IR of XML Documents – A Collective Ranking Strategy

Maha Salem[1], Alan Woodley[2], and Shlomo Geva[2]

[1] Faculty of Electrical Engineering, Computer Science and Mathematics,
University of Paderborn,
Warburger Str. 100, 33098 Paderborn, Germany
MahaSalem@web.de
[2] Centre for Information Technology Innovation,
Faculty of Information Technology,
Queensland University of Technology,
GPO Box 2434, Brisbane Q 4001, Australia
ap.woodley@student.qut.edu.au, s.geva@qut.edu.au

Abstract. Within the area of Information Retrieval (IR) the importance of appropriate ranking of results has increased markedly. The importance is magnified in the case of systems dedicated to XML retrieval, since users of these systems expect the retrieval of highly relevant and highly precise components, instead of the retrieval of entire documents. As an international, coordinated effort to evaluate the performance of Information Retrieval systems, the Initiative for the Evaluation of XML Retrieval (INEX) encourages participating organisation to run queries on their search engines and to submit their result for the annual INEX workshop. In previous INEX workshops the submitted results were manually assessed by participants and the search engines were ranked in terms of performance. This paper presents a Collective Ranking Strategy that outperforms all search engines it is based on. Moreover it provides a system that is trying to facilitate the ranking of participating search engines.

1 Introduction

Modern society unceasingly produces and uses information. To find the relevant information sought within the huge mass of information now available becomes ever more difficult. If information is supposed to be accessible it must be organised [1].

The specific nature of information has called for the development of many new tools and techniques for information retrieval (IR). Modern IR deals with storage, organisation and access to text, as well as multimedia information resources [2].

Within the area of information retrieval, keyword search querying has emerged as one of the most effective paradigms for IR, especially over HTML documents in the World Wide Web. One of the main advantages of keyword search querying is its simplicity – users do not have to learn a complex query language and can issue queries without any prior knowledge about the structure of the underlying data. Since the keyword search query interface is very flexible, queries may not always be precise and can potentially return a large number of query results, especially in large document collections. As a consequence, an important requirement for keyword search is to rank the query results so that the most relevant results appear first.

N. Fuhr et al. (Eds.): INEX 2004, LNCS 3493, pp. 113 – 126, 2005.

Despite the success of HTML-based keyword search engines, certain limitations of the HTML data model make such systems ineffective in many domains. These limitations stem from the fact that HTML is a presentation language and hence cannot capture much semantics. The XML (eXtensible Markup Language) data model addresses this limitation by allowing for extensible element tags, which can be arbitrarily nested to capture additional semantics. Information such as titles, references, sections and sub-sections are explicitly captured using nested, application specific XML tags, which is not possible using HTML.

Given the nested, extensible element tags supported by XML, it is natural to exploit this information for querying. One approach is to use sophisticated query languages based on XPath to query XML documents. While this approach can be very effective in some cases, a disadvantage is that users have to learn a complex query language and understand the schema of underlying XML.

Information retrieval over hierarchical XML documents, in contrast to conceptually flat HTML documents, introduces many new challenges. First, XML queries do not always return whole documents, but can return arbitrarily nested XML elements that contain the information needed by the user. Generally, returning the "deepest" node usually gives more context information, ignoring presentation or other superfluous nodes. Second, XML and HTML queries differ in how query results are ranked. HTML search engines usually rank entire documents partly based on their hyperlinked structure [3]. Since XML queries can return nested elements, not just entire XML documents, ranking has to be done at the granularity of XML elements, which requires complicated computing due to the fact that the semantics of containment links (relating parent and child elements) is very different from that of hyperlinks. As a consequence, traditional information retrieval techniques for computing rankings may not be directly applicable for nested XML elements [4].

This paper presents an approach for effective ranking of XML result elements in response to a user query by considering the results of several other search engines and producing a collective ranking on the basis of some sort of a vote. The hypothesis is that the resulting system will outperform all search engines delivering the results it is based on.

1.1 Overview of INEX

XML retrieval systems exploit the logical structure of documents, which is explicitly represented by the XML markup, to retrieve document components, instead of entire documents, in response to a user query. This means that an XML retrieval system needs not only to find relevant information in the XML documents, but also determine the appropriate level of granularity to return to the user, and this with respect to both content and structural conditions [5]. The expansion in the field of information retrieval caused the need to evaluate the effectiveness of the developed XML retrieval systems.

