Construction of a Test Collection for the Focussed Retrieval of Structured Documents

Gabriella Kazai, Mounia Lalmas, Jane Reid

Department of Computer Science, Queen Mary, University of London, London, E1 4NS
{gabs, mounia, jane}@dcs.qmul.ac.uk

Abstract. In this paper, we examine the methodological issues involved in constructing test collections of structured documents and obtaining best entry points for the evaluation of the focussed retrieval of document components. We describe a pilot test of the proposed test collection construction methodology performed on a document collection of Shakespeare plays. In our analysis, we examine the effect of query complexity and type on overall query difficulty, the use of multiple relevance judges for each query, the problem of obtaining exhaustive relevance assessments from participants, and the method of eliciting relevance assessments and best entry points. Our findings indicate that the methodology is indeed feasible in this small-scale context, and merits further investigation.

1 Introduction

With the widespread use of hypermedia and the rapid adoption of the XML markup language on the Web, there is a growing need to exploit the structural characteristics of documents for the purpose of retrieval. Structure can be found both within an individual document, e.g. a report may contain sections and subsections, and between documents, e.g. Web documents may be connected by hyperlinks. Structured document retrieval (SDR) attempts to exploit such structural information by retrieving documents based on combined structure and content information. This approach has several advantages, including improvement of retrieval effectiveness (e.g. [1], [2], [3], [4], [5], [6]), reduction of user effort (e.g. [7], [8]) and reduction of time and disorientation during the search process (e.g. [9]).

Structural information can be exploited at several stages of the information retrieval (IR) process. Firstly, it can be used at the indexing stage. At this stage, document components are identified and indexed as separate, but related, units. Secondly, structural information can be used at the retrieval stage. There have been three main groups of approaches to SDR. Passage retrieval approaches retrieve documents based on the most relevant passage(s) ([10], [4], [11]). Data modeling approaches employ data models for representation and querying with respect to document content and structure ([12], [13]). Aggregation-based approaches calculate the relevance of document parts based on the aggregation of their own representations and those of their structurally related parts ([14], [15], [6], [16]). Thirdly, structural information can be used at the results presentation stage. This may be achieved by

several different methods. Related objects may be placed together in sub-lists in a traditional-style ranked document list, or grouped together into clusters. Results presentation may be focussed by presentation of selected document components only, rather than all relevant document components. This approach is referred to as *focussed retrieval*. Focussed retrieval is an aggregation-based approach to SDR that combines the browsing and querying paradigms to return the *best entry points* to a structured document. A best entry point (BEP) is a document component from which the user can obtain optimal access by browsing to relevant document components ([9], [17]).

Although SDR systems have already been built, comprehensive evaluation of these systems has not yet been performed[1]. The standard method of evaluating IR systems is by means of a test collection, and the standard measure used is that of retrieval effectiveness ([18], chapter 3). However, traditional test collections (e.g. [19]) are not suitable for evaluating SDR systems because they do not take account of the structural information in the collection, i.e. relevance assessments are made at a document level only. Furthermore, a test collection intended to evaluate an SDR system employing focussed retrieval would also require the ability to evaluate best entry points.

In this paper we discuss the requirements for constructing a structured document test collection for the evaluation of focussed retrieval of structured documents (Section 2). We describe a pilot test of the proposed test collection construction methodology performed on a collection of publicly available Shakespeare plays (Section 3). The outcome of our pilot test is the Shakespeare test collection, available for public use at http://qmir.dcs.qmul.ac.uk/Focus/resources.htm. It comprises 12 XML documents, 43 user queries, relevance assessments and BEPs. The methodology employed in our pilot test collection construction allows us to investigate test collection characteristics and user behaviour during the process of relevance judgement. We evaluate the test collection construction methodology, focussing on the effect of query complexity and type on overall query difficulty, the use of multiple relevance judges for each query, the problem of obtaining exhaustive relevance assessments from participants, and the method of eliciting relevance assessments and BEPs (Section 4). We close with conclusions and future work in Section 5.

2 Structured document test collection requirements

The aim of a test collection construction methodology is to derive a set of queries and relevance assessments for a given document collection. This aim is typically achieved by setting up an experimental study with the document collection and a set of participants. The methodology for constructing a *structured document* test collection has additional, specific requirements relating to the structural information contained in the document collection. Decisions therefore need to be made about several aspects of the methodology before the experiment is performed. The next five sub-sections discuss the requirements for each stage of the structured document test collection construction methodology (Fig. 1) in more detail.

