Wanted: Element Retrieval Users

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

ABSTRACT

Document centric information retrieval is used every day by people all over the world. It is an application well studied, well understood, and of which there is a sound user model. Element retrieval, on the other hand, is a new field of research, with no identified applications, no users, and without a user model.

Some of the methodological issues in element retrieval are identified. The standard document collection (the INEX / IEEE collection) is shown to be unsuitable for element retrieval, and the question is raised – does such a suitable collection exist? Some characteristics of querying behavior are identified, and the question raised – will users ever use structural hints in their queries? Examining the judgments and metrics, it is shown that the judgments are inconsistent and the metrics do not measure the same things.

It is suggested that identifying an application of element retrieval could resolve some of these issues. Aspects of the application could (and should) be modeled, resulting is a more sound field of element retrieval. Alternatively, whatever it is, users don't want it, judges can't judge it, and the metrics can't measure it.

1. INTRODUCTION

INEX [3] was introduced in 2002 as a forum for the evaluation of element retrieval from XML documents. Since then there has been considerable discussion on relevance ranking algorithms, exactly as expected. Unexpectedly, there has also been considerable discussion on element retrieval methodology.

Element retrieval differs enormously from traditional document retrieval. The chunk of retrieval is a document element, not a document. Since elements might overlap within a document, this raises the issue of identifying the "best" element to return (from a given path). In effect the search engine can increase the exhaustivity (E) of a result by returning elements close to the root of a document tree, or can increase the specificity (S) of a result by returning elements close to the leaves. The search engine must balance these two to identify the most appropriate elements to return to the user. At INEX E and S scores are marked on a 4 point scale (0 = not, 1 = marginally, 2 = fairly, 3 = highly), and written EnSm.

As structure exists within the documents, users are able to use that structure in their queries. INEX identifies two types of queries, those that contain structure (Content and Structure (CAS) queries) and those that do not (Content Only (CO) queries).

Already, the difference between document retrieval and element retrieval is apparent. With element retrieval the user might use structural hints in their query, the search engine must interpret these, identify the most appropriate elements to return, and return a ranked list of elements (out of context) to the user.

As element retrieval is radically different from document retrieval, it has proven difficult to transfer prior experience to this new area. So it is time to start over – to address element retrieval as a new field, and to address the issues in the context in which they lie.

The first vital move is to identify an application of element retrieval.

Arguments are given here that show the current "model" is unsound: the document collection is inappropriate [18], the method of query is inappropriate, and the metrics are inappropriate.

Once users of an element retrieval system are identified, a sound model can be built and methodological issues can be resolved with reference to the model.

2. DOCUMENT COLLECTION

Intuitively element retrieval is important. Given a large collection of large documents marked up in XML [1] (or any other element based markup language such as SGML [7]) it's obvious that document components are a better result than whole documents. After all, the documents are large and the information relevant to the need is likely to be only part of the document.

In the case of a collection of books, each book might be a separate document. A document centric query to the collection would result in a ranked set of books. The user is then forced to wade through each book to find the relevant information (only a few pages might prove useful). Element retrieval might be used to identify only those relevant pages – surely presenting snippets of a book is more valuable to a user than presenting whole books.

Problematically, "a few pages" is not an element. An element retrieval system would return a book chapter, section or subsection, and not the said pages. Returning "a few pages" is passage retrieval [10]. Returning elements imposes a restriction on the passages: each passage must be a complete element (or at best a series of elements). Returning elements gives disregard to the most appropriate information to return and gives regard only to the most appropriate element to return.

Worse, the example of a collection of books assumes those books are divided into chapters, sections, and subsections. This is not the case for a many novels. Such work is a continuous flow of paragraphs from start to end. Markup would be used to identify the colophon, the title, and the author. The body might be a single element, or just a sequence of paragraphs. There are, after all, no other document components to mark up.

Surely element retrieval is useful for the IEEE document collection? This is a collection of 12,107 articles published by the IEEE computer society between 1995 and 2002. The documents were taken from 12 magazines and 6 transactions. They are primary and secondary academic scientific literature with some conference calls and news articles. This collection is used at INEX.

