

Investigating the Exhaustivity Dimension in Content-Oriented XML Element Retrieval Evaluation

Paul Ogilvie
Language Technologies Institute
Carnegie Mellon University
Pittsburgh PA, USA
pto@cs.cmu.edu

Mounia Lalmas
Department of Computer Science
Queen Mary, University of London
England, UK
mounia@dcs.qmul.ac.uk

ABSTRACT

INEX, the evaluation initiative for content-oriented XML retrieval, has since its establishment defined the relevance of an element according to two graded dimensions, exhaustivity and specificity. The former measures how exhaustively an XML element discusses the topic of request, whereas specificity measures how focused the element is on the topic of request. However, obtaining relevance assessments is a costly task. In XML retrieval this problem is exacerbated as the elements of the document must also be assessed with respect to the exhaustivity and specificity dimensions. A continuous discussion in INEX has been whether such a sophisticated definition of relevance, and in particular the exhaustivity dimension, was needed. This paper attempts to answer this question through extensive statistical tests to compare the conclusions about system performance that could be made under different assessment scenarios.

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Measurement, Standardisation

Keywords

XML evaluation, relevance, statistical tests, INEX

1. INTRODUCTION

In recent years there has been an explosion in the amount of research and development for the retrieval of structured documents. This has been in large part an investigation of XML retrieval systems aimed at supporting content-oriented

retrieval. Instead of retrieving whole documents, many XML retrieval systems aim at retrieving document components, i.e. XML elements, of varying granularity that fulfill the user's query. As the number of XML retrieval systems increases, so does the need to evaluate their effectiveness. The INitiative for the Evaluation of XML retrieval (INEX)¹, established in 2002, is providing an infrastructure and methodology to evaluate these XML retrieval systems.

The typical approach to evaluate a system's retrieval effectiveness is with the use of test collections constructed specifically for that purpose. A test collection usually consists of a set of documents, topics, and relevance assessments. As many related elements within a document may be relevant (and may be retrieved by the XML retrieval system), the relevance assessments must reflect which elements are better to retrieve than others. To help solve this difficulty, INEX defines relevance according to two dimensions, *exhaustivity* and *specificity* [6]. Exhaustivity (*e*) measures how exhaustively an element discusses the topic. Specificity (*s*) measures how focused the element is on the topic. That is, it discusses no other, irrelevant topics.

Although there have been arguments against this separation at the INEX workshops, this solution is believed to provide a more stable measure of relevance than if assessors were asked to rate elements on a single scale. One reason is that assessors are likely to place varying emphasis on these two dimensions when assigning a single relevance value. For example, one assessor might tend to rate highly specific elements as more relevant, while another might to be more tolerant of lower specificity and prefer high exhaustivity.

In addition to the dimensions of exhaustivity and specificity, assessments at INEX have historically been done using *multiple grades* on each dimension. The use of a graded scale was deemed necessary to reflect the relative relevance of an element with respect to its sub-elements. For example, an element may be more exhaustive than any of its sub-elements alone given that it covers all of the aspects discussed in each of the sub-elements. Similarly, sub-elements may be more specific than their parent elements, given that the parent elements may cover multiple topics, including irrelevant ones.

However, obtaining relevance assessments is a very tedious and costly task [9]. A continuous discussion in INEX has been whether such a sophisticated definition of relevance, and in particular the exhaustivity dimension, was needed. The ultimate aim of an evaluation is to be able to state

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-433-2/06/0011 ...\$5.00.

¹<http://inex.is.informatik.uni-duisburg.de/>

that system A performs consistently better than system B. A simpler definition, e.g. using one dimension, reducing the scale, etc., would be less costly to obtain, and an analysis of the results may arrive at the same conclusion. In particular, it was always felt that having to assess the exhaustivity of an element could be avoided. More precisely, assessors have felt that gauging exhaustivity was a cognitively difficult task to perform, and that the extra burden has led to less consistent assessments [14] compared to those obtained at TREC.

In 2005, as described in Section 2.4, a new assessment procedure was adopted, including a reduced scale for the exhaustivity dimension. The latter was adopted after a thorough investigation carried out on the INEX 2004 data, which is reported in detail in Section 4.1. A separate investigation by Piwowarski et al. [10] shows that these changes led to better assessments with respect to the specificity dimension. It is also shown that higher rates of agreement were obtained with the specificity dimension than with the exhaustivity dimension, even with its reduced scale, thus still questioning the benefit of the exhaustivity dimension.

This paper attempts to provide an answer to the question of whether there is value added by the multi-graded scale for exhaustivity. The analysis in this paper examines this question through the use of quantisation functions to simulate different assessment procedures. These quantization functions map the exhaustivity and specificity dimensions to a single dimension which is used as input to evaluation measures. There is naturally some risk that changing the assessment procedure will change assessor’s behavior with respect to the specificity dimension. However, we feel that these simulations enable a more informed decision. This paper presents a thorough statistical analysis of the results when using the different quantisations. From this analysis, we make recommendations about the use of the exhaustivity dimension in XML element retrieval.

We first introduce the evaluation measures and data sets used at INEX (Section 2). Section 3 discusses the statistical significance tests, which we apply to the INEX content oriented retrieval tasks of 2004 and 2005 in Section 4. Section 5 summarizes the implications and concludes the paper.

2. INEX DATA SETS AND METRICS

Our investigation of the exhaustivity dimension is based on the INEX 2004 [6] and 2005 [7] data sets and metrics.

2.1 Documents

In 2004, the document collection consisted of the full-text of 12,107 articles, marked up in XML, from 12 magazines and 6 transactions of the IEEE Computer Society publications, covering the period of 1995-2002, and totaling 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. The overall structure of a typical article consists of a frontmatter (containing e.g. title, author, publication information and abstract), a body (consisting of e.g. sections, sub-sections, sub-sub-sections, paragraphs, tables, figures, lists, citations) and a backmatter (including bibliography and author information). In 2005, the collection was extended with total of 4,712 new articles from the IEEE Computer Society, giving a total of 16,819 articles, in size to a total of 764Mb, and around 11 millions elements.

2.2 Topics

The evaluation in this paper focuses on *Content-only (CO)* topics, which are traditional IR topics written in natural language and constrain the content of the desired results, e.g. only specify what an element should be about without specifying what that element is. As in TREC, an INEX CO topic consists of the standard title, description and narrative fields. Only the title was used as input to XML systems in both 2004 and 2005. INEX 2004 and 2005 have 29 and 34 topics with relevance assessments, respectively.

2.3 Retrieval task

In any IR evaluation campaign, participants are asked to submit runs that would satisfy specified retrieval tasks. The retrieval task investigated in this paper is the ad hoc retrieval for CO topics, which can be described as a simulation of how a library might be used. This task involves the searching of a static set of XML documents using a new set of CO topics. It is left to the retrieval system to identify the most appropriate XML elements to return.

