Report on the ad-hoc track of the INEX 2005 workshop

Mounia Lalmas and Gabriella Kazai
Queen Mary, University of London
mounia,gabs@dcs.qmul.ac.uk

1 Introduction

The INitiative for the Evaluation of XML retrieval (INEX) has, since 2002, been working towards the goal of establishing an infrastructure, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems. In 2005, 47 organizations registered to participate in INEX. Throughout the year a number of groups dropped out due to resource requirements, while 11 further groups joined. INEX 2005 concluded with a total of 41 active groups. In addition to the main ad-hoc retrieval track, six further research tracks were included in INEX 2005, each studying different aspects of XML information access: interactive, relevance feedback, heterogeneous, natural language processing, and two new tracks for 2005, multimedia and document mining tracks.

INEX 2005 ended with a workshop, held in Shloss Dagstuhl, Germany, on 28-30 November 2005. A total of 50 participants from 25 organizations attended. The workshop was organised into 9 paper sessions, 3 discussion sessions, and one reporting session. In total, 32 papers were presented. This report describes the work carried out as part of the ad-hoc retrieval track which were presented at the workshop (Section 3). All papers summarised here can be found in [1]. The test collection, which is a result of the collaborative effort of the INEX participants, is also described, together with the employed evaluation measures (Section 2). Further details on these can be found in [4] and [3], respectively. Finally, the outcome of the working group discussions at the workshop are reported in Section 4.

2 The INEX 2005 test collection

Test collections, as traditionally used in IR, consist of three parts: a set of documents, a set of information needs called topics and a set of relevance assessments listing the relevant documents for each topic. Although a test collection for XML IR consists of the same three parts, each component is rather different from its traditional IR counterpart. These three components of the INEX test collection are described next.

2.1 Documents

This year the collection of documents that forms the INEX ad-hoc test collection has been extended with further publications donated by the IEEE Computer Society. A total of 4 712 new articles from the period of 2002-2004 have been added to the previous collection of 12 107 articles, giving a total of 16 819 articles. This extension raised the number of retrievable XML elements above 10 million and the size of the document collection to a total of 764Mb.

2.2 Topics

As in previous years, INEX 2005 distinguished two types of topics: Content-Only (CO) topics and Content-And-Structure (CAS) topics. These topic types reflect two types of users with varying levels of knowledge about the structure of the searched collection. The first type simulates ignorant users who do not have any

knowledge of the document structure or who choose not to use such knowledge. The latter type of user aims to make use of any insight about the document structure that they may possess. Building on these basic types, INEX 2005 defined and investigated various extensions and interpretations of topic types.

In an effort to investigate the usefulness of structural constraints treated as hints, the CO topics were extended to Content-Only+Structure (CO+S) topics. The aim was to enable the comparison of system performance across two retrieval scenarios (on the same topic): when structural hints are taken into account and when these hints are ignored.

The aim of the CAS topics was to support investigations on the effect of different interpretations of structural query constraint on retrieval effectiveness. CAS topics have been defined as topic statements that contain explicit references to the XML structure, and explicitly specify the contexts of the user's interest (e.g. target elements) and/or the contexts of certain search concepts (e.g. containment conditions). Each structural constraint in a CAS query can be considered as a strict (must be matched exactly) or vague (simply as hints) criterion. Four combinations of vague and strict interpretations of the structural constraints are then possible, depending on how the target elements and/or the containment conditions are treated:

- VVCAS: where the structural constraints in both the target elements and the support elements are interpreted as vague.
- SVCAS: where the structural constraints in the target elements are interpreted as strict and the structural constraints in the support elements are interpreted as vague.
- VSCAS: where the structural constraints in the target elements are interpreted as vague and the structural constraints in the support elements are interpreted as strict.
- SSCAS: where the structural constraints in both the target elements and the support elements are interpreted as strict.

Both CO+S and CAS topics are made up of several parts, each representing the same information need, but for different purposes. These include a topic title and a CAS title, which are short summaries of the information need and represent purely content and combined content and structure conditions, respectively. CAS titles are expressed in the query language of NEXI [7]. An example of a CAS title is: `//article[about(.,interconnected networks)]//p[about(., Crossbar networks)]`. Further topic parts include a topic description, which is a one or two sentence natural language definition of the information need, and a narrative, which is a detailed explanation of the information need and a description of what makes a document/component relevant or not.