Science has organized itself to allow ideas, no matter of what size, to be published and cited. There are conference posters of only a couple of pages, conference articles of half a dozen pages, journal contributions (short and full), and books of varying size and structure.

Science has organized itself to allow these items to be cited. Articles are cited as a whole, books are either cited as a whole or with additional page references. Citers are expected to have read, in its entirety, the work being cited.

Page ranges in books has already been discussed – that isn't element retrieval!

A consequence of how articles are written and cited is that they are atomic. Each is the smallest indivisible unit of information that makes sense in its entirety. As such, element retrieval is a mismatch – there are no parts that make sense out of context. Even if there were, entire articles must be cited and as such read – reaffirming the atomicity of the articles.

There is but one part of an article that might make sense on its own – the front matter (author, title, and abstract). It is written to be read in isolation and to lie in isolation. There exist databases of millions of such "abstracts" (e.g. Medline). Using element retrieval to extract abstracts from documents is overkill.

For element retrieval to be useful it is necessary to identify the environment in which it might be used. Given there is a user who wants the said technology, it's possible to identify the characteristics of the document collection they are using.

The document collection must be in a markup language that contains elements. This might be XML, SGML or any other mark-up language.

The documents must contain several disparate parts (elements) that, while atomic in themselves, are also atomic in the context of the document.

O'Keefe [18] defines coupling as the association of an element to its context. He identifies elements that have low coupling as those suitable for element retrieval. Specifically, newspaper articles are given as an example of low coupling elements in a larger document (the newspaper). He also identifies extracting chapters from text books as a possible application of element retrieval.

If each newspaper story is held as a separate document then element retrieval is not necessary. Such is the case with the TREC [5] Wall Street Journal collection. Searching books is identified above as an inappropriate use of the technology.

The document elements must be large enough to contain relevant information, while at the same time small enough to be sub-documents. Kamps *et al*. [9] analyses the probability of an XML element being relevant given its length (in the IEEE collection) and identifies that although most elements are small (mean 29 terms), most relevant elements are not (mean over 1000 terms). In short, there is a mismatch between how elements are used and what information is relevant.

Element retrieval falls in the middle ground between question-answering and document retrieval. The units of retrieval are too large to be question answers, and too small to be documents. But the technology might be used in either.

Identifying elements that contain question answers might result in a reduced amount of natural language processing to obtain answers.

For document retrieval it might be used to identify where (within a document) the most relevant information lies for highlighting [22].

Perhaps element retrieval is a technology not at all appropriate for text. To retrieve elements is to pluck information from its context and present it as atomic. Such a technology should be applied where atomic information is strung together in an essentially random mix. Identifying such a place has proven hard.

Popular radio broadcast consists of segments of news, idle chatter and music interspersed with advertising. Stations give different news stories on the hour and half past the hour (to avoid repeating the stories every half hour). Perhaps element retrieval could be used to extract news stories on a given topic from the idle chatter and advertising. The same principle applies to "magazine" television broadcast such as MTV and E!.

Even within this context, it is not clear why element retrieval is necessary. Surely the atomic chunks can be separated from each other and stored as separate documents? If so, then element retrieval is not necessary. If not then the chunks are not atomic.

2.1 Discussion Point

A document collection for element retrieval must consist of documents that have a low coupling to their elements, while at the same time the coupling must be strong enough to bind the elements to the documents (or else why should the documents be maintained as such). The elements must be large enough to contain information that satisfies an information need yet elements tend to be very small chunks of text.

It is not obvious whether such a text collection exists. A suitable document collection might be found by changing the focus from text documents to audio or video.

3. QUERYING

Several pages from a book is identified above as "not an element". The IEEE collection is a collection of atomic elements that should not be broken into pieces. Element retrieval is a technology that is waiting for an application.

In the IEEE collection tags are used for two purposes: presentation and structure.

Presentation tags include those for italics, bold, and bold-italics (<it>, <b> and <bi>). Such tags are used, for example, to identify Latin words that are traditionally written in italics in English text. These tags are not the target of element retrieval.

Structure tags are used to mark paragraphs, subsections, sections and to separate the body from the front matter and back matter. It is these tags that are the target of element retrieval.