In 2005, two CO sub-tasks were investigated, built on assumed user behaviors taking into account how the results may be presented. In the *focused* sub-task the aim was for systems to find the most exhaustive and specific element on a path within a given document and return to the user only this most appropriate unit of retrieval. In the *thorough* sub-task, the aim was for systems to estimate the relevance of potentially retrievable elements in the collection. This is task that most systems performed up to 2004 in INEX. Due to space constraints, our investigation is based on the thorough sub-task (although all conclusions drawn from analysis on this sub-task hold for the focused sub-task).

During INEX 2004 and 2005, participating organizations evaluated each topic set on the document collections in 2004 and 2005, respectively, and produced a list of XML elements as their retrieval results for each topic. The top 1,500 elements returned for each topic’s retrieval results were then submitted to INEX. 70 and 55 runs were submitted in 2004 and 2005, respectively, which are used in our investigation.

2.4 Assessments

In INEX 2004, exhaustivity and specificity were both measured on a four-point scale with degrees of highly (3), fairly (2), marginally (1), and not (0) exhaustive/specific. For example, (2, 3) denotes a fairly exhaustive and highly specific element, which means that it discusses many aspects of the topic of request and the topic of request is the only theme of the component.

Assessors were asked to provide an exhaustivity value and a specificity value for all elements forming the pool to assess². To ensure complete assessments, for any element assessed as relevant ($e, s > 0$), its parent, its children and its sibling also had to be assessed. This assessment procedure led to a very laborious assessment task, with questions regarding the quality of the assessments; e.g. lower agreement between assessments was found at INEX compared to those reported for TREC [14].

While the definition of the relevance dimensions was the same in INEX 2005, their scales were revised. The scale for exhaustivity was changed to 3 + 1 levels: highly exhaustive

²As in TREC, INEX uses a pooling method, to elicitate which elements to assess [9].

(2), somewhat exhaustive (1), not exhaustive (0) and ‘too small’ (?). The latter category of ‘too small’ was introduced to allow assessors to label elements, which although containing relevant information were too small to sensibly reason about their level of exhaustivity. This change in the exhaustivity scale resulted in part from the investigation detailed in Section 4.1. Specificity was measured on a continuous scale with values in $[0, 1]$, where 1 represents a fully specific component (i.e. contains only relevant information). For example, $(2, 0.72)$ denotes a highly exhaustive element, 72% of which is relevant content.

The assessment procedure used in INEX 2005 was a two-phase process. In the first phase, assessors highlighted text fragments containing only relevant information. The specificity value of an element was calculated as follows: a completely highlighted element had a s value of 1, whereas a non-highlighted element had a s value of 0. For all other elements, s was defined as the ratio (in characters) of the highlighted text (i.e. relevant information) to the element size. In the second phase, assessors assigned one of the 3 + 1 levels of the exhaustivity scale to those elements that intersected with any of the highlighted text fragments.

Although this new assessment procedure reduced the time needed to assess, and higher levels of agreement were obtained [8]³, it was still felt that similar results in terms of comparing systems effectiveness could be obtained without the exhaustivity dimension. This would mean that the assessment procedure would only include the highlighting phase, which is believed would greatly simplify the assessment task, both in terms of time and quality.

2.5 Quantisations

Given that INEX employs two graded relevance dimensions, the evaluation metrics used in INEX (see Section 2.6) require for these two to be combined. The quantization functions aim to do just that, by providing a relative ordering of the various combinations of (e, s) values and a mapping of these to a single relevance scale in $[0, 1]$. Tables 1 and 2 list the quantisation functions used on the INEX 2004 and 2005 data sets, respectively. Some were used officially in INEX as a means to model assumptions regarding the worth of retrieved elements to users. Others have been defined for this work as a means to test various assessment scenarios.

Strict₄ is used to evaluate retrieval methods with respect to their capability of retrieving highly exhaustive and highly specific elements in INEX 2004. The generalized (Gen₄) and the specificity-oriented generalized (SOG) functions credit elements according to their degree of relevance, hence allowing modeling varying levels of user satisfaction gained from not fully specific and highly exhaustive elements, such as less relevant elements or near-misses. The difference between Gen₄ and SOG is that the former shows slight preference toward the exhaustivity dimension, while the latter assumes that more specific elements are of greater value to the user. The Any Rel quantisation is used to evaluate retrieval methods that return any relevant element, regardless of their exhaustivity and specificity value (as long as $e > 0$ and $s > 0$). This quantisation allows the investigation of an assessment procedure where an assessor is only required to mark which elements contain relevant text.

Both Strict₅ and Gen₅ have the same intent as Strict₄

³The agreement level was calculated on very few topics, so any results should be taken as indications.

Name	Function
Strict ₄	$f(e, s) := \begin{cases} 1 & \text{if } e = 3 \text{ and } s = 3, \\ 0 & \text{otherwise.} \end{cases}$
Gen ₄	$f(e, s) := \begin{cases} 1 & \text{if } (e, s) = (3, 3), \\ 0.75 & \text{if } (e, s) \in \{(2, 3), (3, 2), (3, 1)\}, \\ 0.5 & \text{if } (e, s) \in \{(1, 3), (2, 2), (2, 1)\}, \\ 0.25 & \text{if } (e, s) \in \{(1, 2), (1, 1)\}, \\ 0 & \text{if } (e, s) = (0, 0). \end{cases}$
SOG	$f(e, s) := \begin{cases} 1 & \text{if } (e, s) = (3, 3), \\ 0.9 & \text{if } (e, s) = (2, 3), \\ 0.75 & \text{if } (e, s) \in \{(1, 3), (3, 2)\}, \\ 0.5 & \text{if } (e, s) = (2, 2), \\ 0.25 & \text{if } (e, s) \in \{(1, 2), (3, 1)\}, \\ 0.1 & \text{if } (e, s) \in \{(2, 1), (1, 1)\}, \\ 0 & \text{if } (e, s) = (0, 0). \end{cases}$
Any Rel	$f(e, s) := \begin{cases} 0 & \text{if } e = 0 \text{ and } s = 0, \\ 1 & \text{otherwise.} \end{cases}$

Table 1: Quantisations used for INEX 2004 analysis.

