Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003

Norbert Fuhr, Saadia Malik
University of Duisburg-Essen, Germany

Mounia Lalmas
Queen Mary University of London, United Kingdom

1. INTRODUCTION

The widespread use of the extensible Markup Language (XML) in scientific data repositories, digital libraries and on the web, brought about an explosion in the development of XML retrieval systems. These systems exploit the logical structure of documents, which is explicitly represented by the XML markup, to retrieve document components, the so-called XML elements, instead of whole 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.

Evaluating the effectiveness of XML retrieval systems requires a test collection (XML documents, tasks/queries, and relevance judgements) where the relevance assessments are provided according to a relevance criterion that takes into account the imposed structural aspects. A test collection as such has been built as a result of two rounds of the Initiative for the Evaluation of XML Retrieval (INEX 2002 and INEX 2003). The aim of this initiative is to provide means, in the form of a large testbed and appropriate scoring methods, for the evaluation of content-oriented retrieval of XML documents.

This paper presents an overview of INEX 2003. In section 2, we give a brief summary of the INEX participants and their systems. Section 3 outlines the retrieval tasks. Section 4 provides an overview of the INEX test collection along with the description of how the collection was constructed. Section 5 briefly reports on the submission runs for the retrieval tasks. Section 6 describes the relevance assessment phase. Section 7 discusses the different metrics used. Section 8 summarises the evaluation results. The paper finishes with some conclusions and outlook for INEX 2004.

2. PARTICIPATING ORGANISATIONS

In response to the call for participation issued in March 2003, around 40 organisations registered from 18 different countries within six weeks. Throughout the year, the number of participants decreased due to insufficient contribution while a number of new groups joined later at the assessment phase. The active participants are listed in Table 1.

The participating groups used a broad variety of approaches for performing XML retrieval. We tried to categorise them into two approaches [Fuhr & Lalmas 04]:

Model-oriented approaches (MO) were based on established information retrieval (IR) models; e.g. vector space model, language model, logistic regression or Bayesian inference model.

System-oriented approaches (SO) focused more on systems aspects; e.g. adding an XML-specific post-processing step to a normal text retrieval engine, using a relational database system for query processing, performing retrieval in distributed environment.

Participants and their corresponding approaches (i.e. MO vs. SO) are shown in Table 1.

3. THE RETRIEVAL TASKS

In INEX 2003, we focused on ad hoc retrieval. This task has been described as a simulation of how a library might be used, where the collection of documents is known while the queries to be asked are unknown [Voorhees & Harman 02]. Three ad hoc retrieval sub-tasks were defined in INEX 2003: the CO (content-only), SCAS (strict content-and-structure) and VCAS (vague content-and-structure) ad-hoc retrieval of XML documents. Within the CO task, the aim of an XML retrieval system is to point users to the specific relevant portions of documents, where the user's query contains no structural hints regarding what the most appropriate granularity of relevant XML elements should be. Within the SCAS task, the aim of a retrieval system is to retrieve relevant nodes that strictly match the structural conditions specified within the query. In the VCAS task, the goal of a system is to retrieve relevant nodes that may not exactly conform to the structural conditions expressed within the user's query, but are structurally similar. CO and (S/V)CAS are discussed in Section 4.2.

4. THE TEST COLLECTION

Like most IR collections, the INEX test collection is composed of three parts: the set of documents, the set of topics and the relevance assessments.

4.1 Documents

The document collection was donated to INEX by the IEEE Computer Society. It consists of the full-text of 12,107 articles, marked up in XML, from 12 magazines and 6 transactions of the IEEE Computer Society's publications, covering the period of 1995-2002, and totalling 494 MB in size, and 8 millions in number of elements. The collection contains scientific articles of varying length. On average, an article contains 1,532 XML nodes, where the average depth of the node is 6.9. More details can be found in [Gövert & Kazai 03]

Organisations	Retrieval approach	no of runs submitted	Assessed topics
University Of Otago	SO	2	68 100 101
LIP 6	MO	3	82 116
Carnegie Mellon University	MO	3	75 113
University of California, Berkeley	MO	6	70 102
Tarragon Consulting Corporation	MO		88 105
Queensland University of Technology	SO	11	89 124
RMIT University	SO	6	86 117
Nara Institute of Science and Technology	MO	5	65 125
doctronic GmbH & Co. KG	SO	4	107 108
University of the Saarland	MO	4	69 79
University of Amsterdam	MO	9	71 103 104
University of Helsinki	MO	3	111 112
University of Bayreuth	SO	9	95 96
University of California, Los Angeles	MO	3	92 98
IBM, Haifa Research Lab	MO	9	85 90
University of Minnesota Duluth	MO	2	87 121
University of Tampere	MO	6	64 93
Royal School of LIS	MO	3	62 97
Institut de Recherche en Informatique de Toulouse	SO	9	73 91 94
Cornell University	MO	1	80 81 123
University of Rostock	MO	0	61 115 122
University of Michigan	MO	2	77
University of Twente and CWI	SO	5	74 109 110
Hebrew University	MO	6	72 119
Universität Duisburg-Essen	MO	9	66 99
Organisations joining at the relevance assessments phase:			
Waterloo University			76
Oslo University College			63 67
Seoul National University			78 126
Czech Technical University			83 84
Illinois Institute of Technology			118

