"Rock Bottom", p. 605 et al The Lives of John Lennon by Ray
"Making Magic", p. 580 The Lives of John Lennon by Ray Albe
</snippet>

Fig. 4.12: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 11

### Experiment 12

This experiment is identical to experiment 10 other than it uses the scoring method specified in equation 4.2 to rank the sentences. Fig. 4.13 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet rsv="1890.00" doc-id="2023257">
John Lennon (2006)Â· Chapter 27 (2008) Â· The Killing of John Lennon (2008)
Family Julia Lennon Â· Alfred Lennon Â· Mimi Smith Â· Cynthia Lennon Â· Julian
Lennon Â· Yoko Ono Â· Se
</snippet>

Fig. 4.13: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 12

### Experiment 13

This experiment is identical to experiment 11 but uses the scoring method specified in equation 4.2. Fig. 4.14 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1.

<snippet rsv="1890.00" doc-id="2023257">
The Lives of John Lennon 2023257 244259489 2008-10-10T00:14 John Lennon (2006)Â· Chapter 27 (2008) Â· The Killing of Jo John Lennon Studio albums John Lennon/Plastic Ono Band Â· I
</snippet>

Fig. 4.14: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 13

### Experiment 14

This experiment is similar to experiment 1. It uses the child focusing strategy to generate elements. However, in this experiment, instead of using text from all the focused elements, we use only the text from the highest-ranked focused element. We then split the text into sentences, rank them in decreasing order and merge them. The first 180 characters form the snippet. Fig. 4.15 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet rsv="1890.00" doc-id="2023257">
Goldman implies that Mark David Chapman's murder of John Lennon may have been part of a conspiracy by Fundamentalist Christians. According to Goldman, on the day Lennon was murde
</snippet>

Fig. 4.15: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 14

### Experiment 15

This experiment is identical to experiment 14 but scores the sentences based on the scoring method of equation 4.2. Fig. 4.16 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet rsv="1890.00" doc-id="2023257">
Goldman implies that Mark David Chapman's murder of John Le According to Goldman, on the day Lennon was murdered he was In the sketch another subject of Goldman, Elvis Presley (pl
</snippet>

Fig. 4.16: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 15

## Experiment 16

This experiment is identical to experiment 14 but uses the correlation focusing strategy. Fig. 4.17 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1).

<snippet rsv="1890.00" doc-id="2023257">
Rock Bottom, p. 605 et al The Lives of John Lennon by Ray Albert Goldman (1988, William Morrow and Co. The Lives of John Lennon by Ray Albert Goldman (1988, William Morrow and Co
</snippet>

Fig. 4.17: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 16

## Experiment 17

This experiment replicates experiment 15 but uses the correlation focusing strategy. Fig. 4.18 shows a sample snippet generated for document 2023257 and query 2012001 (Fig. 4.1 using this experiment.

```
<snippet rsv="1890.00" doc-id="2023257">
 Rock Bottom, p. 605 et al The Lives of John Lennon by Ray A The Lives
of John Lennon by Ray Albert Goldman (1988, Willi You Can't Catch Me,
p. 517 et al The Lives of John Lennon b
</snippet>
```

Fig. 4.18: Sample Snippet Generated for Document 2023257 and Query 2012001 (Fig. 4.1) by Experiment 17

## 4.4 2012 INEX Runs

Of these 17 experiments, only 11 experiments were submitted to INEX. We observed that the snippets formed by taking the first 60 characters of top three sentences did not produce a cohesive result. Hence we decided not to submit those runs. Table 4.1 shows the experiment and corresponding names of the runs submitted to INEX.

| Experiment | Title of the run submitted to INEX |
|---|---|
| Experiment 1 | UMD_focused_child_1 |
| Experiment 5 | UMD_focused_child_2_bleu |
| Experiment 3 | UMD_focused_correlation_1 |
| Experiment 7 | UMD_focused_correlation_2_bleu |
| Experiment 12 | UMD_snippets_article_bleu_180 |
| Experiment 9 | UMD_snippets_article_first_180 |
| Experiment 10 | UMD_snippets_article_score_180 |
| Experiment 15 | UMD_snippets_child_score_3_100_Single |
| Experiment 14 | UMD_snippets_child_score_300_Single |
| Experiment 17 | UMD_snippets_correlation_score_3_100_Single |
| Experiment 16 | UMD_snippets_correlation_score_300_Single |

