What does Shakespeare have to do with INEX?

User Queries, Assessment Behaviour and Best Entry Point Selection Strategies in XML Retrieval

Gabriella Kazai
Microsoft Research Cambridge
gabkaz@microsoft.com

Elham Ashoori
Queen Mary, University of London
elham@dcs.qmul.ac.uk

ABSTRACT

Since 2002, the INitiative for the Evaluation of XML Retrieval (INEX) has been building an XML test collection for the evaluation of content-oriented XML search systems. In 2006, INEX extended its range of investigated user tasks to include the Best in Context task, where systems are required to return Best Entry Points (BEPs) to the user. In this paper we take a look back at a small user study conducted at Queen Mary, University of London, which resulted in the construction of the Shakespeare XML test collection. This test collection includes - in addition to the standard components of documents, user queries and relevance assessments - BEP judgments, where BEPs were defined as optimal points for browsing a document's structure to access relevant information. We examine some of the findings of topic author and assessor behaviours in the Shakespeare study and draw comparisons to findings reported at INEX. In addition, we provide a detailed analysis of users' BEP selection strategies and review related user studies with the aim to help guide efforts at INEX.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

XML test collection, Best Entry Points

1. INTRODUCTION

The Shakespeare user study [9] involved 11 English and Drama students at Queen Mary, University of London and resulted in the construction of a small XML test collection[1] (10MB). Each XML document in the collection is a Shakespeare play consisting of the original text of the plays and the XML markup. The markup follows the logical structure of the plays with the following main structural components: PLAY (root nodes), ACT, SCENE, SPEECH (composite nodes), and LINE or STAGEDIR (leaf nodes).

Participants of the study were asked to come up with user queries for 3 plays of their choice, to provide relevance assessments using binary relevance scale and a yellow-marker design (the same assessment procedure which is to be adopted at INEX this year), and finally to provide BEP assessments.

We report on some of the observations of the study regarding the types of user information needs in the context of XML retrieval (Section 2) and user behaviour during relevance assessments (Section 3). In Section 4, we examine the relationship of relevance assessments at INEX 2005 to semantic units. In Section 5 we provide detailed analysis of users' BEP selection strategies.

2. QUERIES

Based on participants' familiarity, 12 plays were selected (out of 37) for the study. Participants were asked to formulate queries addressing real information needs, and covering topics related to their chosen plays that were of interest to them. It was desirable to obtain queries of varying complexity, and two main types were identified in this context:

- Factual questions, where it is likely that a small number of short passages will provide the answer, e.g. "How old is Juliet?"
- Essay topics, where it is likely that reference will have to be made to many complex passages, e.g. "The character of Lady Macbeth".

2.1 CAS vs. CO queries

Participants were not told about possibilities to query using structural constraints, for example, to limit the context of possible answers. Any decision to impose such structural constraints within a query was left up to the participants. The aim was to obtain an unbiased query set where it could be observed whether there is in fact a real world need for the different query types, i.e. content-only (CO) and content-and-structure (CAS).

A total of 215 queries were submitted (average 18 per play and 19.5 per participant). Table 1 shows their distribution across the different query categories. 43% of the queries were CAS and 57% were CO queries. This shows that both CO

[1]Available at http://qmir.dcs.qmul.ac.uk/Focus/resources.htm

SIGIR 2006 Workshop on XML Element Retrieval Methodology
August 10, 2006, Seattle, Washington, USA.

	CO	CAS	Total
Factual question	54	15	69
Essay topic	68	78	146
Total	122	93	215

Table 1: Distribution of the original 215 submitted queries across query types

and CAS queries are naturally needed and used by novice users when searching structured documents.

In contrast to the above methodology, INEX explicitly instructs topic authors to create both CO and CAS queries, while the complexity of the query is left unspecified. A number of studies, however, showed that INEX topics can also be classified as specific (narrow) or general (broad) topics [5, 14]. It could be argued that users' requests for more specific information is related to the assumed advantage of XML IR over traditional IR: the ability to locate exact relevant fragments within documents. Given that such relevant fragments may intuitively be thought of as smaller, more focused components, this could inadvertently influence users in requesting more specific information and hence ask more factual questions. This has been raised by [20, 12] commenting that INEX participants struggle to come up with queries that can take advantage of the structure of the collection (and make sense at the same time).