O'Keefe [18] identifies the majority of INEX topics in 2003 and 2004 targeting <article> or <sec> elements

Table 1: Target elements and number of times each is requested as a target element in an INEX 2003 / 2004 topic

Element	Occurrences	Proportion
sec	26	0.406
article	17	0.266
* & (p|fgc)	5	0.078
abs	4	0.063
p	4	0.063
bb	2	0.031
vt	2	0.031
bdy	1	0.016
bib	1	0.016
fig	1	0.016
fm	1	0.016
Total	64	1.000

Table 1 presents the list of target elements from 2003 and 2004. It can be seen that the majority of topics (67%) target <article> and <sec>. Only 11 unique tags were chosen from a possible 192 in the DTD. None of the tags were formatting tags.

Only 6% of the tags were identified as useful by topic authors. All those tags were structural. In these topics there is a 67% likelihood of the target being either <article> or <sec>.

One interpretation of this is that only <article> and <sec> tags are useful in this document collection. Searching for <article> elements is synonymous with whole document retrieval (a wanted but absent track at INEX). Searching for <sec> elements might be occurring because topic authors are required to submit structured topics, and there is no other useful element in the collection. The document collection only has two tags that might be used for retrieval, <article> and <sec>.

An alternative interpretation is that topic authors don't know (or don't want to know) the intricacies of the DTD. This is an argument often tabled for not using structural queries (the so-called CAS topics in INEX).

O'Keefe and Trotman [19] identify that not only use of structure is problematic; but also the syntax and semantics of structured queries. Of the 30 CAS topics at INEX 2003, topics written by IR researchers, 19 (63%) contained errors. This error rate appears to have dropped as a consequence of the introduction of the query language NEXI [26], but in practice it has not. Trotman and Sigurbjörnsson [27] identify that the online NEXI syntax parser was used 635 times for 84 CAS topics, or 7.5 times per query!

Experiments into the nature of interactive searching of elements were conducted at INEX 2004 [24]. On the comparative task the average number of search terms was 3.4, and on information foraging tasks (the background task), the average length of a query was 3.0 terms (including stop-words). In some participating groups the overall average was lower [15].

Tombros *et al*. [24] report that none of the queries contained use of the "+" or "-" operators. Quoted phrases appeared in less than 10% of queries. Only 50% of the log was analyzed due to software issues – so their result is partial.

Accepting that these queries could not contain structural hints; there still remains no evidence to suggest that structural hints would be used if they could be used. If mastery of "emphasis" and "negative emphasis" operators is beyond the scope of an interactive user, then the complexities of NEXI [26] certainly are.

Although there is no definitive evidence yet, it appears as though including structural hints in a query increases precision [20]. Experiments comparing structural with non-structural queries (of the same information need) are being conducted as part of INEX 2005 [23].

This leads to a conundrum. If structural hints increase precision, and users won't use them, then what to do? Use natural language queries [30]? Of 3.4 words on average? Of which one term is structural?

The enormous gulf between the searching behavior of database users and search engine users may account for the lack of use of the sophisticated search techniques. Again; identifying the target audience of element retrieval is essential. This will shed light on the patterns of query use by the audience. This, in turn, will help identify how queries should be formulated.

3.1 Discussion Point

The evidence predicts that users won't use structural hints in a query. It is the view of the author that early interactive experiments at INEX will show a slight contradiction to this. This contradiction is expected because structural hinting will appear novel in the mind of the user, who will experiment with it. Use during subsequent experiments is expected to match that of the "emphasis" operator.

If users won't use structural hints, then the INEX CAS task is futile. Concentration on identifying the target element (the CO task) is far more valuable than the CAS task.

If experiments conducted at INEX 2005 show no significant difference in the performance of CO and CAS runs then CAS should be dropped along with syntax for writing queries containing structural hints.

4. MEASUREMENT

With the introduction of the Cranfield methodology a line was drawn between the user and the search engine. On one side of the line lies the human / computer interaction (HCI) issues of social computer science, while on the other lies an experimental science of search engine design. In the early TREC experiments the line was mean average precision.

The documents, topics, and judgments are a model of interaction. The documents are a model of the type of information a user might find. A topic is a model of the type of query a user might have. The judgments are a model of how useful each document is to that user for that topic.