Table 4.1: 2012 UMD's Submitted Runs

## 4.5 Results

For the 2012 Snippet Retrieval Track, both snippet-based and document-based assessment are used to evaluate the submitted runs. For each topic, the assessor reads the details of the topic, then reads through each snippet and determines whether or not the underlying document is relevant to the topic. After snippet-based assessment is complete, he/she then assesses each document for relevance based on it's full text [19]. Table 4.2 displays the evaluation measures of UMD's runs submitted to 2012 INEX Snippet Retrieval Track. Table 4.3 displays the ten top ranked runs and their scores for the 2012 INEX Snippet Retrieval Track.

| Run | MPA | MNPA | Recall | NR | GM | PA | NA |
|---|---|---|---|---|---|---|---|
| UMD_focused_child_1 | 0.7443 | 0.6844 | 0.5071 | 0.8476 | 0.6121 | 0.5292 | 0.7685 |
| UMD_snippets_article_score_180 | 0.7014 | 0.6706 | 0.4873 | 0.8338 | 0.5798 | 0.4954 | 0.6878 |
| UMD_focused_child_2_bleu | 0.7643 | 0.6920 | 0.4424 | 0.9111 | 0.5466 | 0.4447 | 0.7987 |
| UMD_focused_correlation_1 | 0.7429 | 0.6481 | 0.4566 | 0.8617 | 0.5319 | 0.4708 | 0.7517 |
| UMD_snippets_article_bleu_180 | 0.6986 | 0.6669 | 0.4552 | 0.8869 | 0.5305 | 0.4306 | 0.7396 |
| UMD_focused_correlation_2_bleu | 0.6800 | 0.6144 | 0.3853 | 0.8714 | 0.5007 | 0.4108 | 0.7057 |
| UMD_snippets_child_score_3_100_Single | 0.6900 | 0.6332 | 0.3441 | 0.8752 | 0.4991 | 0.3774 | 0.7172 |
| UMD_snippets_correlation_score_3_100_Single | 0.7143 | 0.6492 | 0.3668 | 0.8954 | 0.4927 | 0.4162 | 0.7311 |
| UMD_snippets_child_score_300_Single | 0.7157 | 0.6179 | 0.4080 | 0.8257 | 0.4706 | 0.4463 | 0.6952 |
| UMD_snippets_correlation_score_300_Single | 0.6743 | 0.5836 | 0.3986 | 0.7854 | 0.4268 | 0.4078 | 0.6713 |
| UMD_snippets_article_first_180 | 0.7500 | 0.6164 | 0.3092 | 0.9405 | 0.4063 | 0.3726 | 0.7899 |

Table 4.2: 2012 Snippet Retrieval Evaluation Measures of UMD Runs

| Rank | Run | Score |
|---|---|---|
| 1 | UMD_focused_child_1 | 0.6121 |
| 2 | QUT_2012_Focused | 0.6096 |
| 3 | UMD_snippets_article_score_180 | 0.5798 |
| 4 | QUT_2012_Focused_Split | 0.5712 |
| 5 | TheCNGL_DCU_SnippetTrack2012_SRReferenceRun02 | 0.5648 |
| 6 | TheCNGL_DCU_SnippetTrack2012_SRReferenceRun03 | 0.5510 |
| 7 | UMD_focused_child_2_bleu | 0.5466 |
| 8 | SR2012-Baseline | 0.5431 |
| 9 | TheCNGL_DCU_SnippetTrack_2012_SRun04 | 0.5390 |
| 10 | UMD_focused_correlation_1 | 0.5319 |

Table 4.3: INEX 2012 Snippet Retrieval Track Run Ranking

The GM values of the runs indicate that the snippets generated by the 2012 experiments produced better GM values than those generated in 2011. Table 4.3 shows that we were successful in improving the recall values. Table 4.3 shows that UMD occupied ranks 1,3,7 and 10 positions among the top ten in the 2012 INEX Snippet Retrieval Track results. The runs submitted can be broadly categorized into three groups, i.e., group-1, group-2 and group-3. The runs produced using text from all the elements generated by a focusing strategy are categorized as group-1(i.e., experiments 1, 3, 5 and 7). Runs produced by using text from the articles are categorized as group-2(i.e., experiments 9, 10 and 12). Runs produced by using text from the top-ranked focused element are categorized as group-3(i.e., experiments 14, 15, 16 and 17). Analysis of each of these experiments by group is as follows.

