Testing the polyrepresentation principle by performing fusion of INEX results

Marianne Haugen

MSc Thesis - Department of Information Studies

University of Tampere

Spring 2007

Table of contents

List of tables, figures and equations

Tables

Figures

Equations

Abstract

This thesis seeks to test the theory of polyrepresentation by performing data fusion. The theory provides a framework for investigating overlap between different representations. Further, it hypothesises that overlap between these contain a higher number of relevant documents than one representation alone. In this case, the representations are different result sets which are combined in order to examine the new results in comparison with each original result set. Four different result sets for various XML retrieval approaches were selected from INEX. In evaluation, performance measures such as normalised cumulated gain, effort-precision/gain-recall, and average values for these measures were calculated. Statistical tests were carried out on the mean average values for the results. The results show that improved performance is achieved, especially for one of the approaches used in INEX for element retrieval.

1.0 Introduction

The field of Information Retrieval (IR) is constantly concerned with finding ways to improve retrieval effectiveness. As the amount of information available through online sources increases, the task of retrieving relevant information in order to fulfill a user's information need is a challenging one.

The concept of *data fusion* by utilising multiple forms of representations to increase the quality of results has been investigated by a number of researchers over the past years (such as Hsu and Taksa, 2005; Lee, 1997). During recent years, there has been an increased interest in IR to combine cognitive and functional representations in an attempt to improve search results. Essentially this is the idea behind what Ingwersen has labelled *polyrepresentation* in IR (Ingwersen and Järvelin, 2005). The polyrepresentation principle can be exploited by performing data fusion, but unlike data fusion it also includes cognitive representations which are important in this theory.

1.1 The research problem

The aim of this thesis is to test the principle of polyrepresentation on results from XML retrieval. This is done by means of data fusion of result sets. It is believed that overlaps between different result sets indicate a higher degree of relevance. Therefore, it is hypothesised that when constructing a new result list based on overlaps between best-performing results, this combined list will contain more relevant elements than is found in each of the result sets.

The data used in this study was selected from the Initiative for the Evaluation of XML Retrieval (INEX)[1] from 2005. This made it possible to examine elements rather than documents which has traditionally been done in IR. According to Smeaton (1998), results used for combination should originally have achieved fairly good performance. Therefore, the best performing results were selected for this study.

Elements from different result sets were fused in order to compare the performance effectiveness of the new constructed result sets with the original results. In simple terms, the principle explored in this study state that if an element is retrieved by several search

[1] Webpage of INEX: http://inex.is.informatik.uni-duisburg.de

algorithms for a certain topic, it is thought to be more relevant than an element which is retrieved by only one search algorithm.

The aim of the fusion of result sets is to obtain a superior combined result set with relevant elements receiving a high rank. This is important to the constant improvement of the retrieval results, i.e. rank more relevant information higher in the result set.

1.2 The outline of the thesis

This thesis will start by introducing relevant background literature within the field explored in the research study (Section 2.0). Initially, the theory of polyrepresentation will be presented and explained. In addition, the idea behind data fusion will be covered as this is an example of one way of combining multiple forms of representations to improve the results. Research studies which are considered relevant for this field will be included. Furthermore, XML will be shortly explained along with XML retrieval as the data used for this study is in XML format. This data is available through INEX which will be explained in the end of Section 2.0.

In Section 3.0, the data and methods used in this research will be outlined and explained, followed by Section 4.0 presenting the results referring to the research questions of interest. Furthermore, possible explanations for tendencies in the results will be discussed in Section 5.0. In addition, suggested further research within the field is incorporated into this chapter. Towards the end of the paper, the main findings from this research project are summarised in Section 6.0.

2.0 Review on earlier research

According to a user's information need, an information system aims at finding relevant information and presenting this to the user. In addition, the more relevant a document is, the higher rank it should receive and the higher it should be situated in the result list (Järvelin & Kekäläinen, 2002). Finding ways to accomplish these goals is a real challenge to IR researchers. The same information need will retrieve different sets of documents when used on different IR systems even though retrieving from the same collection of documents. Furthermore, different actors (such as the author of a text, the indexer, and the user) may possess different interpretations of information objects.

In 1992, Robertson and Hancock-Beaulieu claimed that the field of IR had seen significant changes during the preceding 10 years (Robertson and Hancock-Beaulieu, 1992). Without doubt, this has also been the case for the years following their article as researchers are constantly trying to find more effective ways of finding relevant information in electronic information systems. As IR research has developed, the user and the cognitive state of the user has received increased attention in the process of information seeking (Robertson and Hancock-Beaulieu, 1992). This interest in the cognitive space of the user is also apparent in the polyrepresentation principle.

2.1 Polyrepresentation

The theory of polyrepresentation was developed through the 1990s. Ingwersen (1994) reports that earlier studies within the exact match setting have found that by combining natural language representations with index representations (meta-data), the retrieval results are better in comparison with performing the search with only one type of representation (Ingwersen, 1994).

To illustrate the theory of polyrepresentation, one can consider the following simple example. An author of a text may be of the opinion that a certain term is very relevant in the paper he/she has written and include this in the title. During indexing, the same term may be considered highly relevant by the indexer. Furthermore, when a user uses a specific term or terms to fulfil his/her information need, s/he may choose exactly this term. This is an example of what the theory of polyrepresentation seeks to take advantage of, namely different representations. These representations encompass not only the ones constructed by different actors (as illustrated above), but also other representations of an information need. This example is a simple case scenario and it may of course be the case that various actors choose different terms in this process. However, the overlap between these representations indicates that the term is likely to be important.

Different representations may then represent the same information object, just in different ways. Essentially, it is this idea that Ingwersen wanted to explore when he introduced the theory of polyrepresentation within information retrieval (Ingwersen, 1994). This theory exploits the overlap between various representations with the assumption that the overlap will contain documents of higher relevance.

According to Larsen and Ingwersen (2002), overlaps between cognitive representations of both users' information situations and documents can be utilised in order to diminish uncertainties which are naturally present in IR, consequently improving the performance of IR systems. The main idea behind the theory of polyrepresentation is to utilise both functionally and cognitively different representations of an information need to improve retrieval results. More specifically, the hypothesis behind the theory is stated as the following:

> "…the more interpretations of different cognitive and functional nature, based on an IS&R [Information Seeking & Retrieval] situation, that point to a set of objects in so-called cognitive overlaps, and the more intensely they do so, the higher the probability that such objects are *relevant* (pertinent, useful) to a perceived work task/interest to be solved, the information (need) situation at hand, the topic required, or/and the influencing context of that situation" (Ingwersen and Järvelin, 2005, p. 208).

The overlap obtained from these different representations is believed to be evidence of higher relevance than the non-overlapping representations. Representations by different actors on the basis of their interpretations are considered as cognitively different representations, whereas the functionally different ones are abstractions or meta-data which fullfill various functions in the paper (Ingwersen, 1994).

To exemplify, cognitive different representations may be descriptors provided by different indexers. Their resulting index will differ as this is made on the basis of each actor's cognitive space. However, if examining the indexing terms for a specific piece of text made individually by several different people, it is natural to think that the term which is indexed by all is important in describing the content of the text. Functionally different ones can be representations such as title, abstract and references. These representations are usually made by the same actor, the author of the text.

The different types of representations that may contribute to forming the overlap, are illustrated in the figure below (Larsen, Ingwersen and Kekäläinen, 2006).

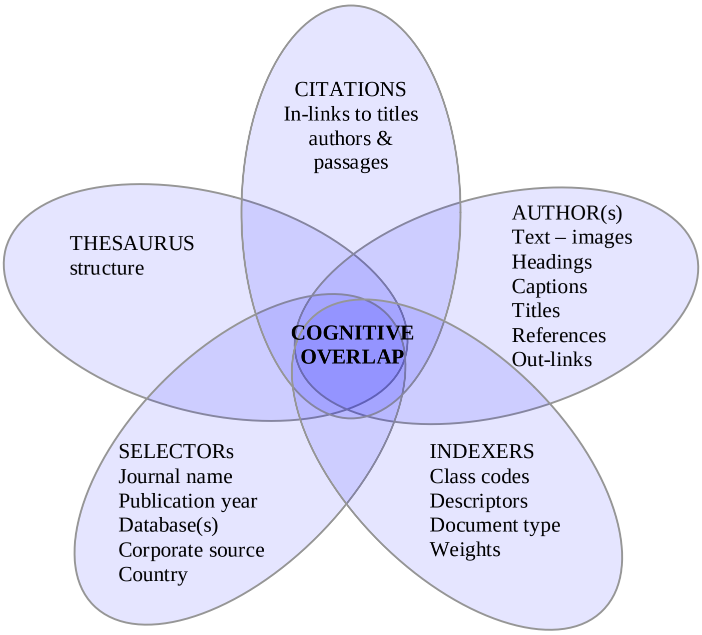

Figure 1: The polyrepresentation principle illustrated through the overlap of cognitively and functionally different representations

To examine how retrieval effectiveness can be improved, it is desirable to look at different representations to discover how they overlap. The fact that they overlap is thought to be an indication of relevance.

In a paper outlining the theory of polyrepresentation, Hjørland (2006) states that numerous studies indicate that only in about 23 percent of occasions do searchers and indexers agree on search terms. This statement further stresses the importance of the possible increased value by combining representations from different actors.

2.1.1 Previous research on polyrepresentation

Skov and colleagues (2004) performed a research study examining the principle of polyrepresentation on the Cystic Fibrosis test collection in a best match setting. Because this test collection contains both functionally and cognitively different characteristics, it is claimed to be well suited for examining inter and intra-document characteristics of polyrepresentation (Skov, Larsen and Ingwersen, 2006). Different functional representations like titles (TI), abstracts (AB) and references (RF) were explored in combination with cognitively different ones such as major (MJ) and minor (MN) MeSH (Medical Subject Headings in Medline) descriptors created by the indexer. Two types of queries were used in the study. Queries constructed from natural language (bag of words) and highly structured queries were tested. The four representations TI/AB, MJ, MN and RF resulted in 15 different overlap combinations which could be examined. The documents retrieved by one representation only are also included in these overlap combinations. The findings support the use of cognitively different representations (such as TI/AB and MeSH titles) and functionally different representations (such as references). From the results, it is apparent that the principle of polyrepresentation gains support in this study (Skov, Pedersen, Larsen and Ingwersen, 2004).

The previous study was elaborated on when Skov and colleagues performed re-ranking tests on the 15 overlap combinations (Skov et al., 2006). The results from these overlaps were combined into one search result. Weighing of terms was carried out according to 1) precision received for highly structured queries and 2) precision received for highly relevant documents. These runs were compared to natural language queries and the run where no weighting was applied. Natural language queries, referred to as bag-of-words queries, were found to be superior in terms of best overall performance obtained when the top 15 ranks were examined. However, for the following 15 results, the run based on polyrepresentation (highest precision overlaps) outperform InQuery's weighting of natural language queries. Re-ranking was also carried out based on citation impact, assuming that documents that have been cited are of higher quality. However, the result for citation impact was found to slightly decrease retrieval performance in comparison with the best run of weighted overlaps (Skov et al., 2006).

In general, higher precision was obtained with overlaps created from three or four representations in comparison with overlaps created from two or one representation(s). In addition, highly structured queries performed better in terms of higher precision compared to queries in natural language. Cited reference was found to be an important representation for high precision. Skov and colleagues (2006) found that when functionally different representations such as references were part of the overlap, precision increased. The results support the idea behind the cognitive theory in IR of incorporating both cognitively different representations and functionally different ones (Skov et al., 2006).

In her thesis, Lund (2005) describes a study testing the hypothesis in polyrepresentation by performing fusion of the results by different information retrieval algorithms from TREC-5. The best-performing systems were chosen for further investigation in her study. The reason for eliminating some systems was the fear of these poorer results negatively influencing the fusioned data set. Altogether, twelve different systems were selected.

Lund investigated the results by grouping documents into four categories according to number of systems where a document is found. In other words, the documents were grouped according to the number of systems they were retrieved by. The following three groups were used: 2 to 5 systems, 6 to 9 systems, and 10 to 12 systems. However, most of the tests were performed on the four best systems. Two of the systems are so-called functionally different ones, whereas the two others are cognitively different from the others and one another.