Table 1: List of INEX 2003 participants

4.2 Topics

The topic format and guidelines were based on TREC guidelines, but were modified to accommodate the two types of topics used in INEX: CO and CAS topics:

Content-and-Structure (CAS) queries are topic statements that allow the query conditions to explicitly refer to XML document structure by restricting either the context of interest or context of certain search concepts.

Content-Only (CO) queries are requests that ignore document structure and contain only content-related conditions.

4.2.1 Topic format

The topic is made up of four parts: topic title, topic description, narrative and keywords. The DTD of the topic is shown in Figure 1.

As in TREC, the topic title is a short version of the topic description and usually consists of a number of keywords identifying the user need. CO topics are the same as the standard TREC topics for ad hoc retrieval tasks. CAS topic title may contain structure and content related conditions. In INEX 2003, the format of the title part of CAS topic was based on an enhanced subset of XPath. A concept of "aboutness" in the form of *about(path,string)* was added. The about function usually applies to a context element (CE) that can be described by the syntax "CE[about(path,string)]". For example //article[about(//sec,'"XML retrieval"')] represents the request to retrieve articles, article being the context element, that contain within them a section about "XML retrieval". The string parameter in the about condition may contain a number of terms separated by a space, where a term can be a single word, or a phrase encapsulated in double quotes. Furthermore symbols +,- maybe used to express additional preference regarding the importance of some terms, where + can be used to prioritise terms while - can be used to mention unwanted terms.

The topic description consists of one or two sentences in natural language describing the information need. The narrative is the detailed explanation of the topic statement and description of what makes a document or component relevant. The keyword component contains the set of terms separated by comma that were collected during the topic development process (see Section 4.2.2).

The attributes of the topic are: topic_id (which ranges from 61 to 126), query_type (with value CAS or CO) and ct_no, which refers to the candidate topic number (which ranges from 1 to 120). Examples of both types of topic can be seen in Figure 2 and Figure 3.

4.2.2 The topic development process

The topics were created by participating groups. Each participant was asked to submit up to 6 candidate topics (3 CO and 3 CAS). A detailed guideline was provided to the participants for the topic creation [Kazai et al. 04b]. Four steps were identified for this process: 1) Initial Topic Statement creation 2) Collection Exploration 3) Topic Refinement and 4) Topic Selection. The first three steps were performed by the participants themselves while the selection of topics was decided by the organisers.

During the first step, participants created their initial topic statement. These were treated as a user's description of his/her information need and were formed without regard to system capabilities or collection peculiarities to avoid artificial or collection biased queries. During the collection exploration phase, participants estimated the number of relevant documents/components to their candidate topics. The HyREX retrieval system [Fuhr et al. 02] was provided to participants to perform this task. Participants had to judge the top 100 retrieved results and were asked to record the relevant document/component XPath paths in the top 25 retrieved components/documents and the number of relevant documents/components in the top 100. We were interested in topics that would have at least 2 relevant documents/components and less than 20 documents/components in the top 25 retrieved elements. In the topic refinement stage, the topics were finalised ensuring coherency and that each part of the topic can be used in stand-alone fashion.

After the completion of the first three stages, topics were submitted to INEX. A total of 120 candidate topics were received, of which 66 topics (36 CO and 30 CAS) were selected. The topic selection was made on the basis of a combination of criteria such as 1) balancing the number of topics across all participants, 2) eliminating topics that were considered too ambiguous or too difficult to judge and 3) uniqueness of topics. Table 2 shows some statistics on the INEX 2003 topics.

5. SUBMISSIONS

Participants processed the final set of topics with their retrieval systems and produced ranked lists of 1500 result elements in a specific format. Details of the submission format and procedure were given in [Kazai et al. 04a]. For the CO task, they were asked to submit up to 3 runs per topic and for the two CAS sub-tasks, SCAS and VCAS, up to 3 runs for each could be submitted per topic. In total 120 runs were submitted by 24 participating organisations. Out of the 120 submissions, 56 contained results for the CO topics, 38 contained results for the SCAS topics and 26 contained results for the VCAS topics. For each topic, the top 100 results (of 1,500) from all the submissions for that topic were merged to create the pool for assessment. Table 3 shows the pooling effect on the CAS and CO topics.

