User Behaviour in the Context of Structured Documents

Karen Finesilver and Jane Reid

Open and Distance Learning Unit / Department of Computer Science
Queen Mary, University of London, London, E1 4NS
karen@odl.qmul.ac.uk
jane@dcs.qmul.ac.uk

Abstract. This paper describes a small-scale experimental study examining user behaviour in the context of structured documents. Two variants of the same interface to support information seeking were implemented, one highlighting relevant objects and one highlighting best entry points (BEPs). BEPs are intended to support users' natural information seeking behaviour by providing optimal starting points for browsing to relevant objects. Analysis of the results from the comparative study of these two interfaces shows that the BEP interface was strongly preferred to the relevant object interface, and that appropriate usage of BEPs can lead to improved task performance. However, the study also highlighted shortcomings related to the inconsistent nature of BEPs and to BEP interface design.

1 Introduction

Document collections often display structural characteristics. Structure can be internal, e.g. sections within an individual document, or external, e.g. hyperlinks between web documents. Structured document retrieval (SDR) aims to combine structural and content information in order to improve retrieval effectiveness (e.g. [1], [2], [3], [4], [5]), support users in identifying relevant information quickly and efficiently (e.g. [6], [7]), and relieve problems caused by distance between related document components in results interfaces (e.g. [8]).

Structural information can be exploited at all stages of the information retrieval process: indexing ([9], [10]), retrieval and results presentation. This paper focusses on the last two stages, specifically: the use of aggregation-based approaches to SDR, which calculate the relevance of document components based on the aggregation of their own representations and those of their structurally related components ([11], [12], [5], [13]); and results presentation methods designed to support information seeking behaviour in the context of structured documents. There are several methods of employing structural information in results presentation, e.g. fisheye views to enable effective browsing of large documents [14]; expand-collapse functionality to support focus on, and movement between, particular structural elements of documents [15]; and use of clustering or sub-lists of related objects by web search engines (Google, Northern Light). Results presentation may also be focussed by presentation

F. Sebastiani (Ed.): ECIR 2003, LNCS 2633, pp. 104-119, 2003.

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 acknowledges the importance of users' natural browsing behaviour, and combines the browsing and querying paradigms to return *best entry points* to structured documents. A best entry point (BEP) is a document component from which the user can obtain optimal access, by browsing, to relevant document components ([8], [16]). The use of BEPs instead of relevant document components as the basic units of the results list is thus intended to support information seeking behaviour, and enable users to gain more effective and efficient access to relevant information items.

This paper describes a small-scale experimental study that aims to examine:
Usage and effectiveness of BEPs and relevant objects in the context of task performance
User behaviour in the context of structured documents

Section 2 presents the experimental methodology employed. Section 3 analyses the main results of the study in terms of questionnaire data, composition of objects chosen during searching, information seeking behaviour, and task performance. Section 4 discusses correlations between these different elements, the relationship of our results to previous research on BEPs, and implications of our results for interface design for focussed SDR systems. We close with conclusions and further work in Section 5.

2 Experimental Methodology

In order to examine the usage and effectiveness of BEPs and relevant objects in the context of task performance, we required access to pre-defined queries, relevance assessments and BEP judgements. This section describes the focussed structured document test collection used in the study (Section 2.1), the participants (Section 2.2), the system developed for the study (Section 2.3), and the experimental design (Section 2.4).

2.1 Test Collection

The test collection used was the Shakespeare test collection, publicly available at http://qmir.dcs.qmw.ac.uk/Focus/resources.htm. The test collection is based on a document collection of 12 Shakespeare plays, which forms a subset of the complete collection of Shakespeare plays marked up in XML by Jon Bosak (available from the web at http://www.ibiblio.org/bosak/). Every individual object in the collection has a unique object ID, which is stored as an attribute of the object's XML tags. The main structural objects are PLAY, ACT, SCENE, SPEECH and LINE (Fig. 1). In addition, there is a small number of other tags, e.g. Persona, Stagedir (stage directions).

The test collection also contains 43 queries, each of which relates to one of the 12 plays. The queries are grouped on two dimensions [17]: query type (content-only or content-and-structure) and query complexity (factual or essay-topic).