The top 100 documents were examined for each of the systems. To test the theory of polyrepresentation, Lund combined the four best-performing systems into what she named FUS4. The overlapping result of these four systems could then be examined and further evaluated. Additionally, a combination of all the twelve systems was carried out (named FUS12), as well as combinations of the two best systems (FUS2) and three best systems (FUS3). After re-assigning weights, FUS4 will contain less than 100 documents, so the best-performing documents from FUS3 were included in the rankings to examine if this combination could improve performance compared to all the individual systems alone and the combinations between them. Lund hypothesised that this combination, which she labels as a super-system, would be superior in finding relevant documents and give these the highest ranks.

Furthermore, Lund was interested in how the topics selected affected the performance of the different combinations examined. Therefore, the thirty best topics were arranged into three different groups based on number of relevant documents : the top ten with the largest number of relevant documents, the next ten topics with the second largest number of relevant documents, and the last ten topics with the third largest number of relevant documents. Using the polyrepresentation principle, the overlaps between the top 100 ranked documents for each of the systems were used to create a new list of overlapping documents. (Lund, 2005)

The results for the combinations show that the combination of four systems (FUS4) achieved the highest precision followed by two of the combinations of three systems and then some of the combinations of two systems. Precision increases for documents that are retrieved by more systems. The combination of all twelve systems, FUS12, was found to perform considerably poorer in comparison with most other combinations of the four best systems.

Lund concluded that a few systems in combination may perform better than twelve different systems. In addition, for most of the performance measures the super-system performs better than the FUS4, FUS12 and each of the systems individually. This can then be used as a benchmark for other systems to examine how they have ranked their relevant documents in comparison to this super-system.

In her thesis, Lund also noticed the problem of multiplying the weight with a certain number according to how many systems retrieved the document. For instance, if a document found by two systems with very high ranks is compared to a document found by all four systems at low ranks, the first one will receive a lower combined score in comparison to the second. In her thesis, Lund therefore weighted documents by multiplying with 1, 5, 10, 20, …, 90 and 100 for FUS12. However, the comparison was done so that the ones retrieved by four runs were ranked first, then the ones retrieved by three runs, if there were not enough of the first ones to fill the top 100.

Based on the polyrepresentation principle, Christoffersen (2004) developed and tested a method to identify relevant core documents within a subject domain. He acknowledges the strengths and weaknesses of some common search rationales, and suggests combining these and sorting out the information objects retrieved by at least two methods. The idea of this filtering process is to keep the relevant documents and eliminate the irrelevant ones.

In particular, Christoffersen (2004) emphasises the importance of citation data as this may provide a more comprehensive impression of the document content than solely the texts used in the document. According to Christoffersen (2004, p.389), citation data can be utilised by selecting "seed documents" or "terms derived from the titles of cited documents". One of the reasons for the importance of cited documents is that they are chosen by the author herself. As mentioned elsewhere in this thesis, citation information has also been found to be important by other researchers (e.g. Larsen, 2002; Skov et al., 2004). Three different databases were used in the study and high precision was obtained when exploiting both term and citation searching. The results were found to be statistically significant (Christoffersen, 2004). This method is claimed to be useful in rapidly identifying core documents related to a subject, as well as being easy to comprehend.

It is claimed (Skov et al., 2006, p.99) that "polyrepresentation in the true sense of the concept cannot be achieved with weakly structured queries in natural language". In a best match IR system, highly structured queries obtain superior results due to the Boolean features inherent in polyrepresentation. Structured queries identify overlapping documents where all search terms are present in the different representations. On the other hand, queries in natural language only require match of one search key in the retrieved documents. Therefore, overlaps from these queries will contain documents with less relevance to the information need (Skov et al., 2006).

Larsen (2002; Larsen and Ingwersen, 2002) presents an extension of the theory of polyrepresentation by adding what the author refers to as the Boomerang effect in retrieval of scientific documents. An important goal of the Boomerang effect is to utilise link and citation information to improve performance, while simultaneously allowing for natural language queries. Larsen (2002, p.155) explains this in terms of a 'cycling strategy' which does the following:

> "starting with documents retrieved by a subject search, wherefrom new documents are identified automatically by following the network of citations in scientific papers backwards and forwards in time".

In this way, potentially relevant documents can be accessed and returned to the user using the principle of polyrepresentation by examining overlaps between the results. Larsen and Ingwersen (2002) explains the Boomerang effect by including three steps. Firstly, natural language queries are used in the retrieval of sets of documents. Secondly, the references are extracted creating a pool for each of these sets. Overlaps between these pools can then be identified. These overlapping references are used further to discover and retrieve documents where these are cited. This is done for each overlap between the pools of references. Overlaps between the documents which cite the references can be found and the documents can be partially ordered in a number of overlap levels. These documents are thought to be potentially relevant. (Larsen and Ingwersen, 2002)

The Boomerang effect, as explained above, follows a Boolean approach and consequently the documents within an overlap are not ranked. A best match boomerang effect that uses weighting of references was therefore proposed (Larsen and Ingwersen, 2002). Larsen (2004) tested this by using various weighting schemes based on frequency of occurrence of cited documents. The results show that the Boomerang effect did not increase performance, and using only polyrepresentation was found to be somewhat better.

The research carried out by Lund (2005) is similar to the one that is described in this thesis. However, there are important differences such as the fact that Lund considered document overlap whereas this study investigates element overlap. In addition, the method of combining results is different in the two studies. Consequently, the data and methods used differ and were chosen according to the test data.

2.2 Data fusion

Data fusion is not a new field within information retrieval and it has been investigated by a number of researchers during the last years (e.g. Hsu and Taksa, 2005; Lee, 1997; Lund, 2005; Wu and McClean, 2006a; 2006b). The idea behind data fusion is that different systems will retrieve different sets of results for the same information need. This is also the case for different retrieval strategies within the same system. Combining these multiple result sets can improve performance effectiveness. This approach is what researchers refer to as data fusion.

If one system retrieves a certain document, it is an indication of that document being relevant. Within data fusion, the term multiple evidence then refers to a combination of these single evidences. Furthermore, taking advantage of these multiple evidences are thought to improve results. A document is therefore more likely to be relevant if it is found by more than one retrieval system or retrieval strategy. Multiple evidence can also be found within one system, for instance when using different strategies for retrieving documents.

2.2.1 Data fusion in relation to polyrepresentation

It may be difficult to understand the difference between the concept of data fusion and the concept of polyrepresentation. One main difference is that data fusion is more practically-oriented in comparison to polyrepresentation which is an abstract explanation of the process behind multiple evidence. Within data fusion, researchers deal with methods to combine result sets from different systems or within one system to increase performance. On the other hand, polyrepresentation is a more abstract model/phenomenon seeking to explain what is going on in this process. This can be taken advantage of in practice when performing data fusion.

Lund (2005) believes that data fusion is a concept within the area of polyrepresentation. More specifically, polyrepresentation covers a wider aspect as it also incorporates the user and its cognitive space (Lund, 2005). Therefore, data fusion can be seen as an example of utilising the idea behind polyrepresentation. However, few researchers discuss these concepts together in their research. At the same time, both of them are described as multi evidence by researchers representing each field (Ingwersen, 1994; Ingwersen and Larsen, 2002; Lee, 1997). As described by Ingwersen (1994), polyrepresentation can encompass any kind of representation of the information need provided by different actors.

2.3 Previous research on data fusion

When combining multiple evidences it is necessary to use one or more combination algorithms (Lee, 1997; Fox and Shaw, 1994). This is required in order to give a particular result a new rank or position in a list. There are different methods to calculate these and the methods value various aspects in the combination process. For instance, a method may value a result retrieved by a great number of retrieval systems more than the the scores/rank these result received in the original lists. A common way is to use some kind of combination of number of systems a document is retrieved by, and score/rank that the result achieved originally. It is suggested that “the more runs a document is retrieved by, the higher the rank that combining functions should assign to the document” (Lee, 1997, p.269).

Fox and Shaw (1994) have proposed six different combination methods within data fusion. These are outlined below in Table 1.

Name	*Combined Similarity =*
CombMIN	Minimum of individual similarities
CombMAX	Maximum of individual similarities
CombSUM	Summation of individual similarities
CombANZ	CombSUM / number of nonzero similarities
CombMNZ	CombSUM * number of nonzero similarities
CombMED	Median of individual similarities

Table 1: Combining functions by Fox and Shaw (1994)

To further explain these methods, the reasoning behind the first one, CombMIN, was to reduce the probability of ranking a non-relevant document at a high position whereas CombMAX was aimed at reducing amount of relevant documents receiving poor ranks. These two methods are targeted at specific problems without considering how other retrieved documents are affected by this. According to Fox and Shaw (1994, p.246) “the CombMIN combination method will promote the type of error that the CombMAX method is designed to minimize, and vice versa.” As a response to this issue, the CombMED combination method is proposed as a simple way of dealing with this. Both situations are avoided by determining the median similarity value (Fox and Shaw, 1994).

The combination methods explained in the above paragraph can be critisized in the way that they do not consider all the different ranks, or similarity values, a document has received. Rather, the methods select a single value from different runs or result sets without taken into consideration the others. To deal with these issues, Fox and Shaw (1994) proposed three additional combination methods. The CombSUM method summarises the similarity values from different runs, whereas CombANZ takes the average of non-zero similarity values. The last combination method, CombMNZ, summarises the non-zero similarity values and gives higher weights to documents retrieved by several retrieval methods. With the exception of this combination method, the others will not be any further explained in this paper.

Rather than examining multiple evidence within one system, Lee (1997) performed research with the aim of investigating data fusion between different retrieval systems. He wanted to explore the reasons behind improved retrieval effectiveness when combining several representations of an information need. The argument for doing this was that up until then a number of researchers had concluded that retrieval effectiveness could be improved by utilising multiple representations, however none of these had attempted to find out why (Lee, 1997).

Representations can be various document or query representations, or different retrieval techniques. As mentioned in the above section, the idea behind data fusion is to combine the retrieval results from multiple representations to achieve superior retrieval effectiveness in comparison to a single representation or one particular retrieval technique. By utilising data fusion, Lee (1997) wanted to explore the number of relevant documents and non-relevant documents in an overlap between retrieval results. He sought to examine various combination methods to discover to what extent retrieval effectiveness is affected.

In Lee’s study, six different retrieval results from the TREC3 ad-hoc track were examined (Lee, 1997). Two and two systems (referred to as pair-wise combination) were examined together. All systems were combined pair-wise in this way. More specifically, combinations were done for two and two result sets (set 1 with 2, set 1 with 3 and so on), but also for three and three result sets (set 1 with 2 and 3, set 1 with 2 and 4, set 1 with 2 and 5, set 1 with 2 and 6; set 2 with 3 and 4, and so on), and for the various combinations of four and five result sets.

In Lee’s study (1997), the combination algorithm CombMNZ was found to perform better than CombSUM in terms of retrieval effectiveness, independent of number of documents considered in the result set. To be more precise, various cut-off values of documents were examined. From the results, it was apparent that if there was a greater overlap of relevant documents than non-relevant documents among the results, this had a positive effect on data fusion (Lee, 1997).

According to Wu and McClean (2006) scores from different retrieval systems may vary greatly and it is not possible to compare these scores directly. Therefore, some kind of score normalisation is necessary (Lee, 1997; Wu and McClean, 2006). This may also be the case for different result sets as it is in this thesis study. For the research study in this thesis, it is also necessary to perform normalisation on the different results in order to be able to rank these in a new list of combining elements. A well known score normalisation method is the linear

[0,1] normalisation function proposed by Lee (1997). In this method, the maximal score is given as 1 and the mimum score is given as 0, while all other scores are altered linearly to scores between 0 and 1 (Lee, 1997).

In some cases, score information may not be available and ranking information may then be used when combining results in data fusion. An example of this can be Web search engines which rarely have scores associated to the retrieved Web documents (Wu and McClean, 2006). However, for the research carried out in this thesis, score information was available and this was then used when performing fusion of result sets.