In 2005, a total of 87 topics (40 CO+S and 47 CAS) were added to the INEX test collection.

2.3 Retrieval tasks

Within the ad-hoc retrieval track, INEX 2005 distinguished several retrieval strategies, each based on different assumptions regarding a search system's output. These strategies were investigated using the CO+S topics based on both retrieval scenarios, i.e. when structural hints were taken into account and when these hints were ignored.

- Focussed: This strategy was intended for approaches concerned with the focussed retrieval of XML elements, i.e. aiming at targeting the appropriate level of granularity of relevant content that should be returned to the user for a given topic. The aim was for systems to find the most relevant element on a path within a given document containing relevant information and return to the user only these most appropriate units of retrieval.

- Thorough: This strategy was intended for XML retrieval approaches that do not deal with the dependence between retrieved elements (returning a section and its paragraph) when generating their output list for the evaluation, but consider this a presentation issue at the user interface level. The aim here was for systems to find and rank all relevant elements within the collection.
- Fetch & Browse: This strategy was intended for XML retrieval approaches that are based on a mixture of document retrieval and element retrieval strategies. The aim of the fetch and browse retrieval strategy was to first identify and rank relevant articles (the fetching phase), and then to identify and rank the relevant elements within the fetched articles (the browsing phase).

As described in section 2.2, different interpretations of CAS topics on the basis of target elements and containment conditions resulted in the four sub-tasks of VVCAS, SVCAS, VSCAS and SSCAS. In these sub-tasks, the aim was to retrieve all relevant elements with respect to the topic of request, following the Thorough strategy. An analysis of the outcome of the CAS sub-tasks can be found in [6].

2.4 Relevance dimensions and scales

Relevance assessments were given according to two relevance dimensions: exhaustivity (e), which describes the extent to which the document component discusses the topic of request; and specificity (s), which describes the extent to which the document component focuses on the topic of request. While the above definition of the relevance dimensions has remained unchanged since 2003, the scale that these dimensions were measured on has been revised in 2005. The scale for exhaustivity was changed to 3 + 1 levels: highly exhaustive ($e = 2$), somewhat exhaustive ($e = 1$), not exhaustive ($e = 0$) and "too small" ($e =?$). The latter category of "too small" was introduced to allow assessors to label document components, which although contained relevant information were too small to sensibly reason about their level of exhaustivity. Specificity, for the first time, was measured on a continuous scale with values in [0, 1], where s = 1 represents a fully specific component (i.e. contains only relevant information).

INEX 2005 also introduced a new assessment procedure based on a yellow-marker design, where assessors were asked to first highlight text fragments that contained only relevant information and then to judge the exhaustivity level of any XML elements that had highlighted parts. This allowed specificity to be measured automatically based on what ratio of a document component has been highlighted by the assessor.

The relevance degree of an assessed component was then given by the combined values of exhaustivity and specificity, denoted as (e, s), where e $\in \{?, 0, 1, 2\}$ and s $\in$ [0, 1]. For example, (2, 0.72) denotes a highly exhaustive component, 72% of which is relevant content.

2.5 Evaluation measures

In INEX 2005, a new set of measures, the eXtended Cumulated Gain (XCG) measures [3], were introduced with the aim to provide an evaluation framework, where the dependency among XML document components can be taken into account. In particular, two aspects of dependency were considered: 1.) near-misses, which are document components that are structurally related to relevant components, such as a neighbouring paragraph or a container section, and 2.) overlap, which regards the situation when the same text fragment is referenced multiple times, as in the case when a paragraph and its container section are both retrieved.

The XCG measures are an extension of the Cumulated Gain based measures proposed in [2]. These measures were chosen as they have been developed specifically for graded relevance values and with the aim to allow IR systems to be credited according to the retrieved documents' degree of relevance. The motivation for the XCG measures was to extend the CG metrics for the problem of content-oriented XML IR evaluation, where the dependency of XML elements is taken into account. The extension lies partly in the way the gain value for a given document component is calculated via the definition of so-called relevance value (RV) functions, and partly in the definition of the ideal recall-bases. The former allows to consider the

dependency of result elements within a system's output, while the latter regards the dependency of elements within the test collection's recall-base[1].