2.2 Influence of a semantic unit

A closer look at the CAS queries of the Shakespeare study reveals that the most commonly used structural constraint is simply to limit the context of the query to the level of PLAY (e.g. "How is Sebastian feminised in the play?") and even to a specific play (e.g. "Trickery and treachery in Much Ado about Nothing."). 80% of the CAS queries were of this type. Only 18 of the 93 CAS queries contained explicit structural references to ACT, SCENE or even LINE elements.

The fact that the majority of the CAS queries only specifies the unit of the whole play as structural constraint suggests that these semantically coherent and independent units of information represent the default context for users. Some users may then go further and explore the inner structure of the individual documents, but the unit of the documents themselves are typically identified first. These findings go hand in hand with the investigations of the FERMI project on multimedia IR [2], and provide support for the fetch and browse strategy proposed there: retrieving whole documents (fetch) then focusing the users attention to the most specific components within the documents (browse).

It is necessary to note, however, that the findings of the Shakespeare study were heavily biased due to the experimental setup, whereby participants were asked to come up with queries for their selected plays. So naturally, the obtained queries tended to be limited to the scope of a given play. The same can be said for CO queries, most of which were also meant with specific plays in mind, even though the context of the query was not explicitly stated. For example the query "To what extent is Hamlets madness a pretence?" implicitly assumes that the context is the play Hamlet.

Studies of the INEX topics have shown similar results, highlighting users' natural association of complete ARTICLE elements as overall semantic units (27% of INEX 2003/2004 CAS topics targets ARTICLEs) [12, 20]. Although it has also been argued that this is due to forced pressure on topic authors to introduce structure into their information needs. An analysis of the relevance assessments for CO topics, on the other hand, revealed that users do generally prefer to be returned smaller, more specific elements [15, 16] regardless of the existance of an article level semantic unit. The combination of these findings again supports the fetch and browse strategy, which is explored since 2005 at INEX, as an intuitive user task.

3. RELEVANCE ASSESSMENTS

From the original pool of 215 queries submitted by the participants of the Shakespeare study, 43 were selected into the final set for which relevance assessments were collected. A binary relevance scale was employed and assessments were collected from multiple judges. Following the yellow-marker design, participants were provided with printed versions of the plays and queries and were asked to highlight relevant passages on the printed documents by hand. Relevant passages were described as those that they would consult (read or reference) in order to answer a given query.

The highlighted passages were then converted into assessments on structured documents, where the derived set of relevance assessments consists of all the leaf nodes that contain highlighted parts. The obtained 117 sets of relevance assessments (from the 11 participants for the 43 queries) lead to a total of 6,296 relevant leaf level XML elements. The multiple sets of relevance assessments were then merged for each query to form the final set of assessments for the test collection. After merging, a total of 4,898 unique leaf level XML elements were obtained in 43 query sets (average 114 leaf XML elements per query).

3.0.1 Assessor agreement

Assessor agreement was measured as the size of the intersection of the different relevant sets, obtained by the different participants for the same query, divided by the size of the union of the relevant sets [21].

Since assessments were collected at the leaf node level, in order to investigate assessor agreement at higher structural levels (e.g. SPEECH or SCENE), an optimistic relevance propagation strategy was employed [19]. According to this, a container element is judged relevant if at least one of its contained elements is relevant.

The resulting assessor agreement data can be found in Table 2. It shows that agreement increases consistently with higher structural levels. This leads to the overall conclusion that while participants are likely to disagree about the exact location of the relevant information, they tend to agree on the general area in which the answers to a query can be found. The results also show that query type and complexity do not have a strong effect on assessor agreement, although factual queries show slightly higher agreement at most structural levels.

The above agreement levels are (expectedly) much superior

	Leaf	SPEECH	SCENE	ACT	PLAY
Factual question	35	43	59	84	100
Essay topic	27	30	68	76	100
CO	29	35	65	80	100
CAS	30	30	63	73	100
Average	31	35	64	78	100

Table 2: Average assessor agreement (as %) for the different query types at various structural levels

to those reported for INEX (e.g. 0.27 for INEX 2003 and 0.39 for INEX 2004 data [13, 16]) due to the implicit selection of a PLAY as the context of a query.