Fox and Shaw (1994) performed fusion of five different retrieval runs in a study. Since research was performed at retrieval time, the similarity of each document for each run was accessible in comparison to only having the top 1000 documents for each run. The researchers created a document vector file when indexing a document collection (Fox and Shaw, 1994). Furthermore, they combined retrieval runs from both this and P-norm type queries when determining the probability of relevance for a certain document in a collection. Overall, P-norm queries achieve better results in comparison to vector queries. Evaluation of results was performed for each of the nine collections used in the study. Results from each of the collections were merged and the result list was presented to the user. Six different combination algorithms (see Table 1) were examined. The runs were examined individually and in combination.

According to Fox and Shaw (1994), there are two types of errors that an information retrieval ranking method aims to avoid, namely to give a rather high rank to a non-relevant document and to give a rather low rank to a relevant document. Furthermore, they claim that when one retrieval system assign a high rank to a non-relevant document, another one is likely to assign the same document a considerably lower rank. Therefore, a method which takes both of these two methods into consideration, could reduce the likelihood of this occurring (Fox and Shaw, 1994).

In their study, Fox and Shaw (1994) analysed the different combination methods and compared them with the best individual single run. In this analysis, the CombSUM method was found to perform considerably better. However, results for both runs using CombANZ and CombMNZ shows that the combinations perform better than the individual runs. A combination of all five runs was found to improve retrieval effectiveness in comparison to each individual run. It is concluded that the more query representations which are taken into account, the better results achieved. This gives support to the idea that fusioning improves retrieval effectiveness. However, Fox and Shaw (1994) also noticed that this better effectiveness is not necessarily achieved by fusioning two or three runs. Furthermore, this finding justifies the selection of four different result sets for the research questions examined in the study described in this thesis.

Recently, Wu and McClean (2006b) performed a research study with the aim to discover variables that may affect data fusion by using multiple regression in the analysis of component results taken from TREC. More specifically, they wanted to do so by using a great number of combinations on several different data fusion algorithms. It was desirable to identify in which situations data fusion can improve performance. CombSUM, CombMNZ and Round-robin were the fusion algorithms examined in this study. CombSUM and CombMNZ are well-known combination algorithms (as explained earlier in this section), while Round-robin combines result sets by taking one document from each list and removes the document if it has occurred earlier. The results are merged, however no weighting for the documents’ ranking is given. Therefore, the order of a document is solely dependent on its original position in one of the results (Wu and McClean, 2006b).

Variables considered in the study by Wu and McClean (2006) included number of results, the overlap rate between the results, the mean average precision of the results, as well as the standard deviation of the mean average precision of results. All of these variables were found to be highly significant in influencing performance of data fusion. Furthermore, a prediction analysis of performance of data fusion methods was carried out. Nearly all the combined results (using either CombSUM or CombMNZ) were found to perform better in comparison with the average performance of component results. A high percentage of the combined results were even found to achieve higher performance than the best component result. The overlap-rate variable was found to be highly important in predicting performance. The prediction of performance of data fusion methods was found to be fairly accurate. The outcomes of this research can be utilised in improving data fusion algorithms (Wu and McClean, 2006b).

CombMNZ performs well in different research settings performed by various people. Based on the results of other researchers, this algorithm was chosen when combining the results examined in this research study. This is described in the next chapter of this thesis.

2.4 Extensible Markup Language

The Extensible Markup Language (XML) is a simple, but flexible markup language derived from SGML (Standard Generalized Markup Language) (W3C, 2006). An increasing amount of information is formatted according to the W3C standard for information repositories; XML. It is claimed to become “a standard document format on the Web, in Digital Libraries and Publishing” (INEX, 2006[2]). It can be used to represent data in both the document-centric and the data-centric view. The former is used to mark up the structure of documents, while the latter is used to model and represent data items (Baeza-Yates and Lalmas, 2006). The focus of INEX is on document-centric XML issues.

Below is an example of an XML document illustrating different elements and their starting- and ending-tags. This example is taken from the result files from INEX and therefore includes topic and result information, as well as other information related to the runs.

```
<inex-submission participant-id="12" run-id="VSM_Aggr_06"
task=”CO.Focussed” query=”automatic”>
      <description>
            Using VSM to compute RSV at leaf level combined with
            aggregation at retrieval time, assuming independence and
            using augmentationweight=0.6.
      </description>
      <collections>
            <collection>ieee</collection>
      </collections>
      <topic topic-id="01">
            <result>
                  <file>tc/2001/t0111</file>
                  <path>/article[1]/bm[1]/ack[1]</path>
                  <rsv>0.67</rsv>
            </result>
            <result>
                  <file>an/1995/a1004</file>
                  <path>/article[1]/bdy[1]/sec[1]/p[3]</path>
                  <rsv>0.1</rsv>
            </result>
            [ ... ]
      </topic>
      <topic topic-id="02">
            [ ... ]
      </topic>
```

[2] No page number available for this quote as it is taken from an online source

```
          [ ... ]
     </inex-submission>
```

Table 2: Example of an XML document from INEX

2.5 XML retrieval

One of the main features of XML in relation to retrieval, is the separation of layout and structure of a document. This implies that retrieval of XML data is concerned with retrieval of elements rather than documents (Baeza-Yates and Lalmas, 2006). This is thought to be particularly beneficial to the user as the most relevant elements of a document can be retrieved and returned, thus decreasing the effort the user has to spend examining non-relevant information in search of the relevant parts of a document (INEX, 2006). In this way, users can take advantage of the internal structure of a document by accessing specific elements.

Due to the nature of marking up information in XML, some issues are necessary to take into consideration to facilitate retrieval of these elements. One of them is the overlap of components when retrieving arbitrary document elements. These can not always be regarded as independent units which has been the case in traditional IR (Baeza-Yates and Lalmas, 2006).

As an example, a section and one or more of its paragraphs may be returned in the result list at different ranks. Therefore, evaluation metrics for XML data need to provide means that consider the overlap among the result elements (Kazai, Lalmas and de Vries, 2004a). This study will use data both with and without overlapping elements. This is further explained in Section 3.2.

Another issue within retrieval of elements is that the retrieved elements may be of varying granularity and therefore the size of these elements cannot be expected to be the same (Kazai et al., 2004b). An additional problem is how to present the elements to the user, and one method is to cluster elements according to which document they are taken from.

2.6 Initiative for the Evaluation of XML retrieval

INEX is an annual evaluation effort which started in March 2002. It was established as a natural response to the fast-growing amount of XML data and the need to evaluate and improve XML retrieval systems. In other words, it was *and is* necessary to continuously work on evaluating retrieval on such systems in an attempt to improve retrieval effectiveness on XML data.

The constant contribution of the participating groups in INEX is crucial on the way towards this goal. The following is stated on the INEX web page as the main intention by starting this:

> "The aim of the INEX initiative is to establish an infrastructure and provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems." (INEX, 2006)

In other words, INEX is comparable to other well-known IR collections such as the Text Retrieval Conference (TREC[3]) and the Cross Language Evaluation Forum (CLEF[4]). However, the setting and environment is adapted to XML retrieval evaluation.

INEX participants from all over the world contribute to the test collection by submitting topics and performing relevance assessment on these topics. In 2006, participants from 87 institutes were involved in developing a test collection consisting of Wikipedia documents marked in XML. From the start of INEX in 2002 until 2005, the test collection contained 12 107 articles from 18 IEEE-journals from the period between 1995 and 2002. The size of the test collection used in INEX has grown rapidly since the start, and in 2006 the size of the collection made from English documents from Wikipedia was 4.6 Gigabytes, excluding images (INEX, 2006).

In INEX, the tasks are primarily ad-hoc retrieval thus performed similarly as a user would perform a search task in the library searching in a static set of documents. However, differences for INEX are that the documents are marked in XML, the query may contain additional structural constraints and the retrieval results consist of XML elements.

There are mainly two types of topics to be specified in INEX. The first type of topics, Content-only (CO), is related to the content of the text to be retrieved and does not contain any structural constraints. The other one, Content-and-structure (CAS) topics, contain structural conditions according to where the information should be located (INEX, 2006).

[3] TREC web page: http://trec.nist.gov/
[4] CLEF web page: http://www.clef-campaign.org/

These topics may reflect different types of users according to how familiar the user is with the document structure (Malik et al., 2006). The topics are used in the creation of queries.

In INEX, relevance assessments are performed along two dimensions; exhaustivity (E) and specificity (S). Exhaustivity refers to how well the element discusses the topic of request, whereas specificity describes how focused the element is on the topic of request (Malik et al., 2006). Different measures are used in INEX for evaluation of results. From 2005, the extended cumulated gain (XCG) metrics were adopted as official metrics for evaluation in INEX (Pehcevski, Thom and Vercoustre, 2006). Explanation of these metrics will be given in Section 3.4.1.

Elements receive scores within a certain range. Different ranges are used by different participating groups, for different runs, and even for different topics. The score is achieved on the basis of how well the element of examination corresponds to the given topic.

2.7 Justification

An increasing amount of information available on the Web, and also in Digital Libraries and within Publishing, is in XML format (INEX, 2006). Therefore, document retrieval is claimed to be moving towards true information retrieval (Arvola, Junkkari and Kekäläinen, 2005). Within XML, elements are retrieved and returned to the user.

Traditionally, content representation of documents has dominated IR research. However, during recent years, representations of content and layout of a document has been of increased interest (Ingwersen and Järvelin, 2005). The growing interest in and use of XML as a way of representing information, further stresses the importance of these (Baeza-Yates and Lalmas, 2006).

There is a lack of research on data fusion with XML data. Due to different features inherent in XML, a study testing data fusion on XML documents is needed. Combining overlaps between different results of XML retrieval is thought to result in better performance.

2.8 Research questions

The aim of this research is to examine the usefulness in performing data fusion on XML elements. Thus, the following research questions were raised:

> *Research question:* Can better retrieval effectiveness be achieved by using data fusion on several different result sets compared to the individual result sets used in the fusion?

RQ1: For result sets without overlapping elements within the results

- When combining on element level
- When combining elements within overlapping documents

RQ2: For result sets with overlapping elements

Result sets without overlapping elements within the results are in INEX referred to as Focussed, whereas result sets with overlapping elements are referred to as Thorough. This is further explained in Section 3.2. In order to answer the research questions, it is necessary to perform fusion of result sets for these various retrieval approaches from INEX. These combined result sets will be evaluated together with each of the original result sets to examine whether improvement has been achieved with the fusion. The results of the fusion are reported in Section 4.0.

3.0 Data and methods

In this section, the data used for this study will be presented and explained. The procedures for evaluation of topics against the document collection used in INEX 2005 and the construction of the result sets used in this study are described in the subsequent sections. The process of combining results from various result sets from different participating groups was carried out using a well-known combination algorithm, CombMNZ.

3.1 Introduction to INEX 2005

In INEX 2005, there were a total of 41 participating groups from different countries all over the world. The main track in INEX is the ad-hoc retrieval task. Altogether there were seven different tracks or tasks to be performed in INEX 2005. From previous years, the relevance feedback task and the natural query language task were included. These were continued in 2005. Furthermore, the heterogeneous collection track and the interactive track were included from 2004 onwards. In addition, the two new tracks document mining and multimedia started in 2005 (INEX, 2005).

Some of the results from the ad hoc task from 2005 were chosen as data for this study. Recall from the chapter on background literature, that ad hoc retrieval is performed on a static set of documents by using new topics. The test collection used for the ad-hoc track consisted of publications donated by the IEEE Computer Society. The collection used for 2005 contained altogether 16 819 articles equal to the size of 764Mb (Malik et al., 2005).

The *topics* used in INEX represent a set of information needs. The results from the queries generated from the topics, are gathered in a *submission file*. The queries were matched against the document collection and the result elements for all topics make up the submission file. A submission file is used for making the pool for relevance assessments. The submissions are then evaluated using the relevance assessments. A submission file can contain up to 1500 retrieval results for each topic. Different topics are included in one submission file, identified by a topic number (Lalmas, 2005). For the data used in this study, queries were constructed automatically from the topic.

The submission file consists of various information marked up in XML including topic number, document id, element path, in addition to rank and score information for the results.

Each submission file also contains a description of the retrieval approach used to generate the results.