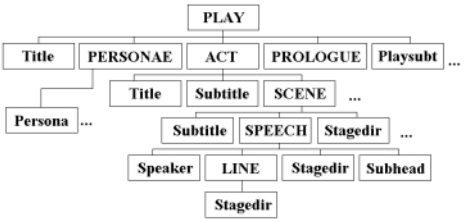

Fig. 1. Part of the Shakespeare collection's XML structure

Queries have associated *leaf-level multi-judge relevance assessments*, which are stored as attributes of the corresponding relevant objects' XML tags. Each query has been judged by 2-4 judges (average 2.75 judges per query), and all relevance assessments are at leaf-level (97% of all relevance assessments are LINE or Stagedir objects).

Each query also has a set of BEPs associated with it, which, like the relevance assessments, are stored as attributes of the corresponding objects' XML tags. BEPs are intended to support users' natural information seeking behaviour by providing entry points to the document structure, from which the user can browse to relevant document components. Unlike relevance assessments, BEPs can be at any structural level (44% of all BEPs in the collection are leaf objects, and 50% are speech objects). It should be noted that BEPs are, therefore, not always objects that have been judged relevant, but are, in some cases, non-relevant container or contained objects. For example, in the case of a long sequence of relevant lines, the first line of the sequence might be chosen as a BEP (*browsing BEP*). Alternatively, if a large proportion of the speeches in a scene were considered relevant, the complete scene might be chosen as a BEP (*container BEP*).

2.2 Participants

Eight undergraduate students were recruited from the Department of English and Drama at Queen Mary, University of London. Each participant was paid £10 on completion of the experiment, on the basis that it would last around 90 minutes. A Masters student was also recruited from the same department to mark the final task outcomes. She was paid £20 on completion of the marking.

2.3 System and Interface Development

The system to be used in the study was built using Borland JBuilder, with a Java interface. A data object model (DOM) was written to parse the XML Shakespeare play files and store them in data trees. These trees were then converted to JTree format (viewable tree structure) by the use of an adapter class. Relevance assessments and BEPs were read from the XML files and stored in arrays.

The user interface (Fig. 2) was designed to support both linear (at the same structural level) and hierarchical (moving between structural levels) information seeking behaviour in the context of structured documents. The interface is split into two main panes. The left-hand pane displays the structure of the play. Any object

that is opened up by the user is displayed in the right-hand pane, e.g. if the user clicks on a speech in the left-hand pane, the contents of that speech are displayed in the right-hand pane. Direct hierarchical information seeking behaviour is supported by use of the left-hand structure pane. Linear information seeking is supported by use of the previous and next buttons in the top right corner of the interface, or by linear use of the left-hand structure pane (e.g. using the keyboard arrow keys).

Queries were pre-entered and chosen from a pull-down menu at the top of the screen. Two variants of the interface were developed, one with the related relevance assessments highlighted (in red), and the other with the related BEPs highlighted. The two interface variants were identical in every other respect.

Software logging was set up, in order to log, for each object selected by the participant, the object ID, the time at which it was selected, its structural level and whether the object was a BEP or a relevance assessment. A separate log was created for each participant session.

2.4 Experimental Design

Two queries were chosen from the 43 queries of the Shakespeare test collection. Since the aim of the experiment was to encourage and examine information seeking behaviour, the queries were chosen to be complex and have a high number of relevance assessments and BEPs. Two essay-topic, content-only queries were, therefore, selected:

Query 1: "To what extent is Hamlet's madness a pretence?" (Hamlet)
Query 2: "Give an analysis of the 'merry war' and witticism that pass between Beatrice and Benedick." (Much Ado About Nothing)

Table 1 shows a breakdown of the BEPs, RJs (relevant objects) and BEP/RJs (objects that appear in both the relevance assessment and BEP sets) for these two queries, organized by structural level. It should be noted that all RJs (and therefore also all BEP/RJs) are leaf objects.

