Report of the INEX 2003 Metrics working group

Gabriella Kazai
Department of Computer Science
Queen Mary University of London
gabs@dcs.qmul.ac.uk

1 INTRODUCTION

This paper summarises the discussions of the metrics working group at the INEX 2003 Workshop, Dagstuhl, Dec 15-17 2003. Members of the group were Djoerd Hiemstra (U. of Twente), Jaap Kamps (ILLC, U. of Amsterdam), Gabriella Kazai (Queen Mary U. of London), Yosi Mass (IBM Haifa), Vojkan Mihajlovic (U. of Twente), Paul Ogilvie (Carnegie Mellon U.), Jovan Pehcevski (RMIT U.), Arjen de Vries (CWI) and Huyen-Trang Vu (LIP 6).

The aim of this workshop was to review the current INEX metrics, collect issues and concerns regarding the suitability of these metrics for the evaluation of content-oriented XML retrieval approaches, and to propose alternative solutions.

The discussions started with a summary of the evaluation objectives and the evaluation considerations to be taken into account (section 2). This was followed by an overview of the current INEX metrics (section 3) and the presentation of proposed new metrics (section 5). The results of the discussions are summarised in sections organised by the topic of the discussion: section 4 summarises the issues, opinions and suggestions with respect to the current metrics, section 6 reflects the comments the proposed metrics received and finally section 7 summarises any other voiced issues.

2 EVALUATION SETUP

2.1 What to evaluate?

INEX'03 defines three tasks: the CO (content-only), SCAS (strict content-and-structure) and VCAS (vague content-and-structure) ad-hoc retrieval of XML documents. Given the different retrieval paradigms these tasks are based on, it is necessary to define the objective of the evaluation separately for all three tasks.

Within the CO task, the aim of an XML retrieval system is to point users to the specific relevant portions of documents, where the user's query contains no structural hints regarding what the most appropriate granularity of relevant XML elements should be. Here the evaluation of a system's effectiveness should hence provide a measure with respect to the system's ability to retrieve components that are both exhaustive and specific to the user's request, where highly exhaustive and highly specific components should be ranked first.

Within the SCAS task, the aim of a retrieval system is to retrieve relevant nodes that strictly match the structural conditions specified within the query. The evaluation criterion should hence only consider a match between a result and a reference element if these conditions have been met.

In the VCAS task, the goal of a system is to retrieve relevant nodes that may not exactly conform to the structural conditions expressed within the user's query, but are structurally similar. The evaluation criteria employed here must therefore allow for a more flexible match between result and reference elements.

Within the workshop, only the evaluation of the CO task was considered in detail.

2.2 What to consider?

The evaluation considerations mentioned here are detailed in [4]. These were mostly just summarised and agreed upon in the workshop, but not discussed in detail.

The first consideration is that a measure of effectiveness within the framework of the INEX initiative must be able to integrate the two dimensions of relevance: exhaustivity and specificity. Second, it was acknowledged that the independence assumption of classical IR, according to which the relevance of a document is independent of the relevance of any other document, does not hold in INEX. This issue was then discussed in more detail when trying to address the problem of overlapping result elements (section 4.1). Another important factor that the group members agreed should be taken into consideration is the varying user effort associated with result elements due to the varying size (length) of returned components. This is already addressed by one of the current INEX metrics (inex-2003), and some of the new proposals have also integrated this parameter within their model (section 5). The final aspect listed was that of linear vs. non-linear output rankings. It was agreed to only concentrate on linear ordering.

3 OVERVIEW OF CURRENT INEX METRICS

This section gives a brief summary of the inex-2002 (aka. inex_eval) and inex-2003 (aka. inex_eval_ng) metrics in order to provide the necessary background information for their discussion in section 4. For a more detailed description of the metrics please refer to [3, 4].

3.1 The inex-2002 metric

The inex-2002 metric applies the measure of *precall* [10] to document components and computes the probability $P(rel|retr)$ that a component viewed by the user is relevant:

$$P(rel|retr)(x) := \frac{x \cdot n}{x \cdot n + esl_{x \cdot n}} \qquad (1)$$

where $esl_{x \cdot n}$ denotes the *expected search length* [1], i.e. the expected number of non-relevant elements retrieved until an arbitrary recall point x is reached, and n is the total number of relevant components with respect to a given topic.