To facilitate research in XML information retrieval systems the **IN**itiative for the **E**valuation of **X**ML Retrieval (**INEX**) has established an international, coordinated effort to promote evaluation procedures for content-based XML retrieval. INEX provides a means, in the form of a large XML test collection and appropriate scoring scheme for the evaluation of XML retrieval systems [6]. The test collection consists

of XML documents, predefined queries and assessments. Topics are the queries submitted to the retrieval systems. Their formats are based on the topics used in Text Retrieval Conference (TREC), however, they were modified to accommodate two types of topics used in INEX: **CO** (Content Only, queries that ignore document structure and only contain content requirements) and **CAS** (Content and Structure, queries whose stipulations explicitly refer to the documents' structure).

The scoring scheme is based upon two dimensions: *specificity* (reflects the relevancy of a particular XML component) and *exhaustiveness* (measures whether a relevant component contains suitable coverage). These values are quantised to the traditional metrics of *precision* (the probability that a result element viewed by a user is relevant) and *recall* (total number of known relevant components returned divided by the total number of known relevant components). In INEX two different quantisation functions are used to calculate the relevancy: $\mathbf{f}_{\mathbf{strict}}$ evaluates whether a component is highly focused and highly relevant. Alternatively, $\mathbf{f}_{\mathbf{generalised}}$ evaluates a component's degree of relevance. These metrics are combined to form a recall/precision curve.

Together they provide a means for qualitative and quantitative comparison between the various competitors participating at INEX. Each year the competitors' systems are ranked according to their overall effectiveness.

2 Ranking of Results

Ranking of results has a major impact on users' satisfaction with search engines and their success in retrieving relevant documents. While searches may retrieve thousands of hits, search engine developers claim their systems place items that best match the search query at the top of the results list.

Since users often do not have time to explore more than the top few results returned, it is very important for a search engine to be able to rank the best results near the top of all returned results. A study conducted by [7] indicates that 80% of users only view the first two pages of results. The user may consider a number of factors in deciding whether or not to retrieve a document. Regardless of relevance-ranking theory, users have an intuitive sense of how well the relevance ranking is working, and a key indicator of this intuitive satisfaction is the number of distinct query words that a document contains. For example, a document containing only two query words from an eight-word query should not be higher ranked than a document containing all eight words [8].

2.1 Collective Ranking

As described before, the INEX workshop is run once a year and is generally based on the following steps:

1. Participating organisations contribute topics (end user queries) and a subset of topics is selected for evaluation.
2. The topics are distributed to participants who run their search engines and produce a ranked list of results for each topic.
3. The results are pooled together (disassociated and duplicates eliminated).

4. The pooled results are individually assessed by the original topic contributors, who act as end users manually assessing the relevance of the results in terms of exhaustiveness and specificity.
5. The search engines are ranked in terms of performance (recall/precision) using several metrics.
6. Results are returned to participants who in turn write up and present their systems and discuss it at the workshop.

During the last two years the execution of step 4 (assessment of topics by human assessors) has emerged as a very time-consuming procedure which led to the idea of a "Collective Ranking Strategy". The idea is to take the entire set of results from all search engines and produce a collective ("committee") ranking by taking some sort of a vote. These approaches are often referred to as "data fusion" or "meta search".

The collectively ranked results are to be evaluated against the assessed pool of results (as determined by the human assessors). The hypothesis is that it may be possible to outperform any single system by taking account of the results from all systems. If this hypothesis is verified, then as a consequence manual assessment of pooled results by human assessors (step 4) may be no longer required. Instead, a relative comparison of submissions with the collective ranking results will be sufficient to derive a ranking of all search engines. Moreover, this would also prove the assumption that you can derive a better performing search engine by solely considering results of several other search engines.

2.2 Strategy

Several strategies were tested and the specifics of the Collective Ranking were determined. During the testing phase it became obvious that the simplest strategy led to best results. The eventually applied Collective Ranking Strategy is described by the following algorithm, which is to be applied separately for both CAS and CO topics:

For each topic $t_1 \ldots t_n$:

For each submission $s_1 \ldots s_m$ for topic t_i:

Take the top x result elements;

Assign a value $\mathbf{p_i}$ (points for ranking position) to each rank $r_i \in [1 \ldots x]$ of the top x submitted result elements applying the following formula:

$$\mathbf{p_i := (x - r_i) + 1;}$$

Compute a total result element score $\mathbf{res\_score_i}$ for each unique submitted result element as follows (m being the number of submissions):

$$\forall \mathbf{x_i} \in \mathbf{result\ elements:}\quad \mathbf{res\_score_i} := \sum_{i=1}^{m} \mathbf{p_i}^{k}$$

Rank the result elements according to the assigned result element score $\mathrm{res\_score_i}$ in descending order;

```
In the format of a submission file write the top 1500 re-
sult elements of the ranked list into the Collective Rank-
ing output file;
```

Within this algorithm, x (≙ number of top ranked result elements taken from each submission for each topic) and k (≙ weighting for $\mathbf{p_i}$) are variables whose optimal values are to be determined according to the best possible results in the testing phase.