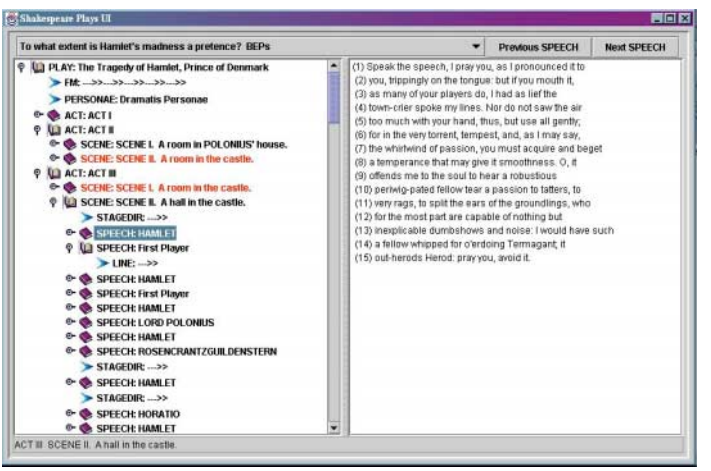

Fig. 2. Screen shot of the information seeking interface

Table 1. Breakdown of BEPs and RJs for the two chosen queries

		Structural level						
		Leaf	Speech	Scene	Act	Play	Other	Total
Hamlet	BEPs	1	23	3	0	0	0	27
	RJs	496	0	0	0	0	0	496
	BEP/RJs	13	0	0	0	0	0	13
	Total	510	23	3	0	0	0	536
Much Ado	BEPs	1	22	3	0	0	0	26
	RJs	377	0	0	0	0	0	377
	BEP/RJs	9	0	0	0	0	0	9
	Total	387	22	3	0	0	0	412

Table 2. Allocation of participants to experimental conditions

	BEP interface	RJ interface
Hamlet	Participants 1-4	Participants 5-8
Much Ado	Participants 5-8	Participants 1-4

Each participant performed both queries, one with the associated BEPs highlighted in the interface, and the other with the associated RJs highlighted (Table 2). Order of interface presentation was counterbalanced.

At the start of the session, each participant was given the chance to become familiar with the interface during a brief training session, using a different Shakespeare play; a help sheet was also issued, which explained the basic elements of the interface. The participant then had 40 minutes to perform each task, which was to write a brief answer to the question (i.e. the query itself), using the interface to find text to back up the answer (indirectly through reading or directly through referencing).

Finally, the participant was asked to fill in a questionnaire, which included questions related to background knowledge, experience of the tasks, and opinions about the interfaces.

After the experiment, the participants' answers were given to the marker, who assigned a percentage score to each answer on the basis of completeness and coverage of content.

3 Results and Analysis

This section describes the results obtained from our user study. There are 4 main sets of results: data from the participant questionnaires, analysis of the composition of the objects chosen by participants during their searching, analysis of the information seeking behaviour of participants during their searching, and task performance results in the form of percentage scores.

3.1 Questionnaire Data

The data elicited by the participant questionnaires was divided into three types: background, query-related and interface-related.

Background information concerns the participants' general familiarity with the works of Shakespeare, and their familiarity with the individual plays used in the experiment (on a scale of 1 to 5, where 1 = very familiar and 5 = not familiar at all). The average general familiarity of participants with the works of Shakespeare was 3.5 (2 participants scored 5, and none scored 1). Participants were, on average, more familiar with Hamlet than Much Ado, and more familiar with the play they used in the RJ interface task than the one they used in the BEP interface task (Table 3). Participants were also asked whether they used the internet and internet search engines. All participants regularly used both.

Query-related information concerns how difficult the participants found the queries to answer (on a scale of 1 to 5, where 1 = very easy and 5 = very difficult). Participants generally found the Hamlet query harder than the Much Ado query, and the query they performed on the BEP interface harder than the query they performed on the RJ interface (Table 4).