Name	Function
Strict ₅	$f(e, s) := \begin{cases} 1 & \text{if } e = 2 \text{ and } s = 1, \\ 0 & \text{otherwise.} \end{cases}$
Fully Spec	$f(e, s) := \begin{cases} 1 & \text{if } s = 1, \\ 0 & \text{otherwise.} \end{cases}$
Gen ₅	$f(e, s) := \begin{cases} e * s & \text{if } e \in \{1, 2\}, \\ 0 & \text{otherwise.} \end{cases}$
Gen Lifted	$f(e, s) := \begin{cases} (e + 1) * s & \text{if } e \in \{1, 2\}, \\ s & \text{if } e = ?, \\ 0 & \text{otherwise.} \end{cases}$
Bin Exh	$f(e, s) := \begin{cases} s & \text{if } e \in \{?, 1, 2\}, \\ 0 & \text{otherwise.} \end{cases}$

Table 2: Quantisations used for INEX 2005 analysis.

and Gen₄, but redefined for the INEX 2005 exhaustivity and specificity scale. Both functions ignore elements assessed as ‘too small’. The Gen Lifted quantization function was introduced to score too small elements as near-misses. Fully Spec and Bin Exh are quantisation functions that reward elements independently of exhaustivity. They simulate an assessment procedure where all that is required is the highlighting of relevant text (Section 2.4). Additional judgments on the exhaustivity of the XML elements would not be required to compute these two quantisations, so they are helpful in the simulation of what conclusions could be made in the absence of the exhaustivity dimension.

2.6 Evaluation Metrics

INEX 2004 and 2005 employ, respectively, *inex_eval* and XCG to measure effectiveness. These measures are used in lieu of the common measures of precision and recall, as they are better suited to handling graded assessments.

2.6.1 The *inex_eval* metric

This metric [6] applies the measure of *precall* [11] to elements and computes the probability $P(\text{rel}|\text{retr})(x)$ that an element is relevant for a given recall level x :

$$P(\text{rel}|\text{retr})(x) = \frac{x \cdot n}{x \cdot n + NNR_{x \cdot n}} \quad (1)$$

where n is sum of the quantisation scores for all relevant elements and $NNR_{x \cdot n}$ is the expected number of non relevant elements retrieved until the recall point x is reached. *pre-*

call is designed to handle multiple elements with the same retrieval status value (RSV). We denote the set of elements c at rank i having the same RSV as $rank(i)$. The rank at which recall level x is achieved at is:

$$r = \min \left\{ j : \sum_{i=1}^j \sum_{c \in rank(i)} f(assess(c)) \geq x \cdot n \right\} \quad (2)$$

where $assess(c) = (e, s)$ is the assessment of the element c and f is one of the quantisation functions in Table 1. That is, r is the minimum rank where the sum of all elements' quantisation scores up to and including that rank is at least $x \cdot n$. We define $nonrel(i)$ as the number of non-relevant elements at rank i :

$$nonrel(i) = \sum_{c \in rank(i)} I(f(assess(c)) = 0) \quad (3)$$

where I is an indicator function which returns 1 when its argument is true and 0 otherwise. We next define the amount of relevance left to be obtained from rank r :

$$left(r) = x \cdot n - \sum_{i=1}^{r-1} \sum_{c \in rank(i)} f(assess(c)) \quad (4)$$

$NNR_{x \cdot n}$ is estimated as:

$$\left(\sum_{i=1}^{r-1} nonrel(i) \right) + \frac{left(r) \cdot nonrel(r)}{r} \quad (5)$$

One can then estimate $P(rel|retr)(x)$ for various recall levels and average across queries, as is done for the common evaluation measure of mean average precision. A main problem with the *inex_eval* metric is that it does not handle overlapping elements in the context of XML retrieval evaluation [4], thus a different measure was adopted in 2005⁴.

2.6.2 The XCG metrics

The XCG measures [5] are an extension of the Cumulated Gain based measures proposed in [3], which were specifically designed for multi-graded relevance. The XCG measures include the user-oriented measures of extended cumulated gain ($nxCG[i]$) and the system-oriented effort-precision/gain-recall measures ($MAep$).

Given a ranked list of elements⁵, the cumulated gain at rank i , denoted as $xCG[i]$, is computed as the sum of the relevance scores up to that rank:

$$xCG[i] := \sum_{j=1}^i f(assess(e_j)) \quad (6)$$

For each query, an ideal gain vector, xCI , is derived by filling the rank positions with $f(assess(c'_j))$ in decreasing order for all assessed elements c'_j . A retrieval run's xCG vector is compared to this ideal ranking by plotting both the actual and ideal cumulated gain functions against the rank position. Normalised xCG ($nxCG$) is:

$$nxCG[i] := \frac{xCG[i]}{xCI[i]} \quad (7)$$

⁴The thorough task investigated in this paper does not leverage ability to penalize overlapping elements of the XCG measures. As such, we present a simplified version of the measure in this paper.

⁵Unlike the *inex_eval* measure, the XCG measures assume a fully ordered list and does not allow ties at a rank.

For a given rank i , $nxCG[i]$ reflects the relative gain the user accumulated up to that rank, compared to the gain he/she could have attained if the system would have produced the optimum best ranking, where 1 represents ideal performance. In INEX 2005 the officially reported cut-offs for $nxCG$ were $i = 10, 25, \text{ and } 50$.

The effort-precision ep at a given gain-recall value gr is defined as the number of visited ranks required to reach a given level of gain relative to the total gain that can be obtained. The measure of effort-precision ep is defined as:

$$ep[r] := \frac{i_{ideal}}{i_{run}} \quad (8)$$

where i_{ideal} is the rank position at which the cumulated gain of r is reached by the ideal curve and i_{run} is the rank position at which the cumulated gain of r is reached by the system run. A score of 1 reflects ideal performance, i.e. when the user needs to spend the minimum necessary effort to reach a given level of gain. The gain-recall gr is calculated as:

$$gr[i] := \frac{xCG[i]}{xCI[n]} = \frac{\sum_{j=1}^i xG[j]}{\sum_{j=1}^n xI[j]} \quad (9)$$

where n is the number of elements c where $f(assess(c)) > 0$.

This method follows the same viewpoint as standard precision/recall, where recall (here gain-recall) is the control variable and precision (here effort-precision) is the dependent variable. As with standard precision/recall, a non-interpolated mean average effort-precision, denoted $MAep$, is calculated by averaging the effort-precision values measured at natural recall-point, i.e. whenever a relevant XML element is found in the ranking.

3. STATISTICAL SIGNIFICANCE TESTS

This section reviews related work in statistical significance testing for IR evaluation and presents the approach used in this work. Section 3.1 briefly reviews the two types of errors in statistical significance testing. Section 3.2 presents an approach to control the rate of incorrectly identifying statistically significant differences when one makes many comparisons. The bootstrap statistical significance test is presented in Section 3.3. That subsection also presents experiments choosing which statistical significance test is most appropriate for the INEX data. Section 3.4 presents and motivates the methodology we use for comparing the effects of using different quantisation functions.

The use of statistical significance testing in IR (Hull [2]) helps researchers make informed decisions about the average performance of two or more retrieval approaches. For example, we may wish to know whether system A, whose score is 0.34 on one topic set and corpus, is significantly greater than system B, whose score is 0.28 on the same topic set and corpus. These data are paired; we have scored results for each system on the same topics. As such, an equivalent question is whether the difference between the score of system A and system B is greater than zero.