Mean average precision is a way of measuring the performance of a search engine with respect to the model. For as long as the performance metric is rigid, it is possible to change either the user model or the search engine. With mean average precision it is possible to determine quantitatively if one search engine is better than another given the user model. Search engine experiments are quantitatively reproducible and ranking functions are comparable.

The user model for element retrieval is not fixed – and the existing metrics do not measure the same thing.

Figure 1 shows the performance of each of the submitted runs at INEX 2004 (as well as those submitted to the LIP6 interface).

Shown is the performance of generalized inex_2002 (RP_g) [11] against generalized NG (RP_ng_o_g) [4]. From visual inspection the best runs scored using RP_g are not a good runs when scored using RP_ng_o_g. At INEX 2003, the University of Otago CO run ranked 1st using generalized ng-o while ranking 34th using generalized inex_2002. Whatever Otago did that year, it was good at NG but bad at inex_2002.

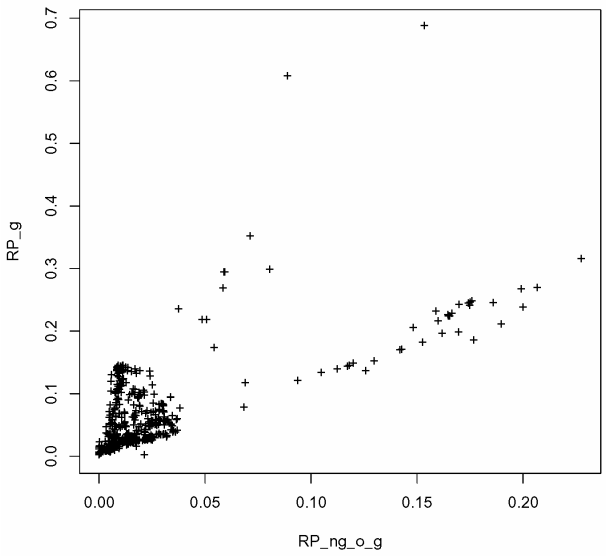

Figure 1: Performance of submitted runs at INEX 2004 (and the LIP6 web site) showing some systems perform well on NG (RP_ng_g) but badly on inex_2004 (RP_g). Image courtesy of Benjamin Piwowarski

Kazai *et al*.[13] recently introduced the metric XCG, an XML extension of the cumulative gain metric [8]. The top ten INEX 2004 runs scored with XCG share only 1 run with the top ten ranked using inex_2002 (and *vice versa*). These metrics measure different things.

Not only do the current metrics measure different things, but some are "hackable". For example, if a given element is known to be relevant then its parent must also be relevant, and its parent's parent, and so on up to the root of the document tree. Each of these elements might not be returning any additional relevant content, but returning such an ancestral list of elements is known to increase the inex_2002 score [13].

This process of "milking" the results to boost the score is commonplace. The two highest scoring CO runs at INEX 2004 show overlap of over 80%. Nine of the top ten show overlap of over 70%. That is, of the elements in the result list, over 70% of the elements have already been seen when scored.

Milking has been defended on two grounds.

If the search engine is trying to identify the most exhaustive and most specific elements (E3S3), then how can milking boost the performance?

Surely there can be no instance of a parent and child in the document tree both being E3S3. If the parent is the "ideal result" then how can a child also be the "ideal result"? In other words, it is simply not possible for both the root of the document tree and any descendant to be E3S3.

The "it can't happen" defense is unfounded. The description of topic 139 states "We wish to identify papers that cite work by authors Bertino or Jajodia that deal with "security models"". A whole document could be classified as E3S3 as could a single citation included in the same document. Of the 361 E3S3 judgments for this topic: 31 are whole documents, 304 are at or below a bibl (a reference).

Examining the judgments for 2004 (version 3[1]) 171 whole documents were judged E3S3, of which 163 have an element beneath the root that is also E3S3. A similar pattern can be seen in the 2003 judgments [20]. In other words, milking will also identify additional E3S3 elements as they, too, exist in paths through the document tree.

The second ground on which milking has been defended is simply that "everyone does it". This defense is untrue – 18 of the CO submissions at INEX 2004 contained no overlapping elements [12].

Milking violates the principle of user modeling. If there is no identifiable end user application of retrieval including milking, it is being done simply to score well on the metrics. In support, the interactive experiments have identified a user disapproval of the practice [15; 24].