To apply the above metric, the two relevance dimensions were first mapped to a single relevance scale by employing a quantisation function, $\mathbf{f}_{quant}(e,s)\colon ES \to [0,1]$, where ES denotes the set of possible assessment pairs (e,s):

$$\begin{aligned} ES = \{&(0,0),(1,1),(1,2),(1,3),\\ &(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}\end{aligned}$$

Two quantisation functions were used: $\mathbf{f}_{strict}$ (Equation 2) and $\mathbf{f}_{gen}$ (Equation 3). The former is used to evaluate retrieval methods with respect to their capability of retrieving highly exhaustive and highly specific document components. The generalised function credits document components according to their *degree of* relevance.

$$\mathbf{f}_{strict}(e,s) := \begin{cases} 1 & \text{if } e=3 \text{ and } s=3,\\ 0 & \text{otherwise.}\end{cases} \qquad (2)$$

$$\mathbf{f}_{gen}(e,s) := \begin{cases} 1 & \text{if } (e,s)=(3,3),\\ 0.75 & \text{if } (e,s)\in\{(2,3),(3,\{2,1\})\},\\ 0.5 & \text{if } (e,s)\in\{(1,3),(2,\{2,1\})\},\\ 0.25 & \text{if } (e,s)\in\{(1,2),(1,1)\},\\ 0 & \text{if } (e,s)=(0,0).\end{cases} \qquad (3)$$

3.2 The inex-2003 metric

A problem with the inex-2002 metric is that it ignores possible overlaps between result elements and rewards the retrieval of a relevant component regardless if it has already been seen by the user either fully or in part.

The inex-2003 metric aims to provide a solution to this problem by incorporating component size and overlap within the definition of recall and precision (Equations 4 and 5). (For the derivation of the formulae based on an interpretation of the relevance dimensions within an ideal concept space [12] refer to [4].) Instead of measuring, e.g., precision or recall after a certain number of document components retrieved, the total size of the retrieved document components is used as the basic parameter, while overlap is accounted by considering only the increment to the parts of the components already seen. The calculations here assume that relevant information is distributed uniformly throughout a component.

$$\text{recall}_o = \frac{\sum_{i=1}^{k} e(c_i) \cdot \frac{|c_i'|}{|c_i|}}{\sum_{i=1}^{N} e(c_i)} \qquad (4)$$

$$\text{precision}_o = \frac{\sum_{i=1}^{k} s(c_i) \cdot |c_i'|}{\sum_{i=1}^{k} |c_i'|} \qquad (5)$$

Components $c_1, \ldots, c_k$ in Equations 4 and 5 form a ranked result list, N is the total number of components in the collection, $e(c_i)$ and $s(c_i)$ denote the quantised assessment value of component c_i according to the exhaustivity and specificity dimensions, respectively, $|c_i|$ denotes the size of the component, while $|c_i'|$ is the size of the component that has not been seen by the user previously. Given a component representation such as a set of (term, position) pairs, $|c_i'|$ can be calculated as:

$$|c_i'| = |c_i - \bigcup_{c \in C[1,n-1]} (c)| \qquad (6)$$

where n is the rank position of c_i in the output list, and $C[1, n-1]$ is the set of components retrieved between the ranks $[1, n-1]$.

Since the inex-2003 metric treats the two relevance dimensions separately, the quantisation functions were also redefined to provide a separate mapping for exhaustivity, $\mathbf{f'}_{quant}(e)\colon E \rightarrow [0,1]$ and specificity, $\mathbf{f'}_{quant}(s)\colon S \rightarrow [0,1]$, where $E = \{0,1,2,3\}$ and $S = \{0,1,2,3\}$. For the strict case, the result of the quantisation was 1 if $e = 3$ or $s = 3$, respectively, and 0 otherwise. For the generalised case, the quantisation function was defined as $\mathbf{f'}_{gen}(e) = e/3$ and $\mathbf{f'}_{gen}(s) = s/3$.