The DTD for the submissions is given below. The *file*, or document id, specifies in which document file the result is found. The *path*, or element path, specifies the element in a certain document which has been assessed. The document id and element path together determine a unique result element. Furthermore, each result element has a rank according to position in the result list, and associated rsv information (retrieval status value).

```
<!ELEMENT inex-submission (description, collections, topic+)>
<!ATTLIST inex-submission
     participant-id  CDATA      #REQUIRED
     run-id          CDATA      #REQUIRED
     task (CO.Focussed | CO.Thorough | CO.FetchBrowse |
          +S.Focussed | +S.Thorough | +S.FetchBrowse |
          VVCAS | VSCAS | SVCAS | SSCAS) #REQUIRED
     query (automatic | manual)      #REQUIRED
>
<!ELEMENT  description     (#PCDATA)>
<!ELEMENT  topic (result*)>
<!ATTLIST  topic
     topic-id CDATA #REQUIRED
>
<!ELEMENT  collections     (collection+)>
<!ELEMENT  collection           (#PCDATA)>

<!ELEMENT  result (in?,file, path, rank?, rsv?)>
<!ELEMENT  in   (#PCDATA)>
<!ELEMENT  file (#PCDATA)>
<!ELEMENT  path (#PCDATA)>
<!ELEMENT  rank (#PCDATA)>
<!ELEMENT  rsv  (#PCDATA)>
```

Table 3: The DTD for the submissions in INEX

The submissions used in making the new combined result lists are in this thesis referred to as result sets even though INEX uses the term submission as described above.

3.2 Result sets

Result sets from both the Focussed and the Thorough approach were utilised when constructing the new result lists. The selection of these original result sets and an explanation for the two approaches are included along with a justification for why these results were chosen.

3.2.1 Focussed result sets

The aim of the *Focussed* approach is to find the most exhaustive and specific element in a path. In the focussed result set, there can only be one element along each path, i.e. no overlapping elements can be present. This can be illustrated through a simple XML tree for an article (see Figure 2). Starting from the root node, an example of a path is when a leaf node is reached [article 1 / section 1 / paragraph 1]. Both a child of a node and a parent of a node are along the same path as the node itself.

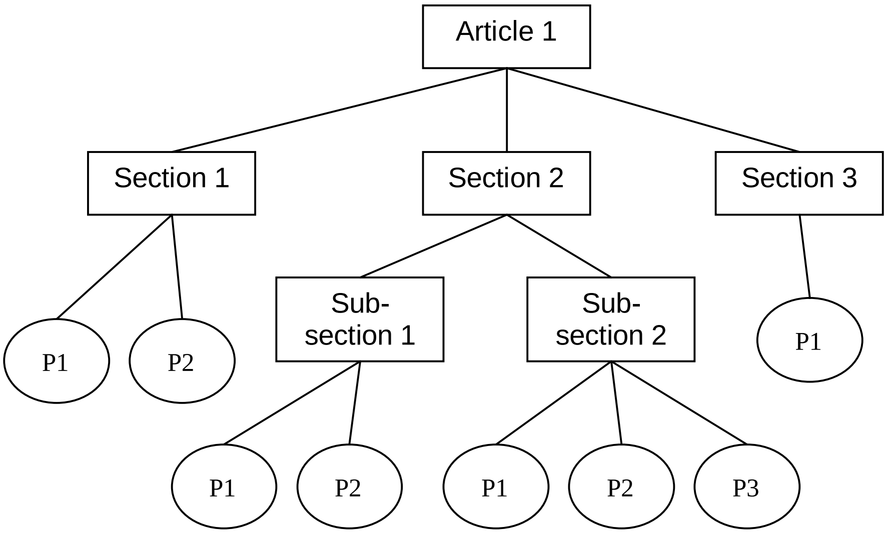

Figure 2: An XML tree showing different paths

In one result set, the following two elements may therefore be present from the same document.

Element 1: [article 1 / section 2 / sub-section 1 / paragraph 1]
Element 2: [article 1 / section 2 / sub-section 1 / paragraph 2]

Even though these elements are both from the same article, the same section and the same sub-section, they are not overlapping as they represent different paths. However, the following element could not be returned in the same result set as the two previous ones:

Element 3: [article 1 / section 2 / sub-section 1]

The reason for this is that this element is overlapping with the others as it is from the same path as the ones above, however in these cases more ‘focussed’ elements are returned.

Results from the Focussed approach were examined in two different ways; on document level and on element level. The first method implies to group the documents which are overlapping in all four result sets, i.e. present in all four sets. The elements for these overlapping documents are then examined and re-ranked. The other way the Focussed results were examined, element level, means to look at element overlap from beginning in order to construct the new result list. It is important to be aware of the algorithm used for constructing the new list, CombMNZ, which does not require elements being present in all four result sets. This means that a situation with the new top-ranked element overlapping in three result sets, is possible.

In the previous paragraphs, two types of overlap are discussed. It is important to be aware of the difference between these two. To specify, when looking at the difference between Focussed and Thorough result sets, one type of overlap is encountered, namely the one which is present in Thorough results. However, this type of overlap is between elements in the same result sets, which means that it has nothing to do with multiple evidence or the polyrepresentation principle. In a simple example below (Table 4), it can be seen that elements from the same result set overlap and are present in the result list. For instance, for section 7 from the first article, different elements within this section are returned as highlighted in the table below:

Document ID	**Element path**
ex/2001/x3066	/article[1]/bdy[1]/sec[7]
ex/2001/x3066	/article[1]/bdy[1]/sec[7]/ss1[1]
ex/2001/x3066	/article[1]/bdy[1]/sec[7]/ss1[1]/ss2[3]
ex/2001/x3066	/article[1]/bdy[1]/sec[3]/ss1[3]/p[1]
ex/2001/x3066	/article[1]/bdy[1]/sec[3]/ss1[1]/p[5]
ex/2001/x3066	/article[1]/bdy[1]/sec[3]/list[1]/item[1]/p[1]

Table 4: Extraction from result set showing overlap between elements

On the other hand, when looking at overlap between result sets, either overlapping elements or documents, the goal is to examine the polyrepresentation principle and finding multiple evidence of an element or a document being relevant. The methods used in this thesis on the result sets from INEX seek to find these overlaps to construct new result sets. This is illustrated in the figure below:

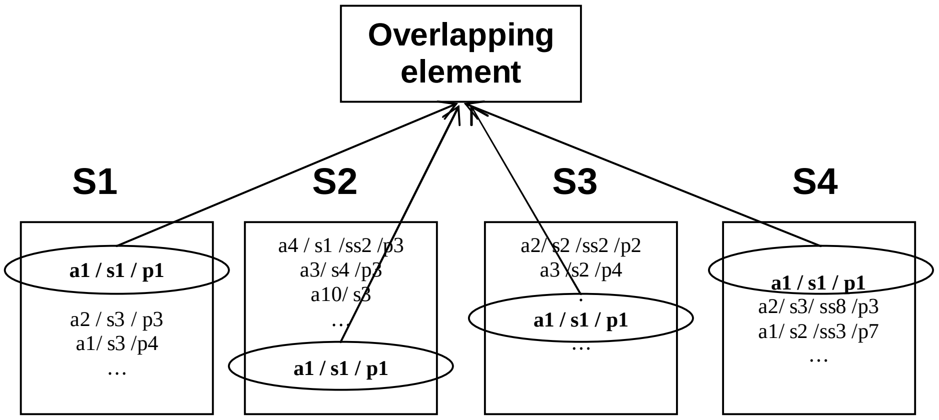

Figure 3: Illustrating overlapping element in four result sets

3.2.2 Thorough result sets

In the *Thorough* approach, on the other hand, the goal is to find all exhaustive and specific elements. Therefore, if a certain element is considered relevant, the parent element of this will also be considered relevant to some degree. Ancestor elements may then also be returned in addition to the element itself and these results will therefore contain overlapping elements. These result sets will contain a great number of overlapping elements. For instance, all of the elements given in the example for Focussed results under Section 3.2.1 would most likely be in the result list. This approach does not address the issue of how to proceed to the final presentation of results to the user.

Examining overlaps for Focussed results can be criticised in the way that important result elements may be left out in the process just because the elements are of varying ‘focussed’ level. To illustrate this point, let us assume that one system (S1) has returned the following element from a certain document:

[article 1 / section 2 / sub-section 1 / paragraph 3]

From the same document, another system (S2) has considered the following element to be relevant and this is the one returned:

[article 1 / section 2 / sub-section 1]

From these results, it is apparent that sub-section 1 in section 2 in this article contains some highly relevant information. From a strict Focussed perspective when performing data fusion, these elements would not overlap and therefore would not appear in a new result list of overlapping elements. This means that one would “lose” important and relevant elements. In other words, the elements in the new result list may not really be the most relevant ones, just because the exact element was not returned by all systems, or in this case all submissions.

Because of these possible issues with Focussed results, it was decided to additionally consider Thorough results in the analysis as these results may be more suitable for fusion of elements. The new constructed lists for Thorough results will naturally contain overlapping elements in the individual result lists. Consequently, several elements along the same path may be present in a result list.

3.2.3 Selecting original result sets

In order to perform fusion of results, it was decided to choose some result sets from INEX 2005 to use as the baseline in this study. Previous research on data fusion suggests that a few result sets or systems used in fusion perform well, however this is not necessarily the case with more sets (e.g. Lund, 2005). Research on polyrepresentation has also found that three or four representations improve performance (Skov et al., 2004). On the basis of this, it was decided to choose four original sets to be used in the fusion.

The result sets were selected based on subjective judgments of the ranked results available through the INEX Web page. These results include ranked lists for all quantisation functions and the selection was based on overall performance for the various result sets. The information available in the submission files also affected the selection of results as some of them were missing score information. To be used in this study, the results needed to have this information available as the scores were used in the re-ranking of elements in the new result lists. The selected result sets for Focussed and Thorough approaches are presented in Table 5 and Table 6 below.

	Participating group	*Name of run*
S1	University of Kaislautern	CO_Pattern_Focussed
S2	IBM Haifa	CO_focused_no_phrase_no_plus_filter_ overlaps_with_clustering_LAREFINEMENT
S3	Queensland University of Technology	CO_3_Focused_highest_VVCAS
S4	University of Oslo	Focussed_co

Table 5: Selected Focussed runs

	Participating group	*Name of run*
S1	Berkeley University	CO_T2FB_PIV50_THR
S2	IBM Haifa	CO_no_phrase_no_plus_LAREFINEMENT
S3	Queensland University of Technology	QUT CO_2_Thorough
S4	University of Kaislautern	CO_Pattern_Thorough_NoERG

Table 6: Selected Thorough runs

The results from the University of Kaislautern are based on the vector-space model with additional use of context patterns to explore patterns in the results and in this way find appropriate result elements (Dopichaj, 2005). In the results from the IBM Haifa group, a component ranking algorithm, referred to as a Lexical Affinity Refinement algorithm, was used (Mass and Mandelbrod, 2005). The selected results (which were the best-performing for this participant) ignored phrases and plus (+) on words. For the Focussed approach, a normal Thorough run is first carried out, followed by a filtering process to get rid of overlaps. The filtering is performed in two stages. In the first filtering process, referred to as “smart filtering”, clusters of highly ranked results are identified and the most relevant element is selected from each of the clusters. In the second stage, a brute-force filtering process is used to remove overlaps which may still be among the results (Mass and Mandelbrod, 2005).

Queensland University of Technology used a fully inverted file structure using XPath in their results. Heuristic algorithms were used to calculate relevance scores for the elements (Geva, 2005). The selected run from Berkeley University is based on probabilistic retrieval algorithms. Blind feedback was utilised with a logistic regression algorithm from TREC2 (Larson, 2005). The results by the University of Oslo calculate the relevance scores at paragraph level by using tf-idf weights.

3.3 The process of combining result sets

In order to have a proper amount of overlap to examine, it was chosen to consider all 1500 results for each topic instead of setting the cut-off value at a lower level. All 40 topics were used in constructing the new lists. This means that the results ready for evaluation are constructed lists for all 40 topics for different runs. The results from INEX 2005 were examined in order to be able to select some of the ones appearing within the top ten in terms of performance for the various evaluation metrics.

As mentioned in the section on background literature (Section 2.0), it is common that different result sets have different range of scores. Even within the same result set, different topics have different types of scores. Consequently, to be able to compare different result sets it is necessary to normalise all scores so that they have a score within the range of 0 to 1. Normalisation was performed on all the result elements giving the top ranked element the score of 1 while the lowest ranked element was assigned 0. This was done for each topic in the result set.