Table 3. Average participant familiarity with the plays

	BEP interface	RJ interface	Overall
Hamlet	3.25	3.50	3.38
Much Ado	1.75	2.50	2.13
Overall	2.50	3.00	2.75

Table 4. Average query difficulty

	BEP interface	RJ interface	Overall
Hamlet	3.00	3.25	3.13
Much Ado	3.25	2.50	2.88
Overall	3.13	2.88	3.00

Participants were also asked whether they had an idea of the answer to the questions before they started searching, and whether they had enough time to complete the task (yes/no). Participants had a prior idea of the answer in less than half of the searches performed, with considerably fewer positive answers for the BEP interface than the RJ interface (Table 5). Very few of the participants (3 of the 16 participant sessions) felt they had enough time to complete their tasks (Table 6).

Interface-related information concerns which interface the participants found easier, faster and more helpful (Table 7), and whether they had any comments on possible differences between the two interfaces. A large majority of participants judged the BEP interface as easier and faster. Most participants perceived both interfaces as being equally helpful.

Table 5. Number of participants with a prior idea of the answer to the query

	BEP interface	RJ interface	Overall
Hamlet	2	3	5
Much Ado	0	2	2
Overall	2	5	7

Table 6. Number of participants who had enough time

	BEP interface	RJ interface	Overall
Hamlet	1	0	1
Much Ado	1	1	2
Overall	2	1	3

Table 7. Number of participants expressing a preference for each interface

	BEP interface	RJ interface	No difference
Easier	6	1	1
Faster	6	0	2
More helpful	2	1	5
Overall	14	2	8

Most of the comments did not shed much light on the participants' reasons for preferring one interface to the other. However, one participant commented "The first interface *(the BEP interface)* prompted me to scan throughout more of the text, to get a better understanding of the text as a whole and a better understanding of the development of themes and relationships. Although the second interface *(the RJ interface)* enabled me to be precise about certain quotes I felt that I was bogged down in certain scenes and I wasn't able to explore the text as a whole."

3.2 Object Composition

In this section, we discuss the average composition of the objects chosen by participants in the course of their searching, analysed by query and interface, object type and structural level (Tables 8 and 9).

Most of the objects chosen by participants, across all query/interface combinations, were speech objects. In the Hamlet/RJ combination, it is noticeable that participants hit on quite a large number of BEPs too, by chance, during their session. This is mostly due to the preference of participants for browsing at speech level (rather than line level), which was also the level of most BEPs (23 of the total 27 BEPs). In the Much Ado/RJ combination, on the other hand, participants examined more leaf level objects and fewer speech objects, thus reducing the number of BEPs they found by chance. Objects of higher structural levels were rarely examined, especially in the RJ interface.

Table 8. Average composition of objects chosen by parabticipants for the Hamlet query

Hamlet		Structural level						
		Leaf	Speech	Scene	Act	Play	Other	Total
BEP interface	BEPs	0 (0%)	11.75 (12.1%)	2 (2.1%)	0 (0%)	0 (0%)	0 (0%)	13.75 (14.2%)
	RJs	2 (2.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (2.1%)
	BEP/RJs	1 (1.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1.0%)
	Others	6 (6.2%)	64 (66.1%)	7.5 (7.8%)	1.5 (1.6%)	0.25 (0.3%)	0.75 (0.8%)	80 (82.7%)
	Total	9 (9.3%)	75.75 (78.3%)	9.5 (9.8%)	1.5 (1.6%)	0.25 (0.3%)	0.75 (0.8%)	96.75 (100%)
RJ interface	BEPs	0.25 (0.4%)	5.75 (8.6%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (9.0%)
	RJs	8.25 (12.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	8.25 (12.3%)
	BEP/RJs	3 (4.5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (4.5%)
	Others	5.5 (8.2%)	39.75 (59.3%)	3.25 (4.9%)	0.5 (0.7%)	0.25 (0.4%)	0.5 (0.7%)	49.75 (74.3%)
	Total	17 (25.4%)	45.5 (67.9%)	3.25 (4.9%)	0.5 (0.7%)	0.25 0.4%)	0.5 (0.7%)	67 (100%)

Table 9. Average composition of objects chosen by participants for the Much Ado query

