Can we at least agree on something?

Andrew Trotman
University of Otago
Dunedin
New Zealand
andrew@cs.otago.ac.nz

Nils Pharo
Oslo University College
Oslo
Norway
Nils.Pharo@jbi.hio.no

Dylan Jenkinson
University of Otago
Dunedin
New Zealand
djenkins@cs.otago.ac.nz

ABSTRACT

During a session of the INEX 2006 workshop in Schloss Dagstuhl the first at-INEX experiment was run. Participants were asked to assess topics in order to increase the number of multiple assessed topics available for analysis (and in order to increase the number of assessors per topic). This contribution presents the experimental set-up, the experiment, and an analysis of the results.

When examining the agreement level across all assessors it is shown that each assessor both brings new material, and disagrees with the there-to consensus. Extrapolation suggests that with 8 assessors, there will be no content that they all agree is relevant, but they continue to agree on which documents are reliant until 19 assessors are present. This suggests that relevance is in the mind of the assessor and not a ground truth.

Additionally examined are several problems encountered in conducting the experiment. These are explained in detail and recommendations for change in the INEX methodology are made.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Information Search and Retrieval – Retrieval models, Search process.

General Terms

Human Factors, Theory.

Keywords

Element retrieval, XML-retrieval, agreement levels.

1. INTRODUCTION

Each year INEX participants travel to Schloss Dagstuhl near Frankfurt in Germany for the annual workshop. For Europe the venue is isolated. There is no airport or railway station. Participants are essentially locked-down, nowhere to go and nothing to discuss except XML and information retrieval. Nothing to do other than present talks, listen to talks, and to participate in lively discussions.

During the 2006 INEX round the decision was made to take advantage of the lock-down in order to conduct an experiment. The INEX workshop participants are a substantial human resource, knowledgeable in the domain of information retrieval, and with the time and motivation to participate in an experiment (while at the workshop).

SIGIR 2007 Workshop on Focused Retrieval
July 27, 2007, Amsterdam, The Netherlands

The nature of such an experiment is dictated by the physical environment, the time available, and the participants. The experiment must require many participants, must be conducted in parallel on each participant, and must require no more than one workshop session. It must also not become overbearing or a disincentive from attending the workshop.

The experiment conducted at INEX has become known as the at-INEX experiment. It was run for the first time in 2006 and is expected to continue as a feature of INEX at future workshops. This contribution outlines the first at-INEX experiment, the motivation behind the experiment, the experiment, and the results.

2. CHOICE OF EXPERIMENT

Two domains were considered for the at-INEX experiment: an interactive experiment, and an assessment experiment.

Unlike a Cranfield methodology laboratory experiment [17], an interactive experiment requires a substantial number of participants (and topics) for statistical significance. The INEX interactive experiment in 2006, for example, had over 80 participants each performing 4 queries selected from a total of 12 [6]. In that experiment each participant was given a total of 15 minutes to fulfill the information need. When the time taken to answer questionnaires before, during, and after the experiment is added to the time it took participants to familiarize themselves with the experimental conditions, and to the four lots of fifteen minutes, a total running time of between 1.5 and 2 hours was needed for each participant.

The at-INEX environment matches the needs of an interactive experiment perfectly. There are many available participants and the time frame is relatively short. Certainly if each participant performed only 2 searches and the questionnaires were kept short then such an experiment could be conducted in just one workshop session.

Assessment experiments (that is, judging topics) require only one participant per topic and can be done by hundreds of people working on different topics in parallel. This is the traditional model used at INEX [7] (and TREC [18]). Assessing a single topic at INEX 2005 took about 11 hours, and at INEX 2006 it took about 7 hours [12] – vastly more time than available at the workshop in Schloss Dagstuhl. On initial inspection an assessment experiment is a bad match to the experimental environment, however this is the nature of the experiment that was conducted.

The time to assess topics for INEX has been of concern and under investigation for many years [12]). INEX assessors are the participants themselves and are not paid for the task. Their reason for participating is their research, not their desire to perform assessment. Assessing is considered by some as a necessary evil

done only to get performance measures for their search engines. Some (including one of the authors) have employed students to assess because the task is dreary and laborious. Much to the surprise of the authors, some INEX workshop participants who participated in the at-INEX experiment had never judged a single document so had managed to duck the task year after year – clearly they did not consider assessing something to look forward to.