The metrics are vitally important and must be treated with respect. Ranking function design is an optimization problem – the object is to optimize the metric score by changing the function. The recent use of Genetic Programming [25] as a method of finding the optimal function demonstrates this. The adoption of milking to boost the score is another. Whatever is necessary, however user-grounded or not, score boosting is rife.

Kekäläinen *et al*. [14] identify the influence of milking on runs at INEX 2004. By removing overlapping elements from their CO run the relative rank of their system dropped from 10th to 45th! It makes the difference from being in the top 10 to being in the top 50 (of 70).

A metric that is user grounded can't be hacked – if it could, it would no longer be modeling the behavior of a user. The current lack of user-grounded metrics at INEX makes the current evaluation round of questionable value. Participants are writing programs to score high on a metric for which there is no evidence of intrinsic value. In other words, they are writing programs that are getting very good at something that is known inherently to be very bad.

Identifying a user application of element retrieval will identify what users want – it is this that should be rewarded.

4.1 Discussion Point

The current collection of metrics is obscuring the problem of increasing performance. With high levels of overlap the current highest scoring systems are effective optimizations of the metric, however not likely to be viable information retrieval systems.

Before an effective metric can be devised it is necessary to stop looking at what might happen, what might be measured, and to

[1] Assessments for topic 127 are excluded as they were corrupt.

take a look at what a user might want. A metric should reflect user behavior and not *vice versa*.

5. JUDGMENTS

Before any metric can be used to compare the performance of two systems, it is necessary to have a stable set of judgments. During INEX 2004 twelve topics were chosen for judging by two judges each [17]. Lalmas *et al*. [16] report that the average level of exact agreement between judges was as low as 3.42% while the level of non-zero agreement was only 12.19%. Two judges non-zero agree if both consider the element to be other than E0S0.

Comparing these results to those of prior multiple judge experiments suggests INEX judge agreement is unacceptably low. Wilbur [29] reports that Saracevic reports that judges are known to agree somewhere between 40% and 70% of the time. At INEX the agreement level is (at best) 12% (according to Lalmas *et al*.).

Table 2: Document level relevance non-zero agreement between judges in the INEX 2004 judgments. P_j is the precision of judge, J_j, against the alternate judge

Topic	J_A	J_B	$\cup$	$\cap$	$\cap/\cup$	P_A	P_B
130	92	18	95	15	0.16	0.16	0.83
133	8	39	42	5	0.12	0.63	0.13
139	43	23	43	23	0.53	0.53	1.00
140	29	216	217	28	0.13	0.97	0.13
143	10	8	11	7	0.64	0.70	0.88
144	5	36	41	0	0.00	0.00	0.00
155	40	30	46	24	0.52	0.60	0.80
165	10	51	51	10	0.20	1.00	0.20
169	26	35	44	17	0.39	0.65	0.49
173	5	23	25	3	0.12	0.60	0.13
175	38	65	79	24	0.30	0.63	0.37
201	27	54	71	10	0.14	0.37	0.19
Total	333	598	765	166			
Mean					0.27	0.57	0.43

The Lalmas *et al*. comparison does not compare like with like.

Examining topic 130 (see Table 2) judge J_A identified 92 relevant elements whereas judge J_B identified 18 relevant elements. Of those 95 were unique, and of those only 15 were considered relevant by both judges.

Table 2 also presents the document level agreement (intersection divided by union) between the two judges. The level of agreement varies from no-agreement on topic 144, to agreement of 0.64 on topic 143. The mean document level agreement between the two judges is 0.27.

Table 3: Agreement levels at TREC and INEX

Evaluation	Agreement ($\cap/\cup$)
TREC-4 P/B	0.49
TREC-4 A/B	0.43
TREC-4 P/A	0.42
TREC-6	0.33
INEX-2004	0.27

Voorhees [28] examines the agreement levels in TREC-4 topics using three judges. Agreement levels of 0.42, 0.43, and 0.49 are seen between two judges and for all three 0.30 is seen. Voorhees reports these levels as high. In the INEX collection the overlap agreement for the 12 topics is 0.27. This is low by comparison to TREC.