Much Ado		Structural level						
		Leaf	Speech	Scene	Act	Play	Other	Total
BEP interface	BEPs	0 (0%)	11 (13.9%)	1 (1.3%)	0 (0%)	0 (0%)	0 (0%)	12 15.2%
	RJs	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	BEP/RJs	0.5 (0.6%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0.5 (0.6%)
	Others	2.5 (3.2%)	61.75 (78.2%)	1 (1.3%)	0.5 (0.6%)	0.5 (0.6%)	0.25 (0.3%)	66.5 (84.2%)
	Total	3 (3.8%)	72.75 (92.1%)	2 (2.5%)	0.5 (0.6%)	0.5 (0.6%)	0.25 (0.3%)	79 (100%)
RJ interface	BEPs	0 (0%)	3.75 (4.5%)	0.25 (0.3%)	0 (0%)	0 (0%)	0 (0%)	4 (4.7%)
	RJs	15.5 (18.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	15.5 (18.4%)
	BEP/RJs	0.75 (0.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0.75 (0.9%)
	Others	2 (2.4%)	61 (72.4%)	0.25 (0.3%)	0 (0%)	0.25 (0.3%)	0.5 (0.6%)	64 (76.0%)
	Total	18.25 (21.7%)	64.75 (76.9%)	0.5 (0.6%)	0 (0%)	0.25 (0.3%)	0.5 (0.6%)	84.25 (100%)

3.3 Information Seeking Behaviour

The user logs were analysed in order to identify types of actions performed during the search process. Six main actions were identified as a result of the analysis:

- JUMPING. Participants often used this action to reach an object that was some distance away from their current object in the play's structure. Common strategies were: jumping to the next BEP or RJ (usually forwards in the linear structure, and at the same structural level); jumping to the next recognizable structural level (either forwards or backwards in the linear structure), e.g. to look at the next act; jumping to a repeat object, i.e. an object revisited by the participant several times in the course of the session; and jumping to an area of the play that the participant already knew, from their previous knowledge, would be useful.
- LINEAR BROWSING. Participants often used linear browsing at a consistent structural level, usually speech level. The majority of this behaviour was forwards, with the occasional backwards browsing to seek contextual information for the current object, often being followed immediately by forwards browsing through previously visited objects. Other common strategies were: browsing of BEPs / RJs only, and browsing using other query-related criteria, e.g. from one speech by a particular speaker to the next speech by the same speaker.
- HIERARCHICAL BROWSING. Participants most often used this action to examine the contents of the current object, e.g. to look at the relevant lines within the current speech, or to examine the container of the current object. It was noticeable that hierarchical browsing almost always involved a single structural level, e.g. browsing from line to speech, rather than multiple levels, e.g. from line to scene.
- SKIMMING. Participants occasionally moved rapidly through the structure of the play by skimming, i.e. without spending any time looking at the content. This action was rare, but was usually employed in sections of the play that consisted of a succession of individual line speeches, thus allowing the participant to gain an overview of the content while still progressing quickly through the structure.
- RANDOM EXPLORATION. This action was rarely used. The aim appeared to be to explore the structure of the play, or perhaps the functionality of the interface, usually at the start of the session.
- PAUSING. Participants sometimes paused on one object for a considerable length of time. Any pause over one minute or so usually indicated a point at which the participant stopped to make notes or record an important quotation.

The amount of time participants spent on their tasks was also analysed (Table 10). There was very little difference between interfaces, as would be expected from the fixed time limit imposed for the tasks. Participants spent, on average, slightly less time on the Much Ado query than the Hamlet query.

Table 10. Average session time (minutes:seconds)

	BEP interface	RJ interface	Overall
Hamlet	29:24	30:04	29:44
Much Ado	25:39	25:50	25:45
Overall	27:32	27:57	27:44

3.4 Task Performance

After the experiment, participants' answers were assigned a percentage score by the marker (Table 11). Scores were tested for significance using the parametric related t-test. There was found to be no effect from either of the independent variables, interface or query, at $p<0.1$. However, it should be noted that participants were, on average, less familiar with the query they were performing on the BEP interface, found the query more difficult, and had less prior idea of the answer. We might, therefore, have expected that participants would perform considerably worse on the BEP interface than the RJ interface; in fact, they achieved a comparable average score.