[1] The first large-scale SDR evaluation initiative, INEX (http://qmir.dcs.qmul.ac.uk/inex/) has just ended.

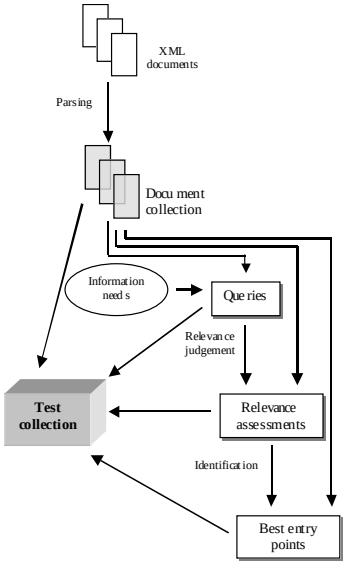

Fig. 1. The structured document test collection construction methodology.

2.1 Documents

There are many different kinds of structure, so the first choice that has to be made is what kind of documents to include in the structured document test collection. The documents could exhibit internal structure (logical structured documents), or external structure (linked Web documents), or a mixture of the two. In the case of linked documents, the links could be either semantic or structural. The nature of the documents chosen also depends on whether a data-centric or document-centric viewpoint is adopted [5]. From the data-centric viewpoint, structured documents serve as containers for data exchange between applications. The document-centric viewpoint, on the other hand, treats documents as traditional textual units, augmented by structural data. To evaluate systems based on the data-centric view, synthetic XML data may be used; however, to evaluate content-based retrieval of structured documents, real-world documents are required.

2.2 Participants

It is normal to recruit participants who are expert in the document collection domain. It is also desirable to choose participants who have real information needs, i.e. who are motivated to take part in the experiment.

2.3 Queries

The format and topic of the queries should be representative of the variety of real user requests that users of the document collection may issue. Queries may take one of several forms, ranging from the actual search statement itself to an expanded version containing supplementary information (e.g. TREC topics [19]). The queries in a structured document test collection should also reflect the additional functionality of structured query languages, i.e. that it is possible to query by structure as well as content. According to this new criterion we can identify the following three types of queries: Content-only, Structure-only, and Content-and-structure.

Content-only queries are the standard type of query in IR. They describe a topic of interest to the user and are represented, in most retrieval systems, by keywords. The need for this type of query in a structured document test collection stems from the fact that users are often unable or unwilling to restrict their search to a specific structural unit. This provides a challenge for SDR systems, since they must not only locate relevant document components, but also identify the appropriate level of granularity.

Structure-only queries do not contain any reference to the content or topic of the information need, but pertain only to the structure of the document collection and/or individual documents. Examples of such queries are “Retrieve the section title and first paragraph of Section 2.3”, and “Retrieve those web pages that are linked from this page”. In this case, retrieval is based on matching between the query requirements and the structural data about the collection (structure index).

Content-and-structure queries combine topical and structural requirements. An example of such a query is “Retrieve the title and the first paragraph of sections about wine-making in the Rhine region”. In this case, retrieval requires matching on both the content index and the structure index of the document collection.

2.4 Relevance assessments

Relevance assessments are then gathered. Two principal decisions need to be made regarding the relevance judgement process: 1) Will there be one judge per query, or more than one? and 2) Will the relevance assessments be binary or multi-valued?

In addition, in a document collection with multiple structural levels, relevance assessments must be derived for each structural level. However, this cannot be achieved by the simple strategy of asking judges to judge each possible structural unit, for two reasons. Firstly, this would be incredibly resource-intensive, especially for large-scale document collections. Secondly, it would be very difficult for judges to assign accurate and consistent relevance assessments in a multiple-layer structure. A choice of structural level for relevance assessments must therefore be made. A possible choice is the smallest structural unit. Relevance assessments at the lowest structural level can allow for the automatic computation of relevance of higher structural levels by a process of relevance propagation. A *pessimistic* propagation strategy would judge a containing element relevant to a given query only if all of its contained elements were relevant. An *optimistic* strategy would judge a containing element relevant to a given query if at least one of its contained elements were relevant. This process could not easily be carried out in the opposite direction, i.e.

given relevance assessments at a higher structural level, it is usually not possible to derive the relevance of lower structural levels.