4 DISCUSSION OF CURRENT INEX METRICS

4.1 Overlapping result elements

A criticism of the inex-2002 metric was that it did not address the problem of overlapping result elements and hence produced better effectiveness results for systems that returned multiple nested components. Evidence to show this effect was given by Benjamin Piwowarski. Figure 1 shows the recall-precision graphs he obtained for different simulated runs, each representing possible retrieval approaches. The graph clearly illustrates that better effectiveness is achieved by systems that return not only the most desired components, but also their parent or ascendant elements. It was agreed that such a system behaviour should not be rewarded, but in fact should be penalised.

A number of suggestions were made as to how the problem of overlapping result elements should be addressed. One recommendation was to remove overlapping results from the submissions prior to the evaluation. This was later rejected as it was thought that such a method would be too lenient while it would also lack the ability to distinguish between systems that, correctly, do not return multiple nested components from those that do. This approach would also provide false effectiveness results given that it changes the actual result lists. An alternative solution is to penalise the retrieval of overlapping result elements. Here the question of how such a penalty-scheme should work was brought up. One suggestion was to only score the first result element that matches a given relevant reference component and regard any additional results that overlap with the same reference element as irrelevant. Two concerns were voiced regarding this proposal. One is that such a method may affect the recall base (i.e. leading to varying recall base), and, second, that it may also prove to be too unstable (i.e. too sensitive to retrieval order). For example, given a section element, $s1$, assessed as $(3,3)$, its article ascendant element, $a1$, assessed as $(3,1)$, and two rankings $r1 = [a1, s1]$ and $r2 = [s1, a1]$, we obtain the following precision values (using the generalised recall and precision calculations of [8] and the generalised quantisation function of Equation 3):

$$P_{r1} = (0.75 + 0)/2 = 0.375$$

$$P_{r2} = (1 + 0)/2 = 0.5$$

It was highlighted that the inex-2003 metric, which already implements a strategy to penalise overlapping results, may be more stable than the above method. This is because contrary to the above method, which only scores the first hit from a number of overlapping results, the inex-2003 metric provides a scoring mechanism that gives partial score to overlapping results, where the score is proportional to the not-yet-seen portion of the component. For example, for the above two rankings, we obtain the following precision values (using Equation 5):

$$P_{r1} = \frac{0.3 \cdot len(a1) + 0 \cdot len(s1 - a1)}{len(a1) + len(s1 - a1)} = 0.3$$

$$P_{r2} = \frac{1 \cdot len(s1) + 0.3 \cdot len(a1 - s1)}{len(s1) + len(a1 - s1)}$$
$$= 1 \cdot 0.1 + 0.3 \cdot 0.9 = 0.37$$

Note that the above calculations assume that the section forms $1/10$-th of the length of the article.

However, a criticism of the inex-2003 metric was that it had separated the two dimensions of relevance while according to the definitions both are required in order to identify the most appropriate units of retrieval. Members of the working group expressed concern regarding the exact meaning of such a measure of recall or precision, which are solely based on the exhaustivity or specificity dimension, respectively. It was agreed that further investigation of this issue would be beneficial.

In summary, preference was given to the inex-2002 metric, although it was agreed that suitable mechanisms should be developed to address the overlap of result elements. The main concerns regarding the inex-2003 metric concerned its separation of the two relevance dimensions and its stability (or sensitivity to small changes in the ranking).

4.2 Quantisation functions

Members of the working group expressed a clear preference towards the use of the strict quantisation functions since the problem of overlapping results presents

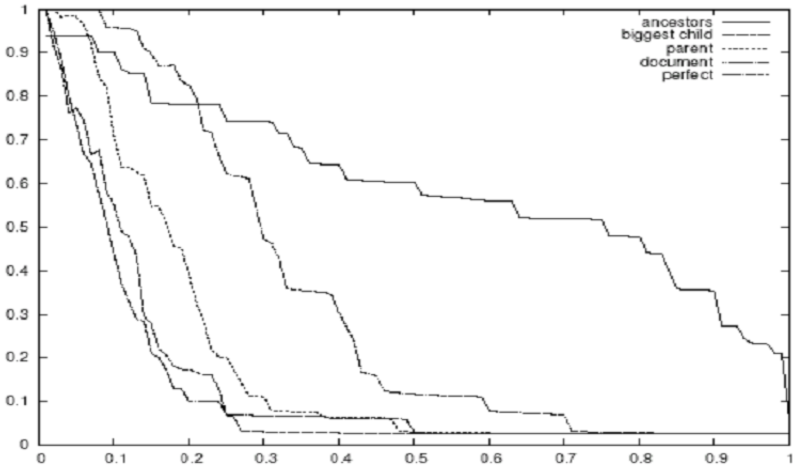

Figure 1: Generalised precision-recall for simulated runs

less of an issue in this case. It was also seen to provide more comprehensible results compared with the generalised quantisation functions. Some members have in fact suggested to base the evaluation solely on the strict assessment criteria.