For the Focussed approach, the results were first examined at document level in order to construct the new result lists. The results were grouped according to document IDs, not considering the element paths. This was done in order to find out which documents were overlapping in all four result sets. Naturally, each document id may appear several times in each result set for different elements within the document. For each of these overlapping documents, it was necessary to re-visit the original files to find the elements which were in the result sets for each of the documents. According to the polyrepresentation principle, the overlapping documents are likely to be more relevant than documents that can be found in one result set only. Because the documents are considered relevant based on this principle, the elements in these documents are consequently also expected to be relevant. Then, the CombMNZ could be used to assign new scores to elements in the combined lists.

For the element level approach of constructing the new lists for Focussed results, the CombMNZ score was calculated for all elements which were then re-ranked in the combined lists. In the same way, the new lists for Thorough results were created.

The CombMNZ method is almost considered a standard within data fusion to rank combined results from for instance several systems or several result sets (Beitzel, Jensen, Chowdhury, Grossman, Frieder and Goharian, 2004). The combined result lists were sorted according to this score for the elements. This means that overlap between elements in all four result sets is not a requirement. Certainly, overlap increases the chance of the element receiving a higher score, but the algorithm also takes into consideration original scores received for the elements, no matter how many result sets retrieved the element.

In Lund's research (2005), on the other hand, the documents overlapping in four search engines were ranked first, then the ones overlapping in three engines, until a list of 100 documents was achieved. Multipliers of 1, 25, 50 and 100 were used to re-assign ranks weights for documents overlapping in one, two, three and four engines respectively. This method may not necessarily suit fusion of elements, as can be seen in the example in the next paragraphs. Then, one might have missed important elements when constructing the new lists.

Below is a small example taken after having constructed the new list for Focussed results based on overlapping documents in the four original sets for one topic (Table 7). The overlapping elements with their original rank and their new score and associated rank are included. It can be seen that the highest ranked element is "only" overlapping in three submissions. However, the high score for this element is due to the high scores received in the original results in all result sets (ranked as 7, 2 and 1 in the submissions 1, 3 and 4 respectively). It is also apparent that some elements receive a high score due to the fact that they were present in all four submissions even though the ranks/scores were relatively low compared to the rank/score they now receive. This demonstrates the nature of the weighting of the CombMNZ algorithm.

Rank	S	Document path	Element path	Original rank	Norm. score	Over-laps	CombMNZ
1	1	tk/2003/k0442	/article[1]/bdy[1]/sec[6]/ip1[1]	7	0.597	3	7.789
	3	tk/2003/k0442	/article[1]/bdy[1]/sec[6]/ip1[1]	2	0.999	3	7.789
	4	tk/2003/k0442	/article[1]/bdy[1]/sec[6]/ip1[1]	1	1	3	7.789
2	1	co/2004/r5026	/article[1]/bdy[1]/sec[6]/p[10]	107	0.276	4	6.249
	2	co/2004/r5026	/article[1]/bdy[1]/sec[6]/p[10]	278	0.354	4	6.249
	3	co/2004/r5026	/article[1]/bdy[1]/sec[6]/p[10]	1030	0.0002	4	6.249
	4	co/2004/r5026	/article[1]/bdy[1]/sec[6]/p[10]	2	0.932	4	6.249
3	1	co/2002/rz077	/article[1]/bdy[1]/sec[2]/p[1]	172	0.208	4	5.934
	2	co/2002/rz077	/article[1]/bdy[1]/sec[2]/p[1]	827	0.129	4	5.934
	3	co/2002/rz077	/article[1]/bdy[1]/sec[2]/p[1]	11	0.984	4	5.934
	4	co/2002/rz077	/article[1]/bdy[1]/sec[2]/p[1]	1023	0.162	4	5.934
4	2	ex/1998/x3040	/article[1]/bm[1]/vt[4]/p[1]	208	0.395	3	5.223
	3	ex/1998/x3040	/article[1]/bm[1]/vt[4]/p[1]	8	0.984	3	5.223
	4	ex/1998/x3040	/article[1]/bm[1]/vt[4]/p[1]	387	0.362	3	5.223

Table 7: Extraction of the new combined list for one topic

If, on the other hand, elements overlapping in all four result sets had been given even higher weights, important elements that only appeared in two or three submissions could fail to receive the rank that they deserved or even be completely excluded from the results. Therefore, it was decided to use the chosen algorithm when constructing the new result lists.

3.4 Evaluation

The combined results were evaluated using measures from INEX 2005. The evaluation was also carried out on the selected original result sets for both Focussed and Thorough. This was done in order to be able to compare the results for various measures between the combined results and the originals for that approach. The aim of doing this comparison was to find out if improvement in performance can be achieved with XML results in any of the approaches investigated in this research study.

3.4.1 Evaluation metrics

The official metrics for the INEX workshop in 2005 were the eXtended Cumulated Gain (XCG) measures. Recall from the INEX part in the background literature (see Section 2.6) that relevance in INEX is assessed on the basis of *exhaustivity (e)* and *specificity (s)*. In 2005, there were 3+1 possible levels for the exhaustivity assessment. These include: highly exhaustive (e=2), somewhat exhaustive (e=1), not exhaustive (e=0), and the last one “too small” (e=?). Values for specificity are given on a continuous scale in [0,1] where s=1 indicates a fully specific element. This implies that the element contains only relevant information. These combined values of exhaustivity and specificity are used in specifying to what degree the element is relevant. (Kazai and Lalmas, 2006a)

Different quantisation functions were used in INEX 2005. The functions are either *strict*, *generalised*, or *generalised lifted*. With the strict function, only fully specific and highly exhaustive elements are regarded and included in the evaluation. On the other hand, the generalised function takes into consideration different levels of relevance of elements. However, elements assessed as “too small” are not considered in either the strict or the generalised functions, i.e. e=? is in these cases treated as e=0. In the generalised lifted function, 1 is added to all exhaustivity values unequal to 0. Thus, the lifted quantisation also takes the elements assessed as “too small” into account. (Kazai and Lalmas, 2006a)

The XCG metrics are extensions of the cumulated gain metrics introduced by Järvelin and Kekäläinen (2002) for evaluation of document retrieval. The metrics were developed to be used for multigraded relevance values (Järvelin and Kekäläinen, 2002). Included in the XCG metrics used in INEX are the normalised extended cumulated gain (nxCG) and the effort-precision/gain-recall measures (ep/gr) (Kazai and Lalmas, 2006a). These were used in the evaluation carried out in this research study and will thus be explained in the following sections.

For the evaluation of results based on the Focussed approach, it is necessary to define an ideal recall base. This ideal recall base is part of the full recall base. However, no overlap between relevant reference elements is tolerated. The ideal recall base consists of ideal answers. In other words, these are the elements which ideally should be returned to the user. (Kazai and Lalmas, 2006a)

In INEX, Thorough results are evaluated using the option overlap=off. This implies that overlap is tolerated during evaluation. However, for Focussed results, the option overlap is set on as there is a penalty for overlapping elements in the results. In addition, retrieval of near-misses is partially rewarded as this may provide access to relevant information which would otherwise be lost (Kazai and Lalmas, 2005). For evaluation of Thorough results there is no ideal recall-base because of overlapping elements being present in the result set (Kazai and Lalmas, 2006a). In this analysis, Thorough results were also evaluated using the overlap=on option. In other words, these results were penalised for having overlapping elements in the results.

Normalised extended cumulated gain (nxCG)

The xCG is defined as a vector of accumulated gain. To further explain this measure, one can consider a ranked list of elements with relevance scores for each of these elements. Then, the cumulated gain at rank [*i*], represented by xCG[i], is given as the sum of the scores up to that rank. This result in the following equation for cumulated gain:

$$xCG[i] := \sum_{j=1}^{i} xG[j]$$

Equation 1: Cumulated gain at rank [i]

As an example, the ranking [3, 2, 0, 0, 1, 2, 3, 2, 0] would give a cumulated gain vector of [3, 5, 5, 5, 6, 8, 11, 13, 13]. (Järvelin and Kekäläinen, 2002)

The ideal gain vector, xI, can be obtained for each query by replacing the rank positions with the relevance scores of all elements in the recall-base. These are filled according to decreasing order of relevance. The cumulated ideal gain vector, xCI, can be obtained in the same way as described above. The normalised xCG (nxCG) measure is found by dividing the runs' xCG vectors with the ideal ranking vectors as illustrated below:

$$nxCG[i] := \frac{xCG[i]}{xCI[i]}$$

Equation 2: nxCG measure

At a certain rank, the associated value of nxCG represents "the relative gain the user accumulated up to this rank, compared to the gain he or she could have attained if the system would have produced the optimum best ranking" (Kazai and Lalmas, 2006b, p.14). Ideal performance is reflected by the normalised value of 1.

Effort precision/gain recall (ep/gr)

The ep/gr measure seeks to measure the amount of effort required by the user to accomplish a given level of cumulated gain when examining a given ranking compared to the ideal ranking. Effort-precision (ep) is defined as the following (Kazai and Lalmas, 2006a):

$$ep[r] := \frac{i_{ideal}}{i_{run}}$$

Equation 3: Effort-precision

In this way, the rank position where the cumulated gain of *r* is achieved in the ideal run, is divided by the rank position where the cumulated gain of *r* is achieved by the run. The ideal score for this measure is 1 representing the minimum effort needed by the user to achieve a given level of gain. Effort-precision is calculated at arbitrary gain-recall points and gain-recall, gr, is defined as the following, where *n* is number of relevant documents:

$$gr[i] := \frac{xCG[i]}{xCI[n]} = \frac{\sum_{j=1}^{i} xG[j]}{\sum_{j=1}^{n} xI[j]}$$

Equation 4: Gain-recall

In the equation, the cumulated gain value is divided by the sum of achievable cumulated gain. The ep/gr measure seeks to calculate the relative effort required by the user when scanning a ranking list compared to the effort an ideal ranking would require to accomplish a given level of gain. Effort in this context means number of visited ranks. (Kazai and Lalmas, 2006a)

Interpolation is used in order to measure effort-precision values at standard gain-recall points, not arbitrary points. The un-interpolated mean average effort-precision, referred to as MAep, is obtained by taking the average of effort-precision values at natural recall-points. When examining the ranking, a natural recall-point is when a relevant XML element is found. Relevant elements which are not retrieved are given the score of 0. It is important to be aware of the requirement of interpolation also for calculating MAep. The reason for this is that a run's natural recall points may not correspond to the ideal ranking's natural recall points. The

MAep value is also claimed to be a main overall performance indicator. (Kazai and Lalmas, 2006a)

3.4.2 Statistical tests

As suggested by Järvelin and Kekäläinen (2002), statistical tests were performed on the normalised average CG vectors in order to find out if the differences obtained when evaluating the results were statistically significant. The results from performing the Wilcoxon test on the Thorough results, and the Friedman test on the Focussed results, are reported in the end of next section.

4.0 Results

In the following section, the results from the evaluation of the various combined results will be reported. These combined results will be compared to the original results to examine whether improvements have taken place. Results are reported according to the research questions outlined in Section 2.8.

To be more precise, combined result lists for both element level and document level are included for Focussed results. For the Thorough approach, results for evaluations using both selections on and off for the overlap option, are presented. The original result sets used in making the combined list for Thorough are hereafter referred to as S1-T, S2-T, S3-T and S4-T. Equally, the result sets for the Focussed approach are referred to as S1-F, S2-F, S3-F and S4-F. For more information about the original result sets, please refer to Table 5 and Table 6 in Section 3.2.3.

A total of 29 topics were used in this evaluation. Even though all 40 topics were used in creating the combined lists, only 29 topics from INEX 2005 have relevance assessments available. Therefore, these were the topics that could be used in evaluation. As mentioned in the previous section, different quantisation functions are used when evaluating the results. Here, only the generalised results are reported hence no limitations were employed when selecting the assessed results for evaluation.

4.1 Focussed results

To answer the first research question, Focussed results based on combining on element level, are examined. These combined result sets are compared to the originals, as well as the Focussed results based on combining at document level. The results for these five sets for the nxCG measure are shown in the following graph.