**Group-1:** If the runs in group-1 are considered, it can be observed from Table 4.3 that runs generated by applying the child focusing strategy ranked 1 and 7 while those generated by applying the correlation strategy ranked 10 and 15. For a given focusing strategy, it can be observed that runs generated using scoring method 1 fared better results than those using scoring method 2. Similarly for a given scoring method, it can be observed that runs generated by applying child strategy produced better results than those produced by applying the correlation strategy. It can also be observed that irrespective of scoring method, runs produced by applying the child strategy produced better results than those generated by applying correlation strategy.

**Group-2:** From the runs in group-2, it can be observed that runs produced by using first 180 characters of the article of the document did not produce better results than any of the submitted runs. Also, when considering the entire article, scoring method 1 produced better result than scoring method 2.

**Group-3:** The following observations can be made from the runs in group-3. For an applied focusing strategy, runs produces using scoring method 2 produced better results than those generated using scoring method 1. For a given scoring method, runs produced by applying the child strategy fared better than those generated by applying correlation strategy . To conclude, it can be said that applying scoring method 2 generated better results for runs in group-3 than group-1 or group-2. As can be clearly seen, runs in group-1 and group-2 produced better results than those in group-3. Hence it can be deduced that scoring method 1 produced better results. Also, it can be said that using all the focused elements or the entire article proved to be more effective than using a single focused element. Keeping experiment 9 in mind, it can also be said that applying some sort of scoring method proved more effective than applying none. It is possible to achieve average NR values close to 0.94, but in these cases the average recall values seem to be lower. This is evident from run UMD_ snippets_ article_ first_ 180 in Table 4.2. Methods can be improved so that recall is high when NR is high, this can ensure high GM values.

# 5 Conclusions and Future Work

As can be seen from the evaluation measures in table 4.2 and 3.1, there is nearly 33 percent increase in our highest GM values compared to the those in 2011. The increase in negative recall values is negligible considering the rise in recall values. Our methods certainly succeeded in achieving the goals set for 2012. Although we succeeded in generating runs that stood among ten top ranked runs in the INEX 2012 Snippet Retrieval Track, we just achieved a 16 percent increase in highest GM value compared to 2011.

If runs can be generated by evaluating snippets in two passes– one aimed at high NR value and– the other aimed at high recall value, we can potentially achieve high GM values. The feasibility of the approach is yet to be determined. Future research should be concentrated on producing snippets that could better determine both relevant and irrelevant documents. Developing new scoring methods can be one possible approach for future research. Also, using all the focused elements or the entire article to generate snippets can produce better results than those produced by using a single focused element.

# References


[1] *2009 Wikipedia Collection*. https://inex.mmci.uni-saarland.de/data/documentcollection.jsp#wikipedia.

[2] *2011 queries*. https://inex.mmci.uni-saarland.de/protected/snippet/inex-snippet-2011-topics.xml.

[3] *2011 Reference Run*. https://inex.mmci.uni-saarland.de/protected/snippet/2011_SR_reference_run.zip.

[4] Bando, L. L., Scholer, F., and Thom, J. "RMIT at INEX 2011 Snippet Retrieval Track". In: *Focused Retrieval of Content and Structure*. Ed. by S. Geva, J. Kamps, and R. Schenkel. Vol. 7424. LNCS. Springer Berlin Heidelberg, 2012, pp. 300–305.

[5] Banhatti, R. "Improving Results for the INEX 2009 Thorough and 2010 Efficiency Tasks". MS. Thesis. Department of CS, UMD, 2011.

[6] Chittilla, S. "Wikipedia and Snippet Generation". MS. Thesis. Department of CS, UMD, 2012.

[7] Crouch, C. J. "Dynamic Element Retrieval in a Structured Environment". In: *ACM TOIS* 24.4 (Oct. 2006), pp. 437–454.

[8] Crouch, C. J., Crouch, D. B., Vadlamudi, S., Cherukuri, R., and Mahule, A. "A Useful Method for Producing Competitive Ad Hoc Task Results". In: *Comparative Evaluation of Focused Retrieval*. Ed. by S. Geva, J. Kamps, R. Schenkel,

and A Trotman. Vol. 6932. LNCS. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 63–70.

[9] Deepak Acquilla, N. "Improving Results for the 2009 and 2010 INEX Focused Tasks". MS. Thesis. Department of CS, UMD, 2011.

[10] *INEX website*. https://inex.mmci.unisaarland.de/.

[11] Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN: 0-262-13360-1.