The XCG measures include the user-oriented measures of normalised extended cumulated gain ($nxCG$) and the system-oriented effort-precision/gain-recall measures (ep/gr). The former is a measure of system performance at a single cutoff value, reflecting the relative gain the user accumulated up to the cutoff rank position, compared to the gain he/she could have attained if the system would have produced the optimum best ranking. The meaning of effort-precision at a given gain-recall value is the amount of relative effort (where effort is measured in terms of number of visited ranks) that the user is required to spend when scanning a system's output compared to the effort an ideal ranking would take in order to reach a given level of gain relative to the total gain that can be obtained. As with the standard IR measures of precision/recall, the non-interpolated mean average effort-precision, denoted as $MAep$, is calculated by averaging the effort-precision values measured at natural recall-points, i.e. whenever a relevant XML element is found in the ranking, and assuming the score of 0 for any non-retrieved relevant elements. Analogue to recall/precision graphs, effort-precision is plotted against gain-recall to obtain a detailed summary of a system's overall performance.

3 Papers on ad-hoc retrieval presented at the workshops

This section provides a brief summary of the approaches developed for the ad-hoc retrieval track (and its various sub-tasks and strategies), which were presented at the workshop. A total of 16 such papers were presented in 4 sessions.

Ad-hoc retrieval session I. Four papers were presented in the first session, chaired by N. Fuhr.

The first paper, *Parameter Estimation for a Simple Hierarchical Generative Model for XML Retrieval* by P. Ogilvie and J. Callan, describes an extension to their previous INEX work in using hierarchical language models for ranking XML elements. The emphasis of the paper is on parameter estimation, which has not been tackled in their previous work. The authors present a parameter estimation method for a simplified version of the hierarchical language models, which is defined by interpolating several language models estimated, respectively, from the text of an element, its parent element, the document element, and its children elements. This simplification allows the application of the generalized expectation maximization algorithm to learn effectively the parameters.

The second paper, *The University of Kaiserslautern at INEX 2005* by P. Dopichaj, a new participant to INEX, proposes a retrieval approach based on the vector space model, enhanced with two XML-specific additions. The first, called element relationship, aims at better supporting queries with vague structural hints by using background knowledge on the document schema to allow matching of similar element types. The second, called context patterns, exploits structural patterns in the retrieval results to find the appropriate results among related elements. Context patterns are based on the observation that the structural properties of retrieval results, like length and position, can provide valuable hints about the importance of the retrieved elements.

The third paper, *Field-Weighted XML Retrieval Based on BM25* by W. Lu, S. Robertson and A. MacFarlane, also first time participants, extends Robertson's field-weighted BM25F approach for document retrieval to element retrieval. The BM25F, which is a probabilistic model, was originally proposed as a means to overcome problems arising from a linear combination of retrieval scores, and was successfully applied to web data. BM25F consists of the linear combination of term frequencies based on BM25 to extend standard ranking functions to multiple weighted fields (e.g. in the context of XML retrieval, different element tag types). The paper shows how to tune weights for selected tag types.

[1]The term recall-base refers to the collection of assessments within the test collection that forms the ground-truth for the evaluation experiments.

The final paper in this session, *Using the INEX Environment as a Test Bed for Various User Models for XML Retrieval* by Y. Mass and M. Mandelbrod, describes an adaptation of the component ranking algorithm presented at previous INEX workshops, which runs each query against different indices where each index contains elements of the same type, to cater for the different INEX 2005 retrieval scenarios. A particular emphasis of the paper is the development of a focussed retrieval algorithm that takes into account the tree structure of an XML document to decide which elements to remove among overlapping elements. The proposed approach is shown to outperform the simple removal of overlapping elements based on their rank comparison only.

Ad-hoc retrieval session II. The second session dedicated to the ad-hoc retrieval track contained four papers and was chaired by G. Kazai.

The first paper in this session was *TopX & XXL at INEX 2005* by M. Theobald, R. Schenkel and G. Weikum who use two different search engines to carry out the various ad-hoc retrieval tasks, namely the XXL search engine and the TopX engine, both following a database approach to XML retrieval. The paper focuses on TopX, which is applied at INEX for the first time. TopX makes use of a pre-computed index list, which is sorted in descending order of appropriately defined scores for individual tag-term pairs and implements a top-k query processor for XML data.