This suggestion has lead to a discussion questioning the validity of the methodology employed for constructing the test collection. The argument was that if the evaluation only makes use of the components assessed as $(3, 3)$ then there should be no reason to justify the currently required effort in collecting such extensive assessments.

The main counter-argument against this proposal was that the definition of the ad-hoc XML retrieval task states that systems should find *all* relevant information, i.e. not just highly relevant information (but should rank highly relevant components first). Therefore, evaluation based on $(3, 3)$ elements only does not provide suitable evaluation criterion in INEX. It was pointed out that systems that do well on retrieving $(3, 3)$ components may not be appropriate for recall-oriented retrieval tasks (this was also the finding of [8]). In addition, it was emphasised that relevant elements assessed other than $(3, 3)$ are not simply a means for the evaluation of near misses, but these components contain relevant information to varying degree, which may be of interest to the user. At this point, Birger Larsen was also invited into the discussion. He further detailed the benefits of graded relevance assessments (see [8, 5, 11]), adding that "Future metrics can make use of the rich data even if we do not yet know how".

Additional arguments against the use of only $(3, 3)$ assessments included points that the recall-base may be too small for reliable evaluation, that assessors would label more elements as $(3, 3)$ due to the lack of alternative relevance degrees, and that no automatic mechanisms could be used to reliably infer the relevance degree of ascendant or descendant relevant components (unless binary relevance is adopted).

As a result of the discussion, it was agreed that it is necessary to consider all levels of relevant components within the evaluation. It was also agreed that due to the overlap problem this criterion is currently not evaluated sufficiently in INEX (which is also believed to be the primary reason why so much emphasis has been attributed so far to the results of the strict evaluation measures).

This has then lead to the agreement that the generalised quantisation functions must also be employed within the evaluation. As mentioned earlier, the aim of the generalised quantisation is to allow the scoring of result elements proportional to their degree of relevance. This viewpoint makes the generalised functions more suitable for the evaluation of content-oriented XML retrieval systems as it closer reflects the evaluation criterion compared with the strict quantisation functions. However, the problem of overlapping result components, which remains so far largely unsolved, does present an issue regarding the output of such an evaluation.

Aiming towards an intermediate solution to the problem, a number of new quantisation functions were defined to be used with the inex-2002 metric. The originating idea here was to find a solution, which like the strict quantisation functions minimises the overlap problem, while at the same time, like the generalised quantisation functions better reflects the evaluation criterion (i.e. finding all relevant elements). Two classes of quantisation functions were defined: specificity-oriented and exhaustivity-oriented functions. The specificity-oriented functions apply strict quantisation with respect to the specificity dimension only, while allow to consider different degrees of exhaustivity. They aim to evaluate systems according to their ability to retrieve the most specific relevant components, where the exhaustivity of the component may vary from marginally and fairly exhaustive to highly exhaustive (Equation 7) or only from fairly to highly exhaustive (Equation 8).

$$\mathbf{f}_{s3\_e321}(e,s) := \begin{cases} 1 & \text{if } e \in \{3,2,1\} \text{ and } s = 3, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

$$\mathbf{f}_{s3\_e32}(e,s) := \begin{cases} 1 & \text{if } e \in \{3,2\} \text{ and } s = 3, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Similarly to the specificity-oriented functions, exhaustivity-oriented quantisation functions were also defined (Equations 9 and 10). Note, however, that these exhaustivity-oriented functions suffer from the same overlap problem as the generalised quantisation functions.

$$\mathbf{f}_{e3\_s321}(e,s) := \begin{cases} 1 & \text{if } e = 3 \text{ and } s \in \{3,2,1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

$$\mathbf{f}_{e3\_s32}(e,s) := \begin{cases} 1 & \text{if } e = 3 \quad \text{and } s \in \{3,2\}, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

In summary, it was agreed that the strict quantisation functions, although are less effected by the overlap problem, are not sufficient alone for the evaluation of XML retrieval. They are useful and necessary, but they reflect a rather strict evaluation criterion according to which only highly exhaustive and highly specific elements are considered relevant for the user. On the other hand, although the generalised quantisations allow a more detailed evaluation, they suffer from the problem of overlapping result elements. As an intermediate solution new versions of the strict quantisation functions were proposed.