[12] Narendravarapu, R. "Improving Results for the INEX 2009 and 2010 Relevant in Context Tasks". MS. Thesis. Department of CS, UMD, 2011.

[13] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318.

[14] Poluri, P. "Focused Retrieval using Exact Methodology". MS. Thesis. Department of CS, UMD, 2009.

[15] Salton, G. *The Smart Retrieval System; Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971.

[16] Salton, G., Wong, A., and Yang, C. S. "A vector Space Model for Automatic Indexing". In: *Commun. ACM* 18.11 (1975), pp. 613–620.

[17] Singhal, A., Buckley, C., and Mitra, M. "Pivoted Document Length Normalization". In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Zurich, Switzerland, 1996, pp. 21–29.

[18] Trappett, M., Geva, S., Trotman, A., Scholer, F., and Sanderson, M. "Overview of the INEX 2011 Snippet Retrieval Track". In: *Focused Retrieval of Content*

*and Structure*. Ed. by S. Geva, J. Kamps, and R. Schenkel. Vol. 7424. LNCS. Springer Berlin Heidelberg, 2012, pp. 283–294.

[19] Trappett, M., Geva, S., Trotman, A., Scholer, F., and Sanderson, M. "Overview of the INEX 2012 Snippet Retrieval Track". In: *INEX'12 Workshop Pre-Proceedings*. Ed. by G. et al. 2012, pp. 68–73.

[20] Vadlamudi, S. "Significance Testing for INEX 2008-09 Ad Hoc Track". MS. Thesis. Department of CS, UMD, 2010.

# A Appendix

Table A.1 and A.2 display the results of significance test [20] run with UMD's runs on ten top ranked runs of 2009 and 2010 INEX tasks respectively. RIC stands for Relevant in Context task, RRIC stand for Restricted Relevant in Context task, and RFocused stands for Restricted Focused task. "*" indicates that there is significant difference in the runs produced by UMD and the corresponding ranked run. "–" indicates that although the runs have different scores, there is no significant difference between the UMD's runs and the corresponding ranked run.

| Runs Ranked | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| UMD Runs | | | | | | | | | | |
| Focused task: correlation strategy | – | – | * | – | * | – | * | – | * | * |
| Focused task: Child strategy | * | * | – | – | * | – | * | * | * | * |
| Focused task: section strategy | – | – | * | – | * | – | – | * | – | * |
| RIC task: correlation strategy | – | * | * | – | * | – | – | * | – | * |
| RIC task: Child strategy | – | – | * | * | * | – | – | – | – | * |
| RIC task: section strategy | * | * | – | – | – | – | * | * | * | * |
| Thorough task | – | – | – | – | – | – | – | – | – | – |

Table A.1: 2009 Significance Results

| Runs Ranked | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| UMD Runs | | | | | | | | | | |
| RFocused task: correlation strategy | – | – | – | – | * | * | * | – | – | – |
| RFocused task: Child strategy | * | * | – | – | – | – | * | * | * | * |
| RFocused task: section strategy | * | * | – | – | * | – | * | * | * | * |
| RRIC task: correlation strategy | – | – | – | – | * | * | * | * | * | * |
| RRIC task: Child strategy | – | – | – | – | * | * | – | – | – | * |
| RRIC task: section strategy | – | * | * | * | * | – | * | – | – | * |
| RRIC task: correlation strategy (FSCORE) | – | – | – | – | – | * | – | – | – | * |
| RRIC task: Child strategy (FSCORE) | * | – | * | * | * | – | – | – | – | * |
| RRIC task: section strategy (FSCORE) | * | * | * | – | – | – | – | – | * | * |
| RIC task: correlation strategy | – | – | – | * | * | – | – | * | – | – |
| RIC task: Child strategy | * | * | – | – | – | – | – | * | – | * |
| RIC task: section strategy | – | – | * | * | * | – | – | – | – | * |
| RIC task: correlation strategy (FSCORE) | – | – | – | * | * | – | – | * | – | – |
| RIC task: Child strategy (FSCORE) | * | * | – | * | – | – | * | – | * | * |
| RIC task: section strategy (FSCORE) | * | * | – | – | – | – | * | * | * | * |
| Efficiency task | – | – | * | – | – | | | | | |

Table A.2: 2010 Significance Results

Table A.3 displays the results of INEX 2009 Ad Hoc Focused Task. As can be seen from the table, (1). Section strategy produced significantly better results than the runs that ranked 3, 5, 8 and 10. (2). Child strategy produced significantly better results than the runs that ranked 2, 5, 7, 8, 9 and 10. (3). Correlation strategy produced significantly better results than the runs that ranked 7, 9 and 10 .

| 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 |

Table A.3: 2009 Ad Hoc Focused Task Top 10 Results

Table A.4 displays the results of INEX 2009 Thorough Task. As can be seen from the table, thorough task results are no different from the runs that ranked better.