3.0.2 *Effect of result presentation assumptions on assessor agreement*

A closer look at the collected assessments reveals a possible reason for low assessor agreement at the lower structural levels. Looking at the patterns of highlighted texts, two clearly identifiable trends emerge: some assessors tended to highlight only the very minimal text fragments which provide the most direct answer to a query, while some assessors followed a different strategy and highlighted large contiguous text fragments. During the interviews it came to light that the latter approach was chosen in order to ensure that contextual information was not missed. This provides direct evidence that relevance assessments can be influenced by assumptions about how information may be returned by a retrieval system to the users. The assumption that relevant information is presented to the user as highlighted text within its context could lead to stricter assessments, where only the most specific fragments are marked relevant. On the other hand, assuming that the user is returned only the highlighted information without its context may encourage assessors to be more liberal regarding their criteria for relevance.

This finding may bear significance when evaluating the various tasks at INEX, given that assessors may be influenced by the actual assessment interface in their assessment task. In particular, the presentation of results grouped by articles and the task of highlighting relevant text fragments provides a close match with the Relevant in Context task. However, it is not clear how the evaluation of, e.g., the Focused task may be influenced by the way relevance assessments are collected.

4. RELEVANT VS. SEMANTIC UNITS AT INEX

Apart from the main semantic unit of a whole document, a document can be considered as a sequence of semantically coherent units or topics. Documents can be semantically decomposed through the application of a topic segmentation algorithm. To this end, we employ the topic segmentation of TextTiling[2] [6], which is based on lexical cohesion where change in vocabulary signifies a topic shift. TextTiling is a linear segmentation algorithm which considers the discourse unit to correspond to a paragraph and therefore subdivides the text into multi-paragraph segments. The algorithm determines the number of segments, referred to as tiles, assigned to each document, by considering segment boundaries to correspond to gaps with depth scores above a certain threshold.

Our aim is to understand how people provide relevance assessments, why and how do they highlight text fragments. We are interested in finding out whether people tend to highlight text fragments which form semantic units, i.e. when strong coupling exist within the fragment which is then loosely coupled to its neighbours, or if they highlight fragments that form only some part of a semantic unit. For example, in the Shakespeare study, we found that some assessors highlighted whole sections, while others highlighted only a couple of important lines from the section.

We investigate whether the text segments produced by TextTiling tend to match up with what is highlighted by the assessors as this would provide some level of evidence that people tend to choose such semantic units over smaller fragments. This in turn could provide evidence that people may be influenced in their assessment task by how they imagine results would be returned by a system. For example, if whole semantic units are more often highlighted, then assessors may assume that users would prefer to see the whole context of relevant information. If text fragments within semantic units are highlighted then assessors may assume that it is more important to point the user's attention to the specific relevant part.

The semantic decomposition of an XML document is used as a basis to calculate the matching between the highlighted passages and the semantic segments based on the relevance assessments v.7 for 29 CO+S and 34 CAS topics of the INEX 2005 data set [11]. We set the TextTiling algorithm's parameters to $W = 20$ and $K = 6$ (recommended values [6]).

We calculate the following measures:

1. Length ratio: length of highlighted text / length of text tiles that completely cover the highlighted text
2. Tile count: average number of text tiles that cover a highlighted text fragment

4.1 Results

For the purpose of our investigation, we consider paragraph elements[3] to be the lowest possible level of granularity of a retrieval unit. Due to the segmentation procedure, out of the 4280 (5942) highlighted passages for CO (CAS) topics, we were only able to use 3309 (4323) passages which start and end somewhere inside a paragraph.

Table 3 shows the calculated statistics for length ratio, passage size and tile count. Results are reported for both CO and CAS topics averaged over all highlighted passages (1st column) and over statistics calculated per query (2nd column).

Comparing the average passage size for both topic set, 768.91 for CO vs. 1463.1 for CAS topic, clearly shows that the highlighted passages for CAS topics are on average larger than

[2]http://elib.cs.berkeley.edu/src/texttiles/

[3]Paragraph elements are any elements of the "para" entity as defined in the INEX document collection DTD (<!ENTITY % para "ilrj|ip1|ip2|ip3|ip4|ip5|item-none|p|p1|p2|p3">).

Table 3: Statistics of the matching between the highlighted relevant passages and the TextTiling semantic segments

	All Passages		Per Query	
	CO	CAS	CO	CAS
Average passage size	768.91	1463.1	1286.86	2127.27
Average tile count	1.90	3.05	2.73	4.56
Average length ratio	33.04	46.70	40.83	49.21
Standard deviation for length ratio	28.98	32.52	12.07	19.03

those for CO topics. This observation is somewhat surprising as one would expect that CAS topics would reflect more specific information needs, associated with shorter relevant snippets. However the finding does accord well with those of the Shakespeare study (Table 1), where most CAS queries were essay topics.