Cormack *et al*. [2] report an experiment in which TREC-6 judgments from NIST were compared to those from judges at the University of Waterloo. In this experiment a mean overlap score of 0.33 is seen and reported (by Voorhees [28]). Again, huge variation is seen in the topics with agreement levels varying from none to total.

In Table 3 a comparison of the TREC-4, TREC-6 and INEX document based agreement levels is presented. From this it is clear the document-centered agreement levels at INEX are comparable to those at TREC. For details of agreement levels in pre-TREC collections see Harter [6].

The performance of a judge can be computed by taking the judgments for that judge and computing precision against the alternate set of judgments (presented in Table 2). Considering whole documents, the mean of precisions for the two judges is 0.57 and 0.43, comparable to that reported by Wilbur for Medline documents [29].

From this comparison it is reasonable to conclude that the performance of non-zero document centric judgments at INEX is consistent with those of TREC-4 and TREC-6. A larger study involving more than 12 topics is needed to confirm this observation. Experiments like those of Voorhees [28] are also needed to determine whether, or not, the disagreement between judgments affects the relative order of different systems.

Table 4: Element level non-zero agreement between judges for the 12 topics double judged at INEX 2004

Topic	J_A	J_B	$\cup$	$\cap$	$\cap/\cup$	P_A	P_B
130	1233	259	1328	164	0.12	0.13	0.63
133	37	451	474	14	0.03	0.38	0.03
139	562	889	1213	238	0.20	0.42	0.27
140	257	2418	2464	211	0.09	0.82	0.09
143	61	48	68	41	0.60	0.67	0.85
144	21	319	340	0	0.00	0.00	0.00
155	496	292	608	180	0.30	0.36	0.62
165	55	697	699	53	0.08	0.96	0.08
169	247	490	586	151	0.26	0.61	0.31
173	60	228	260	28	0.11	0.47	0.12
175	214	1468	1578	104	0.07	0.49	0.07
201	354	618	887	85	0.10	0.24	0.14
Total	3597	8177	10505	1269			
Mean					0.16	0.46	0.27

The INEX topic submission process demands that topic authors submit, along with the topic, a (small) list of elements considered relevant. This list can be compared to the topic assessments for consistency. Some of the with-topic judgments may not be in the judgment pool and *vice versa* so it is possible only to measure the extent to which a judge "changed their mind". That is, of the elements in both the with-topic list and the judgment list, how

many were judged non-relevant. This experiment was not conducted due to time constraints.

Examining the level of non-zero element agreement between judges, (in Table 4) judges do not agree on which elements are relevant. Comparing with the result in Table 2, it is reasonable to conclude that the judges do agree on which documents are relevant, but not on why!

The picture turns sour when E3S3 elements are examined (strict quantization). Table 5 lists the number of documents that are identified as containing E3S3 elements (that is, even if the document is not E3S3, there is an E3S3 element in the document). Judges do not agree on which documents contain the most specific and most exhaustive elements.

The level of E3S3 element agreement is shown in Table 6 where the agreement level is 0.05. There is almost total disagreement on which elements are most specific and most exhaustive.

Table 5: Documents judged to contain E3S3 elements by each judge of the multiple judged topics from the INEX 2004

Topic	J_A	J_B	$\cup$	$\cap$	$\cap/\cup$
130	1	10	10	1	0.10
133	0	0	0	0	0.00
139	38	20	39	19	0.49
140	0	9	9	0	0.00
143	0	0	0	0	0.00
144	0	7	7	0	0.00
155	4	7	9	2	0.22
165	10	3	11	2	0.18
169	3	7	8	2	0.25
173	0	4	4	0	0.00
175	2	3	4	1	0.25
201	0	4	4	0	0.00
Total	58	74	105	27	0.00
Mean					0.12

Table 6: E3S3 element level agreement for the 12 topics double judged at INEX 2004

Topic	J_A	J_B	$\cup$	$\cap$	$\cap/\cup$
130	2	42	42	2	0.05
133	0	0	0	0	0.00
139	361	169	451	79	0.18
140	0	32	32	0	0.00
143	0	0	0	0	0.00
144	0	10	10	0	0.00
155	5	22	24	3	0.13
165	29	15	38	6	0.16
169	10	21	30	1	0.03
173	0	26	26	0	0.00
175	18	5	22	1	0.05
201	0	4	4	0	0.00
Total	425	346	679	92	
Mean					0.05

This result suggests that although judges agree on which documents are relevant, they don't agree on why they are relevant or how relevant those documents are.