Table 11. Average scores (percentages)

	BEP interface	RJ interface	Overall
Hamlet	35.00	43.75	39.38
Much Ado	48.75	37.50	43.13
Overall	41.88	40.63	41.25

It is noticeable that participants performed better on the Hamlet query with the RJ interface, and on the Much Ado query with the BEP interface. This could indicate that the two interfaces supported different types of task to differing extents; this issue is discussed further in Section 4.1.

4 Discussion

We focus on three main areas for our discussion. Firstly, we examine our results for possible correlations between task performance and background data, time spent on searching, and information seeking strategies. Secondly, we discuss links between our results and those of previous, related research. Thirdly, we examine the implications of our results for interface design for focussed SDR systems.

4.1 Correlations

We started by looking for a possible correlation between task performance and background information. We identified the best indicator of performance as whether participants felt they had enough time. The only 3 sessions where the participants felt

they did have enough time were also the 3 top-scoring sessions (1st, 2nd and 3rd=). The second best indicator appeared to be whether participants had a prior idea of the answer to the query before they started searching. In 6 of the 8 top-scoring sessions, the participant stated that they had a prior idea of the answer, compared to only 1 participant in the 8 bottom-scoring sessions. Familiarity with the play did not appear to provide a very good indication of task performance for individual participant sessions, and difficulty of query an even poorer indication.

One noticeable point arising from the questionnaire data was that participants did not necessarily prefer the interface with which they obtained the better score. Of the 8 participants, 2 preferred the interface with which they had performed better on all 3 preference dimensions (easier, faster and more helpful), 3 preferred their better interface on 2 dimensions, 1 judged both interfaces the same on all 3 dimensions, 1 preferred their worse interface on 1 dimension and 1 preferred their worse interface on 2 dimensions. This reminds us that user opinions often arise from a complex mix of experience, not simply as the result of one factor.

Secondly, we looked for a possible correlation between task performance and time spent on searching. Although participants generally performed slightly better on the Much Ado query than the Hamlet query, they actually spent slightly less time on the Much Ado query. However, they did find the Much Ado query, on average, less difficult than the Hamlet query, so this may explain both the lesser amount of time and the better task performance. Participants performed comparably on the BEP and RJ interfaces, both in terms of time spent and task performance. However, they were, on average, less familiar with the query they used with the BEP interface, and found it more difficult, suggesting that the BEP interface may have saved them some time and allowed them to perform better than might have been expected.

Thirdly, we looked for a possible correlation between task performance and information seeking strategies. We attempted to determine good and bad information seeking strategies for each of the different query/interface combinations by ordering our results according to task performance, and identifying the top and bottom participant sessions for each combination. The distribution of queries and interfaces was evenly spread throughout the ordering, so this corresponded to picking participant sessions 1-4 as the top sessions, and 11 and 14-16 as the bottom sessions. Individual participants were also well distributed throughout the ordering, with one participant appearing in both the top and the bottom 4 participant sessions examined.

We started by attempting to identify good and bad strategies for the different interfaces. Good strategies for the BEP interface involved:

- Heavy use of BEPs, especially linear browsing of BEPs
- Browsing at speech level, and using BEPs of higher structural levels mainly for navigation purposes
- Using BEPs as originally intended, i.e. as starting points for browsing to relevant objects

Bad strategies for the BEP interface involved:

- Viewing more objects, but a lower percentage of BEPs
- Less discriminating browsing behaviour, i.e. continuing to browse through non-relevant objects

Good strategies for the RJ interface involved:

- Browsing at speech level, which is more efficient than browsing at line level
- Identifying and browsing long sequences of relevant objects

Bad strategies for the RJ interface involved:

- Duplication of effort by examining both line and container speech objects

We then attempted to identify good and bad strategies for the different queries. This proved to be more difficult, although there were indications that better performance with the Much Ado query was linked to focusing on local areas of the text, while better performance with the Hamlet query was linked to greater breadth of exploration.