6. ASSESSMENTS

The assessments pools were assigned then to participants; either to the original authors of the topic when this was possible, or on a voluntary basis, to groups with expertise in the topic's subject area. Each group was responsible for about two topics. The topic assignments are shown in Table 1. Note that this list excludes topics 105,106,114 and 120 as their relevance assessment process is still in progress.

Two dimensions were employed to define relevance:

Exhaustivity (e-value) measures the extent to which the given element covers or discusses the topic of request.

Specificity (s-value) measures the extent to which the given element is focused on the topic of request.

For both dimensions, a multi-grade scale was adopted. With respect to exhaustivity:

Not exhaustive (0): the document component does not discuss the topic of request at all.

Marginally exhaustive (1): the document component discusses only few aspects of the topic of request.

Fairly exhaustive (2): the document component discusses many aspects of the topic of request.

Highly exhaustive (3): the document component discusses most or all aspects of the topic of request.

With respect to specificity:

Not specific (0): the topic of request is not a theme of the document component.

```
<!ELEMENT inex_topic  (title,description,narrative,keywords)>
<!ATTLIST inex_topic
  topic_id   CDATA  #REQUIRED
query_type CDATA  #REQUIRED
ct_no      CDATA  #REQUIRED
>
<!ELEMENT title (#PCDATA)>
<!ELEMENT description   (#PCDATA)>
<!ELEMENT narrative     (#PCDATA)>
<!ELEMENT keywords      (#PCDATA)>
```

Figure 1: Topic DTD

```
<inex_topic topic_id="76" query_type="CAS" ct_no="81">
<title>
  //article[(./fm//yr = ’2000’ OR ./fm//yr = ’1999’) AND about(.,
  ’"intelligent transportation system"’)]//sec[about(.,’automation
  +vehicle’)]
</title>
<description>
  Automated vehicle applications in articles from 1999 or
  2000 about intelligent transportation systems.
</description>
<narrative>
  To be relevant, the target component must be from an
  article on intelligent transportation systems published in 1999 or
  2000 and must include a section which discusses automated vehicle
  applications, proposed or implemented, in an intelligent
  transportation system.
</narrative>
<keywords>
   intelligent transportation system, automated vehicle,
   automobile, application, driving assistance, speed, autonomous
   driving
</keywords>
</inex_topic>
```

Figure 2: A CAS topic from the INEX 2003 test collection

```
<inex_topic topic_id="98" query_type="CO" ct_no="26">
<title>
   "Information Exchange", +"XML", "Information Integration"
</title>
<description>
   How to use XML to solve the information exchange
   (information integration) problem, especially in heterogeneous data
   sources?
</description>
<narrative>
   Relevant documents/components must talk about techniques of
   using XML to solve information exchange (information integration)
   among heterogeneous data sources where the structures of participating
   data sources are different although they might use the same ontologies
   about the same content.
</narrative>
<keywords>
   information exchange, XML, information integration,
   heterogeneous data sources
</keywords>
</inex_topic>
```

Figure 3: A CO topic from the INEX 2003 test collection

	CAS	CO
no of topics	30	36
avg no of words in title	7	4
no of target elements representing article	13	-
no of target elements representing non-article element	17	-
avg no of words in topic description	16	11
avg no of words in keywords component	5	7

Table 2: Statistics on CAS and CO topics on the INEX test collection

	CAS topics	CO topics
no of documents submitted	30 071	36 113
no of documents in pools	15 077	18 163
reduction	50 %	50 %
no of components submitted	58 828	80 537
no of components in pools	27 633	38 264
reduction	53 %	52 %

Table 3: Pooling effect for CAS and CO topics

Marginally specific (1): the topic of request is a minor theme of the document component.

Fairly specific (2): the topic of request is a major theme of the document component.

Highly specific (3): the topic of request is the only theme of the document component.

The relevance assessment document guideline [Kazai et al. 04c] explaining the above relevance dimensions and how and what to assess were distributed to the participants. This guide also contained the manual to the online assessment tool developed by LIP6 to perform the assessments of the XML documents/components. Features of the tool include user friendliness, implicit assessment rules whenever possible, keyword highlighting, consistency checking and completeness enforcement.

Initially, the collected assessments were with respect to CAS and CO topics. Later, a distinction was made between VCAS and SCAS assessment by filtering elements targeted by the topics from the CAS assessments. Table 4 shows a statistics of the relevance assessments. Figures 4 and 5 show the distribution of relevance for (some of) the elements.

7. EVALUATION METRICS

A number of evaluation metrics were used in INEX 2003.