5 PROPOSED METRICS

Two proposals were presented in detail: the Expected Ratio of Relevant documents (ERR) and the Tolerance to Irrelevance (T_2I) metrics. An additional two proposals, both based on extensions of the Cumulated Gain based metrics [6], were only mentioned during the workshop. This section provides a brief summary of all these proposals.

5.1 Expected Ratio of Relevant

The Expected Ratio of Relevant documents (ERR) was proposed by Benjamin Piwowarski and Patrick Gallinary. This measure provides an estimate of the expectation of the number of relevant document elements (doxels) a user sees when consulting the list of the first k returned doxels, divided by the expectation of the number of relevant doxels a user would see when exploring all the doxels of the database (i.e. the total number of relevant elements for a given topic, denoted by E). The value of ERR for each k between 1 and the total number of retrieved doxels N is given as:

$$\mathrm{ERR} = \frac{\mathbb{E}[N_R|N = k]}{\mathbb{E}[N_R|N = E]} \tag{11}$$

where $N_R|N = k$ represents the total number of relevant doxels the user has access to within the first k elements in the result list, and $N_R|N = E$ represents the total number of relevant doxels within the whole collection.

The actual calculation of this estimate is based on a hypothetical user behaviour, which extends the assumptions used in classical IR, e.g. users browse elements in the list in a linear order, etc., with two additional hypotheses. The first is that the user is assumed to browse through the retrieved document's structure (that is, he/she can ”jump” with a given probability from one element to another within the same document). It is however assumed that users cannot use hyperlinks (i.e. jump to another document). The second hypothesis is that this browsing is influenced by the specificity of the doxels. Based on these assumptions, the parameters within the model are estimated leading to a final estimate of the ERR value.

Further details on this metric are available in [9].

5.2 Tolerance to Irrelevance

Arjen de Vries, Gabriella Kazai and Mounia Lalmas proposed a measure, which is based on an alternative definition of correct results. The main idea is that a user merely needs an entry-point into the document that is ‘close’ to relevant information. Taking this view, a retrieval system produces a ranked list of entry points. The user starts reading the retrieved article from the suggested entry point, giving up when no relevant information is found for some number of words or sentences. So, the user processes the retrieved information until his or her *tolerance to irrelevance* (T_2I) has been reached, at which point the user proceeds to the next system result.

This discourages systems from returning fragments that are too large, since if the entry-point is too far away from the relevant reference component, the user's tolerance to irrelevance will have been exhausted before the relevant information has been reached. The problem with multiple system results intersecting the same reference component is eliminated by extending the definition of irrelevance, according to which a previously seen reference fragment is no longer considered relevant.

T_2I variants of three existing evaluation metrics for system performance are given in [2]. Their common underlying principle is that retrieval systems are ranked on their ability to maximise the number of relevant fragments shown to the user while minimising the amount of user effort wasted on irrelevant information. The tolerance to irrelevance is expressed by a single parameter, τ_{NR}, that represents the maximum amount of non-relevant text the user is expected to read before giving up. The length of retrieved relevant components is ignored, assuming that each result has equal value to the user.

5.3 Cumulated Gain for XML

Two separate proposals were made for the extension of the Cumulated Gain (CG) based evaluation measures [6] for the evaluation of XML retrieval. One by Huyen-Trang Vu and another by Gabriella Kazai.

Huyen-Trang Vu is currently working on a variation of the discounted cumulated gain (DCG) measure, where the discount function employed makes use of a component-length normalisation function. This function is similar to the length normalisation of the inex-2003 metric and takes into account the size of the not-yet-seen part of the retrieved component, where uniform distribution of relevant information within a component is assumed. She is also working on an experimental analysis of the INEX evaluation results with the aim to reach some consensus about evaluation issues raised in INEX such as the overlap problem and the usage of graded assessments. A paper describing the approach is currently in preparation.