We also found several general performance indicators, which spanned all interface/query combinations. Overall indicators of good performance were use of a consistent structural level (usually speech) for browsing, and pausing behaviour in order to make notes during the search. Overall indicators of bad performance were a high percentage of random exploration, backwards browsing, or skipping behaviour. Participants who found a lot of non-relevant material at the start of the session also tended to be less effective.

Finally, we examined the hypothesis, suggested in Section 3.4, that the BEP interface better supported the Much Ado query, while the RJ interface better supported the Hamlet query. Since the same participant group performed both these interface/query combinations, we first examined the participants' background data. This showed that the better-performing participant group had, on average, a lower familiarity with the play, thought the query was more difficult, and had less prior idea of the answer before they started searching. In addition, fewer participants in this group felt they had enough time. The two participant groups spent almost exactly the same amount of time on searching. It is clear, therefore, that background data cannot explain the superiority of one participant group's performance over the other.

We then examined the composition of BEPs and relevant objects selected by participants in the different interface/query combinations. Again, there was nothing unusual that could explain the differences, except perhaps the comparatively high number of BEPs found by chance in the Hamlet/RJ combination.

We turned, lastly, to analysis of the nature of the individual BEPs and RJs for the two queries. It appeared that the BEPs in Much Ado were often *browsing BEPs*, i.e. the first object in a sequence of relevant objects, while in Hamlet, a greater proportion of BEPs were *container BEPs*, i.e. (real or virtual) objects which contained several relevant objects. It is possible, therefore, that the use of BEPs with the Hamlet query did not provide any advantage, since BEPs and RJs were quite similar in composition. In fact, the higher proportion of container BEPs may have encouraged inappropriate browsing behaviour, i.e. participants may have browsed from container BEPs to non-relevant objects. In contrast, the nature of the BEPs in Much Ado differed substantially from that of the RJs, and appears to have succeeded in the original aim, namely to support participants in finding relevant objects through browsing.

4.2 Links with Other Research

This work builds on the Shakespeare user study carried out at Queen Mary, University of London ([17], [18]). In this section, we discuss links with the Shakespeare study and other related research.

Firstly, we examine the concept of BEP itself. In the Shakespeare study, the concept of BEP was shown to be an intuitive one by the comparatively high level of agreement among participants on the choice of BEP in a given context. However, in this study we have shown that the concept of BEP is not always intuitive, since not every participant adopted the "correct" strategy for using BEPs. This could be due to several factors:

- Interface presentation order. Those participants who used the RJ interface first may not have realized that a different strategy was required with the BEP interface.
- The variable nature of BEPs. Participants experiencing container BEPs early on may not have understood that later BEPs were intended as the basis for browsing.
- Participants' individual preferences for browsing (as in a hypermedia interface) or going directly to relevant objects (as in a query-based interface).

Secondly, the related issue of the effectiveness of BEPs is an important one. Although participants performed no better on the BEP interface overall, those participants who made good use of BEPs (mostly with the Much Ado query) did seem to perform better. Participants themselves seemed to recognize the importance of BEPs, since they often returned to them during an individual session; BEPs accounted for 22% of repeat objects in the Hamlet/BEP combination, and 25% of repeat objects in the Much Ado/BEP combination. There also seemed to be a reasonable degree of agreement between participants on which BEPs were important; BEPs accounted for 23% of overlap objects (i.e. objects chosen by at least two participants) in the Hamlet/BEP combination, and 22% of overlap objects in the Much Ado/BEP combination.

Thirdly, we examine the data concerning preference for structural level. A study of information seeking using a document collection of software documentation, the Tess study ([15], [19]), concluded that participants preferred to enter the documentation at the level above that of relevant objects and browse downwards in the structure to examine the relevant objects themselves. The results from the Shakespeare study [18] showed that participants preferred to browse at speech level, rather than leaf level. The data reported in this paper confirm the Shakespeare study results. However, this strategy may have been at least partly due to the interface design, since choosing a container speech object in our interface caused all contained leaf objects to be displayed; there was thus no benefit to our participants in browsing at leaf level. In the Tess interface, on the other hand, only material that belonged exclusively to the container object's structural level was displayed. Because of this difference too, we saw very little evidence in our study of hierarchical browsing behaviour for the purpose of viewing contextual information. However, participants did sometimes use backwards linear browsing, especially from BEPs and relevant objects, in order to view the immediate context.