An ongoing task at INEX is the reduction of the assessment load while at the same time maintaining assessment quality. Considerable advances have been made. From 2002 to 2006 the changes included changing from a two-dimensional graded relevance scale to a one-dimensional continuous scale [12]. Changes from explicit assignment of assessments to each element to the yellow-highlighting method suggested by Clarke [4]. But there remains room for further load reduction. Specifically the at-INEX experiment in 2006 aimed to answer several questions:

1. Do the assessments for a single topic need to be conducted by a single assessor?

The assessment of a single topic might, for example, be split amongst two, three, or more judges, each assessing part of the topic. Advantage might be taken of graduate students studying IR to assess a topic during class as a hands-on method of learning about the process. This could only be done if relevance was the same in the mind of each of there assessors. Conveniently what constitutes a relevant document for a given topic is spelled out in the INEX topic narrative – but it remains open to interpretation.

2. Can the INEX document pool be reduced in size?

INEX uses a round-robin pooling method (called top-n). In this method the top element from each run are collected and the documents from which they come are added to a pool, then the pool is de-duplicated. The process continues for the second element from each run, and so on until eventually the pool contains n (at INEX 2006 n=500) unique documents (see Piwowarski and Lalmas [11] for details).

Investigations into the most appropriate size of n have focused on identifying the remaining number of unidentified relevant documents. Experiments might be conducted to investigate the effect (on relative search engine performance) of reducing the size of the pool. A shallower pool, although leading to a less complete set of relevance assessments, would take less time to assess.

3. How effective are assessments collected with a very short time frame?

The time available to assess at the workshop was limited to one workshop session. If it were possible to reliably assess topics in such a short time frame then many more topics could be assessed in the same time frame. Equally, if the number of topics remained fixed it might be possible to complete the entire assessment task in just a few hours.

3. METHODS OF COLLECTION

In total 41 INEX 2006 workshop attendees participated in the experiment. 15 topics were chosen for re-assessment on the basis that those topics had already been double-judged and further additional assessing of those topics could be used to gain a better understanding of how the concept of relevance crosses a population. There was no order to the manor in which topics were given to assessors. After the assessment process participants answered a short questionnaire containing questions about how they assessed. Table 1 shows the number of assessors that answered questionnaires for each topic.

Table 1: Distribution of topics to assessors

Topic	Assessors	Topic	Assessors
304	3	364	3
310	4	385	3
314	2	403	3
319	2	404	2
321	3	405	3
327	4	406	3
329	3	407	1
355	2	***Total***	***41***

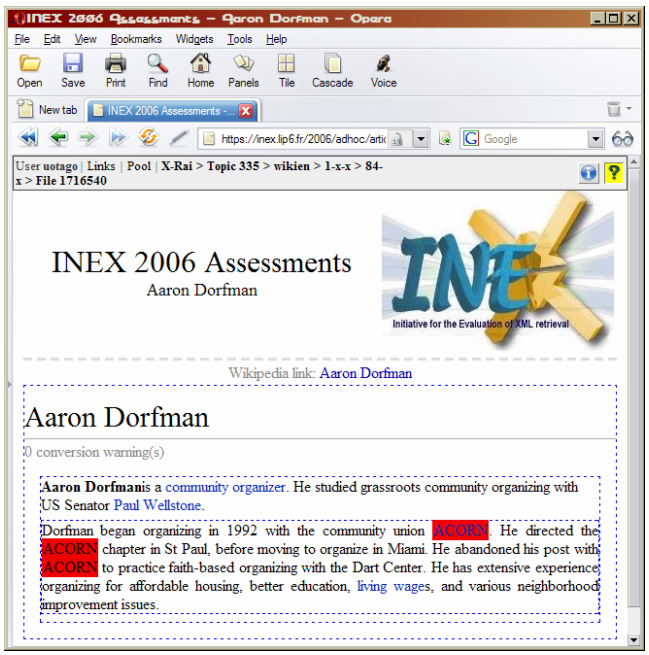

Figure 1: The X-Rai assessment software GUI. In this example the assessor has chosen to highlight keywords.

Assessment was performed using the X-Rai [11] assessment tool built by Benjamin Piwowarski specifically for INEX. Assessors were given a topic to assess, then chose a document from the document-pool to assess, then identified any relevant passages within that document by highlighting them in yellow. Finally they moved on to the next pool document by clicking at the bottom of the window. X-Rai is shown in Figure 1 with the topic keywords highlighted.

The document collection used was the INEX Wikipedia document collection whose details are published elsewhere [5]. Documents were presented to the user in alphabetical order, and not pool order. Presenting out of pool order has the advantage of not

biasing the assessor early (or late) in the experiment – they don't know if the document they are assessing is likely to be relevant or not.

In the at-INEX experiment the time available to assess a topic was limited to one workshop session (1 hours and 20 minutes), but the average time taken to assess a 500 document pool at INEX 2006 was about 7 hours – clearly it was not possible to fully reassess each topic in a single session. To resolve this problem the pool for the topics in Table 1 were reduced to about 100 documents each (that is, the top-n process was stopped after a complete round and 100 or more documents were in the pool).