The central idea of this strategy is to take into account both the number of *occurrences* and the *ranking* position of each result element submitted by the participants' search engines. With reference to the number of occurrences, the summation in the algorithm makes sure that, the more frequently an element occurs in the submitted result lists of various search engines, the higher it is rated and eventually ranked. This becomes evident considering an implication provided by the algorithm: If a particular result element is not returned in a search engine's top 100 results, it receives 0 points for the ranking position ($\mathbf{p_i}$). With respect to the consideration of the ranking position, the definition and incorporation of $\mathbf{p_i}$ (points for ranking position) makes sure that, the higher the ranking position of the same result element in each submitted result list is, the bigger the value $\mathbf{p_i}$ for each occurrence will be.

As the Collective Ranking is derived from a descending list of the top 1500 result element scores, the bigger the value $\mathbf{p_i}$ for each occurrence of a particular result element and as a consequence the bigger the result element score res_score_i (derived from the summation of $\mathbf{p_i}$) is, the better the final ranking position of this particular result element in the Collective Ranking will be.

3 Testing

The Collective Ranking algorithm was tested using different values for the variables x and k, in order to identify the optimal combination of these values, that is, the combination that produced the highest Mean Average Precision (MAP) for our committee submission. In the testing phase it became evident that the bigger the depth value x (≙ number of top ranked result elements taken from each submission for each topic) is, the bigger the applied value for k (≙ weighting for $\mathbf{p_i}$) is supposed to be in order to obtain optimal results. Furthermore, if $k = 0$, the Collective Ranking Strategy is equivalent to the Borda Count voting method discussed in [9].

The algorithm was also tested comparing results when considering all submissions and merely considering the top ten ranked submissions (as determined by INEX 2003), respectively. It became obvious that considering all submissions instead of solely considering those submissions ranked as the top ten in INEX 2003 led to better results while retaining the same values for x and k.

Figure 1 presents an example of the different effect on results when considering all submissions and only top ten submissions of INEX respectively. Figure 2 and 3 show examples of test results with different values for x and constant value for k and vice versa respectively.

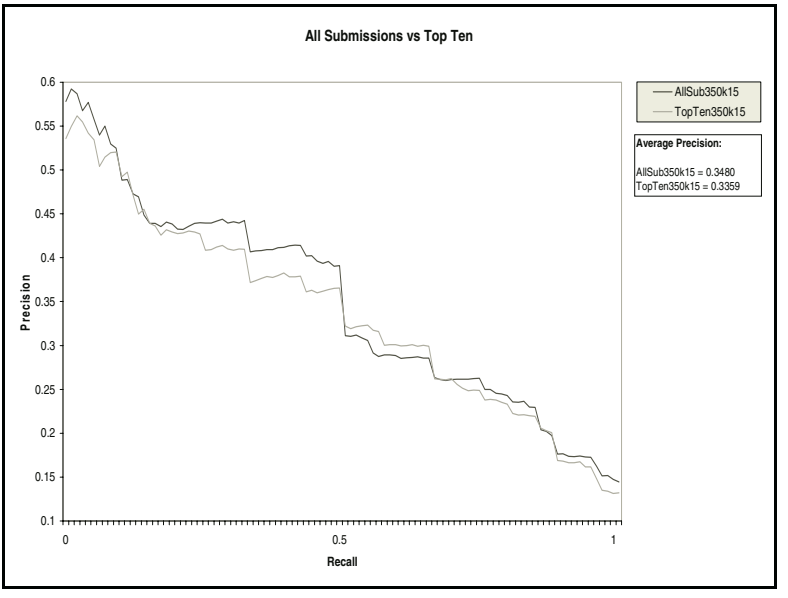

Fig. 1. Comparison of results considering all submissions / only top ten submissions of INEX