Looking at the length ratio, we find that text fragments highlighted by assessors as relevant tend to form only 33% of semantic units for CO topics, compared with 46.7% for CAS topics. This could mean that assessors tend to highlight more context for CAS queries. A counter argument may be that since users are not restricted by the requested structure for CO topics, they are more free to select smaller passages.

Comparing the standard deviation of length ratio for the two averages shows that although in general the length ratio for passages varies highly, smaller deviation exists among assessors when we group passages per query. This suggests that different users follow similar procedures for highlighting passages.

Overall, we found that (for our restricted subset of elements) users highlighted longer passages for CAS topics, where these passages closer matched the semantic segments within the documents.

5. BEP ASSESSMENTS

In the Shakespeare study BEPs have been defined as document components that represent optimal starting points for browsing and accessing relevant information in structured documents.

BEP assessments were solicited by interviewing participants individually. BEPs were identified by consulting the merged relevance assessments collected from all assessors of a query. The selection of BEPs required the use of an interface that allowed participants to browse the document structure and the relevant information within. The purpose of the interface was to show the context of the relevant fragments, and allow the user to form an intuitive understanding of the costs associated with finding relevant information through browsing from potential BEPs. Using the interface, participants were asked to identify BEPs as those document components that they would prefer to be retrieved by a search engine in response to a query.

As a result, a total of 928 BEPs were collected from the 11 participants for the 43 queries (in 117 sets). This number was reduced to 521 by removing duplicates. The average number of BEPs per query was hence 21.58 for non-unique elements and 12.12 for unique elements.

	Leaf	SPEECH	SCENE	PLAY	All
Factual question	63	52	67	-	67
Essay topic	46	62	41	0	57
CO	55	60	45	-	62
CAS	35	59	50	0	53
Average	49	58	51	0	60

Table 4: Average BEP agreement for the different query types at various structural levels (shown as %)

	Leaf	SPEECH	SCENE	ACT	PLAY
Factual q.	58	33	4	0	0
Essay topic	41	53	6	0	0
CO	46	49	5	0	0
CAS	29	57	7	0	0
Average	44	50	5	0	0

Table 5: Distribution of BEPs for different query categories across structural levels (given as %)

5.0.1 BEP assessment agreement

Table 4 shows the results for assessor agreement for BEPs. Compared with assessor agreement for relevance assessment, agreement is much higher for BEPs for all query categories at leaf and SPEECH element levels.

Agreement at higher structural levels is heavily influenced by the sparseness of the sample data (see Table 5), e.g. there are no ACT BEPs, only 5% of all BEPs are SCENE nodes, and only one participant chose the PLAY node as BEP for one query (appears as 0%).

Highest agreement across all levels is for the factual queries. This is likely to be due to the fact that factual queries have a smaller number of relevance fragments (which are also usually tighter clustered) than queries from other categories, so there was less potential for disagreement between participants when choosing BEPs.

Overall, it can be seen that a reasonable level of BEP agreement is achieved for all query categories across all structural levels (with the exception of PLAY), showing that the concept of BEP was found to be intuitive by the participants. These results show, especially given the comparatively low levels of assessor agreement on relevance, that BEPs may provide a more stable basis for retrieval evaluation. A disadvantage is that BEP data tends to be much sparser than relevance data (since one BEP usually represents a whole cluster of relevant nodes), which then has an inverse effect on evaluation stability [1].

5.0.2 BEP selection strategies

The distribution of BEPs in Table 5 shows the overwhelming dominance of leaf and SPEECH level BEPs, which together make up 94% of all BEPs. This suggests that participants generally preferred more specific, focused components as entry points.

A comparison of the different query categories shows that factual queries led to the most specific BEPs, with above average number of leaf level BEPs and below average number of BEPs at higher structural levels. This is likely to be related to the question answering nature of factual queries, where users tend to seek short focused answers. Such answer nodes are also then seen as best candidates for entry points. On the other hand, CAS queries show a trend contrary to this, where SPEECH level BEPs were found the most popular choice for BEPs. This finding is more likely a result of a combination of the structural aspects in CAS queries, i.e. some queries explicitly target SPEECH elements, and the influence of essay topic type queries that make up 10 of the 12 CAS queries, which are often associated with long extents of relevant texts.