As a means of ensuring the validity of the INEX pooling software, new and alternate pooling software was developed for the experiment.

4. RESULTS OF COLLECTION

4.1 The Pooling Process

On average these reduced pools contained 135 documents per topic. Comparison with the official pools showed agreement levels between 92% and 100% with a mean of 98%. That is, for example, for topic 405, 124 of the 135 documents (92%) in the reduced pool were in the official pool. For others all documents were, but on average 98% were.

Investigation into why the reduced pools were not a full subset of the official pools revealed a workflow anomaly that it is hoped will be resolved for future INEX rounds.

The workflow model at INEX proceeds thus: Participants submit topics, the organizers select the final topics from those[1], the final topics are released to participants who submit runs for those topics, the pools are generated, the topics judged, then the performance of each run is determined. Participants can submit both official runs and unofficial runs with only the former being included in the pooling and scoring process.

If a participant submits a run that contains errors (such as an invalid document-ID, an element that does not exist, or a malformed XPath) then the entire run is excluded from the pooling process. However, as different software is used to assess performance as that used to generate the pools, such runs are still scored even though they were not included in the pooling process.

The effect of this appears at first to be negligible because one expects malformed runs to be produced by buggy search engines which will perform badly. However this need not be the case. Should a run contain a simple error, but otherwise be well formed, the top documents in the run will not be assessed unless other runs also identify the same documents in their top ranks. If those documents were to be relevant they would continue to be considered non-relevant because they were not assessed. Such a run performs badly not because it fails to identify relevant documents (or elements), but because the results it does identify are never scored.

Exactly this situation occurred during INEX 2006. A run from University of Granada uniquely identified results in the top few rankings for at least one topic and was excluded from the pooling process – but was then later ranked relative to the other runs. It performed badly and the submitter questioned its official score.

The new software used to generate the reduced pools only parsed parts of the run-file necessary to build the pool. If errors did not occur in those parts of the file then the file was used for pooling[2]. The official pools were built from software that parsed the entire file and rejected all files that contained any errors.

Several changes are recommended to the workflow:

It is not possible to determine from a run whether or not it is official. This could be amended by adding one attribute to the inex-submittion tag. Rejected runs could also be marked in a similar way.

Runs can perform badly because their top ranking results are not in the pool. This can be rectified by ensuring that any run rejected from any part of the workflow is rejected from the entire workflow. One possibility is to fully check every run on submission.

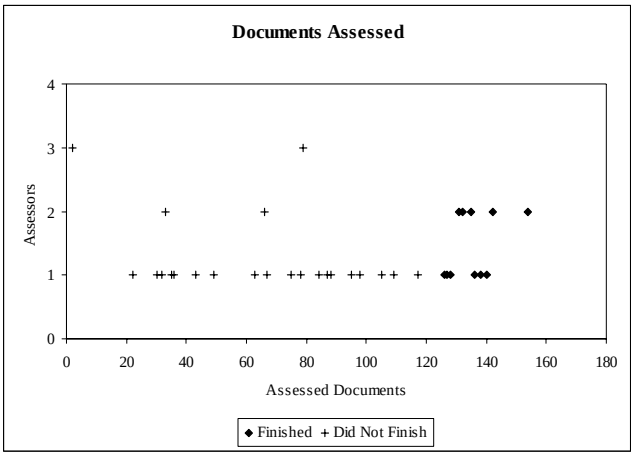

Figure 2: Number of documents assessed in the allotted time

4.2 Workload

Of the participants, 16 completed the assessment task in the allotted time (1 hour and 20 minutes) and 29 did not. Some (4) did not answer the questionnaire. For topics 319 and 355 no assessor completed the task in the allotted time. For topics 314, 327, and 329 two assessors completed the task, for the remainder only one assessor completed the task.

On average 87 documents were assessed in the time period with a minimum of 2 and a maximum of 154 documents assessed. Figure 2 shows the distribution of the number of documents assessed in the time period.

If the relative rank order of the official runs is maintained using the assessments completed in just this short time then the pooling could be stopped at n=100 documents and assessing completed in just an hour and twenty minutes per topic.

A set of assessments for the at-INEX experiment was constructed for the 15 topics by taking the assessment pool with the lowest

[1] All syntactically incorrect, partially completed, and duplicate topics are dropped. All non IR topics are dropped (such as "all papers written by Smith"). There is no formal method other than opinion of the several reviewers.