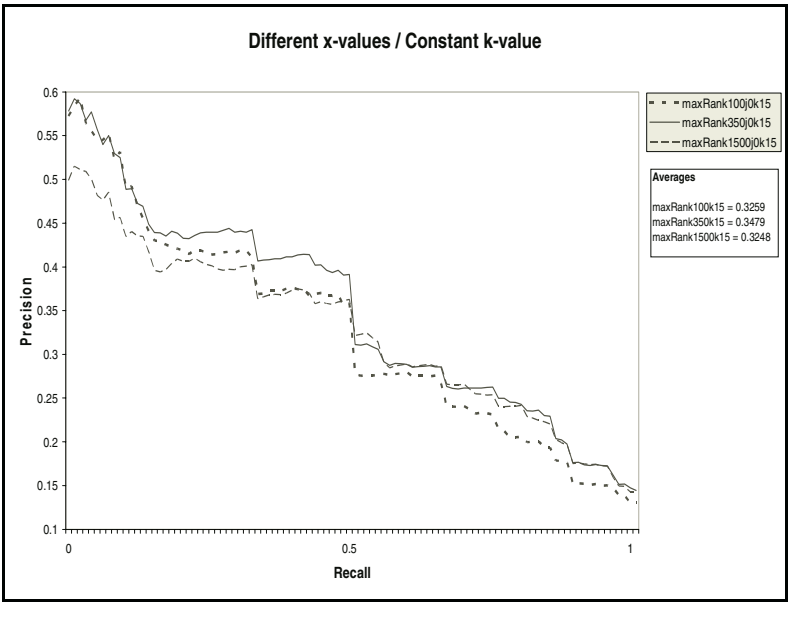

Fig. 2. Example of test results with different values for x and constant value for k

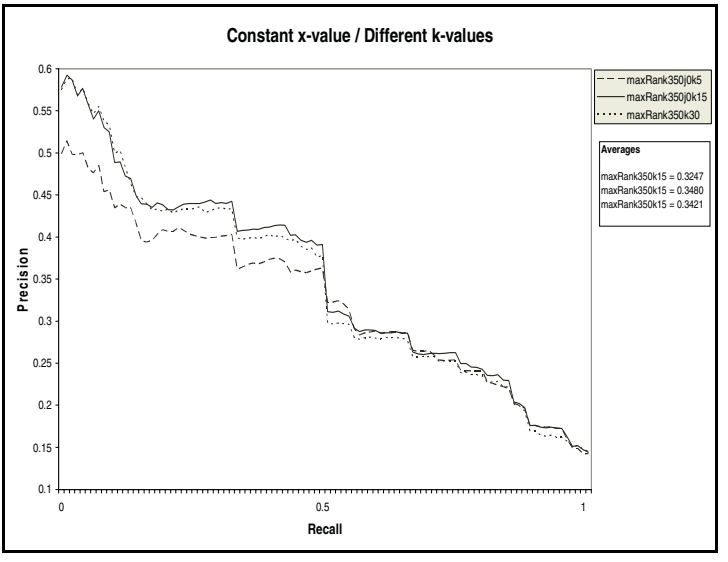

Fig. 3. Example of test results with constant value for x and different values for k

4 Results

After executing the described algorithm for both CAS and CO topics and submitting the obtained Collective Ranking result file to the INEX evaluation software the hypothesis was verified: The recall/precision curve as well as the average precision of the Collective Ranking outperformed all other systems' submissions.

In this context, best results for the different tasks and quantisations were achieved with the following values for x and k:

- CAS strict: $x = 400$ and $k = 18$
- CAS generalised: $x = 1000$ and $k = 30$
- CO strict: $x = 1500$ and $k = 39$
- CO generalised: $x = 1500$ and $k = 39$

Table 1. Comparison of best values of MAP achieved by Collect. Ranking and INEX 2003 participants (*Univ. of Amsterdam [10])

Task	Quantisation	Avg. Precision - Best Value	
		Participants	Collect. Ranking
CAS	Strict	0.3182*	0.3480
CAS	Generalised	0.2989*	0.3177
CO	Strict	0.1214*	0.1339
CO	Generalised	0.1032*	0.1210

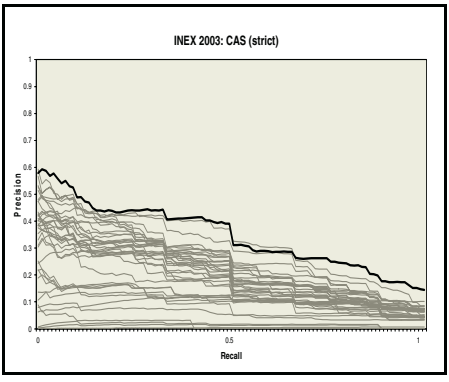

Fig. 4. Results – CAS (strict)

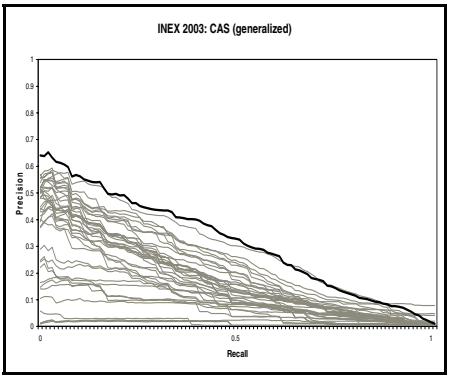

Fig. 5. Results – CAS (generalised)