In this work, we focus on answering this single tailed question. That is, we consider the hypotheses

$$H_0 : \theta_{AB} \leq 0 \text{ versus } H_1 : \theta_{AB} > 0 \quad (10)$$

where θ_{AB} is the true difference of the mean score between systems A and B. This testing procedure assumes that topics are drawn from some probability distribution. The true

mean difference is given by this unknown probability distribution. As we do not know θ_{AB} , statistical significance tests can help us decide whether or not it is appropriate to reject the null hypothesis H_0 and declare (with some level of confidence) that the mean score of system A is greater than the mean score of system B.

One important issue is how one should choose a particular significance test among the myriad of possibilities. Hull [2] outlined a few statistical testing methods for comparing two systems: the paired-t test, the Wilcoxon signed rank test, and the sign test. The paired-t test has the strongest assumptions for the test to be valid: the observed differences should be normally distributed. Wilcoxon’s signed rank test has the less strong assumption that the differences are symmetric at the true mean. The sign test makes the weakest assumptions of the tests: on average half of the system differences should be greater than the true mean.

Generally, with weaker assumptions comes lesser power. The power of a significance test is the probability that the null hypothesis will be rejected. Tests that make stronger assumptions reject the null hypothesis more often. When choosing a statistical significance test, it is desirable to choose the most powerful statistical significance test where the data do not violate the assumptions of the test. Care must be taken so that assumptions of the test chosen closely match those of the data, otherwise many errors in the statistical analysis may occur.

3.1 Errors in Statistical Testing

Statistical significance tests make two kinds of errors. A type I error is a rejection of the null hypothesis when it is in fact true. A type II error results when the null hypothesis is retained when it should be rejected.

When a statistical test is applied then an acceptable type I error rate is chosen (α) and the test is applied at that level. In other words, the probability of a type I error is less than or equal to α . The p-value of a statistical significance test is defined to be the lowest α that the test would reject.

Controlling type I error is generally considered more important than type II error, as we do not want to assert that two systems perform differently when in fact they do not. We are generally willing to miss significant differences with the hope that a rejection of the null hypothesis truly identifies a statistically significant difference.

In related work, Sanderson and Zobel [12] observed that there were more type I errors than desirable when comparing many systems and testing at level α . They further proposed reducing the type I error rate by dismissing the smaller differences between systems as not significant when the statistical test asserts that the differences are significantly different. While this will reduce the type I error rate, statistics has provided more principled methods of controlling type I error during statistical significance testing, which we introduce in the following subsection.

3.2 Controlling Type I Error Rates

When comparing only two systems at level α , one can be fairly confident that the rejection/retention of the null hypothesis was reasonable. However, it is also common to compare several or many systems in a pairwise manner. When this is done, the probability of observing type I errors is larger. As we wish to control type I error to a reasonable level, we must correct for the fact that we are performing

multiple simultaneous hypothesis tests.

A classic way to control this is to control the family wise error rate, which would ensure that the probability of rejecting any of the null hypothesis falsely is less than or equal to α . Hull [2] presented some techniques for testing multiple systems using correction of the family wise error rate. Controlling the family wise error rate provides strong assurances about type I error rates. When one compares more than a few systems, controlling the family wise error rate typically results in very few rejections of the null hypothesis; few system differences are declared significantly different.

To address this problem, one may instead control the false discovery rate, which is the number of type I errors divided by the number of times that the null hypothesis was rejected. Controlling the false discovery rate at level α' gives us the assurance that no more than α' times the number of significant differences identified by the test are type I errors. To gain power over controlling the family wise error rate, we allow that when there are rejections of the null hypothesis, we are willing to accept that $\alpha' \cdot 100\%$ of the rejections may be type I errors. For example, in a retrieval experiment where we compare 10 systems and identify 20 statistically significant pair-wise differences at level $\alpha = 0.05$, we expect that $20 \cdot 0.05 = 1$ difference will be identified in error.

Benjamini and Yekutieli [1] describes a method to control the false discovery rate. This method operates on the p-values of any statistical significance test. This generality is greatly desirable as we may couple this procedure with the most powerful test that does not violate our assumptions about the data.

Let $p_{(1)}, p_{(2)}, \dots, p_{(m)}$ be the p-values resulting from the statistical significance tests in increasing order ($p_{(i)} \leq p_{(j)}$ if $i \leq j$). We define

$$k = \max \left\{ i : p_{(i)} \leq \frac{i}{c_m m} \alpha' \right\} \quad (11)$$

where

$$c_m = \begin{cases} 1 & \text{if the test statistics are independent} \\ \sum_{i=1}^m \frac{1}{i} & \text{otherwise} \end{cases} \quad (12)$$

We then take $p_{(k)}$ as the rejection threshold; we reject all null hypotheses where $p_{(i)} \leq p_{(k)}$. This method for controlling the false discovery rate is powerful when there are many comparisons, as long as the researcher is willing to accept that up to α' of the differences detected may not be true differences. This is a reasonable trade-off for the greatly increased power over the control for family wise error rate. As the tests statistics when comparing system pairs may be dependent (a particular system may be compared many times to other systems), the experiments in this paper use this method under the assumption that the test statistics are not independent.

3.3 Selection of the Significance Test

In this paper we focus our analysis on the outcome of statistical significance tests when comparing all pairs of systems over a variety of quantisation functions. We investigate in this work absolute differences between system scores. In the following sections, we will compare the results of statistical significance tests resulting from the use of different quantisation functions on the INEX retrieval task (Section 2.3). For this work to be as accurate as possible, we wish to choose the statistical significance test that produces the

lowest amount of errors. In addition to the tests described by Hull [2], we also consider the bootstrap.

3.3.1 The Bootstrap Statistical Significance Test

While there are parametric versions of the bootstrap, we will consider the non-parametric bootstrap, as it makes no assumptions about the distribution or continuity of the underlying distributions. The bootstrap simulates b repeated observations of the test statistic by sampling with replacement from the original data (we sample topics in this work). It is often used to estimate confidence intervals, mean values, and variance [13], but it can also be adapted to perform a statistical significance test. To test the null hypothesis these simulated test statistics $\hat{\theta}_{AB}^i$ are sorted in increasing order. Each $\hat{\theta}_{AB}^i$ is viewed as an estimate of the test statistic. If $\hat{\theta}_{AB}^{(\alpha b)} > 0$, then H_0 is rejected and the two systems are declared to have different performance. The p-value of the bootstrap test is computed as

$$\max_{1 \leq i \leq b} \left\{ \frac{i}{b} : \hat{\theta}_{AB}^{(i)} \leq 0 \right\} \quad (13)$$

Another way to think of the p-value for the bootstrap test is that it is the fraction of bootstrap samples where the difference between systems A and B is less than or equal to zero. Generally a large number of simulations should be performed to get reasonable estimates (we take $b = 10,000$).