In the case of XML documents, the smallest structural element corresponds to the last elements of a containment chain [20]. In the case of logically structured documents, a paragraph or a sentence could be set as the lowest level. In the case of a web site, the individual web pages of the site could be considered as the smallest structural units.

2.5 Best entry points (BEPs)

An additional requirement for test collections intended to evaluate focussed SDR systems is to identify BEPs for the given queries. BEP identification should be performed by the same participants who performed the relevance judgement, since they are already familiar with the given queries. The selection of BEPs requires the use of an interface that allows the participants to browse the document structure, including the relevance assessments. The purpose of the interface is to show the context of the relevance assessments, and allow the user to form an intuitive understanding of the costs associated with finding relevant document components from potential BEPs.

The interface should support the following browsing behaviour:

- Next. The user moves to the next sequential unit at the same structural level.
- Previous. The user moves to the previous sequential unit at the same structural level.
- Up. The user moves up a level in the hierarchical structure.
- Down. The user moves down a level in the hierarchical structure.

The next section provides a detailed description of the Shakespeare test collection experiment, taking into account the factors discussed above.

3 The Shakespeare test collection experiment

Our aim was to construct a focussed structured document collection by performing a pilot test, which would take into account the methodological issues discussed in Section 2. In this section, we introduce the basic elements of this pilot test. Section 3.1 describes the document collection, and Section 3.2 discusses the participants who were recruited to take part in the experiment. Section 3.3 describes the test collection construction methodology itself.

3.1 Document collection

The document collection used as the basis of this experiment consists of 37 Shakespeare plays. This material was chosen because of the unusual characteristics of the data. This is in contrast to many of the studies on test collections, which use computer-related data and participants because of their accessibility.

The plays, marked up originally in XML by Jon Bosak, were downloaded from the Web (http://www.ibiblio.org/bosak/). They were then parsed to identify each piece of content enclosed by XML tags as retrievable entities. The parser assigns a unique object identification number to each retrievable XML element and stores this as an attribute of the corresponding XML tag. Figure 2 shows part of the XML document structure. The maximum depth of nested XML elements is 6.

Fig. 2. Part of the Shakespeare collection's XML structure.

A total of 179,689 elements were identified in the 37 plays. Twelve plays were then selected for the final test collection on the basis of participant familiarity. The twelve plays are: Antony and Cleopatra, A Midsummer Night's Dream, Hamlet, Julius Caesar, King Lear, Macbeth, Much Ado About Nothing, Othello, Romeo and Juliet, The Tempest, Troilus and Cressida, and Twelfth Night. On average each of the 12 chosen plays contains 5,096 elements, including 5 acts, 21 scenes, 892 speeches and 3,311 lines.

3.2 Participants

Sixteen students from the undergraduate BA in English Literature and Drama at Queen Mary, University of London originally signed up for the experiment. Fourteen were selected on the basis of their Shakespeare knowledge. However, three dropped out after failing to complete the first task, so the final group of participants consisted of eleven students (five first years, two second years, four third years). The time required to complete the experiment was estimated at approximately seven hours, and payment was fixed at £40, to be paid on completion of all the tasks.

The participants were asked to choose 3 Shakespeare plays with which they were familiar. A questionnaire was administered to all participants to gather data about their interest in Shakespeare, their skill in the use of electronic resources and their familiarity with their chosen plays. Five of the participants were interested in Shakespeare for personal reasons (e.g. they enjoyed the language), five for academic reasons (i.e. reasons related to their current course or their future career) and one participant was interested for both personal and academic reasons. All the students were familiar with the Internet, and all used Internet search engines on a regular basis, but only one had used a full-text poetry database before.

Participants were asked to rate their familiarity with their chosen plays on a five-point scale (1 = very well, 5 = not well at all). Ranked by familiarity on the part of the

participants, the two most confident students gave a rating of 1.00 for all their plays. The least confident student scored an average of 2.83 (across all chosen plays). The average score across all participants was 1.74. Ranked by play, The Tempest was the best known play, with a rating of 1.00 from all the participants who chose it. Troilus and Cressida was the worst known play, with a rating of 3.25 across all the participants who chose it. Data about participants' familiarity with their chosen plays was collated, and 12 plays selected on the basis that 2-3 participants were familiar with each play.