Table A.5 displays the results of INEX 2009 Relevance in Context Task. As can be seen from the table, (1). Section strategy produced significantly better results than

| Participant | MAiP | Rank |
|---|---|---|
| p48-LIG-2009-thorough-3T | 0.2855 | 1 |
| p6-UAmsIN09article | 0.2818 | 2 |
| p5-BM25thorough | 0.2585 | 3 |
| p92-Lyon3LIAmanlmnt | 0.2496 | 4 |
| p60-UJM 15494 | 0.2435 | 5 |
| p346-utCASartT09 | 0.2350 | 6 |
| p72-UMD | 0.2314 | - |
| p10-MPII-CASThBM | 0.2133 | 7 |
| p167-09RefT | 0.1390 | 8 |
| p68-I09LIP6OWATh | 0.0630 | 9 |
| p25-ruc-base-coT | 0.0577 | 10 |

Table A.4: 2009 INEX Thorough Task Results

the runs that ranked 7, 8, 9 and 10. (2). Child strategy produced significantly better results than the run that ranked 10. (3). Correlation strategy produced significantly better results than the runs that ranked 2, 3, 5, 8 and 10.

Table A.6 displays the results of INEX 2010 Ad Hoc Restricted Focused Task. As can be seen from the table, (1). Section strategy produced significantly better results than the runs that ranked 5, 7, 8, 9 and 10. (2). Child strategy produced significantly better results than the runs that ranked 7, 8, 9 and 10. (3). Correlation strategy produced significantly better results than the runs that ranked 7.

Table A.7 displays the results of INEX 2010 Efficiency Task. As can be seen from the table, Efficiency task results are no different from the runs that ranked better except for the run that ranked 3. The run that ranked 3 produced results that are significantly better than ours.

| Participant | MAgP Value | Rank |
|---|---|---|
| University of Minnesota Duluth* (Correlation Strategy) | 0.1889 | - |
| Queensland University of Technology | 0.1885 | 1 |
| University of Otago | 0.1847 | 2 |
| University of Amsterdam | 0.1773 | 3 |
| LIG | 0.1760 | 4 |
| University of Minnesota Duluth* (Section Strategy) | 0.1749 | - |
| University of Tampere | 0.1720 | 5 |
| University of Minnesota Duluth* (Child Section) | 0.1656 | - |
| University of Twente | 0.1188 | 6 |
| Saint Etienne University | 0.1075 | 7 |
| School of Electronic Engineering and Computer Science | 0.1045 | 8 |
| Renmin University of China | 0.1028 | 9 |
| University of Minnesota Duluth** (early run) | 0.0424 | 10 |

Table A.5: 2009 RiC Task Results

Table A.8 displays the results of INEX 2010 RIC(F score) Task. As can be seen from the table, (1). Section strategy produced significantly better results than the runs that ranked 2, 7, 8, 9 and 10. (2). Child strategy produced significantly better results than the run that ranked 2, 4, 7, 9 and 10. (3). Correlation strategy produced significantly better results than the runs that ranked 4, 5 and 8.

| 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 |

Table A.6: 2010 Ad Hoc Restricted Focused Task Results

| Participant | MAiP | Rank |
|---|---|---|
| p167-18P167 | 0.2354 | 1 |
| p4-OTAGO-2010-10topk-18 | 0.2304 | 2 |
| p68-LIP6-OWPCRefRunTh | 0.2196 | 3 |
| p72-UMD | 0.1675 | - |
| p29-ISI2010 thorough | 0.0846 | 4 |
| p98-I10LIA4FBas | 0.0417 | 5 |