3.3.2 Estimating the Error Rate

To choose the statistical significance test with the lowest error rate, we examined the error rate of the paired-t, Wilcoxon signed rank, sign, and bootstrap tests on the thorough task of INEX 2005. To estimate the error rates, we perform an experiment similar to that of Sanderson and Zobel [12] and Voorhees and Buckley [15].

As in previous work, we estimated the error rate using 50 repeated splits of the topic set. On each topic set split, we applied each of the statistical significance tests to the system pairs using the first half of the topics. We took and set aside the set of systems where the null hypothesis was rejected (a difference was found). The error rate was defined to be the percentage of systems in this set where the difference between systems' mean scores on the other half of topics specified by the split was not positive.

When performing this experiment we observed that the bootstrap test made the fewest estimated errors. The error rate for the bootstrap test using the Gen₅ quantisation and the MAep evaluation measure was 8.3% to 14% of error rate observed when using the other significance tests (paired-t, Wilcoxon signed rank, sign). For the Strict₅ quantisation, the error rate of the bootstrap test was 23% to 27% of the other tests. We also observed that the bootstrap test rejected the null hypothesis on both halves of the topic splits more frequently than the other tests. That is, the bootstrap agrees with itself more than the other tests examined.

What is even more remarkable is that the bootstrap test tended to reject the null hypothesis more often than any of the other tests. For the generalized quantisation and the MAep measure, the bootstrap identified on average 2.1 times more significant differences than the other tests. For the strict quantisation, the bootstrap identified on average 64 times more significant differences. For the generalized quantisation applied to the nxCG at 10, 25, and 50 results, the bootstrap test identified on average 20, 8.8, and 3.7 times

as many statistically significant differences (respectively) as the other statistical significance tests.

On these measurements of error, it appears that not only does the bootstrap test have a lower type I error rate, it seems to be a more powerful test. As the test seems to reject more of the null hypotheses than the other tests while still maintaining a lower type I error rate we believe that the bootstrap test is retaining the null hypothesis for fewer of the cases when it should truly be rejected. This leads us to believe that the bootstrap test has a lower type II error rate for the INEX data and evaluation measures. Given these observations, we limit the analysis in the rest of this paper to applications of the bootstrap statistical significance test.

3.4 Comparing Quantisations

Common approaches to compare rankings use measures such as Spearman's and Kendall's rank correlation statistics. We avoid using these measures as they can be misleading. For example, consider the case where we have two groups of systems of equal size, G_1 and G_2 . For two evaluation measures produced using quantisation functions Q_1 and Q_2 , all systems perform about the same within either group. However, for both measures all systems in group G_1 perform better than all systems in group G_2 . Suppose that performing statistical significance tests correctly reject the null hypothesis when comparing systems in G_1 to those in G_2 while also correctly retaining the null hypothesis when comparing systems within a group.

If quantisations Q_1 and Q_2 order the systems within each group in reverse order, then any Spearman's rho and Kendall's tau will both be less than 1. For example, if $n = 20$, then Spearman's rho is 0.5 and Kendall's tau is 0.026. In this case, Spearman's rho does identify a statistically significant correlation between the rankings, while Kendall's tau does not. Similarly, Pearson's correlation would also be less than 1. However, even though Spearman's and the Pearson's tests would probably correctly identify correlation, both tests ignore whether two systems can be distinguished from each other or not. A better analysis would conclude that the two measures are equivalent, because the statistical tests made on the pairwise comparisons for either measure would correctly distinguish the groups.

If we directly compare the outcome of statistical significance tests, we would conclude that the two quantisations are equivalent. It is for this reason we choose to compare the quantisations by examining the results of the statistical tests made when performing pairwise comparisons. Furthermore, comparing the results of the statistical tests will allow us to use easily interpretable statistics, while correlation measures are not generally easy to interpret.

When we compare the statistical decisions made by two different quantisations, we can use the results of the tests performed using one quantisation Q_1 to predict the results of the other quantisation Q_2 . In this case, we treat Q_2 as the 'ground truth'. We thus can use the standard measures of retrieval performance, where $reject(Q)$ is the set of system pairs rejected when applying statistical significance tests to the results when using quantisation Q :

$$recall(Q_1, Q_2) = \frac{|reject(Q_1) \cap reject(Q_2)|}{|reject(Q_2)|}$$

inex_eval

	Strict ₄	Gen ₄	SOG	Any Rel
Recall				
Strict ₄ - 51%	1	0.61	0.60	0.60
Gen ₄ - 79%	0.96	1	0.96	0.96
SOG - 77%	0.92	0.94	1	0.92
Any Rel - 81%	0.96	0.99	0.97	1
F1				
Strict ₄ - 51%	1	0.75	0.73	0.74
Gen ₄ - 79%	0.75	1	0.95	0.97
SOG - 77%	0.73	0.95	1	0.94
Any Rel - 81%	0.74	0.97	0.94	1

Table 3: Quantisation agreement for the INEX 2004 CO Task. Cells correspond to Recall/F1 values given by using the significant differences identified by the quantisation in the row to predict those of the quantisation in the column. Precision may be read from the table by using the quantisation in the column to predict the row. The percentages in the first column are the percent of possible significant differences that were actually identified.

$$precision(Q_1, Q_2) = \frac{|reject(Q_1) \cup reject(Q_2)|}{|reject(Q_1)|}$$

$$F1(Q_1, Q_2) = \frac{2 \cdot recall(Q_1, Q_2) \cdot precision(Q_1, Q_2)}{recall(Q_1, Q_2) + precision(Q_1, Q_2)} \quad (14)$$

These measures give us interpretable measures of correlation between the two rankings. When the results of all of the statistical significance tests on both quantisations agree, recall, precision, and F1 will all be one.

4. ANALYSIS

This section examines the various quantisation functions used at INEX for the CO retrieval tasks of 2004 and 2005. Section 4.1 reviews previously unpublished analysis of the INEX 2004 data that was used by the organizers of INEX to motivate the reduction of the grades in the exhaustivity dimension in 2005. Section 4.2 presents new results from analyzing the INEX 2005 data.

4.1 Analysis of INEX 2004

At the INEX 2004 workshop, assessors commented that multi-graded assessments were time-consuming to perform. There were also concerns that the graded scales may also lead to low inter-assessor agreement. It was generally believed that reducing or eliminating the graded scales would lead to a more consistent and less time consuming assessment procedure. In addition, there was also confusion about whether the quantization functions measured different aspects of the retrieval systems.