To further investigate participants' BEP selection strategies, the relationship of the nominated BEPs to the cluster of relevant information for which they provide an entry point is examined next.

The following tree main types of BEP strategies were identified:

- Container BEP (PBEP): when the parent node of relevant elements is selected as BEP.
- "Start reading here" BEP (SBEP): when a leaf node in a sequence of relevant leaf nodes is selected as BEP. This is usually (but not always) the first node of a sequence that makes up a relevant text fragment. To distinguish between these two cases, BEPs which are the first nodes in a sequence are denoted as SBEP-1, and BEPs which are from somewhere inside the sequence are labeled as SBEP-M. In addition to these, in a number of cases, a BEP was chosen to represent a single relevant leaf node (e.g. when only a single LINE element was highlighted by an assessor), these are denoted as SBEP-SL.
- Combined BEP (CBEP): when a parent node in a sequence of relevant parent nodes was selected, e.g. the first SPEECH node in a sequence of SPEECH elements. Again, usually the first node of a sequence is selected as BEP, these are denoted as CBEP-1, but sometimes nodes in the middle or end of a sequence were picked, these are denoted as CBEP-M.

Table 6 shows the distribution of the different types of BEPs, based on the 521 unique BEPs. Note that micro-averaging was used in these calculations as the sample size for the different BEP types varied widely across queries (hence macro-averaging here would likely lead to skewed averages[4]). By using micro-averaging all BEPs belonging to queries of a given query category were first pooled and then their distribution with respect to the BEP type was examined. This way each BEP was counted with equal importance.

According to the findings, the most popular type of BEPs, with 44.9%, was the "Start reading here" BEP (SBEP), which means that participants in most of the cases simply selected a leaf level entry point, representing the point where they would prefer to be directed to and where they would want to start reading the text. From the SBEP type BEPs (taken as 100%), in the majority of the cases (62%) the BEP chosen was the first leaf node of the sequence of relevant leaf nodes (SBEP-1). Interestingly, however, 10% of SBEP type BEPs were leaf nodes that were selected from somewhere inside the sequence of relevant leaf nodes (SBEP-M). In a few cases this node was the very last node in the sequence. During the interviews, it was explained that such BEPs were usually selected when they contained highly relevant information or for factual questions when they provided the actual answer. One raised point was that once users are directed to these mid-sequence entry points, they can just browse around in the text to read the context if required, but it was more important that the first thing they would see is the relevant information. Finally, a large percentage of the cases concerned SBEP-SL types (28% of all SBEPs), where the single relevant leaf nodes were simply nominated as BEPs themselves. These were usually single LINE nodes that stood relatively separated from any other relevant fragments.

The next most popular BEP type, with 30.8%, was the Container BEP (PBEP). These were nodes at varying levels of the hierarchy: the vast majority being SPEECH nodes (80.12%) then SCENE nodes (19.28%). The remaining 0.06% is a result of the single PLAY BEP chosen by one participant.

Finally, 24.30% of all BEPs were of the combined BEP (CBEP) type, where a SPEECH node is chosen from a sequence of relevant SPEECH nodes. 94% of these were BEPs where the first node is chosen from the sequence. In 6% of the cases, just as for SBEPs a node from the middle of the sequence was chosen, again, for reasons to do with containing highly relevant information.

The following trends can be seen when looking at the breakdown of the distribution of BEPs for the different query categories: the more general queries, i.e. CO and essay topics, have a much larger number of BEPs than the more restricted queries, i.e. factual and CAS. 83.30% of BEPs belong to essay topic queries, compared to the 16.7% belonging to factual queries. This is expected since factual queries tend to have much shorter and more compact relevant fragments, which are typically associated with a single BEP, while more general queries tend to have lots of relevant fragments of various size, distributed over longer streches of texts, which may then be associated with multiple entry points. Similarly, CAS queries (32.65%) tend to focus the relevant information better and hence require less entry points than CO queries (67.35%), where relevant information may be spread over the entire play.