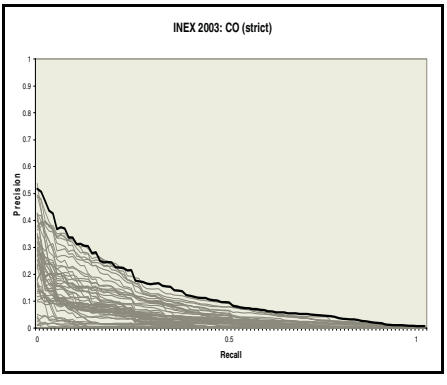

Fig. 6. Results – CO (strict)

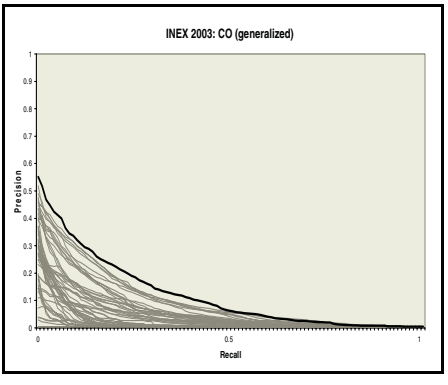

Fig. 7. Results – CO (generalised)

Table 1 displays the best values of average precision achieved by the Collective Ranking in comparison with the best ranked submissions of participants in INEX 2003.

The Precision/Recall curves represented in Figures 4 to 8 demonstrate the performance of the Collective Ranking (displayed in dark bold) in comparison with all other submissions of INEX 2003 (displayed in light grey).

5 Outlook and Future Work

The development and implementation of a Collective Ranking Strategy as presented in this paper and the results obtained establish a basis for copious future work. This chapter gives an overview of challenges and ideas of approaches for further research on this topic.

5.1 Realistic Assessment

In order to identify the extent of possible improvement regarding the Collective Ranking, a programme indicating the notionally maximum performance was implemented, which is based on the following idea:

During the assessment process of the INEX workshop human assessors determine the relevancy of result elements returned by the participants' systems and pooled together in the pool of results in terms of exhaustiveness and specificity. While exploring the XML files of the INEX document collection with respect to the result elements returned for a certain topic, assessors may add elements of the XML files that were not returned by any participating system but, however, are considered relevant to the pooled results. This procedure yields a pool of results referred to as the "Official Perfect Pool of Results", which provides the basis for the INEX Official Assessment Files that are required for the evaluation of the search engines' performance. These assessment files suggest an "idealish" ranking of particular result elements for each topic representing a guideline for the assessment of the actually returned results. This ranking is referred to as "idealish" since elements from some relevant articles might not be included in the top 100 results from any submission, hence are not in the pool of results at all. Adding these elements to the pool would make it theoretically possible to achieve an even better performance. However, due to the fact that some result elements contained in those idealish assessment files are manually added and not returned by any single system, it is not realistic to expect the search engines to actually retrieve these result elements. Therefore, it is equally unlikely that the Collective Ranking system could perform as good as the official ideal results, since it is solely based on the results actually returned by the participants' search engines.

In order to set a more realistic benchmark to identify the (theoretically) best possible performance of the Collective Ranking system, a so-called "Realistic Perfect Pool of Results" is to be established. The appendant programme developed to derive the required Realistic Assessment Files eliminates all result elements not actually submitted by any participant's search engine from the "idealish" Official Perfect Pool of Results.

Figures 8 to 11 display the performance of the Collective Ranking system compared with the precision/recall curves of the "Official Perfect" (displayed as bold grey line) and "Realistic Perfect" Results (displayed as a dotted line). They reveal the remarkably big capability of improvement regarding the Collective Ranking Strategy. Note that for both the strict quantisations, the "Official Perfect" curve is a horizontal line with precision equal to 1 for all recall values.

Surveying these results it is particularly striking to see that the precision/recall curves of the Realistic Perfect Results are remarkably better performing than the Collective Ranking, although the Realistic Perfect "system" avails itself of the same source – solely consisting of result elements returned by INEX participants – that is also available for the Collective Ranking system. This emphasises the crucial importance of successful ranking of returned results and therefore represents a point of origin for further examinations.