The approach taken by Gabriella Kazai is to extend the (D)CG based metrics by separating the model of user behaviour from the actual metric employed. This is achieved via the definition of a set of relevance value (RV) functions implementing scoring mechanisms based on parameters including the relevance degree of a retrieved component, the ratio of already viewed parts, etc. Each such RV function models different possible user behaviours. Within the (D)CG framework, an RV function is then used as a means to calculate the relevance score of a document component within the result list, hence, producing the gain vector G, which forms the basis of the (D)CG calculations. She also proposed different functions for the estimation of a component-part's relevance degree, which moves away from the uniform distribution assumption and is based on the assessment data of the component's child nodes. A paper describing the approach has since been submitted for publication [7].

6 DISCUSSION OF PROPOSED METRICS

All proposals were welcomed by the group. ERR was regarded as an encouraging measure although concerns were raised regarding the use of possibly too many parameters that needed to be estimated. T_2I was assessed as a promising, simple but potentially powerful framework, which however so far lacked implementation details. Both metrics were said to benefit from experiments and analysis of their working.

The CG based metrics were not discussed.

7 OTHER ISSUES

Additional issues raised during the workshop included general problems, such as problems experienced when trying to install the INEX evaluation software. Another criticism was the lack of documentation provided.

The point that systems could not be tuned due to fact that the metrics were not published prior to the task execution was also raised. A related issue concerned the understanding of the metrics and of their workings. A general recommendation was to publish metrics early on within the evaluation round. Another suggestion was to provide effectiveness results for P@5, P@10, P@20 as part of the official evaluation.

Concerns regarding the consistency of assessments due to the increased cognitive load were also expressed. The organisers offered to investigate this issue by providing an analysis of the collected assessments of topics from multiple assessors.

Other issues raised included concerns that article only retrieval was hard to beat. This has lead to questions regarding the quality of the topics used within the test collection and the problem of how to ensure that answer elements were components smaller than

Task	Metric
CO	inex-2002
	inex-2003
	ERR
	T_2I
SCAS	inex-2002
	ERR
	T_2I
VCAS	Extensions of the CO metrics to provide partial score based on structural similarity using distance measures.

Table 1: Tasks and metrics

article elements while maintaining realistic information needs. While no solution was identified, the issue was raised as a concern that should be considered during the topic development process.

The working group ended with a discussion on which metrics can be used for the evaluation of which tasks (i.e. CO, CAS and SCAS). This is summarised in Table 1.

References

[1] W. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1):30–41, 1968.

[2] A. de Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *Recherche d'Informations Assistée par Ordinateur (RIAO 2004)*, Avignon, France, Apr. 2004. To appear.

[3] N. Gövert and G. Kazai. Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2002. In N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors, *Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX). Dagstuhl, Germany, December 8–11, 2002*, ERCIM Workshop Proceedings, pages 1–17, Sophia Antipolis, France, March 2003. ERCIM. http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf.

[4] N. Gövert, G. Kazai, N. Fuhr, and M. Lalmas. Evaluating the effectiveness of content-oriented XML retrieval. Technischer bericht, University of Dortmund, Computer Science 6, 2003.

[5] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In N. Belkin, P. Ingwersen, and M.-K. Leong, editors, *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, Athens, Greece, 2000.

[6] K. Järvelin and J. Kekäläinen. Cumulated Gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (ACM TOIS)*, 20(4):422–446, 2002.

[7] G. Kazai, M. Lalmas, and A. de Vries. The overlap problem in content-oriented XML retrieval evaluation. Submitted for publication, Jan. 2004.

[8] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[9] B. Piwowarski and P. Gallinari. Expected ratio of relevant units: A measure for structured information retrieval. In N. Fuhr, M. Lalmas, and S. Malik, editors, *Proceedings of the Second Workshop of the INitiative for the Evaluation of XML Retrieval (INEX). Dagstuhl, Germany, December 15–17, 2003*, 2004.

[10] V. Raghavan, P. Bollmann, and G. Jung. A critical investigation of recall and precision. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.

[11] E. Sormunen. Liberal relevance criteria of trec - counting on negligible documents? In K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. Myaeng, editors, *Proceedings of the Twenty-Fifth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, 2002.

[12] S. Wong and Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, 1995.