In this section we briefly review our motivation for recommending reducing the scale of exhaustivity from 0,1,2,3 used in 2004 to a scale of 0,1,2. We first examine the official quantisation functions, investigating whether we would draw the same conclusions about systems when using these different quantisation functions (Section 4.1.1). We then investigate whether the 0,1,2,3 exhaustivity/specificity scales were necessary or whether the conclusions could be drawn from 0,1 or 0,1,2 scales (Section 4.1.2).

4.1.1 Comparison of official quantisations

We first review whether the official quantisations used in INEX 2004 identified the same statistically significant differences. From Table 3 we observe that Gen₄ and SOG provide very similar rankings (F1=0.95). There is very high agreement about which systems perform statistically significantly differently from each other between these two quantisation functions. Although these two quantisation functions express different user preferences (Section 2.5), they behave very similarly when ranking systems.

However, the Strict₄ quantisation function identifies a quite different set of statistically significant different system pairs. From Table 3, we see that F1 of the identified system differences of Strict₄ and the Gen₄ and SOG quantisations is 0.75 and 0.73, respectively. This is rather low. Looking at the top part of the table, we see that the Gen₄ and SOG quantisations have high recall of Strict₄ system differences (0.96 and 0.92), but that their precision is quite low (0.61 and 0.60). That is, the differences between systems identified when using the Strict₄ quantisation function are roughly a subset of those identified when using either of the Gen₄ and SOG quantisation functions.

This is also reflected in the fact that the Strict₄ quantisation function tends to identify fewer statistically significant differences (51% of all possible) than the Gen₄ (79%) and SOG (77%). From this, we observe that the Gen₄ and SOG quantisation functions are better at distinguishing systems than the Strict₄ quantisation function.

In summary, the Gen₄ and SOG quantisation functions behave similarly despite the fact that they emphasize different preferences for the exhaustivity and specificity of elements. The Strict₄ quantisation function identifies fewer differences between systems than the other quantisation functions, suggesting that it measures different system behavior. However, the system differences identified using the Strict₄ quantisation function tend to be also identified by the Gen₄ and SOG quantisation functions.

4.1.2 Value of graded exhaustivity and specificity

We next investigated whether multiple grades were necessary. The Gen₄ and SOG quantisations both rely heavily on the graded scales of exhaustivity and specificity. The Strict₄ quantisation requires knowledge of which components are highly exhaustive and highly specific. Here we wanted to know whether we can predict any of the quantisations used in INEX 2004 without leveraging the exhaustivity and specificity scales.

The Any Rel quantisation function was designed to investigate this question. It gives equal merit to any relevant element, regardless of how exhaustive or how specific each element is. Perhaps surprisingly, we see that the Any Rel quantisation tends to predict the same differences as the Gen₄ and SOG quantisation functions (F1 is 0.97 and 0.94). This suggests that for examining the behavior of systems on the INEX 2004 data, graded exhaustivity and specificity is not necessary to identify the system differences that would be identified by the Gen₄ and SOG quantisation functions.

4.1.3 Summary

The analysis on INEX 2004 data resulted in the observation that the Any Rel, Gen₄, and SOG quantisations all identified a similar set of statistically significant different system pairs. However, the Strict₄ quantisation identified a smaller set of differences that were typically identified when

using the other quantisation functions.

The fact that the Any Rel quantisation function does not need graded assessments has positive implications on the assessment procedure that could be used if the ability to identify the differences found when using the Strict₄ is not important. In such a case, the assessor would only have to mark which elements contain relevant text; no further assessment would be required.

Even with the requirement that the Strict₄ quantisation remain a part of the evaluation, one can reduce the exhaustivity and specificity scales. A reduced scale of 0, 1, 2 where 0 corresponds to not exhaustive/specific, 1 to somewhat exhaustive/specific, and 2 to highly exhaustive/specific would be good enough to identify most of the statistically significant system differences for the Strict₄, Gen₄, and SOG, quantisation functions. Given this analysis, INEX chose to reduce the number of grades in the scale for exhaustivity. No recommendation was made for the specificity scale, as the assessment procedure was changed in INEX 2005 (Section 2.4). The new highlighting process allowed for a continuous scale of specificity to be calculated automatically.

4.2 Analysis of INEX 2005

This section aims to analyze: (1) whether considering elements marked as ‘too small’ as relevant has a large impact on which systems are identified as statistically significantly different (Section 4.2.1); and (2) what effect further reducing the exhaustiveness scale to a binary scale would have on the conclusions that could be made (Section 4.2.2).

4.2.1 Too small elements

When assessors were judging an article, they could mark all of the remaining un-assessed elements in an article as ‘too small’ and continue to the next article. The assessors were trusted to use this feature with care, but this was not always the cases. There were situations where large elements (hundreds of words) were marked as ‘too small’. During and after the INEX 2005 workshop, INEX participants expressed concerns that abuse of the ‘too small’ may have resulted in poor system rankings.

The reason was that the generalized quantisation (Gen₅) assigns a score of zero to these elements marked as ‘too small’. To investigate whether this had a large impact on the rankings of systems, the organizers introduced the generalized lifted (Gen Lifted) quantisation, which allowed ‘too small’ elements to be considered relevant, but less relevant than those that had an exhaustivity of one or more.

Table 4 shows that there is reasonably high agreement between the Gen₅ and Gen Lifted quantisations as F1 is 0.9 for both nxCG at rank 25 and MAep. In the interests of saving space, we omit results on nxCG at ranks 10 and 50. The results are similar to those of nxCG at a cut-off of 25 elements, although, nxCG at 10 tends to be more variable and less able to distinguish systems whereas nxCG at 50 tends to identify more statistically significant differences. Despite these differences, levels of agreement are similar to those observed for nxCG at 25 and MAep.

It is also illustrative to examine where the disagreements between the quantisation functions occur. Figures 1a and 1b shows this using the MAep evaluation measure. Figure 1a is a scatter plot of the differences identified as significant using either the Gen₅ or Gen Lifted quantisation. We see that there is high correlation among the differences in

MAep, and that when the two quantizations disagree about whether the systems behave differently or not, the differences in MAep tend to be the smaller differences observed on the plot. Figure 1b presents those differences in MAep using a density plot. In the figure the x-axis corresponds to the difference between system pairs using the Gen₅ quantisation. The figure clearly demonstrates that most of the errors occur with lower absolute differences. The larger the differences in MAep, the more like each other the two quantisations behave. This is encouraging, as the prediction of the Gen Lifted quantisation could be improved with a simple threshold applied to the differences in scores.

In summary, we see that there is reasonably high agreement between which system pairs are identified as significantly different when using the Gen₅ and Gen Lifted quantisations. When they do disagree, the differences in scores tend to be smaller than the differences in scores where the quantisations do agree. From this, we conclude that it does not make a large difference in the rankings whether or not the ‘too small’ elements are considered relevant.