7.1 inex_eval: INEX 2003 metric for CO and SCAS topics

This metric was developed during INEX 2002, and was adapted to deal with the INEX 2003 new dimensions of relevance (i.e. exhaustivity and specificity). inex_eval is based on the traditional recall and precision measures. To obtain recall/precision figures, the two dimensions need to be quantised onto a single relevance value. Quantisation functions for two different user standpoints were used:

- A "strict" quantisation to evaluate whether a given retrieval approach is capable of retrieving highly exhaustive and highly specific document components (e3s3).
- In order to credit document components according to their degree of relevance, a "generalised" quantisation has been used.

Based on the quantised relevance values, procedures that calculate recall / precision curves for standard document retrieval can be directly applied to the results of the quantisation functions. The method of *precall* described by [Raghavan et al. 89] was used to obtain the precision values at standard recall values. Further details are available in [Gövert et al. 03].

7.2 inex_eval_ng: INEX 2003 metric for CO topics

This metric developed for INEX 2003 is for CO topics and is based on the notion of an ideal concept space [Wong & Yao 95]. This metrics considers the size of retrieved elements. Two variants were used, one that does not consider overlaps in the ranking of document components and a second one that considers overlaps within the components of a ranking. Details can be found in [Gövert et al. 03].

7.3 ERR: Expected Ration of Relevant Units

This measure provides an estimate of the expectation of the number of relevant document elements a user sees when he/she consults the list of the first N returned relevant elements divided by the expectation of the number of relevant elements a user would see when he/she explores all the relevant elements in the collection. This measure is based on an hypothetical user behaviour:

1. The user consults the structural context (parent, children, siblings) of a returned document element.
2. The specificity of a relevant element influences the behaviour of the user.
3. The user will not use any hyper-link. More precisely, he/she will not jump to another document. This hypothesis is valid in the INEX corpus but can easily be removed in order to cope with hyper-linked corpora.

Details can be found in [Piwowarski & Gallinari 04].

8. SUMMARY OF EVALUATION RESULTS

As mentioned in Section 5, out of the 120 submissions, 56 contained results for the CO task, 38 contained results for the SCAS task and 26 contained results for the VCAS task. A summary of the

e+s	VCAS		CO		SCAS	
	article level	non-article	article	non-article	article	non-article
e3s3	188	1 389	180	1 316	122	577
e3s2	111	1 269	112	616	28	151
e3s1	186	663	150	635	25	90
e2s3	148	2 417	124	2 105	46	644
e2s2	147	3 110	103	1 779	35	650
e2s1	360	2 159	222	1 358	64	437
e1s3	223	11 135	148	5 029	100	2 701
e1s2	81	5 726	50	3 872	33	493
e1s1	769	17 617	673	8 074	361	1 185
e0s0	8 897	88 816	10 021	70 530	5 652	19 922
All	11 110	134 301	11 783	95 314	6 466	26 850

Table 4: Assessments at article and component levels

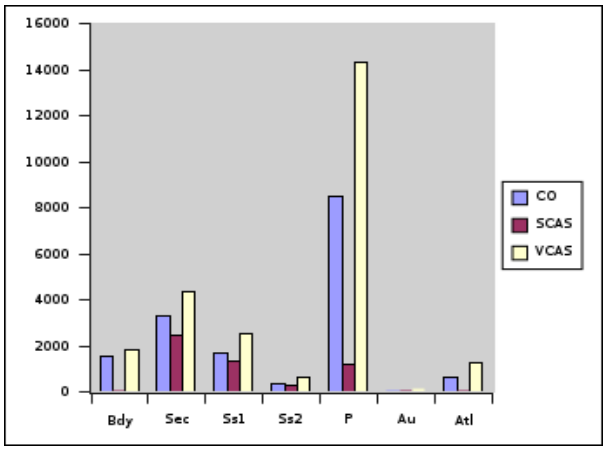


Figure 4: Distribution of relevant elements

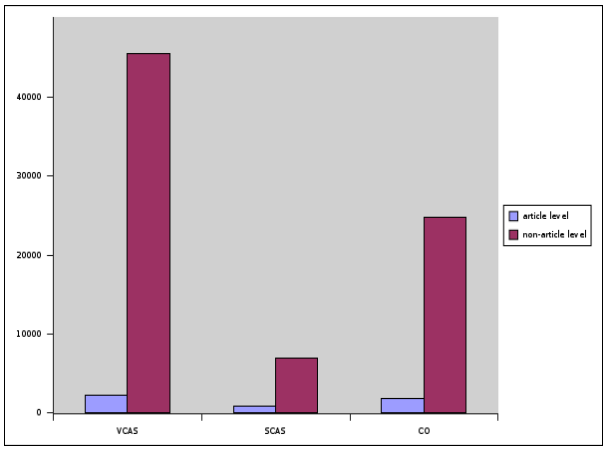


Figure 5: Distribution of relevant article and non-article elements (e > 0 and s > 0)

results obtained with the different metrics is given in the next two sub-sections[1].