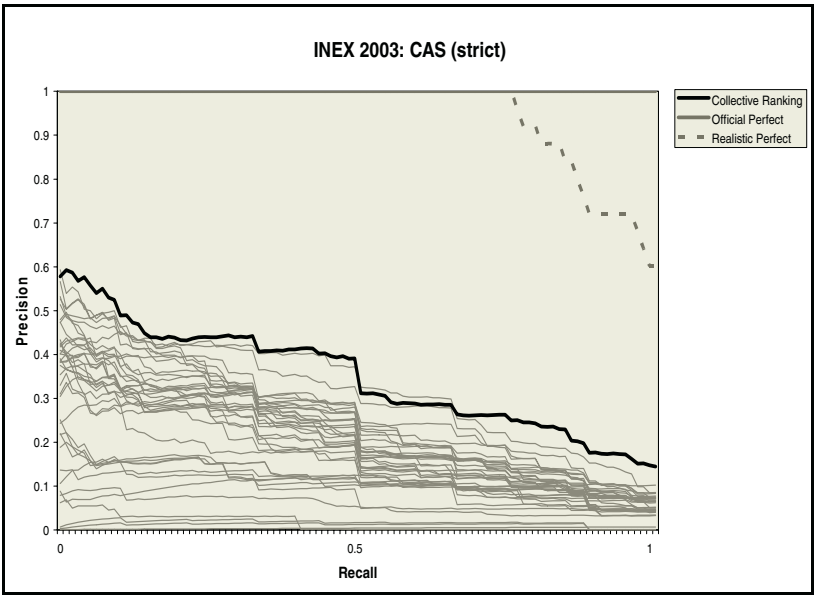

Fig. 8. Collective Ranking compared with "Official Perfect" and "Realistic Perfect" results (CAS – strict)

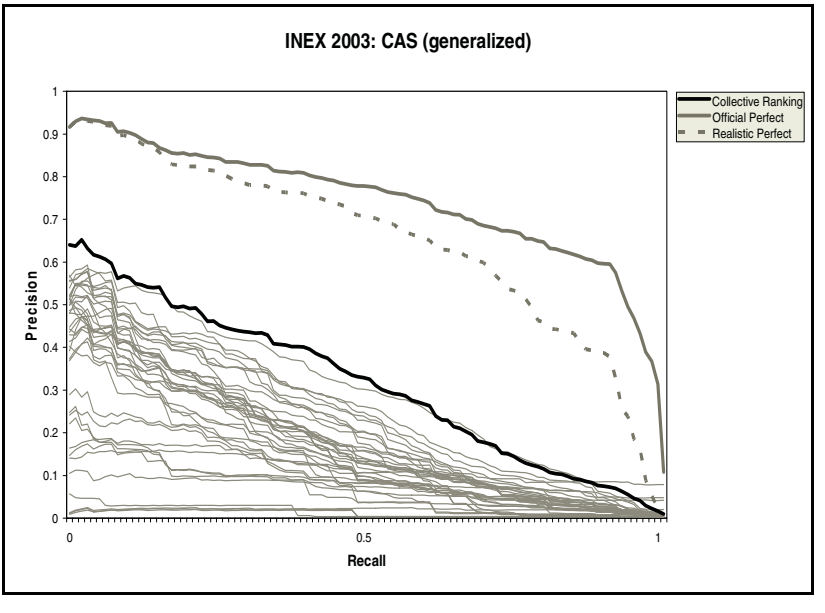

Fig. 9. Collective Ranking compared with "Official Perfect" and "Realistic Perfect" results (CAS – generalised)

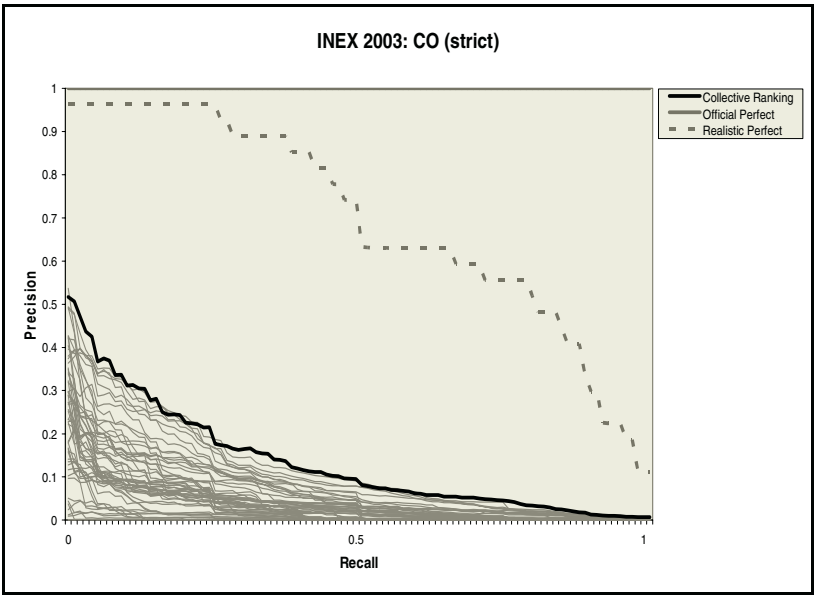

Fig. 10. Collective Ranking compared with "Official Perfect" and "Realistic Perfect" results (CO – strict)