Table A.7: 2010 INEX Efficiency Task Results

Table A.9 displays the results of INEX 2010 RIC(T2I score) Task. As can be seen from the table, (1). Section strategy produced significantly better results than the runs that ranked 3, 4, 5 and 10. (2). Child strategy produced significantly better

| Participant | MAgP Value | Rank |
|---|---|---|
| ENSM-SE | 0.1970 | 1 |
| University of Minnesota Duluth* (Section Strategy) | 0.1838 | - |
| University of Minnesota Duluth* (Correlation Strategy) | 0.1764 | - |
| University of Minnesota Duluth* (Child Strategy) | 0.1753 | - |
| Peking University | 0.1726 | 2 |
| University of Otago | 0.1710 | 3 |
| Renmin University of China | 0.1671 | 4 |
| Radboud University Nijmegen | 0.1623 | 5 |
| RMIT University | 0.1541 | 6 |
| LIA - University of Avignon | 0.1298 | 7 |
| Doshisha University | 0.1122 | 8 |
| Indian Statistical Institute | 0.0693 | 9 |
| Queensland University of Technology | 0.0634 | 10 |

Table A.8: 2010 RiC Task (F-Score) Results

results than the run that ranked 2, 8 and 10. (3). Correlation strategy produced significantly better results than the runs that 4, 5, and 8.

Table A.10 displays the results of INEX 2010 RRIC(F score) Task. As can be seen from the table, (1). Section strategy produced significantly better results than the runs that ranked 1, 2, 3, 9 and 10. (2). Child strategy produced significantly better results than the run that ranked 1, 3, 4, 5 and 10. (3). Correlation strategy produced significantly better results than the runs that 6 and 10.

| Participant | MAgP | Value Rank |
|---|---|---|
| ENSM-SE | 0.1977 | 1 |
| University of Minnesota Duluth* (Section Strategy) | 0.1877 | - |
| University of Minnesota Duluth* (Child Strategy) | 0.1833 | - |
| University of Minnesota Duluth* (Correlation Strategy) | 0.1733 | - |
| Peking University | 0.1615 | 2 |
| LIA - University of Avignon | 0.1588 | 3 |
| Queensland University of Technology | 0.1521 | 4 |
| University of Otago | 0.1436 | 5 |
| Radboud University Nijmegen | 0.1377 | 6 |
| Renmin University of China | 0.1372 | 7 |
| RMIT University | 0.1335 | 8 |
| Doshisha University | 0.1014 | 9 |
| University of Amsterdam | 0.0695 | 10 |

Table A.9: 2010 RiC Task (T2I-Score) Results

Table A.11 displays the results of INEX 2010 RRIC(T2I score) Task. As can be seen from the table, (1). Section strategy produced significantly better results than the runs that ranked 2, 3, 4, 5, 6 and 10. (2). Child strategy produced significantly better results than the run that ranked 5, 6 and 10. (3). Correlation strategy produced significantly better results than the runs that 5, 6, 7, 8, 9 and 10.

| Participant | MAgP Value | Rank |
|---|---|---|
| University of Minnesota Duluth* (Child Strategy) | 0.1534 | - |
| University of Minnesota Duluth* (Section Strategy) | 0.1466 | - |
| University of Minnesota Duluth* (Correlation Strategy) | 0.1277 | - |
| Queensland University of Technology | 0.1064 | 1 |
| LIA - University of Avignon | 0.1053 | 2 |
| Peking University | 0.1030 | 3 |
| University of Otago | 0.0953 | 4 |
| Radboud University Nijmegen | 0.0945 | 5 |
| Doshisha University | 0.0537 | 6 |
| University of Waterloo | 0.0497 | 7 |
| University of Amsterdam | 0.0462 | 8 |
| Indian Statistical Institute | 0.0327 | 9 |

Table A.10: 2010 RRiC Task (F-Score) Results

Detailed descriptions on how to produce these runs can be found in [9, 5, 12]. More information on running the significant tests can be found in [20]. These significance results indicate that our focused elements are better than most of the participating groups and helped us produce best results for 2009 and 2010 INEX tasks.

| Participant | MAgP Value | Rank |
|---|---|---|
| University of Minnesota Duluth* (Child Strategy) | 0.1782 | - |
| University of Minnesota Duluth* (Section Strategy) | 0.1762 | - |
| University of Minnesota Duluth* (Correlation Strategy) | 0.1632 | - |
| Peking University | 0.1580 | 1 |
| LIA - University of Avignon | 0.1541 | 2 |
| Queensland University of Technology | 0.1508 | 3 |
| University of Otago | 0.1436 | 4 |
| Radboud University Nijmegen | 0.1375 | 5 |
| University of Waterloo | 0.0650 | 6 |
| Doshisha University | 0.0600 | 7 |
| University of Amsterdam | 0.0576 | 8 |
| Indian Statistical Institute | 0.0485 | 9 |

Table A.11: 2010 RRiC Task (T2I-Score) Participants