The next paper, *RMIT University at INEX 2005: Ad-hoc Track* by J. Pehcevski, J. A. Thom and S. M. M. Tahaghoghi, discusses the use of a hybrid XML retrieval approach, which combines information retrieval features from Zettair (a full-text search engine) with XML-specific retrieval features from eXist (a native XML database) to analyse the different XML retrieval scenarios of INEX 2005. Different behaviors are observed when looking at the different retrieval scenarios, suggesting that the optimal retrieval parameters are highly dependent on the nature of the XML retrieval task. For instance, results from the evaluation experiments presented in the paper suggest that using structural hints in CO topics leads to more precise search, but only for those retrieval strategies where overlap among retrieved elements is taken into account by the evaluation measure (such as Focussed and Fetch & Browse).

The paper *GPX - Gardens Point XML IR at INEX 2005* by S. Geva describes an approach that uses the same underlying system to perform all the INEX 2005 ad-hoc retrieval tasks. The proposed approach is based on the construction of a collection sub-tree that consists of all elements (nodes) containing one or more of the query terms. Leaf nodes are assigned a score using a $tf \cdot idf$ variant, and scores are propagated upwards in the document XML tree, so that all ancestor XML elements are ranked. Results demonstrated that the approach is versatile and produces consistently good performance across all tasks.

The last paper of this session was *SIRIUS: A Lightweight XML Indexing and Approximate Search System at INEX 2005* by E. Popovici, G. Ménier and P.-F. Marteau, who were new participants to INEX. The paper describes a lightweight indexing and search engine for XML documents called SIRIUS. A simple querying algebra, implemented using fast approximate searching mechanisms for structure and textual-based retrieval, is presented. To evaluate the benefits and drawbacks of the proposed lightweight approach, experiments were carried out using specific data structures dedicated to the indexing and retrieval of information elements embedded within heterogeneous XML data bases. The indexing structures were shown to be well suited to the characterization of various contextual searches, expressed either at a structural level or at an information content level.

Ad-hoc retrieval session III. Four papers were presented in this session, which was chaired by L. Denoyer. The session covered approaches taken for the ad-hoc retrieval track, and approaches taken for other tracks. Only the approaches for the ah hoc retrieval track are reported.

The first paper, *Query Evaluation with Structural Indices* by P. Arvola, J. Kekäläinen and M. Junkkari, describes the retrieval methods used with the TRIX system, which is based on structural indices that make use of the natural tree structure of XML documents. The concept of contextualisation is presented and

experimented with, which re-weights (re-ranks) elements based on related elements. Four contextualisation approaches are considered: combining the weight of an element to that of its root (article element), averaging the weights of the element and its parent, averaging the weights of an element and all its ancestors, and combining parent and root contextualisations.

The paper *TIJAH Scratches INEX 2005: Vague Element Selection, Image Search, Overlap, and Relevance Feedback* by V. Mihajlović, G. Ramírez, T. Westerveld, D. Hiemstra, H. E. Blok, and A. P. de Vries presents an extension of their prototype database system, TIJAH, developed for structured document retrieval. The three levels (conceptual, logical, and physical) of the TIJAH system are enhanced to support (among other tasks) the modeling of vague selection of XML elements for the four CAS sub-tasks. The paper also analyzes existing ways to implement the focussed retrieval strategy and presents a new approach based on a so-called utility function.

The paper *XFIRM at INEX 2005: Ad-hoc and Relevance Feedback Tracks* by K. Sauvagnat, L. Hlaoua and M. Boughanem describes the XFIRM system, which uses a weighted relevance propagation method to estimate the relevance of non-leaf elements based on the index derived for the leaf nodes. The propagation method takes into account the distance between a node and its descendant leaf nodes, and the number of such relevant nodes. Particular emphasis is given to small nodes; their contribution is increased in calculating the relevance of their ancestor nodes during the propagation process.

The last paper of this session, *B^3-SDR and Effective Use of Structural Hints* by R. van Zwol focuses on the use of structural hints as a means to increase retrieval performance. It defines two extensions of an effective model for CO queries based on the GPX (see session II above). The first rewards elements that partly fulfill the structural constraints of the information need, causing them to appear higher in the ranking. The second extension penalizes those elements that contain excessive elements in their path, i.e. by decreasing the relevance score of an element if its path contains additional element tags that are not specified in the information need.

Ad-hoc retrieval session IV. The last paper session was composed of four papers and was chaired by I. Frommholz.