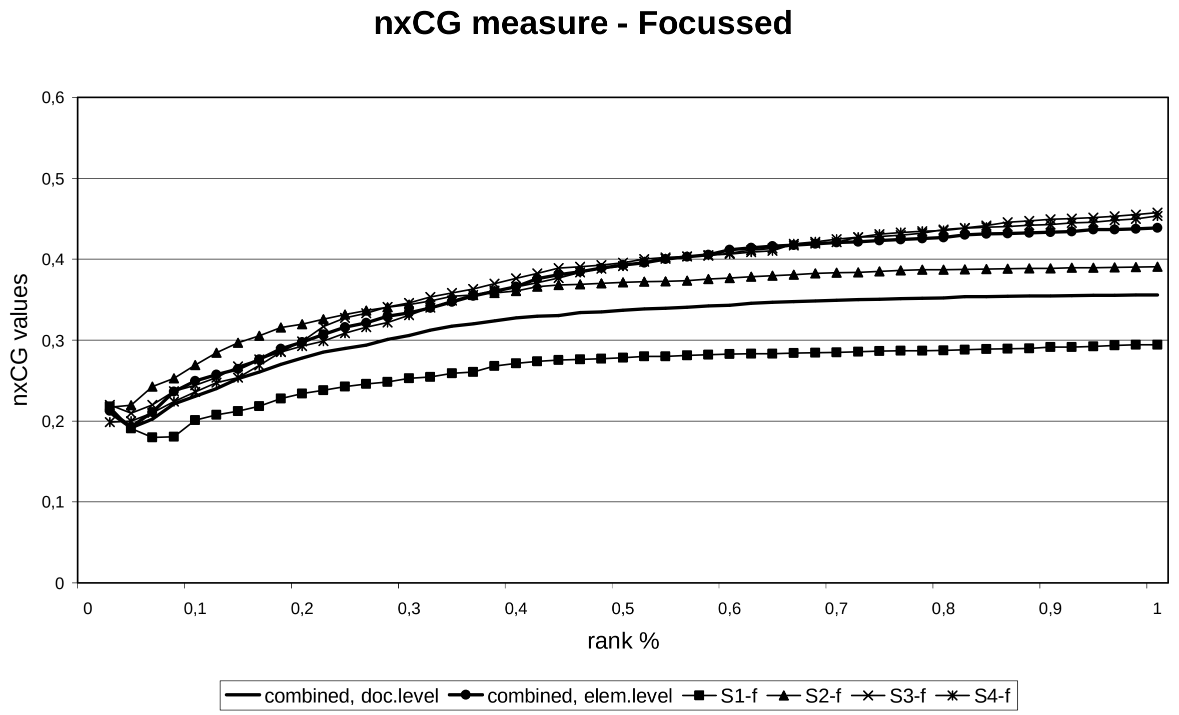

Figure 4: Results for the nxCG measure for Focussed results

The combined results using the approach of comparing at element level receive almost comparable scores as two of the original result sets (S3-F and S4-F) except in the beginning of the examined results. On the other hand, the results from combining using the document level approach obtain poorer scores than all of the others except for one of the original result sets (S1-F).

As with the previous measure, the combined Focussed results for the ep/gr measure are not very different than the original results. Of the results shown in the graph below, the combined results are neither best-performing nor worst-performing, but place themselves somewhere in between the others. Again, the results from combining at element level are better than the results from the document level combination method. At some points, the element level results are better than all the other result sets.

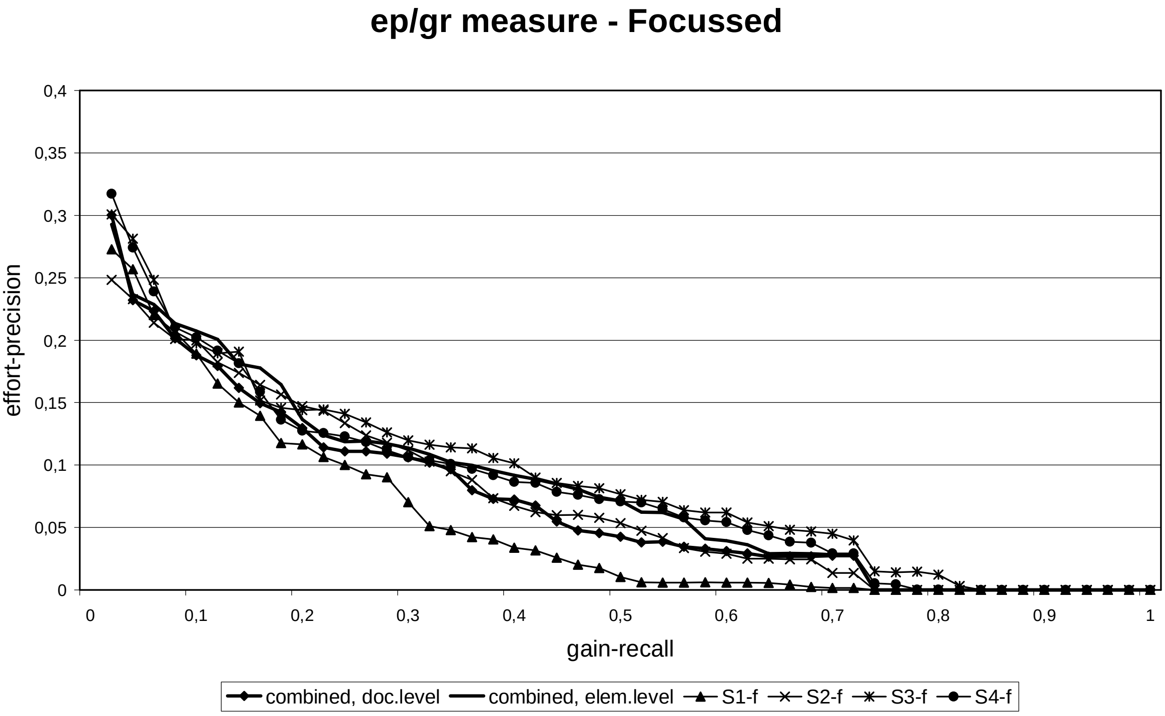

Figure 5: Results for the ep/gr measure for Focussed results

Mean average values for various cut-off points for the nxCG measure show that the combined results, for most cut-off values, are comparable to the originals, however not superior. It is worth noticing that the differences between the two combination approaches are minor, except for larger average values. The results for element level are better than the document level after average values for more than 500 result elements. However, these results are not better than the originals. The results for the two combining approaches are emphasised in the following table.

#MAnxCG@[i]						
	elem. level	***doc. level***	*S1-F*	*S2-F*	*S3-F*	*S4-F*
@1	**0,2276**	**0,2537**	0,2182	0,2517	0,2785	0,3012
@5	**0,2584**	**0,2605**	0,25	0,2556	0,2693	0,2594
@15	**0,241**	**0,2392**	0,2471	0,238	0,2546	0,2374
@25	**0,2311**	**0,2298**	0,2366	0,2293	0,2413	0,2214
@50	**0,2139**	**0,2114**	0,2184	0,2244	0,2251	0,2116
@100	**0,2099**	**0,2056**	0,2007	0,2292	0,2215	0,2095
@500	**0,2785**	**0,2602**	0,2234	0,2978	0,285	0,2728
@1000	**0,3345**	**0,298**	0,2501	0,3339	0,3394	0,3309
@1500	**0,3661**	**0,3163**	0,2631	0,3517	0,373	0,3667

Table 8: Mean average values for nxCG measure for Focussed results

Mean average effort precision values for the Focussed approach are reported in the table below. The combined results are not superior in comparison with each of the original result sets. The results for the element level combination are better than the results of the document level combination. The results of two of the original sets, S2-F and S3-F, are comparable to the results from the element level combination.

#MAep					
elem. level	***doc. level***	*S1-F*	*S2-F*	*S3-F*	*S4-F*
0,089758	**0,081473**	0,067058	0,091733	0,088899	0,081524

Table 9: Mean average effort precision for Focussed results

4.2 Thorough results

The second research question addresses fusion of Thorough result sets. From the evaluation using the nxCG measure, it is obvious that the combined Thorough result performs better in comparison to the four original result sets. This can be seen in the following graph.

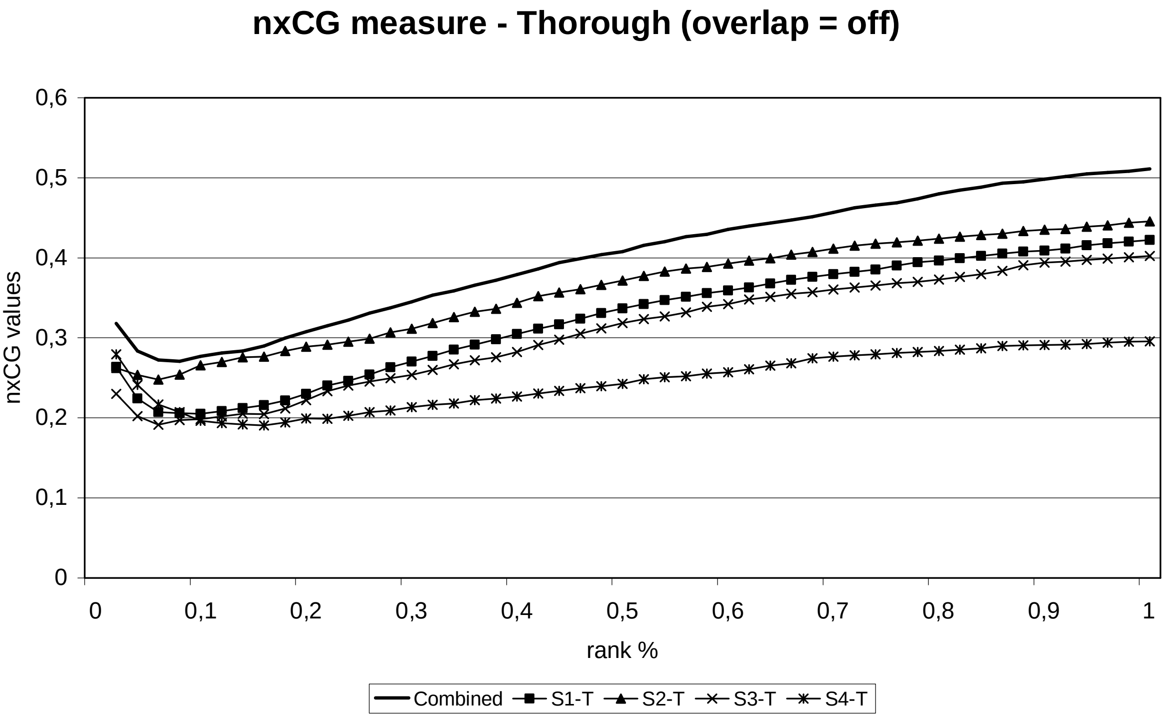

Figure 6: Results for the nxCG measure for Thorough results, overlap=off

As one would expect, the results using the overlap=on option receive lower scores. This is due to the fact that overlapping elements are being penalised in the result set. The combined results perform fairly good also in these cases. However, from the graph below it is apparent that S1-T achieves higher performance and S3-T achieves comparable scores as the combined results throughout the result list.