3.3 The Shakespeare test collection methodology

The experiment was carried out in 3 stages: obtaining queries, gathering relevance assessments and identifying BEPs.

Obtaining queries

Participants were asked to produce queries for each of their plays. They were asked to formulate queries (i.e. search statements) that addressed real information needs, and covered topics that were of interest to them and for which they were motivated to seek the answers. It was desirable to obtain queries of varying complexity, and two main types of queries were identified in this context:

1. Factual queries, where it is likely that a small number of short, simple passages will provide the answer. An example query is "How old is Juliet?"
2. Essay-topic queries, where it is likely that reference will have to be made to many, complex passages. An example query is "The character of Lady Macbeth".

An additional criterion, as discussed in Section 2.3, was that the queries contained a mixture of content-only, structure-only (e.g. 'What is the title of the second scene?"), and content-and-structure queries (e.g. including a structural condition like "at the beginning of the play").

A total of 215 queries were obtained, with an average of 18 per play and 19.5 per participant. Of this pool, 43 queries were finally selected for the latter stages of the study (Table 1). The following selection criteria were employed:

- No more than 4 queries per play, due to the limited number of participants
- A maximum of one factual query per play
- Queries of varying complexity should be selected for each play

Table 1. Distribution of queries across query categories.

	Content-only	Content-and-structure	*Total*
Factual	9	2	*11*
Essay-topic	26	6	*32*
Total	*35*	*8*	*43*

Relevance assessments

In this study we used binary assessments collected from multiple judges. The obvious way of obtaining the relevance assessments would have been by employing the pooling method often used in IR research [21]. This method allows the identification of a smaller, optimal pool of document components for relevance judgement from a large-scale document collection. However, at the time of this study, only one SDR system was available to us, so the decision was taken to provide the participants with printed versions of their plays and associated queries, and ask them to highlight the relevant passages on the printed document by hand. This was considered an acceptable solution in this context, since the students were already familiar with the plays, and the document collection was comparatively small-scale. Relevant passages were described as those that they would consult (read or reference) in order to answer a given query. The participants were given one week to complete this task.

The relevant passages were treated at the lowest structural level, referred to as leaf level elements, as described in Section 2.4. As a result, we obtained 117 sets of relevance assessments, totaling 6,296 leaf level XML elements, from the 11 participants for the 43 queries. The multiple sets of relevance assessments were then pooled for each query to derive the final set of relevance assessments for the test collection. Merging the different sets of relevance assessments, we obtained a total of 4,898 unique leaf level XML elements in 43 query sets. The average number of relevant leaf level XML elements is 114 per query. Since there is only one relevant play for each query, and given that a play contains, on average, 5,096 elements, the relevant elements for a given query represent 2.23% of the play.

Best entry points (BEPs)

BEPs were solicited by interviewing the participants individually. An interview lasted approximately 2 hours, and was divided into 3 stages. Stage 1 (10 – 15 minutes) involved the completion of a questionnaire regarding their own background knowledge and interests, together with some questions about the tasks (Section 4). Stage 2 (20 – 30 minutes) involved the participants explaining how they had interpreted a given query and why they had judged particular texts as relevant. Stage 3 (75 – 90 minutes) involved the participants choosing best entry points for each of their queries. The BEPs were identified by consulting the pooled relevance assessments of all the participants assigned to that individual query. It should be noted that the BEPs did not have to be elements that had been judged relevant, but were, in some cases, non-relevant container or contained elements. The participants were aided in their selection of BEPs by the use of a user interface (Fig. 3) that explicitly showed both the structure and content of the plays, and clearly highlighted the elements that had been marked relevant by at least one participant. They were asked to identify the BEPs as elements that they would prefer to be retrieved by a search engine in response to a query.

Each play was viewed in an expandable / collapsible tree view. The queries were presented in a drop-down list at the top of the screen. Users could either select a query from this list, or type it into the text box directly. Once a query was entered, the tree view section of the window was updated to display the appropriate play and relevance assessments. Each higher level structural element, such as SCENE or SPEECH, could be expanded to view its lower level child nodes, or could be collapsed to hide its child

nodes. By default, all non-relevant elements appeared collapsed and all relevant elements appeared expanded. Relevant elements were marked with a red arrow. Users could also scroll up and down the text.