8.1 inex_eval and inex_eval_ng mertics

The submissions have been ranked according to the average precision. The top ten submissions, according to average precision, for each task and each quantisation function are listed in Table 5 (inex_eval) and in Table 6 (inex_eval_ng).

When comparing the rankings for the two different quantisation functions and two different user standpoints (considering overlap and ignoring overlap) it becomes evident that they are quite similar. A regression analysis based on average precision values for the submissions shows a strong linear correlation between results obtained using the strict quantisation and results obtained using the generalised quantisation, and result obtained by ignoring and by considering overlap between the retrieved components. Figure 6 shows the scatter plots for the SCAS and CO tasks and the respective regression lines. For the SCAS task the correlation coefficient is 0.9515, and for the CO task, it is 0.7347. Figure 7 shows the scatter plot for the CO task by considering component overlap and by ignoring component overlap for the two quantisations. For strict quantisation, the correlation coefficient is 0.8775, and for generalised quantisation it is 0.9174.

8.2 ERR metric

Table 7 shows a summary of the evaluation results obtained using the ERR metric. The rankings of the submissions were done according to a specific rank (10,100,1500) and averaged over all values. The top ten submissions are shown in Table 7.

9. CONCLUSION AND OUTLOOK ON INEX 2004

INEX 2003 was a success and showed that XML retrieval is a challenging new field within IR research. In addition to learning more about XML retrieval approaches, INEX 2003 has made further steps in the evaluation methodology for XML retrieval. In addition to the presentation of retrieval approaches, four working groups were formed to discuss issues regarding the evaluation of content-oriented XML retrieval approaches: topic format, relevance definition and assessment, online assessment tool, and metrics.

INEX 2004 will start in March of this year, and in addition to the standard ad-hoc retrieval tasks, has 4 new tracks:

Interactive track focusing on interactive XML retrieval, considering also navigation through the hierarchical structure,

Heterogeneous collection track comprising various XML collections from different digital libraries, as well as material from other computer science-related resources,

Relevance feedback track dealing with relevance feedback methods for XML,

Natural language track where natural language formulations of CAS queries have to be answered.

10. ACKNOWLEDGEMENTS

We would like to thank the IEEE Computer Society for providing us the XML document collection. Special thanks go to Shlomo Geva for the set up of the WIKI server, Norbert Goevert for providing the evaluation metrics, Gabriella Kazai for helping with the various guideline documents, and Benjamin Piwowarski for providing the on-line assessment tool. Finally, we would like to thank the participating organisations for their involvement in INEX.

[1]All evaluation results have been compiled using the assessment package version 2.5 and evaluation package version 2003.007.

11. REFERENCES

Fuhr, N.; Lalmas, M. (2004). Report on the INEX 2003 Workshop. *SIGIR Forum 38(1)*.

Fuhr, N.; Gövert, N.; Großjohann, K. (2002). HyREX: Hyper-media Retrieval Engine for XML. In: Järvelin, K.; Beaulieu, M.; Baeza-Yates, R.; Myaeng, S. H. (eds.): *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval*, page 449. ACM, New York. Demonstration.

Fuhr, N.; Gövert, N.; Kazai, G.; Lalmas, M. (eds.) (2003). *INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the First INEX Workshop. Dagstuhl, Germany, December 8–11, 2002*, ERCIM Workshop Proceedings, Sophia Antipolis, France. ERCIM. `http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf`.

Gövert, N.; Kazai, G. (2003). Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In [Fuhr et al. 03], pages 1–17. `http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf`.

Gövert, N.; Kazai, G.; Fuhr, N.; Lalmas, M. (2003). *Evaluating the effectiveness of content-oriented XML retrieval*. Technical report, University of Dortmund, Computer Science 6.

Kazai, G.; Lalmas, M.; Gövert, N.; Malik, S. (2004a). INEX Retrieval Tasks and Run Submission Specification. In: *Proceedings of INEX 2003*.

Kazai, G.; Lalmas, M.; Malik, S. (2004b). INEX Guidelines for Topic Development. In: *Proceedings of INEX 2003*.

Kazai, G.; Lalmas, M.; Piwowarski, B. (2004c). INEX Relevance Assessment Guide. In: *Proceedings of INEX 2003*.

Piwowarski, B.; Gallinari, P. (2004). Expected ratio of relevant units: A measure of structured information retrieval. In: *Proceedings of INEX 2003*.

Raghavan, V. V.; Bollmann, P.; Jung, G. S. (1989). Retrieval System Evaluation Using Recall and Precision: Problems and Answers. In: *Proceedings of the Twelfth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68. ACM, New York.