Experiments to determine if this disagreement affects the relative ranking of search engines is yet to be performed. At INEX 2004 two judgment sets were made available. Each contained judgments for 60 topics, however they only differed in the 12 topics discussed herein. That is, they were 75% identical because only 25% of topics were judged by more than one judge. Not surprisingly the relative performance of systems was relatively stable – no doubt because the judgment sets were essentially the same.

There could be many reasons why judges disagree on what constitutes a relevant element. Studies on whole document retrieval have identified a plethora of such factors [21]. With element retrieval there is at least one additional contributing factor – there is no agreed user model as there is no example application.

Identifying an application of element retrieval will help reduce disagreement levels. Judges will be aware of a common model and consequently will be able to refer to the model in case of uncertainty.

Of course, with an appropriate document collection and suitable queries such levels of disagreement may simply vanish. Disagreement levels may be a reflection of the collection and topics, not inherent in element retrieval.

5.1 Discussion Point

Judges agree on relevant documents at levels comparable to TREC. They agree less so on relevant elements, and less so again on relevance levels. This disagreement in relevance levels suggests quantization functions based on relevance levels will prove unsound.

The quantization functions rely on a judge's fine grained ability to identify the relevance level of a given element. It appears as though judges do not agree on this – if this is the case then developing ranking functions that utilize this is futile.

6. CONCLUSIONS

That element retrieval has methodological issues is evidenced by the INEX Element Retrieval Methodology Workshop. It is argued here that these problems stem from one cause, the lack of user grounding.

The identification of an (existing) application of element retrieval may resolve many of the issues.

A document collection containing elements that make sense as atomic retrieval results is needed. Should an application be identified then a document collection mirroring the collection in use could be built – the very collection in use might be used. With no application it is proving hard to identify even the distinguishing characteristics of a suitable collection. It is, however, proving possible to demonstrate that the characteristics of the existing collections make them unsuitable.

Given an application, the queries entered by users can be studied and suitable languages and querying methods can be identified. At present it appears as though even the simplest query operators are beyond the use of typical users. Given this, research into how to improve such searching strategies will have little or no measurable effect on performance.

The metrics used for measuring the performance of element ranking strategies have proven to be open to practices identified by users as of negative value. Again, if an application of element retrieval can be identified then the nature of a good result set can be identified. The metrics should reflect good user practice.

At present there is no identified application of element retrieval. There is no practical model and consequently no theoretical model. This has lead to multiple interpretations of the task and continued debate on what the search engine is trying to identify. In essence, each INEX participant has their own retrieval model.

Identifying an application of element retrieval is a vital first step. If it isn't possible to identify such an application, such an application may not exit. Unless the community can collectively identify such an application methodological issues will continue plague the research.

In summary, element retrieval methodological issues arise from one problem – the lack of a user model. To move beyond this, a real-world application must be identified and a model derived that is based on this use. In this way the identified element retrieval issues would be resolved against a user model.

7. REFERENCES.

[1] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., & Cowan, J. (2003). Extensible markup language (XML) 1.1 W3C proposed recommendation. The World Wide Web Consortium. Available: http://www.w3.org/TR/2003/PR-xml11-20031105/.

[2] Cormack, G. V., Palmer, C. R., To, S. S. L., & Clarke, C. L. A. (1997). Passage-based refinement (multitext experiements for TREC-6). In *Proceedings of the 6th Text REtrieval Conference (TREC-6)*, (pp. 171-186).

[3] Fuhr, N., Gövert, N., Kazai, G., & Lalmas, M. (2002). INEX: Initiative for the evaluation of XML retrieval. In *Proceedings of the ACM SIGIR 2000 Workshop on XML and Information Retrieval*.

[4] Gövert, N., Kazai, G., Fuhr, N., & Lalmas, M. (2003). *Evaluating the effectiveness of content-oriented XML retrieval*: University of Dortmund, Computer Science 6.