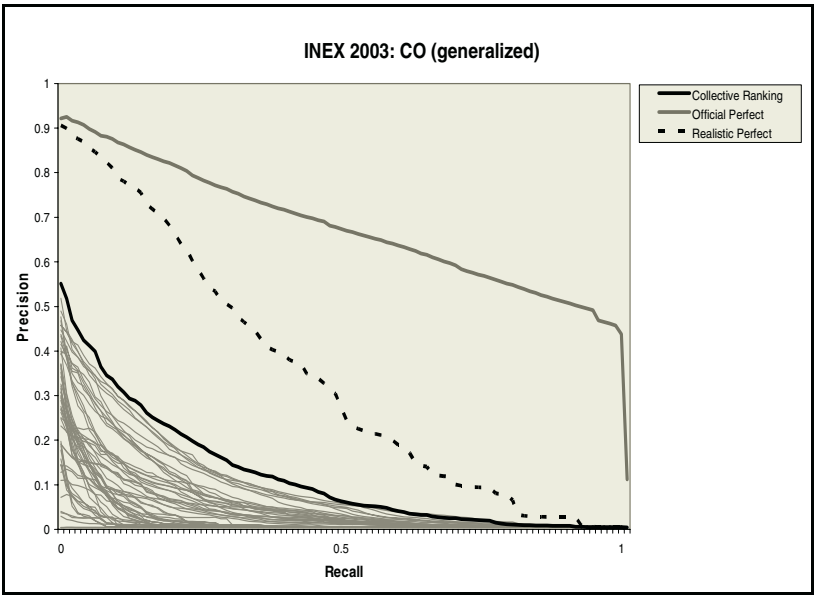

Fig. 11. Collective Ranking compared with "Official Perfect" and "Realistic Perfect" results (CO – generalised)

5.2 Modification of Algorithm

A possible approach to improve the performance of the Collective Ranking system is the modification of the algorithm applied for the implementation of the Collective Ranking Strategy. In this context two practical ideas are described as follows:

Quality Factor: The main idea is the introduction and implementation of a so-called Quality Factor which represents an iterative assignment of a value $\mathbf{q_i}$ (0, 1] to each submission depending on its performance in relative comparison with the Collective Ranking. In this regard the definition of the result element score res_score_i (currently derived from the summation of $\mathbf{p_i}$ only) would be the following:

$$\forall \mathbf{x_i} \in \textbf{result elements: } \mathbf{res\_score_i} := \sum_{\mathbf{i=1}}^{\mathbf{m}} (\mathbf{p_i}^{k} * \mathbf{q_i}^{j}) \quad (1)$$

Initially, for the first run $\mathbf{q_i}$ equals 1 for every submission. After this initial run, a first ranking of submissions can be derived from relative comparison with the Collective Ranking and an individual value for $\mathbf{q_i}$ (0, 1] can be assigned for each submission applying the following formula (with m = number of submissions and sr_i = rank of submission i according to submission ranking derived from previous run compared with Collective Ranking):

$$\mathbf{q_i := (m - sr_i + 1) \ / m} \quad (2)$$

This means for example, if there are ten submissions, the submission ranked first achieves the value ($\mathbf{q_i}$ = 1) for its individual quality factor whereas the submission ranked tenth will be assigned a quality factor value of ($\mathbf{q_i}$ = 0.1) only. Consequently, as the Collective Ranking is derived from a descending list of the top 1500 result element scores res_score_i, the bigger the value $\mathbf{p_i}$ and the bigger the value $\mathbf{q_i}$ for each occurrence of a particular result element is, the better the final ranking position of this particular result element in the Collective Ranking will be.

Implementing the idea of a quality factor $\mathbf{q_i}$ would emphasise the impact of better performing submissions and as a consequence might lead to a better performance of the Collective Ranking system.