Voorhees, E. M.; Harman, D. K. (eds.) (2002). *The Tenth Text REtrieval Conference (TREC 2001)*, Gaithersburg, MD, USA. NIST.

Wong, S. K. M.; Yao, Y. Y. (1995). On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst. 13(1)*, pages 38–68.

rank	avg precision	organisation	run ID
1.	0.3182	U. of Amsterdam	UAmsI03-SCAS-MixedScore
2.	0.2987	U. of Amsterdam	(UAmsI03-SCAS-ElementScore
3.	0.2601	Queensland University of Technology	CASQuery_1
4.	0.2476	University of Twente and CWI	LMM-ComponentRetrieval-SCAS
5.	0.2458	IBM, Haifa Research Lab	SCAS-TK-With-Clustering
6.	0.2448	Universität Duisburg-Essen	scas03-way1-alias
7.	0.2437	RMIT University	RMIT_SCAS_1
8.	0.2419	RMIT University	RMIT_SCAS_2
9.	0.2405	IBM, Haifa Research Lab	SCAS-TDK-With-No-Clustering
10.	0.2352	RMIT University	RMIT_SCAS_3

a) SCAS task; strict quantisation

rank	avg precision	organisation	run ID
1.	0.2989	U. of Amsterdam	UAmsI03-SCAS-MixedScore
2.	0.2456	U. of Amsterdam	UAmsI03-SCAS-ElementScore
3.	0.2451	U. of Amsterdam	UAmsI03-SCAS-DocumentScore
4.	0.2399	IBM, Haifa Research Lab	SCAS-TDK-With-No-Clustering
5.	0.2378	IBM, Haifa Research Lab	SCAS-TK-With-Clustering
6.	0.2222	IBM, Haifa Research Lab	SCAS-TDK-With-Clustering
7.	0.2212	University of Twente and CWI	LMM-ComponentRetrieval-SCAS
8.	0.2050	Queensland University of Technology	CASQuery_1
9.	0.1934	Universität Duisburg-Essen	scas03-way1-alias
10.	0.1893	Queensland University of Technology (QUT)	scas_ps

b) SCAS task; generalised quantisation

rank	avg precision	organisation	run ID
1.	0.1214	U. of Amsterdam	UAmsI03-CO-lamda=0.20
2.	0.1144	U. of Amsterdam	UAmsI03-CO-lambda=0.5
3.	0.1102	U. of Amsterdam	UAmsI03-CO-lambda=0.9
4.	0.1001	Universität Duisburg-Essen	factor 0.2
5.	0.0952	IBM, Haifa Research Lab	CO-TDK-With-No-Clustering
6.	0.0929	LIP 6	local-okapi-element,list,ef
7.	0.0915	Universität Duisburg-Essen	difra_sequential
8.	0.0780	Carnegie Mellon University	LM_context_TDK
9.	0.0708	Universität Duisburg-Essen	factor 0.5
10.	0.0688	University of Bayreuth	_co_second

c) CO task; strict quantisation

rank	avg precision	organisation	run ID
1.	0.1032	U. of Amsterdam	UAmsI03-CO-lamda=0.20
2.	0.1009	U. of Amsterdam	(UAmsI03-CO-lambda=0.5
3.	0.0962	IBM, Haifa Research Lab	CO-TDK-With-No-Clustering
4.	0.0960	U. of Amsterdam	UAmsI03-CO-lambda=0.9
5.	0.0881	LIP 6	local-okapi-element,list,ef
6.	0.0839	Carnegie Mellon University	LM_context_TDK
7.	0.0740	University of Bayreuth	_co_second
8.	0.0691	University of Bayreuth	CO-third
9.	0.0687	Universität Duisburg-Essen	factor 0.2
10.	0.0676	Universität Duisburg-Essen	difra_sequential

d) CO task; generalised quantisation

Table 5: Ranking of submissions w. r. t. average precision
using inex_eval metric

rank	avg precision	organisation	run ID
1.	0.1626	IBM, Haifa Research Lab	CO-TDK-With-No-Clustering
2.	0.1575	University of Minnesota Duluth	01
3.	0.1483	Universität Duisburg-Essen	factor 0.2
4.	0.1464	U. of Amsterdam	UAmsI03-CO-lamda=0.20
5.	0.1429	IBM, Haifa Research Lab	CO-TDK-With-Clustering
6.	0.1409	Universität Duisburg-Essen	difra_sequential
7.	0.1403	University Of Otago	CO4
8.	0.1380	University of Twente and CWI	LMM-CLengthModifie
9.	0.1374	U. of Amsterdam	UAmsI03-CO-lambda=0.5
10.	0.1328	doctronic GmbH & Co. KG	1