Fig. 3. User interface for best entry point selection.

A total of 928 BEPs were collected from the 11 participants for the 43 queries, in 117 sets. This number was reduced to 512 by removing duplicate elements. The BEPs for each query, as judged by each participant, were then combined to form the final set of BEPs; only elements judged as BEPs by the majority of the participants were included. This was to avoid the problem of multiple BEPs representing the same cluster of relevant elements, e.g. two individual participants choosing two different lines of the same speech as best entry points. The average number of BEPs per query in the final set was 21.58 for non-unique elements and 12.12 for unique elements.

4 Analysis

In our analysis, we focus on an evaluation of the methodology employed in the Shakespeare user study.

Firstly, we examined the effect of query complexity (factual vs. essay-topic) and query type (content-only vs. content-and-structure) on the participants' assessment of the difficulty level of the queries. As mentioned in Section 3.3, we administered a questionnaire to the participants, in which we asked them to rate each query they judged with respect to two dimensions, using a five-point scale (1 = very easy, 5 = very difficult). The two dimensions were: 1) How easy it was to understand the query, and 2) How easy it was to find the answers.

We obtained an estimate of the overall difficulty of each query by averaging the scores for the two dimensions over all the participants who judged that query. Scores

were then averaged across all queries belonging to an individual query category (Table 2).

Table 2. Query difficulty for different query categories.

Query category	Ease of understanding	Ease of finding answer	*Average difficulty*
Factual	1.20	1.55	*1.37*
Essay-topic	1.83	2.39	*2.11*
Content-only	1.60	2.06	*1.83*
Content-and-structure	1.96	2.67	*2.31*
Overall average	*1.66*	*2.18*	*1.92*

We can see that the participants generally found it easier to understand the queries than to find the answers, despite the fact that most participants reported a high level of familiarity with the plays they were using. The ordering of the query categories was the same for both dimensions, and the average difficulty reflects this. Factual queries were found to be easiest, as might have been expected from the small number of relevant objects generally required for these queries. Content-and-structure queries were found to be the most difficult, as they require both content and structural constraints to be fulfilled; this implies an increased amount of effort in identifying relevant objects.

Secondly, we analysed the feasibility of involving multiple participants in assessing each individual query by examining the degree of agreement among the multiple sets of relevance assessments and BEPs. Several studies have examined agreement between relevance assessors (e.g. [22]); however, BEP agreement has not yet been studied. Furthermore, few test collections have employed multiple relevance judges; one exception to this is [23]. We therefore measured the overlap for both the relevant object and BEP sets, where overlap was defined as the size of the intersection of the relevant sets divided by the size of the union of the relevant sets [24].

Since BEPs could be of any structural level, it was possible to examine BEP agreement directly at different structural levels; BEP agreement was also calculated across *all* levels. It should be noted that there were no BEPs at ACT level, and only one at PLAY level. However, it was not possible to make this direct comparison at different structural levels for relevance assessments, since those were made at leaf level only (97% were LINE objects). We therefore created *extrapolated relevance assessments* at higher structural levels by assuming relevance at the structural level above that of the relevance assessment on which it was based. An optimistic relevance extrapolation strategy was used [16]; for example, if one line was marked as relevant, relevance was extrapolated to the (complete) speech containing that line, and so on.

The resulting relevant object and BEP agreement data can be found in Tables 3 and 4, respectively. The data shows that (extrapolated) relevance agreement increases consistently with structural level, except for content-and-structure queries at speech level. This exception may be due to the fact that the location of relevant material is already constrained by the structural element of the query, so agreement does not show improvement at the higher structural level. Overall, participants may not always

agree on the exact context of the relevant object, but tend to agree on the general area in which the relevant objects can be found. The results also show that query type and complexity do not have a strong effect on relevance agreement, although factual queries show slightly higher relevance agreement at most structural levels.

Table 3. Average relevance agreement for different query categories across structural levels.