The first paper *Probabilistic Retrieval, Component Fusion and Blind Feedback for XML Retrieval* by R. R. Larson describes a fusion approach for multiple probabilistic searches against different XML components using different probabilistic retrieval algorithms. A new approach to combining and weighting the elements using only logistic regression-based algorithms for retrieval is also presented. This logistic regression parameters for different components of the INEX document collection are estimated using relevance assessments from the INEX 2003 data set. The paper also describes a blind relevance feedback method within this fusion framework. All these approaches are implemented with the Cheshire II XML/SGML search engine.

The second paper *Machine Learning Ranking and INEX'05* by J.-N. Vittaut and P. Gallinari presents a machine learning based ranking model for XML element retrieval. The ranking algorithm combines features characterizing the elements to be ranked, which depend on the element itself, its parent element, and the document containing that element. The ranking algorithm learns to combine these different features in an optimal way according to a loss function using a set of examples composed of query and element pairs (based on relevance assessments). The proposed model is shown to improve the performance of a baseline IR system (OKAPI adapted to XML retrieval).

The third paper of this session, *The Effect of Structured Queries and Selective Indexing on XML Retrieval* by B. Sigurbjörnsson, J. Kamps and M. de Rijke, makes use of a language modeling approach for XML retrieval. The paper addresses three different research questions building on experiences obtained in previous INEX participations. First, it looks at the contribution of structural constraints in querying XML documents and shows that improvement can be achieved at early precision. Second, it experiments with several selective indexing strategies (e.g. indexing elements above a certain length or of a given type) as a means to increase efficiency by reducing the index size. Third, it looks at the automatic creation of structured queries as a

form of query expansion using blind feedback.

The final paper of the session was *Searching XML Documents - Preliminary Work* by M. Hassler and A. Bouchachia, who are new participants at INEX. The paper concentrates on the design of an architecture that reconsiders several aspects of traditional IR, applied on flat documents. The paper also describes an adaptation of the standard vector space model to implement the various INEX 2005 retrieval strategies.

4 Working groups report

As in previous INEX workshops, a number of working group sessions were organised dedicated to discussing the evaluation methodology adopted in INEX 2005. Three such sessions occurred at the workshop, the outcome of which are reported here. Two rapporteurs were assigned to each working group, and this section is based on their summary.

Ad-hoc retrieval tasks working report. In this session, the INEX 2005 ad-hoc retrieval tasks and strategies (CO, CO+S, and CAS) were discussed and suggestions for the INEX 2006 retrieval tasks and strategies were formulated. The rapporteurs for this session were G. Kazai and J. Kamps. The session started with a report on the Interactive Track experiences in 2005 (papers presented in Interactive track session are not reported here, but can be found in [1]). The main observation was that the grouping of retrieval results by articles, implemented within the INEX 2005 test system, was widely appreciated by test persons. This was in sharp contrast with the INEX 2004 test system, which displayed ranked-lists of elements, receiving criticism. This was followed by a lengthy discussion on the Fetch & Browse retrieval strategy, which also requires retrieved elements to be clustered by article. The main debate was on whether to return all relevant elements per article, or just to return the best elements per article. At the end, the consensus was to regard these as separate tasks, and to distinguish between finding best entry-point(s), and between finding all relevant elements per article.

Additional to the Fetch & Browse task, the Thorough retrieval strategy (i.e. finding all relevant elements in the collection), the Focused retrieval strategy (i.e. returning the most relevant non-overlapping elements as a ranked-list), and the CO+ sub-task (formulating a more precise query by including structural hints) were discussed. It was suggested to collect various meta-data on the candidate topics and their authors (especially on the nature of the information need, and on the expected types of results). This would allow for making meaningful sub-divisions of the whole topic set, and may help to explain some of the diverging results presented at the workshops (e.g. using structural hints could either increase or decrease effectiveness).

Relevance assessments and online assessment interface working group report. This session was chaired by B. Piwowarski and S. Geva. The first concern was related to the exhaustivity dimension. In 2005, the exhaustivity scale had only 2 non-null (some relevance) values and a special "too small" value. It was argued that going binary would be the only way to speed up the assessment process[2]. The "too small" value, which was meant to be used when the assessor could not possibly judge an XML fragment which was a part of a relevant answer, was found to be misused by assessors to state that they wanted the tool to do the rest of the assessments automatically. This was identified as a problem regarding the consistency of the relevance assessments within INEX.