One of the most important characteristics of BEPs is that they represent an entry point to relevant information. This next analysis hence aims to examine the different BEP types with respect to the proportion of relevant information (measured at leaf node level) that is accessible from a given BEP. This is calculated as the percentage of relevant leaf nodes included in the cluster of the relevant text that the BEP represents, to the total number of leaf nodes contained within the cluster. For example, if the BEP is the first node in

[4]For example one query with a single BEP may distort the averages over the 43 queries as it would contribute 1/43-th of the overall statistics

	SBEP				PBEP	CBEP			Total
	-1	-M	-SL	Total		-1	-M	Total	
Factual	5.57	1.30	2.78	9.65	3.71	3.15	0.19	3.34	16.70
Essay topic	22.26	3.34	9.65	35.25	27.09	19.66	1.30	20.96	83.30
CO	19.11	3.90	6.68	29.69	20.78	15.58	1.30	16.88	67.35
CAS	8.72	0.74	5.75	15.21	10.02	7.24	0.19	7.42	32.65
CO.Factual	4.27	0.93	2.41	7.61	2.23	1.11	0.19	1.30	11.14
CAS.Factual	1.30	0.37	0.37	2.04	1.48	2.04	0	2.04	5.56
CO.Essay	14.84	2.97	4.27	22.08	18.55	14.47	1.11	15.58	56.21
CAS.Essay	7.42	0.37	5.38	13.17	8.54	5.19	0.19	5.38	27.09
All	27.83	4.64	12.43	44.90	30.80	22.81	1.49	24.30	100

Table 6: Distribution of BEPs according to BEP types for different query categories

the sequence of 5 relevant, 2 non-relevant, 1 relevant, 8 non-relevant and 4 relevant LINE nodes, then its proportion of relevant accessible leaf nodes is (5+1+4)/(5+2+1+8+4) = 50%. When the BEP is at a higher structural level, all leaf nodes that are contained within the higher level node are counted. For example, if the BEP is the second node in a sequence of 10 SPEECH nodes, then the total number of relevant leaf nodes within the sequence is divided by the total number of leaf nodes contained in the 10 SPEECH elements.

Table 7 shows the resulting scores (as percentages). At a first glance, the most salient finding is the overall difference between the ratio of relevant information accessible by the different type of BEPs: SBEPs are the most focused with 90% of the contained leaf nodes being relevant, while in general a third of the content of BEPs at higher structural levels is irrelevant.

Combining this information with the distribution of BEPs, one can conclude that the majority of users have a strong preference to the most specific, most focused components that contain the most amount of relevant information and the least amount of irrelevant content. This is since 44.9% of all BEPs contain only 10% irrelevant content, 30.8% contain 33% irrelevant content and 24.3% contain 38% irrelevant content.

Of the SBEP type entry points, the most focused nodes are those selected for factual queries and in particular for CAS.Factual queries, which is the most restrictive as to the location of relevant information.

It is interesting that when container nodes are voted as BEPs, participants tend to be more liberal with the inclusion of irrelevant content. Remember that Table 6 showed that such higher level nodes were usually chosen as BEPs for more general queries, e.g. essay topic and CO, where the inclusion of contextual information may contribute to the understanding of the content, rather than being strictly irrelevant. Based on this observation, an expectation here would be that when container nodes are selected as BEPs for factual queries, they would be more focused than those for essay topic queries. This is however not the case, in fact the findings show quite the opposite. For factual queries, on average, half of the container BEPs' content is actually irrelevant. Again, since the data here is based on a small sample size (a total of 20 nodes for the factual set and only 12 for the CO.Focussed set), outliers do have a larger impact on the overall results. Such an outlier is a `SPEECH` BEP of query no. 19, which contains 2 relevant and 30 irrelevant leaf nodes. However, the data does contain other BEPs whose irrelevant content is in the region of 80-30%. It is not clear why such BEPs were indeed chosen. A possible reason is that the contained relevant content's degree of relevance is not that different from the rest of the node's content. For example, if the highlighted fragments were not actually very relevant, participants may have felt that this did not justify an entry point just by itself.

Unlike PBEPs, the findings for combined BEPs follow the intuition and factual queries are characterised by more focused BEPs: 73% of the content being relevant. Another anomalie, however, is that CO.Factual queries have a higher score (79%) than CAS.Factual queries (69%). This again may be due to small sample size (there are only 2 CAS.Factual queries) or other currently unknown factors of the user behaviour that cannot be further analysed here.

Note that the surprising score of 17% for SBEP-SL in the CAS essay topic query category is a result of sparse data: there are only 2 samples in this set, both with atypical characteristics. Their effect on the overall scores is, however, neglegible. Other odd results are the less than 100% scores for SBEP-SL. This is due to a couple of strange BEP selections, where single non-relevant nodes were nominated as BEPs. This seems more of an issue related to disagreement between judges about relevance assessments.