[5] Harman, D. (1993). Overview of the first TREC conference. In *Proceedings of the 16th ACM SIGIR Conference on Information Retrieval*, (pp. 36-47).

[6] Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science,* 47(1), 37-49.

[7] ISO8879:1986. (1986). *Information processing - text and office systems - standard generalised markup language (SGML)*.

[8] Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *Transactions on Information Systems,* 20(4), 422-446.

[9] Kamps, J., Rijke, M. d., & Sigurbjörnsson, B. (2004). Length normalization in XML retrieval. In *Proceedings of the 27th ACM SIGIR Conference on Information Retrieval*, (pp. 80-87).

[10] Kaszkiel, M., & Zobel, J. (1997). Passage retrieval revisited. In *Proceedings of the 20th ACM SIGIR Conference on Information Retrieval*, (pp. 178-185).

[11] Kazai, G. (2003). Report of the INEX 2003 metrics working group. In *Proceedings of the INEX 2003 Workshop*.

[12] Kazai, G., Lalmas, M., & Vries, A. d. (2004). Reliability tests for the xcg and inex-2002 metrics. In *Proceedings of the INEX 2004 Workshop*, (pp. 60-72).

[13] Kazai, G., Lalmas, M., & Vries, A. P. d. (2004). The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th ACM SIGIR Conference on Information Retrieval*, (pp. 72-79).

[14] Kekäläinen, J., Junkkari, M., Arvola, P., & Aalto, T. (2004). Trix 2004 - struggling with the overlap. In *Proceedings of the INEX 2004 Workshop*, (pp. 127-139).

[15] Kim, H., & Son, H. (2004). Interactive searching behavior with structured XML documents. In *Proceedings of the INEX 2004 Workshop*, (pp. 424-436).

[16] Lalmas, M., Fuhr, N., Malik, S., Szlavik, Z., & Trang, V. H. (2004). *Some statistics for INEX 2004* (PDF of Presentation Slides). London: Queen Mary University of London.

[17] Malik, S., Lalmas, M., & Fuhr, N. (2004). Overview of INEX 2004. In *Proceedings of the INEX 2004 Workshop*, (pp. 1-15).

[18] O'Keefe, R. A. (2004). If INEX is the answer, what is the question? In *Proceedings of the INEX 2004 Workshop*, (pp. 54-59).

[19] O'Keefe, R. A., & Trotman, A. (2003). The simplest query language that could possibly work. In *Proceedings of the 2nd workshop of the initiative for the evaluation of XML retrieval (INEX)*.

[20] Pehcevski, J., Thom, J. A., Tahaghoghi, S. M. M., & Vercoustre, A.-M. (2004). Hybrid XML retrieval revisited. In *Proceedings of the INEX 2004 Workshop*, (pp. 153-167).

[21] Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology,* 29, 3-48.

[22] Sigurbjörnsson, B. (2005 - to appear). Focused information retrieval from semi-structured documents (abstract). In *Proceedings of the 28th ACM SIGIR Conference on Information Retrieval*.

[23] Sigurbjörnsson, B., Trotman, A., Geva, S., Lalmas, M., Larsen, B., & Malik, S. (2005 - to appear). INEX 2005 guidelines for topic development. In *Proceedings of the INEX 2005 Workshop*.

[24] Tombros, A., Larsen, B., & Malik, S. (2004). The interactive track at INEX 2004. In *Proceedings of the INEX 2004 Workshop*, (pp. 410-423).

[25] Trotman, A. (2005). Learning to rank. *Information Retrieval,* 8(3), 359-381.

[26] Trotman, A., & Sigurbjörnsson, B. (2004). Narrowed Extended XPath I (NEXI). In *Proceedings of the INEX 2004 Workshop*, (pp. 16-40).

[27] Trotman, A., & Sigurbjörnsson, B. (2004). NEXI, now and next. In *Proceedings of the INEX 2004 Workshop*, (pp. 41-53).

[28] Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management,* 36(5), 697-716.

[29] Wilbur, W. J. (1998). A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task. *Journal of the American Society for Information Science,* 49(6), 517-529.

[30] Woodley, A., & Geva, S. (2004). Nlpx at INEX 2004. In *Proceedings of the INEX 2004 Workshop*, (pp. 382-394).