[2] The error was subtle, some paths in some documents were missing instances. That is, in the file /article[1]/body[1]/emph3 is seen whereas /article[1]/body[1]/emph3[1] was needed.

pool-id for those topics that were complete in the allotted time, and for those that no assessor completed, the pool with the most assessed documents.

A reduced set of official assessments was taken by excluding all topics except the 15 multiple-assessed topics.

The performance of the All-In-Context[3] runs submitted to INEX in 2006 was scored using the INEX assessment tool. The metric MAgP was used. Two scores were generated, one against the reduced assessment set of only 15 topics, the other against the full assessment set, but for only the 15 topics.

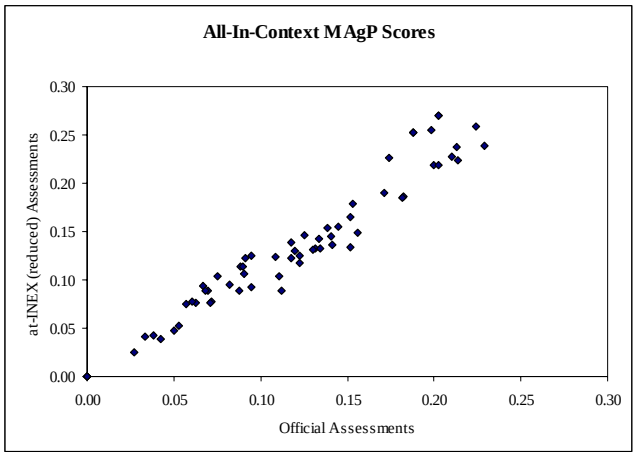

Figure 3: Run performance against the two assessment sets

Figure 3 shows the performance of each run against the two sets of assessments. The average amount of time needed to complete the assessment of a single topic (across all topics, not just the 15) for the official assessments was 6 hours and 51 minutes. The maximum time allotted to the at-INEX assessment task was 1 hour and 20 minutes. The Spearman's rank correlation coefficient for the relative performance of the search engines is 0.97. There is a strong positive correlation of one to the other.

The subset of runs that ranked in the top 10 against either set of assessments contains 13 runs. The Spearman's rank correlation for just those runs is -0.03, that is, there is a very weak negative correlation for the top performers. In an hour and twenty minutes of assessing the top runs can be separated from the bottom runs, but the relative performance of the top runs cannot be determined.

It can be seen from Figure 3 that the performance of a run against the at-INEX assessments is often better than the performance against the official assessments. One possible reason is that the average amount of relevant material per document in the at-INEX assessments is larger than that of the official assessments (1833 vs. 1059) so any fixed length passage from a run is more likely to intersect a relevant passage in the at-INEX set. An alternative and more likely reason is that the number of relevant documents in the at-INEX set is smaller than that in the official set (22 vs. 60) so one point of generalized recall (1/gR) is larger in the at-INEX assessments than in the official assessments.

[3] Trotman *et al*. [16] examined XML-IR use cases and consider this to be the most viable task examined.

5. RESULTS OF QUESTIONNAIRE

5.1 Factors Influencing Assessments

In the questionnaire the assessors were asked to state the factors that helped them to decide what would make a passage relevant. The factors were categorized and in all 14 different categories were found. Some were very idiosyncratic, e.g. "geographical facet" or "discourse", and others were very common. Three factors appeared to be much more influential than the others; titles, content and keywords.

The study showed (see Table 2), not surprisingly, that content was the most important factor. Many (12 assessors) also used keywords collected from the task description, either by system highlighting or browser enabled searching in the article. The third most important factor was the titles of documents, sections and sub-sections which were used by 11 assessors. The most common other factors were context, links, introductory text and bibliographic references, each being indicated by three assessors as being important.

Table 2: Factors affecting passage choice at INEX

Factors	Titles	Content	Keywords	Other
Assessors	11	30	12	21

From studies of information searching behavior (e.g. Barry [1]) it is known that there are many diverse factors influencing readers' relevance judgments of information sources. In fact, Barry's study showed that "every respondent mentioned factors beyond the topical appropriateness of documents during their evaluation" [1]. In the study herein a similar variety of criteria is not seen, perhaps due to the heterogeneity of the information sources, all being Wikipedia articles. Another, and more interesting, theory is that the assessors (all being information scientists) may have a very rationalistic set of factors for determining the relevance. This is in-line with Pharo and Järvelin's [10] findings relating to the relationship between information scientists and end-users perspectives on web information searching, pointing out the mismatch of viewing the searcher as a very rational individual when prescribing information searching procedures when in fact the searcher to a very large degree is looking for satisfaction (see e.g. Prabha *et al*., [13]) during information retrieval.