A solution for this issue (unless the exhaustivity scale is binary), would be to ask assessors to highlight "atomic" units of retrieval, defined as "passages which cannot be broken down to smaller units without becoming useless". In INEX 2005, assessors were asked to highlight passages containing only relevant material and then to give the exhaustivity value of all the XML elements the passage intersected with.

[2]In INEX the relevance assessments are performed by the participants. A guideline explaining the relevance dimensions and how and what to assess is distributed to the participants. This guide also contains the manual to the online assessment tool, referred to as X-RAI, developed by B. Piwowarski.

With the "atomic" assessment style, assessors would be asked to recursively break these passages into smaller parts until they are atomic. The drawback of this approach is that a highlighted passage means two different things: it can be an atomic or non-atomic unit. Another approach, bottom-up and more elegant, would be to directly highlight atomic units. This would also allow a one-step process by using different colors for different exhaustivity levels. No consensus were reached about the best way to proceed, although some statistical analysis is currently being carried out by P. Ogilvie to determine, among others, whether it would be sufficient to use a simple binary scale for the exhaustivity dimension for evaluating XML retrieval performance.

Metrics working group report. This working group, which was chaired by A. de Vries and D. Hiemstra discussed the a Fetch & Browse task description and metric proposal, and made some recommendations. The suggested Fetch & Browse user scenario is that a user would want to submit a query to a search engine. The search engine would print the top N documents, which the user will take on a train trip to read. The system highlights the parts of the document that are of interest. So, a good system would order the best articles on top, and highlight only the best elements. The setup could be considered similar to that of summarisation.

This suggests that the evaluation should consider only the top N articles, and measure the gain of highlights inside the article as some combination of "precision" and "recall". To address a concern on results with only one highly relevant highlighted sentence, and take into account the role of the answer context, people agreed that graded assessments were desirable, but, it would suffice to have graded assessments at the document level. The discussion did not conclude whether these grades should be based on the document's exhaustivity, or just be defined as 'highly relevant'.

A measure that suits these considerations could be based on the macro- or micro-averaged document-level F-measures. This would be equivalent to cumulative gain $CG[N]$ with gain defined as F-measure, or, using the scoring proposed in HiXEval [5] for each document independently. To handle overlap, an alternative would be to consider defining gain as tolerance to irrelevance (T2I) [8].

The main recommendations of the working group were to investigate if graded element assessments are (1) really worth the effort, and (2) outweigh the risk of assessor disagreement/imprecision. It is also suggested to choose one ad-hoc task only, modelled after Fetch & Browse retrieval strategy; and, if possible, to define a single evaluation metric that is known at the data release time. As for a CAS task, it was suggested to consider a known item task, which would result in cheap evaluation, while still allowing database-oriented researchers to enter INEX.

Acknowledgements

We would like to thank all workshop participants for the lively discussions. We would also thank the sessions chairs and the rapporteurs. We hope that everybody benefited as much as we did from attending the workshop and that we will meet again to discuss work related to XML retrieval and evaluation.

INEX is an activity of the DELOS Network of Excellence on Digital Libraries. We gratefully thank the organisers of the various tasks and tracks who did a superb job, their work is greatly appreciated.

References

[1] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai. (eds) Advances in XML Information Retrieval and Evaluation: Proceedings of the Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Schloss Dagstuhl, 28-30 November 2005, *Lecture Notes in Computer Science, Vol 3977, Springer-Verlag*, 2006.

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

[3] G. Kazai and M. Lalmas. INEX 2005 evaluation metrics. In [1].

[4] S. Malik, G. Kazai, M. Lalmas, and N. Fuhr. Overview of INEX 2005. In [1].

[5] J. Pehcevski and J.A. Thom. HiXEval: Highlighting XML Retrieval Evaluation. In [1].

[6] A. Trotman and M. Lalmas. The Interpretation of CAS. In [1].

[7] A. Trotman and B. Sigurbjörnsson. Narrowed extended XPATH I (NEXI). In *Advances in XML Information Retrieval, Third Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004), Dagstuhl, Germany, December 6-8, 2004, Revised Selected Papers, Lecture Notes in Computer Science, Vol 3493,Springer-Verlag*, page 16–40, 2005.

[8] A. de Vries, G. Kazai and M. Lalmas. Tolerance to Irrelevance: A User-effort Oriented Evaluation of Retrieval Systems without Predefined Retrieval Unit RIAO 2004 Conference on Coupling approaches, coupling media and coupling languages for information retrieval, University of Avignon (Vaucluse), France, April 2004.