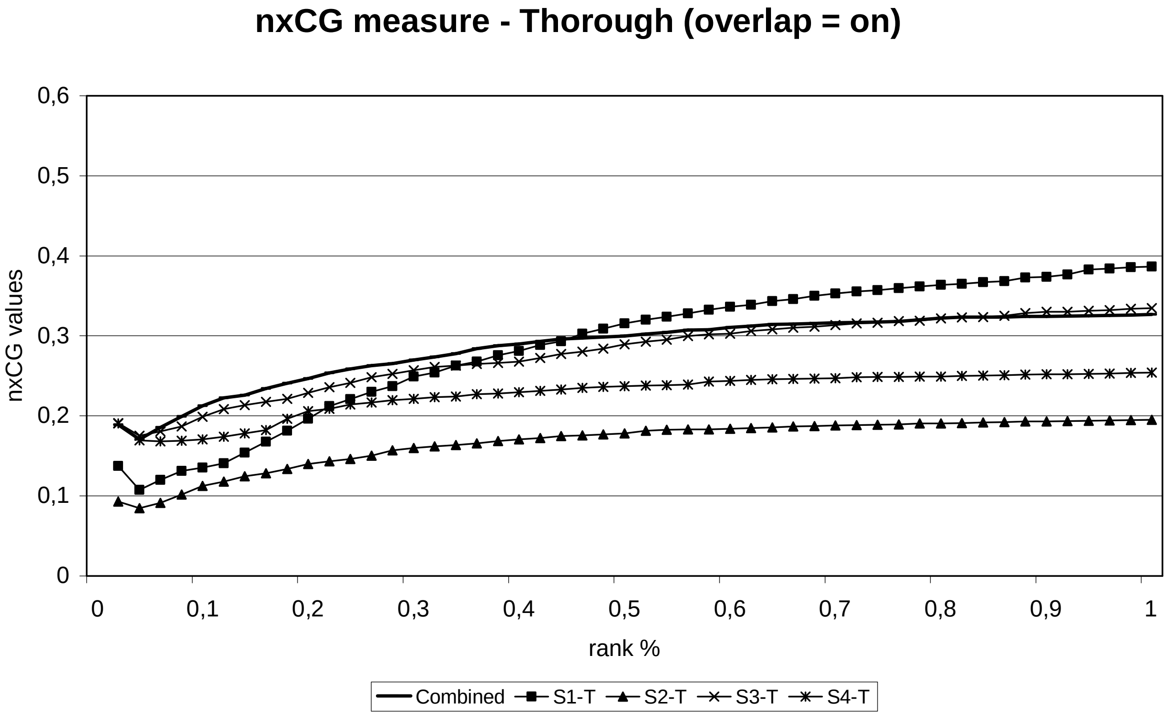

Figure 7: Results for the nxCG measure for Thorough results, overlap=on

When examining the results from the ep/gr measure, the combined Thorough results obtain superior scores in comparison with the original result sets. As can be seen from the graph below, the combined result set is continuously better than all originals.

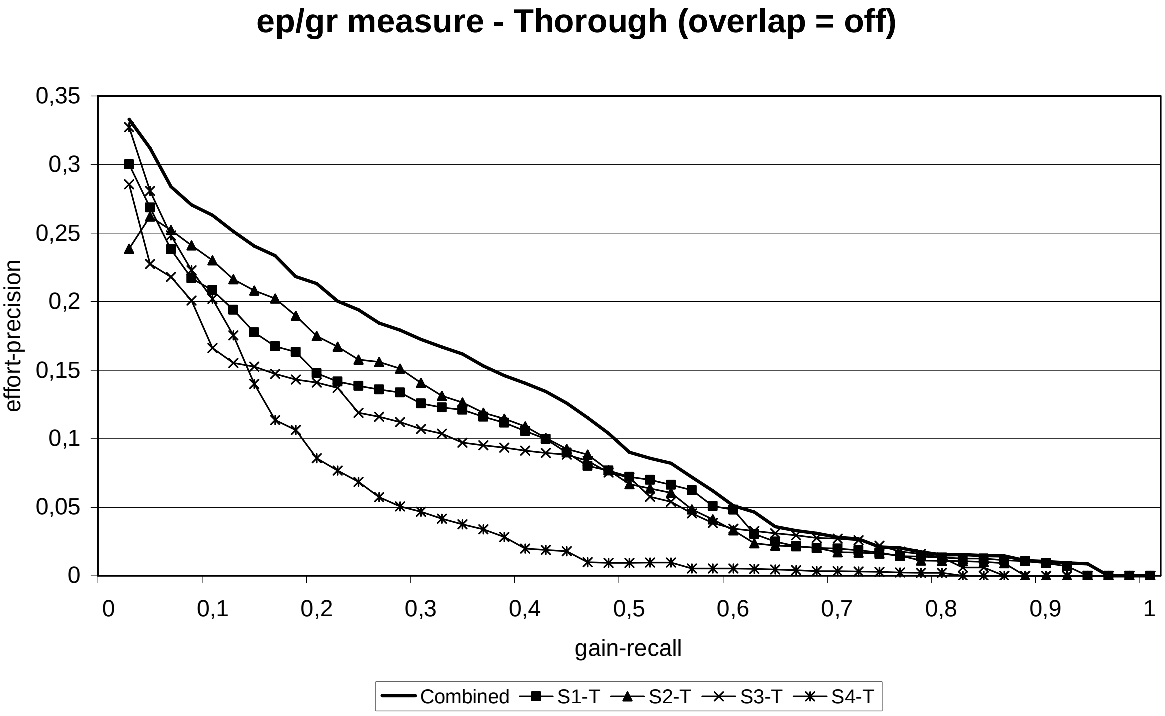

Figure 8: Results for the ep/gr measure for Thorough results, overlap=off

For the same measure for Thorough when there is a penalty at the evaluation phase for overlapping elements, the combined results do not perform better than the others. However, the results are at a comparable level to the best-performing original results.

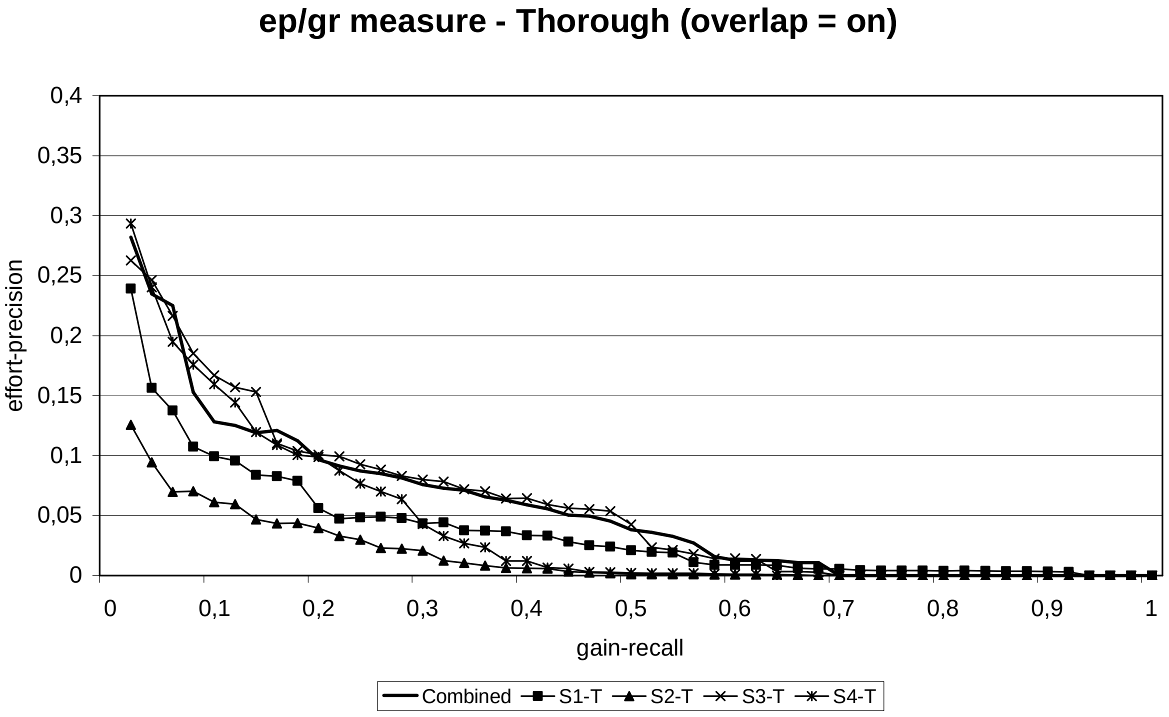

Figure 9: Results for the ep/gr measure for Thorough results, overlap=on

When looking at mean average values for various cut-off points for the nxCG measure, the results from evaluating the combined Thorough results are very promising. These results with the option overlap=off are shown in the table below. The scores for the combined results are emphasised.

MAnxCG[i], overlap = off					
	Combined	*S1-T*	*S2-T*	*S3-T*	*S4-T*
@1	**0,3492**	0,2816	0,1888	0,2918	0,3073
@5	**0,3365**	0,2868	0,2277	0,2748	0,3245
@15	**0,3271**	0,2825	0,2458	0,2552	0,3083
@25	**0,3241**	0,2756	0,2531	0,247	0,2986
@50	**0,3121**	0,2597	0,2554	0,2295	0,2772
@100	**0,2934**	0,2357	0,252	0,2122	0,2502
@500	**0,3071**	0,2363	0,2814	0,2234	0,2117
@1000	**0,358**	0,2853	0,3259	0,2694	0,2278
@1500	**0,4004**	0,3238	0,3598	0,3065	0,2473

Table 10: Mean average values for nxCG measure for Thorough results, overlap=off

The same results from evaluation with overlap=on shows that the combined set is performing well. However, the differences between the combined and the originals are not as noteworthy as above. When comparing the different results, some of the values for the originals are comparable or slightly better than the combined results. This is not the case for the results when overlap=off.

MAnxCG[i], overlap = on					
	Combined	*S1-T*	*S2-T*	*S3-T*	*S4-T*
@1	**0,2803**	0,2409	0,1949	0,2452	0,29
@5	**0,2678**	0,2112	0,1537	0,2402	0,2771
@15	**0,2452**	0,1753	0,132	0,224	0,2412
@25	**0,2247**	0,1613	0,1176	0,2125	0,2227
@50	**0,1987**	0,14	0,1022	0,1954	0,198
@100	**0,1906**	0,1281	0,0963	0,1877	0,1832
@500	**0,234**	0,1823	0,1295	0,2214	0,1962
@1000	**0,2673**	0,2467	0,1537	0,2547	0,2166
@1500	**0,2855**	0,2872	0,1663	0,2777	0,2279

Table 11: Mean average values for nxCG measure for Thorough results, overlap=on

Table 12 shows mean average effort precision values for Thorough runs with overlap = off. As with most of the other measures, the Thorough results for the combined set are better than each of the original result sets.

#MAep, overlap = off				
combined	*S1-T*	*S2-T*	*S3-T*	*S4-T*
0,102586	0,068674	0,086744	0,070564	0,051114

Table 12: Mean average effort precision for Thorough results, overlap=off

Furthermore, for the overlap=on option the same results show that the combined set performs better than the original sets.

#MAep, overlap = on				
combined	*S1-T*	*S2-T*	*S3-T*	*S4-T*
0,072523	0,046010	0,029123	0,066334	0,064787

Table 13: Mean average effort precision for Thorough results, overlap=on

4.3 Statistics

From the results reported in the preceding sections, it is apparent that improved results are achieved with combining Thorough result sets. The evaluation of Focussed results demonstrates more varying results and the trend of retrieval improvement is not present.

When examining the Focussed results based on the element level approach, the fused result performs significantly better than one of the original result sets, S1. However, compared to the other result sets there are no significant findings.

	#MAnxCG@1500	Difference to Fused-F-e
Fused-F-element	0,3661	
S1-T-off	0,2631	0,103**
S2-T-off	0,3517	0,0144
S3-T-off	0,373	-0,0069
S4-T-off	0,3667	-0,0006

Legend: * p< 0.05, ** p<0.01

Table 14: #MAnxCG@1500 values for Focussed on element level

For the document level approach, the results are more varying than the previous results. Here, two of the original result sets, S2 and S3, receive higher scores than the fused result set, and for one of these, S3, the difference is significant. This means that the original result set performed significantly better in comparison to the fused result set. Nevertheless, the fused result set is significantly better than S1.

	#MAnxCG@1500	Difference to Fused-F-d
Fused-F-document	0,3163	
S1-T-off	0,2631	0,0532**
S2-T-off	0,3517	-0,0354
S3-T-off	0,373	-0,0567*
S4-T-off	0,3667	-0,0504

Legend: * p< 0.05, ** p<0.01

Table 15: #MAnxCG@1500 values for Focussed on document level

From performing the Wilcoxon test for Thorough result sets where overlap is set off, it can be seen that all the differences between the originals and the fused result set are significant (Table 16). In other words, the fused result set performed significantly better than each of the original result sets.

	#MAnxCG@1500	Difference to Fused-T-off
Fused-T-off	0,4004	
S1-T-off	0,3238	0,0766**
S2-T-off	0,3598	0,0406**
S3-T-off	0,3065	0,0939**
S4-T-off	0,2473	0,1531**

Legend: * p< 0.05, ** p<0.01

Table 16: #MAnxCG@1500 values for Thorough, overlap=off

For Thorough results when overlapping elements are penalised during evaluation, the Wilcoxon test show significant differences between two of the original sets and the fused result set. This means that the fused result set performed significantly better than two of the original sets.