4.2.2 Value of graded exhaustivity

We next consider whether we can simulate the behavior of either generalized quantisation and the Strict₅ quantisations using only the specificity dimension of relevance. If this were the case, it would eliminate the need for assessors to judge elements with respect to the exhaustivity dimension.

We first ask whether we can simulate the behavior of either the Gen Lifted or Gen₅ quantisation without the use of the exhaustivity dimension. The quantisation we examine for this is binary exhaustivity (Bin Exh), which simulates an assessment procedure where the user highlights all relevant text (Section 2.5). Table 4 shows that Bin Exh agrees highly with Gen Lifted. F1 is again quite high for both the nxCG at rank 25 and MAep with values of 0.87 and 0.93, respectively. Scatter plots and density plots for these comparisons look similar to those in Figures 1a-b, but have been omitted to save space. We can conclude that Bin Exh function simulates the behavior of Gen Lifted.

However, Bin Exh is not an ideal predictor of the Gen₅ quantisation function. This could be in part a side effect of treating the ‘too small’ elements as relevant. We considered a variant of Bin Exh which did not treat ‘too small’ elements as relevant ($f(?, s) = 0$). With this variant resulted in an F1 of 0.88 and 0.93 for the nxCG at rank 25 and MAep evaluation measures. This is encouraging, and we hypothesize that a simple thresholding based on the element size to automatically filter out most ‘too small’ elements would gain similar levels of agreement with the Gen₅ quantisation. We leave this to future work.

The ability to reasonably simulate Gen Lifted without the use of the exhaustivity dimension is similar to our findings with the Any Rel quantisation function on the INEX 2004 data (Section 4.1.2). Again we see that a graded scale for exhaustivity is not necessary for simulating rankings that treat somewhat exhaustive elements as relevant.

We now look whether Strict₅ can be simulated. In Section 4.1.3 we outlined our previous recommendation to preserve a graded scale for exhaustivity. Doing so allowed the continued use of a strict quantisation function, which only considers highly exhaustive and highly specific text as relevant. We investigate whether the Fully Spec quantisation (Section 2.5), which considers only the highly specific el-

nxCG25						MAep					
	Strict ₅	Fully Spec	Gen ₅	Gen Lifted	Bin Exh		Strict ₅	Fully Spec	Gen ₅	Gen Lifted	Bin Exh
Recall						Recall					
Strict ₅ - 15%	1	0.27	0.31	0.31	0.25	Strict ₅ - 35%	1	0.52	0.55	0.55	0.51
Fully Spec - 53%	0.98	1	0.69	0.74	0.78	Fully Spec - 63%	0.93	1	0.76	0.84	0.81
Gen ₅ - 42%	0.88	0.54	1	0.86	0.78	Gen ₅ - 59%	0.92	0.71	1	0.89	0.83
Gen Lifted - 47%	0.98	0.66	0.96	1	0.87	Gen Lifted - 61%	0.95	0.81	0.92	1	0.95
Bin Exh - 46%	0.82	0.68	0.86	0.87	1	Bin Exh - 60%	0.86	0.77	0.85	0.92	1
F1						F1					
Strict ₅ - 15%	1	0.42	0.45	0.47	0.39	Strict ₅ - 35%	1	0.67	0.69	0.70	0.64
Fully Spec - 53%	0.42	1	0.61	0.70	0.72	Fully Spec - 63%	0.67	1	0.73	0.82	0.79
Gen ₅ - 42%	0.45	0.61	1	0.90	0.81	Gen ₅ - 59%	0.69	0.73	1	0.90	0.84
Gen Lifted - 47%	0.47	0.70	0.90	1	0.87	Gen Lifted - 61%	0.70	0.82	0.90	1	0.93
Bin Exh - 46%	0.39	0.72	0.81	0.87	1	Bin Exh - 60%	0.64	0.79	0.84	0.93	1

Table 4: Quantisation agreement for the INEX 2005 thorough sub-task. Format of the table is the same as Table 3.

ements ($s = 1$) as relevant, can be used to simulate the Strict₅ quantisation function. Unfortunately, the Fully Spec is not a good predictor of Strict₅ (Table 4).

The Fully Spec quantisation is able to recall a large portion of the differences identified when using the strict quantisation, but it also predicts many more differences. Figures 1c-d illustrate where the erroneously predicted differences occur with respect to system differences. In these plots, the Fully Spec quantisation is used to predict the Strict₅ quantisation applied to MAep. We see from these plots that there is a large degree of overlap between the region corresponding to correctly identified differences and those that Fully Spec identified in error. Applying a threshold on the differences in score will not improve the predictions made. We conclude from this analysis that it is difficult to simulate the behavior of the Strict₅ quantisation without the use of the exhaustivity dimension.

4.2.3 Summary and Discussion

We investigated two questions on the INEX 2005 data. We first studied whether considering items assessed as ‘too small’ as relevant would greatly change rankings. We found that ‘too small’ elements did not have a large impact.

We also investigated the value of the graded exhaustivity scale. We first found that a graded exhaustivity scale is not necessary to simulate the results of the Gen Lifted quantisation. We also found that knowledge of which elements are ‘too small’ is necessary for good simulation of the Gen₅ quantisation, but we hypothesize this could be approximated with a simple threshold based on the element’s length in characters.

As with the Strict₄ quantisation of INEX 2004, we were unable to simulate the Strict₅ quantisation without the use of a graded exhaustivity scale. However, we argue that the Strict₅ quantisation function is not a very useful measure of system performance on the INEX 2005 data. Table 4 shows that the nxCG at rank 25 only identified 15% of all possible system differences as significant, while using MAep, it identified 35% of all possible differences as statistically significant. Both numbers are very low, and very few systems are distinguishable when using the Strict₅ quantisation function. This makes results analysis very difficult, as few conclusions can be drawn with any certainty.

Given that the Strict₅ quantisation proved to be unable to distinguish retrieval systems, we recommend that it be omitted from future INEX evaluations. Doing this would open

the door to further simplification of the assessment process used at INEX. With only requiring assessors to highlight the most specific text, INEX could make use of the binary exhaustivity (Bin Exh) quantisation function, which only leverages specificity. Since this quantisation function reasonably simulates the behavior of the Gen Lifted quantisation function currently used at INEX, there would be consistency in the task from INEX 2005 to the coming years.

5. CONCLUSIONS

INEX defines relevance according to a specificity and an exhaustivity dimensions, themselves defined on a graded scale. Obtaining relevance assessments is very costly, in particular in the context of XML retrieval, where elements in addition to documents have to be assessed. For the purpose of comparing retrieval effectiveness, it has been argued in INEX that such a complex definition of relevance was not needed. This paper provides an answer to this argument, through extensive statistical tests.