Query category	Leaf-level	Speech	Scene	Act
Factual	35%	43%	59%	84%
Essay-topic	27%	30%	68%	76%
Content-only	29%	35%	65%	80%
Content-and-structure	30%	30%	63%	73%
Overall average	*31%*	*35%*	*64%*	*78%*

Table 4. Average BEP agreement for different query categories across structural levels.

Query category	Leaf-level	Speech	Scene	Act	Play	*All levels*
Factual	63%	52%	67%	---	---	*67%*
Essay-topic	46%	62%	41%	---	0%	*57%*
Content-only	55%	60%	45%	---	---	*62%*
Content-and-structure	35%	59%	50%	---	0%	*53%*
Overall average	*49%*	*58%*	*51%*	---	*0%*	*60%*

Agreement is better for BEPs than relevance assessments for all categories at leaf and speech level. Agreement then deteriorates at higher structural levels, except for factual queries. This exception may be due to the fact that factual queries have a lower number of relevant objects than queries from other categories, so there was less potential for disagreement between participants. The general deterioration may be heavily influenced by the reduced number of BEPs at higher levels. Another, related reason for this result might be the optimistic method of relevance extrapolation employed. This implies that there will be more relevant objects at higher structural levels, and the number of BEPs at higher structural levels may thus appear artificially low in comparison.

Overall, we can see 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 is an intuitive one for our participants. However, relevance agreement is rather low, especially for leaf-level elements. Although comparative evaluation of retrieval systems has proved robust in the face of quite large differences between relevance judges [24], these results show that BEPs would clearly provide a more stable basis for retrieval.

Thirdly, we examined the issue of eliciting exhaustive, rather than merely selective, relevance assessments from the participants, in order to explore whether this might explain the relatively low relevance agreement. Participants were asked directly, in the course of the interview (Section 3.3, Best Entry Points), to state, for each query they judged, whether they had made exhaustive or selective relevance assessments. Percentage exhaustiveness of relevance assessment sets was then calculated for each query, over all participants who judged that query. The results for

different query categories can be seen in Table 5. It should be noted that BEPs were chosen after a full review of the associated relevance judgements and in discussion with the interviewer, and may, therefore, safely be regarded as exhaustive.

Table 5. Exhaustiveness of relevance assessments for different query categories.

Query category	Exhaustiveness
Factual	48%
Essay-topic	65%
Content-only	60%
Content-and-structure	62%
Overall average	*60%*

Most of the query categories show a similar level of exhaustiveness, with the exception of factual queries. This exception can be explained by the fact that participants often stopped searching for further relevant passages once they felt they had found the answer to a factual query. These results confirm that the low level of relevance agreement may have been partially due to selective relevance assessments. This indicates that use of the pooling method for obtaining relevance assessments is strongly recommended in order to identify an optimal subset of documents for relevance judgement. Given this modification, we can conclude that the collection of relevance assessments and BEPs from multiple judges should, indeed, prove feasible in practice.

Finally, we assessed the effect of soliciting relevance assessments at leaf-level only, in order to explore whether users are influenced into choosing relevant objects and BEPs at this lowest structural level. We examined two factors, for different query categories:

- For relevance assessments, the number of full speeches considered relevant as a proportion of the number of speeches of which at least one line was considered relevant (Table 6).
- For BEPs, the proportion of BEPs at different structural levels (Table 7).

We can see from these results that relevance assessments usually consist of complete speeches, rather than single lines. Over all query categories, only 22.3% of the full speeches considered relevant were found to consist of a single line only. This shows that participants did not feel pressurised into choosing single lines as relevant objects. In fact, the most natural structural level for relevance assessments, from the users' viewpoint, was clearly speech level.

Although the total number of BEPs differs considerably according to query category, their relative distribution across structural levels is rather similar for all query categories. Overall, the majority of BEPs were selected from leaf or speech levels, together accounting for 94% of all BEPs. Speech level was the most common, with the exception of factual queries, for which leaf-level BEPs were most common. This slightly different pattern for factual queries can be explained by the nature of the queries themselves, which involve a question-answering, rather than an evidence-gathering process. The "answer" to factual queries is, therefore, likely to be contained in fewer, lower-level contexts.

Table 6. Proportion of speeches considered relevant for different query categories.

Query category	Proportion of speeches considered completely relevant
Factual	95%
Essay-topic	92%
Content-only	93%
Content-and-structure	93%
Overall average	*93%*

Table 7. Distribution of BEPs for different query categories across structural levels.