5.2 Dynamics of Relevance Assessments

This study also examined how the learning effect affected the relevance assessments. It is known from studies of information searching that often searchers will have an unclear formulation of the information need, which may become clearer throughout the search process as they start interacting with potential information sources. Would a similar development be spotted among assessors?

In response to the question of whether they had changed their mind during the assessment process 17 persons said they had changed their mind whereas 22 persons said they had not (two assessors did not answer this question). The main reason given by those who had changed their mind was, in fact, related to the learning process, they had to get acquainted with the topic, the document type or the assessment software. Thus it is seen that a learning effect is also involved in cases where the search task is formulated and where the goal of the information searcher clearly is directed at full recall, which is the case in this type of

experiment. This also suggests that the assessments would benefit from assessor training before assessment. They might even benefit from reassessment of documents judged at an early stage of the assessment process.

5.3 Size of Relevant Passages

Half of the searchers said they preferred to use a standard size for marking relevant passages; the other half disagreed and claimed that the size differed. This might be related to characteristics of the tasks, but a closer inspection of the individual searchers does not reveal any systematic connection between tasks and passage preference. Of the assessors favoring specific sizes the large majority had a preference for small or smallish (e.g. one paragraph) passages.

Table 3: Preference of element size

Elements	**< 1**	**1**	**2**	**3+**
Assessors	25	24	10	15

The assessors were also asked to mark the correspondence of their selected passages to article elements (see Table 3). The results strengthen the notion that searchers prefer smaller elements. More than half of the assessors choose to select passages equivalent to one element (24 assessors) or less of length (25 assessors). More than one out of three assessors, however, chooses to use passages covering more than two elements. Some assessors pointed out that the tasks they performed very much suggested a specific size of passage to be marked.

Assessors were also asked for their preference with respect to Best Entry Points (BEP), i.e. the element they recommended as the best place in the document from which to start reading. Only eight assessors stated a definitive need for more than one good entry point per article. 22 assessors rejected such an option whereas a few stated they sometimes would have liked to add more than one BEP.

6. AGREEMENT LEVELS

6.1 Agreement Levels

Search engine performance is often measured against a gold standard set of assessments produced by a single individual. Wherever possible at INEX the topics are assessed by the original topic author, thus the assessments can be considered the "right answer" in the mind of the person with the information need.

However, Spink *et al*. [14] show that (on the web at least) many queries will be seen repeatedly, and are issued by many different individuals. It is not clear if each individual has the same definition of relevance, and if they do not then how this affects the relative performance of search engines.

Trotman [15] and Pehcevshi and Thom [9] examined agreement levels for whole documents and for elements at various rounds of INEX. They initially showed very low agreement levels. Their work resulted in changes to the assessment methods. These changes in turn resulted in improvements in both the assessor load and agreement levels.

In the at-INEX experiment multiple judges were available to assess each topic. Computing the agreement levels with multiple assessors provides insights into how different users view the relevance of the same documents with respect to the same query. That is, it is possible to identify those parts of the document everyone agrees are relevant and those that only some agree are relevant.

At INEX 2006 assessors identified passages of relevant text using a yellow-highlighting method. From these passages the relevant elements were automatically deduced. To compute the agreement levels for this data it is therefore necessary to examine the passages in the assessment files and not the elements. There are complications.

6.2 Reading INEX Assessment Files

X-Rai [11] produces XML files containing three kinds of assessments: relevant passages, relevant elements, and best entry points. The passages are those parts of a document highlighted by an assessor. The elements are those document elements crossing a relevant passage. The BEPs are separately identified by the assessor. These are grouped by document (file) and stored in separate files for each topic.

Passages are identified in an assessment file in the following way:

```
<passage start="/article[1]/body[1]/section[
14]/normallist[1]/item[25]/text()[1].0" end=
"/article[1]/body[1]/section[14]/normallist[
1]/item[25]/text()[1].19" size="20"/>
```

This passage starts before the 1st character of the first text node of the 25th item of the 1st normallist of the 14th section of the 1st body of the 1st article in the file. The passage is 20 characters[4] (not bytes as the files are encoded in UTF-8) in length and finished after the 19th character of the same text node in the document tree.

Elements are identified in the following way:

```
<element path="/article[1]/body[1]/section[1
4]/normallist[1]/item[25]" exhaustivity="2"
size="34" rsize="20"/>
```

The path is specified in XPath [3]. For 2006 assessments the exhaustivity is redundant, the size is the size of the element and the rsize if the quantity of relevant text. 20 of the 34 characters in the element were highlighted as relevant making the specificity 20/34=0.59.