a) CO task; strict quantisation; overlapping considered

rank	avg precision	organisation	run ID
1.	0.1500	University Of Otago	CO4
2.	0.1489	University of Twente and CWI	LMM-CLengthModified
3.	0.1447	University of Twente and CWI	LMM-Component
4.	0.1365	University of Minnesota Duluth	01
5.	0.1113	IBM, Haifa Research Lab	CO-TDK-With-Clustering
6.	0.1110	IBM, Haifa Research Lab	CO-TDK-With-No-Clustering
7.	0.1091	IBM, Haifa Research Lab	CO-T-With-Clustering
8.	0.1063	U. of Amsterdam	UAmsI03-CO-lamda=0.20
9.	0.1051	doctronic GmbH & Co. KG	1
10.	0.1011	Carnegie Mellon University	LM_context_TDK

b) CO task; generalised quantisation; overlapping considered

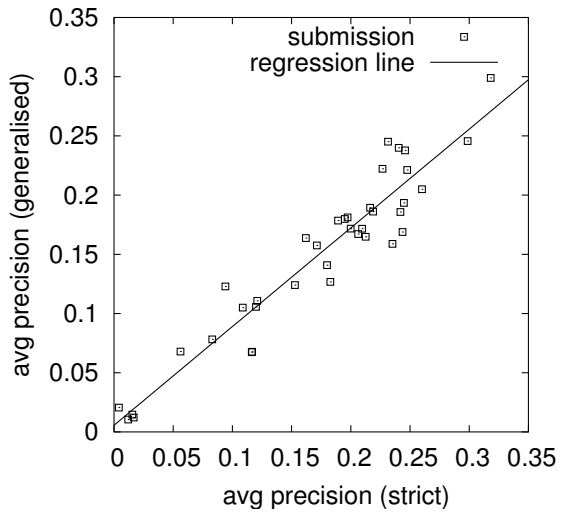
rank	avg precision	organisation	run ID
1.	0.1915	U. of Amsterdam	UAmsI03-CO-lamda=0.20
2.	0.1780	University of Twente and CWI	LMM-CLengthModified
3.	0.1755	U. of Amsterdam	UAmsI03-CO-lambda=0.5
4.	0.1707	University of Twente and CWI	LMM-Component
5.	0.1674	Carnegie Mellon University	LM_context_TDK
6.	0.1631	U. of Amsterdam	UAmsI03-CO-lambda=0.9
7.	0.1627	IBM, Haifa Research Lab	CO-TDK-With-No-Clustering
8.	0.1332	LIP 6	local-okapi-element,list,ef
9.	0.1312	University of Minnesota Duluth	01
10.	0.1281	IBM, Haifa Research Lab	CO-TDK-With-Clustering

c) CO task; strict quantisation; overlapping ignored

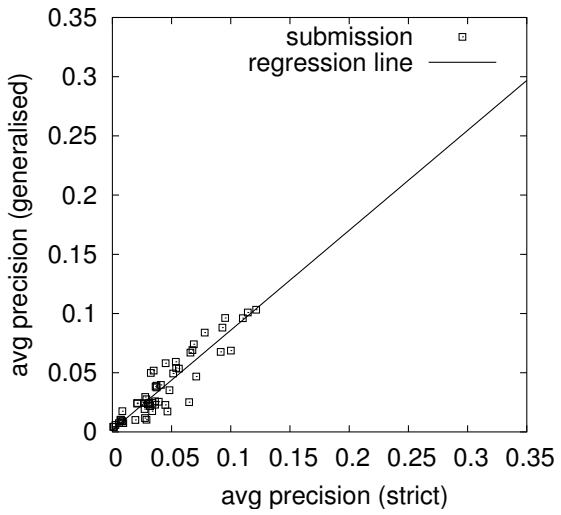
rank	avg precision	organisation	run ID
1.	0.1809	University of Twente and CWI	LMM-CLengthModified
2.	0.1749	University of Twente and CWI	LMM-Component
3.	0.1570	U. of Amsterdam	UAmsI03-CO-lamda=0.20
4.	0.1462	IBM, Haifa Research Lab	CO-TDK-With-No-Clustering
5.	0.1403	Carnegie Mellon University	LM_context_TDK
6.	0.1376	U. of Amsterdam	UAmsI03-CO-lambda=0.5
7.	0.1363	University Of Otago	CO4
8.	0.1269	U. of Amsterdam	UAmsI03-CO-lambda=0.9
9.	0.1268	University of Minnesota Duluth	01
10.	0.1231	Queensland University of Technology	co_ns

d) CO task; generalised quantisation; overlapping ignored

Table 6: Ranking of submissions w. r. t. average precision
using inex_eval_ng metric