Query category	Leaf-level	Speech	Scene	Act	Play	Other	*Total*
Factual	2.93 (58%)	1.67 (33%)	0.18 (4%)	0 (0%)	0 (0%)	0.30 (6%)	*5.09 (100%)*
Essay-topic	3.66 (41%)	4.73 (53%)	0.52 (6%)	0 (0%)	0.01 (0%)	0.03 (0%)	*8.95 (100%)*
Content-only	3.95 (46%)	4.21 (49%)	0.45 (5%)	0 (0%)	0 (0%)	0.06 (1%)	*8.67 (100%)*
Content-and-structure	1.39 (29%)	2.79 (57%)	0.35 (7%)	0 (0%)	0.03 (1%)	0.29 (6%)	*4.86 (100%)*
Overall average	*3.48 (44%)*	*3.95 (50%)*	*0.44 (5%)*	*0 (0%)*	*0.01 (0%)*	*0.10 (1%)*	*7.97 (100%)*

These results show that the participants were not influenced by the choice of leaf-level as the basis for relevance assessments. This means that the strategy of choosing a lowest structural level, with a view to propagating relevance to higher structural levels at a later stage, is feasible as well as desirable, since it reduces the complexity of the methodology as well as the time taken to perform the experiment. Further support for the relevance judgement process could be provided in the form of an interface similar to that used during the BEP phase of this study (Section 2.5).

5 Conclusions

This paper proposes a methodology for the construction of structured document test collections. We address the additional requirements imposed by structured document retrieval, and by focussed retrieval in particular, over standard IR. We carried out a pilot test of the proposed methodology, which resulted in the construction of the Shakespeare test collection. In our analysis of the resulting data, we focussed on an evaluation of the methodology employed in the user study.

Firstly, we found that factual queries were considered the easiest, and content-and-structure queries the most difficult, due to the combination of content and structural constraints that have to be fulfilled.

Secondly, we discovered that (extrapolated) relevance agreement increases consistently with structural level for most query categories, with factual queries

showing a slightly increased relevance agreement at most structural levels. BEP agreement is higher than relevance agreement at lower structural levels, but usually deteriorates slightly at higher levels. The low level of relevance agreement, compared to BEP agreement, may be at least partially due to participants employing selective, rather than exhaustive, relevance assessments. However, if the pooling method is used to obtain relevance assessments, it is anticipated that a more satisfactory level of agreement will be achieved. We therefore conclude that the collection of relevance assessments and BEPs from multiple judges is, indeed, feasible in practice, and that the use of BEPs will provide a stable basis for focussed SDR.

Lastly, our analysis showed that, in fact, relevance assessments almost always consist of complete speeches, rather than single lines. The majority of BEPs were also selected from speech level. The apparent preference for speech level is further supported by analysis of information seeking behaviour from this study [25] and from a follow-on, small-scale user study [26]. We can conclude, therefore, that participants were not unduly influenced by the choice of leaf-level as the basis for relevance assessments. This means that the strategy of choosing a lowest structural level, with a view to propagating relevance to higher structural levels at a later stage, is a sensible and feasible one.

Recent work has built on the results reported in this paper. The methodology was modified and used successfully in the INEX Initiative [27]. This involved the construction of a large-scale test collection based on a document collection of more than 12,000 scientific articles provided by the IEEE Computer Society. Ongoing work aims to identify what further adaptation is necessary to use the standard test collection evaluation methodology in the context of SDR, e.g. adaptation of recall and precision measures. Finally, further work will focus on an in-depth examination of the characteristics of factual queries, which appear to yield different results from other query categories for many of the factors we examined.

6 Acknowledgement

This work was carried out under EPSRC grant number GR/N37612.