Best entry points are identified in the following way:

```
<best-entry-point path="/article[1]/body[1]/
section[14]/normallist[1]/item[25]/text()[1]
.0"/>
```

In this instance, the best entry point is before the first character of the same text node identified in the element description above.

6.2.1 Discrepancies

X-Rai requires all text that can be highlighted by an assessor to be in separate leaf nodes of the document tree. Unfortunately, the document collection is not structured in this way so a series of

[4] More accurately it is 19 in length, but the assessments state 20.

simple transformation are applied to the documents before assessment starts. For example[5]

```
<a>some.text<b>.and.some.other.with.spaces.after.</b>...</c>
```

becomes

```
<a><xrai:s>some.text</xrai:s><b>.and.some.other.with.spaces.after....</b></c>
```

Because the assessors are assessing against a transformed document collection and not the original, the assessments do not always match the original document structure. For example, the topic 310 assessments for document 2545650 contain the passage:

```
<passage start="/article[1]/name[1]" end="/article[1]/body[1]/section[4]/section[3]/normallist[1]/item[2]/text()[1].33" size="14581" />
```

The end point is 33 characters into the first text node of the given path. That contains the text (delineating quotes added for clarity):

```
", designed by Richard Loomis"
```

which isn't 33 characters in length (its 28 characters in length). In this case the extra white-space occurring after the element has been included in the element for X-Rai which identified the end of the highlighting as occurring 33 characters into the transformed element.

Another way in which the transformation can cause discrepancies is with passage lengths, the length of a passage can be larger then the amount of text between the start and end points in the original XML files.

Runs submitted to INEX are generated against the original untransformed document collection. Given the assessments can indicate more content per element than exists; it may not be possible to submit a perfect run.

6.3 Agreement Level Algorithm

Assessment discrepancies make it difficult to compute agreement levels for multiple assessors that will agree with future results published by other researchers for the same assessments – unless the algorithms are stated up front. The approach taken for the work described in this contribution is:

For each topic

  For each relevant document

    Load and parse the original XML document

    Replace each character in all text nodes with '0'

    For each assessor

      For each passage

        Locate the start point, start

        Locate the end point, end

        Increment each character between start and end

All end points are truncated to at most the length of the element in which they terminate. In the example above, it would be truncated at 28 characters.

6.4 Assessment Subset

Not all assessors completed the assessment task. The assessments from those assessors who completed less that 50% of the task were discarded from the analysis.

As the results from the INEX 2007 double-assessment experiment were also available they were included in the analysis, as were the official INEX assessments.

The assessors of these two sets did not all assess the same documents for two reasons: first, different pools were used; second, some assessors did not complete the task. The documents used in the analysis were those that all assessors assessed. Table 4 shows the number of assessors per topic and the number of documents they all assessed on common for that topic. In total 60 assessors assessed 1,471 documents across 15 topics (an average of 98 documents and 4 assessors per topic).

Table 4: Pool sizes and number of assessors used for analysis

Topic	Documents	Assessors
304	135	3
310	91	4
314	130	4
319	78	4
321	132	3
327	78	5
329	86	5
355	83	3
364	56	5
385	87	4
403	113	4
404	104	4
405	99	4
406	67	5
407	132	3
Total	**1,471**	**60**

6.5 Results

Figure 4 shows the mean number of documents considered relevant as the number of assessors is increased. As a different number of assessors assessed each topic several lines are shown, each being the mean of only those with at least *m* given assessors where *m* is the number of points on the line (that is, for the line with 3 points all topics were used to generate the means).

[5] This example is lifted directly from private communication with B. Piwowarski.

The figure shows that as the number of assessors is increased from 1 to 5 the assessors continue to find further relevant documents (the union increases). It also shows that the number documents they all agree are relevant decreases (the intersection decreases).

Taking the case where the number of assessors was 4, and fitting a logarithmic trend line to the intersection-curve resulted in an R-squared of 0.991. This line was extrapolated and it crosses the x-axis at 19 assessors. That is, if the trend continued then with 19 assessors there would be no single document that all assessors agree relevant to any information need. Fitting a logarithmic line to the union, R-squared value is 0.999 and at 19 assessors 33 documents would be identified.

Figure 5 shows the mean number of characters of text that are considered relevant as the number of assessors is increases. A similar pattern to that of documents is seen. Fitting logarithmic lines to the intersection and union (of 4 assessors) resulted in an R-squared of 0.958 and 0.997 respectively. Extrapolating to 8 assessors and there are no characters in common, but a total of 64,167 characters of relevant content.

Even though the 8 assessors would not agree on relevant content within a document they will all agree that some documents are relevant. Care should be taken with this conclusion because of the inherent inaccuracy of extrapolating such a small number of points over such a long distance.