	#MAnxCG@1500	Difference to Fused-T-on
Fused-T-on	0,2855	
S1-T-on	0,2872	-0,0017
S2-T-on	0,1663	0,1192**
S3-T-on	0,2777	0,0078
S4-T-on	0,2279	0,0576**

Legend: * p< 0.05, ** p<0.01

Table 17: #MAnxCG@1500 values for Thorough, overlap=on

5.0 Discussion

In this chapter, the results from the previous section will be investigated. Possible explanations for the results achieved in this thesis study will be presented and examined.

From the evaluation, it is apparent that superior results for the combined sets are achieved for Thorough results. On the other hand, for Focussed results this improvement can not be observed. For most of the measures, the combined Focussed sets position themselves somewhere in between the best- and worst-performing originals. In comparison with the combined results at document level, the element level ones continuously receive better scores, with the exception of a couple of the early mean average values for the nxCG measure (see Table 8).

The graphs for the nxCG measure for Thorough results demonstrate that the fused results with overlap off perform better than with overlap on in comparison to the original result sets. As expected, the scores in the latter case will be lower than those received in the former case. The reason for this is that overlapping elements are penalised during evaluation when overlap is on. However, the results from this evaluation show not only lower scores, but also that the overall performance of the combined results is not better than all of the original results.

For the Focussed approach, the fused results for the element level combination constantly perform better than the other fused results. A possible reason for this may be that, in some cases, important elements are missed when combining elements from documents overlapping in all original result sets. If there for instance is an element which can be found in three original result sets for one document with fairly good scores, it will not be overlapping in all result sets (if the document has no other elements from which each of the runs have retrieved an element). However, if considering the results at element level, this element will most likely receive a high score due to the nature of the CombMNZ algorithm and how it weights the elements in the combined result set.

As with the nxCG measure, the Thorough combined results receive higher scores for the ep/gr measure than each of the original result sets when overlap is off in the evaluation. Again, when penalising overlapping elements, the trend for the combined results is not so clear. However, the results are still comparable to the best-performing original result set S3-T.

These results indicate that there might be a relatively large amount of overlapping elements in the combined result set.

As mentioned in the data and methods chapter (see Section 3.0), the Thorough results were included in the fusion due to certain drawbacks with combining Focussed results. Overlap between elements from different result sets is smaller because the Focussed approach can only have one element along each path. When combining elements to construct a new result list, relevant elements may then receive a lower score because the specific elements are not overlapping in the different result sets. This will not be the case with results from the Thorough approach as overlapping elements can be present in the result lists.

Overall, the results from the analysis in this thesis are very promising. Clearly, improved results can be obtained when performing fusion on result sets with XML elements.

5.1 Suggestions for further research

For future studies, it would be useful to select different combinations of original result sets than the ones used in this study. As highlighted by Lund (2005), a poor performing result set can influence the combination negatively and the combined result list is affected accordingly. A more thorough selection process could therefore be the basis for selecting best performing result sets to examine if these produce better combined result sets. In this selection process, results from various evaluation metrics could be analysed.

Furthermore, various combining algorithms, other than the CombMNZ used in this study, could be used in the construction of new result sets. Elements overlapping in all four result sets could be ranked higher than the ones appearing in three result sets, independent of the ranks of the element in each of the original result set. In this thesis, four result sets were selected, however future research could explore various number of result sets to see how this affects performance in fusion of XML elements.

6.0 Conclusion

This thesis study has investigated the theory of polyrepresentation by performing data fusion of XML results. Result elements from four different result sets were combined. The combined results and the original result sets were evaluated using well-known measures from INEX. Both Focussed and Thorough result sets from INEX 2005 were used in the fusion. In addition, the combined results and the original results were evaluated and compared to each other.

The results show that the combined result sets for Thorough results perform significantly better than the original result sets when overlap is tolerated in the evaluation. Improved results are also achieved for some measures when overlapping elements within the combined result sets are penalised. In other words, improved effectiveness is achieved by utilising the polyrepresentation principle stating that overlap between different representations, in this case result sets, will contain a higher number of relevant documents, in this case elements, than in each of the original result sets.

Generally, the trend of improved effectiveness for Thorough results was not present when analysing the Focussed results. As highlighted earlier in this thesis, the reason for this could be that one is missing out on important result elements when combining Focussed results without overlapping elements. This is not the case for Thorough results as these can contain more than one element along each path. Therefore, it can be beneficial to carry out the combination process for these results and then perform some operations in order to “focus” the results so that they can be presented to a user.

The results from this study show that the polyrepresentation principle is valid also for XML elements and that fusion of these can improve retrieval effectiveness. This finding is beneficial for IR research where XML retrieval has received increased attention during recent years. The attempt to constantly improve retrieval effectiveness and present the most relevant texts to the user can benefit from the results obtained in research on the theory of polyrepresentation.

7.0 References

Arvola, P., Junkkari, M. and Kekäläinen, J. (2005) Generalized contextualization method for XML information retrieval, *Proceedings of the 14th ACM international conference on Information and knowledge management*, Oct 31-Nov 05, 2005, Bremen, Germany.

Baeza-Yates, R. and Lalmas, M. (2006) XML Information Retrieval, *Tutorial at the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, Aug 6, Seattle, USA.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O. and Goharian, N. (2004) Fusion of effective retrieval strategies in the same information retrieval system, *Journal of the American Society for Information Science and Technology*, 55(10), pp.859-868.

Christoffersen, M. (2004) Identifying core documents with a multiple evidence relevance filter, *Scientometrics*, 61(3), pp 385-394.

Dopichaj, P. (2005) The university of kaiserslautern at INEX 2005, In Fuhr et al. (eds) *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Dagstuhl*, 28-30 November 2005, Lecture Notes in Computer Science, vol. 3977, Springer-Verlag , pp.196-210.

Fox, E.A. and Shaw, J.A. (1994) Combination of multiple searches, *Proceedings of the 2nd text Retrieval Conference (TREC-2)*, National Institute of Standards and Technology Special publication, pp.243-252.

Geva, S. (2005) GPX – Gardens Point XML IR at INEX 2005, In Fuhr et al. (eds.) *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Dagstuhl*, 28-30 November 2005, Lecture Notes in Computer Science, vol. 3977, Springer-Verlag , pp.240-253.

Hjørland, B. (2006) Polyrepresentation, available online: http://www.db.dk/bh/lifeboat_ko/CONCEPTS/polyrepresentation.htm [accessed 27.09.2006]

Hsu, F.D. and Taksa, I. (2005) Comparing rank and score combination methods for data fusion in information retrieval, *Information Retrieval*, 8(3), pp.449-480.

INEX (2006) *Initiative for the evaluation of XML retrieval*, available online: http://inex.is.informatik.uni-duisburg.de/2006/ [accessed 30.11.2006]

INEX (2005) *Initiative for the evaluation of XML retrieval*, available online: http://inex.is.informatik.uni-duisburg.de/2005/ [accessed 12.04.2007]

Ingwersen, P. (1994). Polyrepresentation for information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In: *Croft, W. B. & van Rijsbergen, C. J. (eds.). Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval*, 3-6 July 1994. Dublin, Ireland. Lindon: Springer-Verlag, pp. 101-110.

Ingwersen, P. and Järvelin, K. (2005) *The Turn: Integration of Information Seeking and Retrieval in Context,* Dordrecht, The Netherlands: Springer.

Järvelin, K. and Kekäläinen, J. (2002) Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems,* 20(4), Oct. 2002, pp.422-446.

Kazai, G. and Lalmas, M. (2005) Notes on what to measure in INEX, *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, Glasgow, July 2005.

Kazai, G. and Lalmas, M. (2006a) INEX 2005 Evaluation metrics, Advances in XML Information Retrieval and Evaluation: Fourth Worskshop of the Initiative for the evaluation of XML Retrieval (INEX 2005), Schloss Dagstuhl, 28-30 Nov, Springer-Verlag, *Lecture Notes in Computer Science*, vol.3977, pp.16-29, 2006.

Kazai, G. and Lalmas, M. (2006b) Extended cumulated gain measures for the evaluation of content-oriented XML retrieval, *ACM Transactions on Information Systems*, 24(4), Oct, pp.503-542.

Kazai, G., Lalmas, M. and de Vries, A. (2004a) The overlap problem in content-oriented XML retrieval evaluation, In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.72-79.

Kazai, G., Lalmas, M., Fuhr, N. and Gövert, N. (2004b) A report on the first year of the initiative for the evaluation of XML retrieval (INEX 2002), *Journal of the American Society for Information Science and Technology,* 55(6), pp.551–556.

Lalmas, M. (2005) INEX 2005 Retrieval task and result submission specification, *INEX*, June 20.

Larsen, B. (2002) Exploiting citation overlaps for information retrieval – generating a boomerang effect from the network of scientific papers, *Scientometrics*, 54(2), pp.155-178.

Larsen, B. and Ingwersen, P. (2002) The boomerang effect: retrieving scientific documents via the network of references and citations, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* New York, ACM, pp. 397-398.

Larsen, B., Ingwersen, P. and Kekäläinen, J. (2006) The polyrepresentation continuum in IR, In: *Proceedings of the first symposium on Information Interaction in Context (IIIiX)*, 18-20 Oct, 2006, Copenhagen, Denmark.

Larson, R.R. (2005) Probabilistic retrieval, component fusion and blind feedback for XML retrieval, In Fuhr et al. (eds) *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Dagstuhl*, 28-30 November 2005, Lecture Notes in Computer Science, vol. 3977, Springer-Verlag , pp.225-239.

Lee, J.H. (1997) Analysis of multiple evidence combination, In N.J.Belkin, A.Desai Narasimhalu, P.Willett, and W.Hersh (eds.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* New York, ACM, pp.267-275.

Lund, B.R. (2005) Polyrepræsentation og datafusion: test af teorien om polyrepræsentation gennem forsøg med fusion af TREC-5 resultater, Danmarks Biblioteksskole (MSc Thesis).

Malik, S., Kazai, G., Lalmas, M. and Fuhr, N. (2006) Overview of INEX 2005, In: *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Dagstuhl*, 28-30 November 2005, Lecture Notes in Computer Science, vol. 3977, Springer-Verlag , pp.1–15.

Mass, Y. and Mandelbrod, M. (2005) Using the INEX environment as a testbed for various user models for XML retrieval, In Fuhr et al. (eds.) *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Dagstuhl*, 28-30 November 2005, Lecture Notes in Computer Science, vol. 3977, Springer-Verlag , pp.187-195.

Pehcevski, J., Thom, J. and Vercoustre, A.-M. (2006) XML retrieval evaluation revisited: a comparison of metrics, Schloss Dagstuhl, Dec 19.

Robertson, S.E. and Hancock-Beaulieu, M.M. (1992) On the evaluation of IR systems, *Information Processing & Management*, 28(4), pp.457-466.

Skov, M., Larsen, B. and Ingwersen, P. (2006) Inter and intra-document contexts applied in polyrepresentation, In: *Proceedings of the first symposium on Information Interaction in Context (IIiX)*, 18-20 Oct, 2006, Copenhagen, Denmark, pp.97-101.

Skov, M., Pedersen, H., Larsen, B. and Ingwersen, P. (2004) Testing the principle of polyrepresentation, In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, ACM, pp.47-49.

Smeaton, A.F. (1998) Independence of contributing retrieval strategies in data fusion for effective information retrieval, In: *Proceedings of the 20th BCS-IRSG Colloquium*, France, 1998, Bonn: Springer-Verlag, Electronic Workshops in Computing, pp.268-278.

Wu, S. and McClean, S. (2006a) Improving high accuracy retrieval by eliminating the uneven correlation effect in data fusion, *Journal of the Americal Society for Information Science and Technology,* 57(14), pp.1962-1973.

Wu, S. and McClean, S. (2006b) Performance prediction of data fusion for information retrieval, *Information Processing & Management,* 42(4), pp.899-915.

W3C (2006) Extensible Markup Language (XML), *W3C*, available online: http://www.w3.org/XML/ [accessed 01.12.2006]