7 References

1. Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine. In: 7th WWW Conference, Brisbane, Australia (1998)
2. Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E., Ziviani, N.: Link-Based and Content-Based Evidential Information in a Belief Network Model. In: 23rd ACM-SIGIR, Athens (2000)
3. Géry, M., Chevallet, J-P.:Toward a Structured Information Retrieval System on the Web: Automatic Structure Extraction of Web Pages. In: Pre-Proceedings of the International Workshop on Web Dynamics, London (2001)
4. Wilkinson, R.: Effective Retrieval of Structured Documents. In: 17th ACM-SIGIR, Dublin (1994) 311-317

5. Kotsakis, E.: Structured Information Retrieval in XML documents. In: Proceedings of the 17th ACM Symposium on Applied Computing (SAC'02), Madrid, Spain (2002)
6. Myaeng, S., Jang, D.H., Kim, M.S., Zhoo, Z.C.: A Flexible Model for Retrieval of SGML Documents. In: 21st ACM-SIGIR, Melbourne, Australia (1998) 138-145
7. Roelleke, T.: POOL: Probabilistic Object-Oriented Logical Representation and Retrieval of Complex Objects - A Model for Hypermedia Retrieval, Ph.D. Thesis, University of Dortmund, Verlag-Shaker (1999)
8. Fuhr, N., Großjohann K.: XIRQL: A Query Language for Information Retrieval in XML Documents. In: 24th ACM-SIGIR, New Orleans (2001) 172-180
9. Chiaramella, Y., Mulhem, P., Fourel, F.: A Model for Multimedia Information Retrieval, Technical Report Fermi ESPRIT BRA 8134, University of Glasgow (1996)
10. Callan, J.: Passage-Level Evidence in Document Retrieval. In: 17th ACM SIGIR, Dublin (1994) 302-310
11. Salton, G., Allan, J., Buckley, C.: Approaches to Passage Retrieval in Full Text Information Systems. In: 16th ACM SGIR, Pittsburgh (1993) 49-58
12. Burkowski, F.J.: Retrieval Activities in a Database Consisting of Heterogeneous Collections of Structured Texts. In: 15th ACM SIGIR, Copenhagen (1992) 112-125
13. Navarro, G., Baeza-Yates, R.: A Language for Queries on Structure and Content of Textual Databases. In: 18th ACM-SIGIR, Seattle (1995) 93-101
14. Frisse, M.: Searching for Information in a Hypertext Medical Handbook. Communications of the ACM 31 (1988) 880-886
15. Lalmas, M., Moutogianni, E.: A Dempster-Shafer Indexing for the Focussed Retrieval of a Hierarchically Structured Document Space: Implementation and Experiments on a Web Museum Collection. In: 6th RIAO Conference on Content-Based Multimedia Information Access, Paris (2000)
16. Roelleke, T., Lalmas, M., Kazai, G., Ruthven, I., Quicker, S.: The Accessibility Dimension for Structured Document Retrieval. In: 24th European Conference on Information Retrieval Research (ECIR'02), Glasgow (2002)
17. Kazai, G., Lalmas, M., Roelleke, T.: A Model for the Representation and Focussed Retrieval of Structured Documents based on Fuzzy Aggregation. In: String Processing and Information Retrieval (SPIRE 2001), Laguna De San Rafael, Chile (2001)
18. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)
19. http://www.trec.nist.gov. TREC web site.
20. Chinenyanga, T.P., Kushmerick, N.: Expressive Retrieval from XML Documents. In: 24th ACM-SIGIR, New Orleans (2001) 163-171
21. Harman, D.K.: The TREC Conferences. In: Kuhlen, R., Rittberger, M. (eds.): Hypertext - Information Retrieval - Multimedia: Proceedings of HIM 95, Konstanz, Germany (1995) 9-28
22. Janes, J.W.: Other People's Judgments: A Comparison of Users' and Others' Judgments of Document Relevance, Topicality and Utility. Journal of the American Society of Information Science 45 (1994) 160-171
23. Shaw, W.M., Wood, J.B., Wood, R.E., Tibbo, H.R.: The Cystic Fibrosis Database: Content and Research Opportunities. Library and Information Science Research 13 (1991) 347-366
24. Vorhees, E.M.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J. (eds.): 21st ACM-SIGIR, Melbourne (1998) 315-323
25. Lalmas, M., Reid, J., Hertzum, M.: Information Seeking Behaviour in the Context of Structured Documents. In preparation.
26. Finesilver, K., Reid J. User behaviour in the Context of Structured Documents. To appear in: 25th European Conference on Information Retrieval Research (ECIR'03), Pisa (2003)
27. Fuhr, N., Goevert, N., Kazai, G., Lalmas, M. (eds.): INEX Proceedings, Schloss Dagstuhl (2002)