Pehcevski [8] reports that in the INEX 2005 interactive experiment participants agreed at the extreme ends of the old INEX multi-grade relevance scale (i.e. highly-relevant and not-relevant) but not in the middle. A similar result can be seen in the Cystic Fibrosis collection [2]. Figure 4, Figure 5, and the extrapolations strongly suggest that relevance is in the mind of the assessor and not a universal truth. These two results are not contradictory – assessors can, in general, agree where within a document the relevant content is found even if some don't.

If, indeed, each assessor continues to identify new relevant documents, and there is no one document that every assessor agrees is relevant then it is not clear that Cranfield experiments are meaningful for XML-IR and Passage Retrieval. Further investigation is required.

For the 4 topics that have 5 assessors, Figure 6 shows the proportion of the union that was identified by at least 1, 2, 3, 4, and 5 assessors (mean over all possible assessor groupings). In each case a decrease is seen suggesting that each time a new assessor is added, they will disagree on an otherwise commonly held belief.

Figure 7 shows the same but for documents. Particular note should be taken of topic 327 in which four assessors agree on a relevant document, but not where within that document the relevant material can be found. This is exactly as predicted by Figure 4 and Figure 5.

7. CONCLUSION

This investigation examined the first at-INEX experiment and reports on the results. Several methodological problems were encountered and suggestions made to tighten the practice.

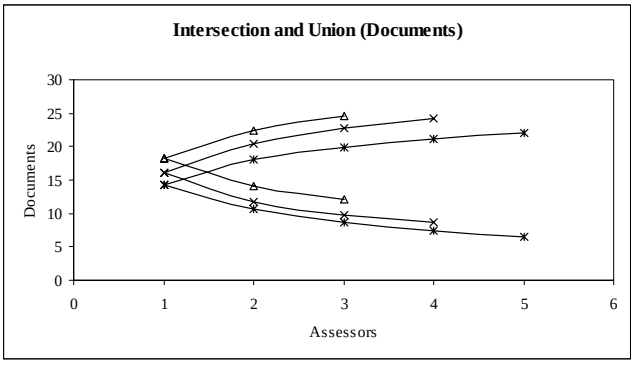

Figure 4: Cross-assessor intersection and union of documents

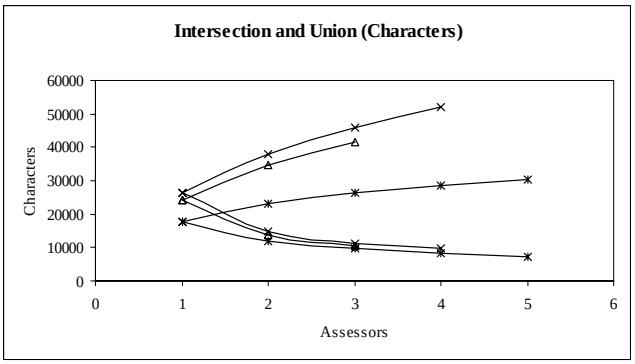

Figure 5: Cross-assessor intersection and union of characters

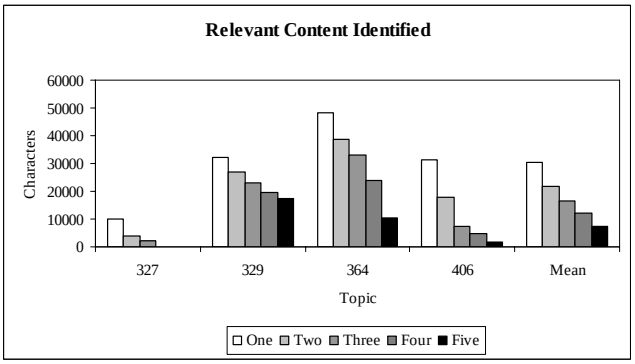

Figure 6: Decreasing agreement of relevant text as the number of assessors increases

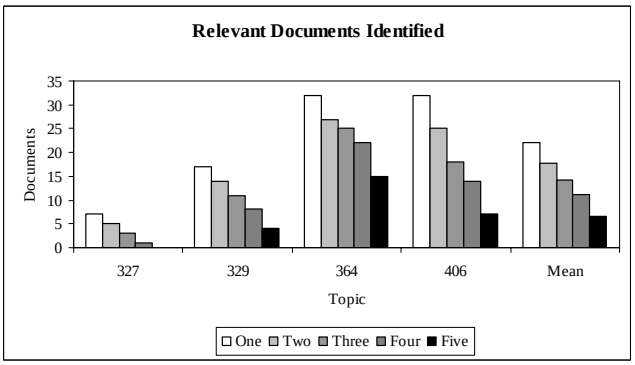

Figure 7: Decreasing agreement of relevant documents as the number of assessors increases

A program was written to generate pools for assessment, that program identified a different set of documents that those in the official pool. This was because the runs included in the pool were different for each program. In turn this is because it is not possible to know from a given run if, or not, that run is official or not. It is also not possible to know if it was rejected for some reason.