Improvement of Participants' Ranking: As the Realistic Pool results have revealed that most of the result elements contained in the official INEX assessment files have actually been submitted by participants and as a consequence must be accessible for the Collective Ranking, it becomes obvious that an improved ranking of results for both the INEX submissions and the Collective Ranking could be the key for noticeable improvement of performance. However, at present it is not quite clear yet how this idea can be translated into successful methods.

5.3 Automatic Testing

At this stage, values identified best for *x* and *k* applied in the Collective Ranking programme are based on results derived from experimental testing. However, since possible values for *x* can range from 1 to 1500 and appropriate values for *k* can theoretically range from 0 to infinite, it was not possible to test all possible combinations of these two values. Therefore it is conceivable that better "optimal" combinations may

be identified by using automated testing methods which in turn requires the assignment of an adequate implementation.

5.4 Automatic Assessment

At the present time the INEX Assessment Files that are used for the evaluation of submissions are derived from assessments conducted by human assessors who work through the INEX document collection to identify relevant result elements. Since this has emerged as a very time-consuming procedure, future work and development with respect to the Collective Ranking could benefit the INEX workshop at such a rate that human assessments might eventually be replaced by assessment and ranking of submissions derived from a relative comparison of those submissions with the Collective Ranking. For this purpose, however, Automatic Assessment Files are to be established within the scope of further research and testing.

6 Conclusion

The results achieved within the scope of this research project by the development and implementation of a Collective Ranking Strategy may benefit the future procedure of the INEX workshop since – although not yet a suitable substitute for human assessments of results – a ranking of participating search engines can now be derived without manual assessment.

The hypothesis stated at the beginning of this project, suggesting that it may be possible to outperform any single system by taking account of the results from all systems was verified. Moreover it was proven that an outperforming search engine can be developed on the basis of other search engines' results. However, the results derived from the implementation of the Realistic Pool Assessment Programme revealed that there is still much room for improvement. Therefore, ample research on the reasons for the performance of the Collective Ranking system will be required in order to identify means to improve the current results.

However, the baselines at INEX are relatively low at present, and it is questionable whether the Collective Ranking Strategy will still lead to the same results after bringing up the baselines. There exists some work revealing that meta search or data fusion methods do not seem to provide extra benefit when the systems being combined all work very well [11].

These conclusions will provide a basis for further research on this topic, especially for the automatic assessment and ranking of search engines, and may be considered a starting point for the exploration of new challenges regarding ranking strategies within this area of modern Information Retrieval.

References

[1] B. C. Vickery. "The Need for Information". In Techniques of Information Retrieval, p 1, London, 1970.

[2] G. G. Chowdhury. "Basic concepts of information retrieval systems". In Introduction to Modern Information Retrieval, pp 1-2, London, 2004.

[3] S. Brin, L. Page. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, WWW Conf., 1998.
[4] L. Guo, F. Shao, C. Botev, J. Shanmugasundaram. “Ranked Keyword Search over XML Documents”, p.1, San Diego, CA, June 9-12, 2003.
[5] N. Fuhr and S. Malik. “Overview of the Initiative for the Evaluation of XML Retrieval (INEX) 2003”. In INEX 2003 Workshop Proceedings, pp 1-11, Schloss Dagstuhl, Germany, December 15-17, 2003.
[6] N. Fuhr, N. Gövert, G. Kazai and M. Lalmas. “Overview of the Initiative for the Evaluation of XML Retrieval (INEX) 2002”. In Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX), pp 1-15, Schloss Dagstuhl, Germany, December 9-11, 2002.
[7] Jansen, J. Bernard, A. Spink, J. Bateman, T. Saracevic. “Real Life Information Re-trieval: a Study of User Queries on the Web”. In SIGIR Forum 32 No. 1, pp. 5-17, 1998.
[8] M. B. Koll. “Automatic Relevance Ranking: A Searcher's Complement to Indexing” . In Indexing, Providing Access to Information: Looking Back, Looking Ahead, Pro-ceedings of the 25th Annual Meeting of the American Society of Indexers, pp 55-60, Alexandria, VA, May 20-22, 1993. J. A. Aslam and M. Montague. “Models for Metasearch”. In Proc. of
[9] the 24th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, pp 276-284, 2001.
[10] B. Sigurbjornsson, J. de Rijke, M. Kamps. An Element-based Approach to XML Retrieval. In INEX 2003 Workshop Proceedings, pp 19-26, Schloss Dagstuhl, Ger-many, December 15-17, 2003.
[11] S. Beitzel, A. Chowdhury, O. Frieder, N. Goharian, D. Grossman, and E. Jensen. “Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Informa-tion Retrieval Stategies”. In ACM Eighteenth Symposium on Applied Computing (SAC), Melbourne, Florida, March, 2003.