This paper introduced several statistical tools useful for the evaluation of retrieval systems. In particular, we introduced the approach in [1] to control the false discovery rate for multiple test correction. This method is well suited for comparisons of very many retrieval systems.

This paper also introduced a framework for investigating the potential impact of changing an evaluation measure (such as using a different quantisation function) on the statistical conclusions that can be made about system effectiveness. Specifically, we demonstrated the application of this approach to investigating the impact of reducing the grades in a scale and the omission of the exhaustivity scale for XML component retrieval.

Using these tools, we performed analysis of the exhaustivity and specificity dimensions used in INEX 2004 and 2005. In this analysis, we found that the strict quantisations do not distinguish systems as well as the generalized quantisations. As such, it is a more difficult evaluation standard. Coupled with our findings that many of the generalized class of quantisations can be simulated without the use of assessments of exhaustivity, we feel that it is prudent to drop the strict quantisations from evaluation at INEX.

Another positive side effect of dropping assessment of exhaustivity would be more topics with assessments. Indeed, INEX has only 34 and 29 topics assessed for the 2004 and 2005 CO tasks, respectively, whereas a typical evaluation at TREC has 50 topics assessed. Voorhees and Buckley [15]

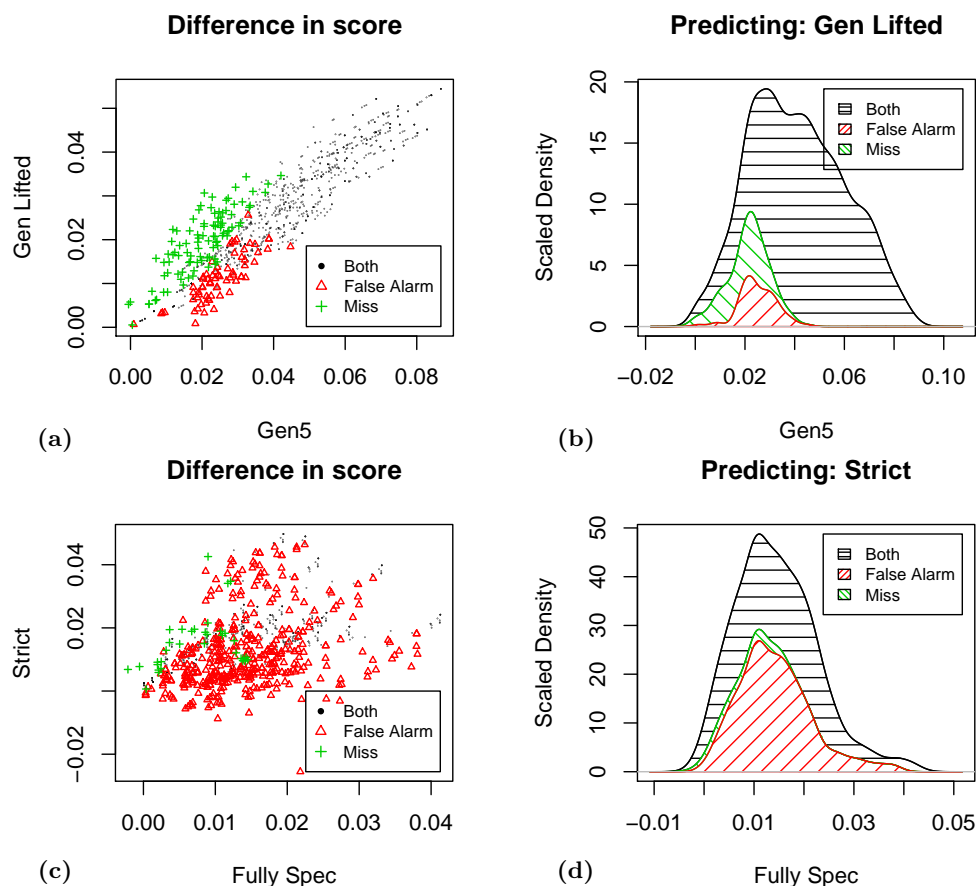


Figure 1: Scatters plot (left) of pairwise system differences identified as significant by the quantisation functions labeled on the axes for the thorough sub-task of INEX 2005. On the right are stacked kernel density estimates of the same data, using the differences of the quantization labeled on the x-axis. A ‘miss’ is a rejection of the null hypothesis identified by the quantisation presented on the y-axis but a retention of the null hypothesis by the quantisation used on the x-axis. A ‘false alarm’ is a retention of the null hypothesis by the quantisation of the y-axis when the null hypothesis was rejected when using the quantisation of the x-axis.

and Sanderson and Zobel [12] have observed that more systems are distinguishable when using more topics. Having more topics assessed at INEX would have a similar effect, enabling researchers to draw more reliable conclusions about their research from the evaluation at INEX.

Acknowledgements

The INEX initiative is an activity of DELOS, a network of excellence for digital libraries. Paul Ogilvie was funded in part by NSF grant IIS-0534345. Any opinions, findings, conclusions or recommendations expressed in this paper are the authors’, and do not necessarily reflect those of the sponsor.

6. REFERENCES

- [1] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- [2] D. Hull. Using statistical testing in the evaluation of retrieval experiments. *SIGIR*, 1993.
- [3] K. Järvelin and J. Kekäläinen. Cumulated Gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.
- [4] G. Kazai, M. Lalmas and A. P. de Vries. The Overlap Problem in Content-Oriented XML Retrieval Evaluation. *SIGIR*, 2004.
- [5] G. Kazai and M. Lalmas. INEX 2005 Evaluation Metrics. *INEX 2005 Proceedings*, 2006, In press.
- [6] S. Malik, M. Lalmas and N. Fuhr. Overview of INEX 2004. *INEX 2004 Proceedings*, 2005.
- [7] S. Malik, G. Kazai, M. Lalmas and N. Fuhr. Overview of INEX 2005. *INEX 2005 Proceedings*, 2006.
- [8] J. Pehecvski and J. A. Thom. Hixeval: Highlighting XML retrieval evaluation. *INEX 2005 Proceedings*, 2006.
- [9] B. Piwowarski and M. Lalmas. Providing Consistent and Exhaustive Relevance Assessments for XML Retrieval Evaluation. *CIKM*, 2004.
- [10] B. Piwowarski, A. Trotman and M. Lalmas. Sound and Complete Relevance Assessments for XML Retrieval. Submitted, 2006.
- [11] V.V. Raghavan, P. Bollmann and G.S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *TOIS* 7(3):205–229, 1989.
- [12] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. *SIGIR*, 2005.
- [13] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *IP&M* 33(4):495–512, 1997.
- [14] A. Trotman. Wanted: Element retrieval users. *INEX 2005 Workshop on Element Retrieval Methodology*, 2005.
- [15] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. *SIGIR*, 2002.
- [16] L. Wasserman. *All of Statistics*. Springer, 2004.