Changes to the submission and acceptance process are recommended: Official runs should be marked as such within the run. To avoid confusion runs should be verified at the submission point and no run should be accepted if it is possible for it to later be rejected.

The number of documents assessed in the allotted time period varied greatly – that is, we can't agree how many documents can be assessed in an hour and twenty minutes. The mean was 87 documents.

When examining the assessments, the passages in the assessment files did not match the elements in the documents. This was because changes had been made to the original documents in order to use the X-Rai assessment tool. It is not clear how this effects the assessment overall (further investigation is required). This problem might be rectified in two possible ways. First, the translation to move all content into leaf node might be changed to avoid moving the relative location of text (even though it is just white-space). Second, the leaf-node requirement might be removed from the assessment tool.

Measuring the performance of each search engine against the two sets of assessments showed a strong positive correlation for the runs, but a weak correlation for the top performing runs. In answer to questions 2 and 3 in Section 2, the assessments collected over one hour and twenty minutes (per topic) are effective at separating good from bad runs, but in order to separate good from very good runs the pool cannot be reduced in size (to about 100 documents).

To answer to question 1 in Section 2, the agreement level of assessors was measured as the number of assessors was increased. Only about 8 assessors are needed before they stop agreeing which parts of a document are relevant, but 19 assessors are needed before they disagree on which documents are relevant. Relevance is in the mind of the assessors and assessors do not agree with each other.

When deciding on relevance, assessors do not agree on which factors are important, some think the content, while others think titles and keywords. The size of a relevant passage also varies across assessors with some identifying whole elements as relevant and others non-elemental passages.

It is pertinent to ask if we can at least agree on something. In answer: yes. We agree which runs performed well even though we don't agree on how we decided this.

8. ACKNOWLEDGEMENTS

Funded in part by a University of Otago Research Grant.

9. REFERENCES

[1] Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3), 149-159.

[2] Berkeley. (2005). Cystic fibrosis reference collection. Available: http://www.sims.berkeley.edu/~hearst/irbook/cfc.html [2005, 25 February].

[3] Clark, J., & DeRose, S. (1999). XML path language (XPath) 1.0, W3C recommendation. The World Wide Web Consortium. Available: http://www.w3.org/TR/xpath.

[4] Clarke, C. (2005). Range results in XML retrieval. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 4-5).

[5] Denoyer, L., & Gallinari, P. (2006). The wikipedia XML corpus. In *Proceedings of the INEX 2006 Workshop*.

[6] Malik, S., Tombros, A., & Larsen, B. (2006). The interactive track at INEX 2006. In *Proceedings of the INEX 2006 Workshop*.

[7] Malik, S., Trotman, A., Lalmas, M., & Fuhr, N. (2006). Overview of INEX 2006. In *Proceedings of the INEX 2006 Workshop*.

[8] Pehcevski, J. (2006). Relevance in XML retrieval: The user perspective. In *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology*, (pp. 35-42).

[9] Pehcevski, J., & Thom, J. A. (2005). HiXEval: Highlighting XML retrieval evaluation. In *Proceedings of the INEX 2005 Workshop*.

[10] Pharo, N., & Järvelin, K. (2006). "irrational" searchers and IR-rational researchers. *Journal of the American Society for Information Science and Technology*, 57(2), 222-232.

[11] Piwowarski, B., & Lalmas, M. (2004). Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of the 13th ACM conference on Information and knowledge management*, (pp. 361-370).

[12] Piwowarski, B., Trotman, A., & Lalmas, M. (2007 (un-published)). Sound and complete relevance assessments for XML retrieval.

[13] Prabha, C., Connaway, L. S., Olszewski, L., & Jenkins, L. R. (2007). What is enough? Satisficing information needs. *Journal of Documentation*, 63(1), 74-89.

[14] Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 53(2), 226-234.

[15] Trotman, A. (2005). Wanted: Element retrieval users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, (pp. 63-69).

[16] Trotman, A., N.Pharo, & Lehtonen, M. (2006). XML-IR users and use cases. In *Proceedings of the INEX 2006 Workshop*.

[17] Voorhees, E. M. (2001). The philosophy of information retrieval evaluation. In *Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, (pp. 355-370).

[18] Voorhees, E. M. (2003). Overview of TREC 2003. In *Proceedings of the 12th Text REtrieval Conference (TREC-12)*.