5.0.3 *Related studies of BEP selection strategies*

The study in [3] identified similar BEP types: browsing BEPs (equivalent to SBEPs here) and container BEPs (equivalent to PBEPs). In a separate study in [17, 18], similar and more detailed investigations were carried out (although most statistics were calculated using macro-averaging). The aim of the analysis there was to investigate aspects of BEP that could then be used for automating BEP identification. The work in [18] defined an additional three BEP types: relevance judgment BEP (which is essentially the same as SBEP-SL), combination BEPs (same as CBEPs above) and context BEP, which are non-relevant nodes that are intended to provide contextual information for a relevant fragment. These were not separately identified in this study as they were too rare to provide sufficient sample data for analysis.

In addition to the analysis of BEP data obtained from the Shakespeare study, [18] also investigated the results of a

	SBEP				PBEP	CBEP			Total
	-1	-M	-SL	Total		-1	-M	Total	
Factual	94	99	93	95	53	73	71	73	81
Essay topic	86	80	100	89	68	60	64	60	75
CO	88	91	97	91	69	61	60	61	76
CAS	86	59	100	90	65	64	100	65	76
CO.Factual	93	99	92	93	46	80	71	79	82
CAS.Factual	98	100	100	99	63	69	0	69	78
CO.Essay	87	88	100	90	70	60	58	60	75
CAS.Essay	83	17	100	88	65	62	100	63	76
All	87	86	99	90	66	62	65	62	76

Table 7: Ratio of relevant and total leaf nodes accessible from a BEP, broken down by BEP types and query categories (given as %)

small study conducted on the INEX 2002 test collection. Their analysis showed that combination and container BEPs were hardly used by subjects participating in this test. In fact they claim that 55% of BEPs were of two new types: partial relevance judgement BEPs and so-called new BEPs. The former, accounting for 50% of BEPs, where defined as sub-parts of a relevance judgement. The example mentioned is that of a participant choosing a paragraph as BEP from a relevant section. This, however, seems to point to a methodological issue within the evaluation. The BEP types defined based on the Shakespeare data built on the notion of a smallest unit, i.e. the leaf nodes. BEPs hence could not be chosen at a lower level than this. The analysis of the INEX data, however, it seems was based on different principles, which raises the question whether the two studies could actually be compared reasonably. Unfortunately, no other studies exist yet that could provide an insight into what aspects may characterise a BEP within the INEX test collection.

6. CONCLUSIONS

The Shakespeare test collection, aimed for the evaluation of focussed retrieval approaches to structured document retrieval (SDR), was constructed based on the methodology described in [9] and resulted in a small (around 10MB) test collection (with around 180 000 XML elements). The test collection has proved especially suitable for experimentations regarding user's search behaviour in a focussed SDR environment [3, 10, 17, 18].

A finding concerning the analysis of the collected queries is their wide variation of complexity: from the simplest factual questions, through more general essay topics, to complex queries that are closer in nature to actual user tasks than search topics. The main result of this study regarding user queries is the evidence that both CO and CAS queries are in fact types of queries that are needed by real users in real information seeking situations. The use of structural constraints in queries appears as natural to novice users as the traditional use of CO queries. At the same time the use of CO queries confirms the need for their support by XML IR systems.

Conclusions regarding assessor agreement showed that while participants were likely to disagree about the exact location of the relevant information, they did in fact agree on the general area in which the answers to a query were to be found. The observed agreement statistics at leaf level were slightly worse than those reported for TREC in [21, 22]. In general, factual queries showed highest agreement and CAS queries the lowest.

A closer look at the relevance assessments revealed that the low level of agreement was partly due to assessors' varying implicit assumptions about how the retrieval results may be presented to users.

The BEP assessments were investigated with the aim to derive conclusions regarding users' preferences in what they consider would be the best document components that an XML IR system should return to them in response to a query. Assessor agreement results for BEPs showed that the concept of BEP is an intuitive one. The evidence found suggests that users prefer to be pointed directly at the most specific relevant information. If there are key relevant fragments, these are preferred as users would then browse around to obtain any necessary contextual information. BEPs were usually chosen at the level of the relevance assessments, i.e. leaf level, or one level up in the hierarchy. Similar findings were reported in the Tess study [7, 8], where in the majority of the tasks, users preferred entry points into the documentation that was equal to a relevant item.