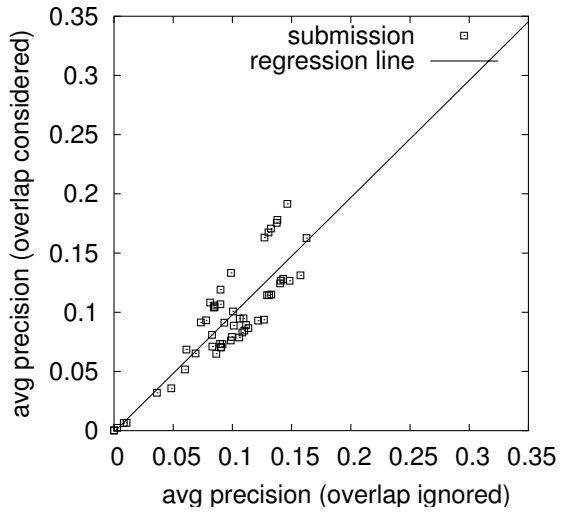


a) SCAS

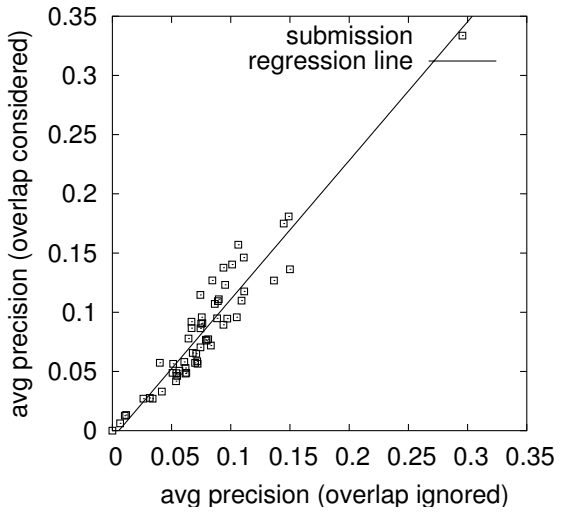


b) CO

Figure 6: Scatter plots and regression lines for average precision of submissions, using strict and generalised quantisation



c) CO: strict quantisation



d) CO: generalised quantisation

Figure 7: Scatter plots and regression lines for average precision of submissions, considering component overlap and ignoring component overlap

rank	avg	organisation	run ID
1.	49.9	IBM, Haifa Research Lab	CO-TDK-With-No-Clustering
2.	46.8	IBM, Haifa Research Lab	CO-TDK-With-Clustering
3.	45.2	Universität Duisburg-Essen	factor 0.2
4.	43.6	Universität Duisburg-Essen	difra_sequential
5.	42.0	U. of Amsterdam	UAmsI03-CO-lambda=0.5
6.	41.9	LIP 6	local-okapi-element,list,ef
7.	41.0	Carnegie Mellon University	LM_context_TDK
8.	40.2	U. of Amsterdam	UAmsI03-CO-lambda=0.9
9.	39.8	U. of Amsterdam	UAmsI03-CO-lamda=0.20
10.	39.5	IBM, Haifa Research Lab	CO-T-With-Clustering

a) CO task

rank	avg	organisation	run ID
1.	48.1	U. of Amsterdam	UAmsI03-SCAS-MixedScore
2.	47.4	U. of Amsterdam	UAmsI03-SCAS-ElementScore
3.	42.3	U. of Amsterdam	UAmsI03-SCAS-DocumentScore
4.	35.7	University of Bayreuth	first_scas
5.	35.7	Universität Duisburg-Essen	scas03-way3-noalias
6.	35.7	University of Bayreuth	cas_third
7.	34.5	Queensland University of Technology	CASQuery_1
8.	33.5	IBM, Haifa Research Lab	SCAS-TDK-With-Clustering
9.	33.5	University of Bayreuth	second_scas
10.	32.9	Queensland University of Technology	QUTscas_st

b) SCAS task

rank	avg	organisation	run ID
1.	40.9	U. of Amsterdam	UAmsI03-VCAS-NoStructure
2.	37.6	U. of Amsterdam	UAmsI03-VCAS-TargetFilter
3.	33.0	IBM, Haifa Research Lab	VCAS-TDK-With-No-Clustering
4.	32.4	IBM, Haifa Research Lab	VCAS-TK-With-Clustering
5.	32.2	IBM, Haifa Research Lab	VCAS-TDK-With-Clustering
6.	29.2	University of Twente and CWI	LMM-ComponentRetrieval-VCAS
7.	28.2	University of Bayreuth	second_vcas
8.	28.0	Universität Duisburg-Essen	vcas03-way2-alias
9.	28.0	University of Bayreuth	first_vcas
10.	27.9	University of Bayreuth	vcas_third

c) VCAS task

Table 7: Ranking of submissions w. r. t. average using ERR metric