In comparison, a study of BEPs for the INEX test collection, reported in [17, 18], showed that section nodes were most often preferred. It is however not clear if this may be due to the generality of the used queries or the different nature of the collection, where more context may be required. In addition, the different definition of BEP types in this study makes the comparison of the results across the different studies questionable.

7. REFERENCES

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR'00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, New York, NY, USA, 2000. ACM Press.

[2] Y. Chiaramella, P. Mulhem, and F. Fourel. A model for multimedia information retrieval. Technical Report FERMI ESPRIT BRA 8134, University of Glasgow, 1996.

[3] K. Finesilver and J. Reid. User behaviour in the context of structured documents. In F. Sebastiani,

editor, *ECIR*, volume 2633 of *Lecture Notes in Computer Science*, pages 104–119. Springer, 2003.

[4] N. Fuhr, M. Lalmas, S. Malik, and Z. Szlavik, editors. *Advances in XML Information Retrieval. Third Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2004), Schloss Dagstuhl, 6-8 December 2004*, volume 3493 of *Lecture Notes in Computer Science*. Springer, 2005.

[5] K. Hatano, H. Kinutani, M. Watanabe, Y. Mori, M. Yoshikawa, and S. Uemura. Keyword-based xml fragment retrieval: Experimental evaluation based on inex 2003 relevance assessments. In N. Fuhr, M. Lalmas, and S. Malik, editors, *INEX*, pages 81–88, 2003. http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf.

[6] M. A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[7] M. Hertzum and E. Frøkjær. Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Trans. Comput.-Hum. Interact.*, 3(2):136–161, 1996.

[8] M. Hertzum, M. Lalmas, and E. Frøkjær. How are searching and reading intertwined during retrieval from hierarchically structured documents? In *INTERACT '01: Proceedings of the IFIP TC 13 International Conference on Human- Computer Interaction*, pages 537–544, Amsterdam, 2001. IOS Press.

[9] G. Kazai, M. Lalmas, and J. Reid. Construction of a test collection for the focussed retrieval of structured documents. In F. Sebastiani, editor, *Advances in Information Retrieval, Proceedings of the 25th European Conference on IR Research, Pisa, Italy*, volume 2633 of *Lecture Notes in Computer Science*, pages 88–103. Springer, April 2003.

[10] M. Lalmas and J. Reid. Automatic identification of best entry points for focused structured document retrieval. In *CIKM '03: Proceedings of the 12th international conference on Information and knowledge management*, pages 540–543, New York, NY, USA, 2003. ACM Press.

[11] S. Malik, G. Kazai, M. Lalmas, , and N. Fuhr. Overview of inex 2005. volume 3977 of *Lecture Notes in Computer Science*. Springer-Verlag, 2006.

[12] R. A. O'Keefe. If inex is the answer, what is the question? In Fuhr et al. [4], pages 54–59.

[13] J. Pehcevski and J. A. Thom. Hixeval: Highlighting xml retrieval evaluation, 2006.

[14] J. Pehcevski, J. A. Thom, S. M. M. Tahaghoghi, and A.-M. Vercoustre. Hybrid xml retrieval revisited. In Fuhr et al. [4], pages 153–167.

[15] B. Piwowarski and M. Lalmas. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 361–370, New York, NY, USA, 2004. ACM Press.

[16] B. Piwowarski, A. Trotman, and M. Lalmas. Sound and complete relevance assessments for XML retrieval. 2006. Submitted for publication.

[17] J. Reid, M. Lalmas, K. Finesilver, and M. Hertzum. Best entry points for structured document retrieval - part i: Characteristics. *Inf. Process. Manage.*, 42(1):74–88, 2006.

[18] J. Reid, M. Lalmas, K. Finesilver, and M. Hertzum. Best entry points for structured document retrieval - part ii: Types, usage and effectiveness. *Inf. Process. Manage.*, 42(1):89–105, 2006.

[19] T. Rölleke, M. Lalmas, G. Kazai, I. Ruthven, and S. Quicker. The accessibility dimension for structured document retrieval. In F. Crestani, M. Girolami, and C. Rijsbergen, editors, *Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research, Glasgow, Scotland*, volume 2291 of *Lecture Notes in Computer Science*, pages 284–302. Springer, 2002.

[20] A. Trotman. Wanted: Element retrieval users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 58–64, 2005.

[21] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, New York, NY, USA, 1998. ACM Press.

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