4.3 Implications for Interface Design

Our results have two main implications for interface design for focussed SDR systems. Firstly, there is evidence to suggest that participants sometimes failed to identify the correct strategy for using BEPs because of confusion over their dual nature as starting points for browsing and containers of relevant objects. In order to foster development of good information seeking strategies, the nature of BEPs should, therefore, be consistent within a particular interface.

Secondly, it was found that participants only looked at approximately half the BEPs (51% of the 27 BEPs for the Hamlet query and 46% of the 26 BEPs for the Much Ado query, on average). If participants had consistently adopted the approach of browsing from BEPs, this proportion could have been even lower. It therefore becomes crucial to find methods of supporting users in identifying the most important BEPs, e.g. by grouping or ordering the BEPs in some way.

5 Conclusions and Further Work

The main conclusions from our work are that participants strongly preferred the BEP interface to the relevant object interface, and that *appropriate* usage of BEPs can lead to improved task performance. However, our study has also highlighted shortcomings related to the inconsistent nature of BEPs and to BEP interface design.

Future work will examine the issue of what should constitute a BEP, particularly which of the suggested types (starting point for linear browsing or container object of relevant objects) provides more intuitive support for users' information seeking behaviour. It will also focus on identifying what factors might affect the choice of using one or other type of BEP in a particular interface, e.g. the type of queries being supported [17], or the target group of users. Results from this work can also feed into ongoing work on automatic identification of BEPs from relevance assessments [20].

Work on interface design to support information seeking in the context of structured documents is already in progress. One approach is to provide more explicit support for users in exploiting the structural information of the document collection. A current student project is evaluating the effect on information seeking behaviour and task performance of different methods of presenting the hierarchical structure of documents [21]. Another approach is to employ some form of ranking in order to support participants in identifying the most important BEPs. A current student project is evaluating the effect on information seeking behaviour and task performance of two methods of ranking: (standard) relevance ranking and ranking by informativeness. Informativeness is a subjective measure that acknowledges the prime importance of presentation order by determining the "ideal" ordering of documents [22].

Finally, many of our results are, inevitably, heavily influenced by the characteristics of the document collection. Further work will, therefore, need to be carried out on different document collections, and on collections from different domains, in order to determine the extent to which we can generalize from our results.

Acknowledgement

This work was carried out by the first author as the final year project component of a BSc in Computer Science at Queen Mary, University of London during academic session 2001/02.

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] Wilkinson, R.: Effective Retrieval of Structured Documents. In: 17th ACM-SIGIR, Dublin (1994) 311-317
[4] Kotsakis, E.: Structured Information Retrieval in XML documents. In: Proceedings of the 17th ACM Symposium on Applied Computing (SAC'02), Madrid, Spain (2002)
[5] 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
[6] 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)
[7] Fuhr, N., Großjohann K.: XIRQL: A Query Language for Information Retrieval in XML Documents. In: 24th ACM-SIGIR, New Orleans (2001) 172-180
[8] Chiaramella, Y., Mulhem, P., Fourel, F.: A Model for Multimedia Information Retrieval, Technical Report Fermi ESPRIT BRA 8134, University of Glasgow (1996)
[9] Tenopir, C., Ro, J.S: Full Text Databases. Greenwood Press (1990)
[10] Cleveland, D.B., Cleveland, A.D., Wise, O.B.: Less than full text indexing using a non-Boolean searching model. JASIS 35(1984):19-28
[11] Frisse, M.: Searching for Information in a Hypertext Medical Handbook. Communications of the ACM 31 (1988) 880-886
[12] 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)
[13] 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)
[14] Furnas, G.W.: The fisheye view: A new look at structured files. In Card S.K., Mackinlay J.D., Shneiderman B. (eds.): Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann (1999), pp.312-330. (Reprinted from The fisheye view: A new look at structured files, Bell Laboratories Technical Memorandum #81-11221-9, October 12, 1981)

[15] Hertzum, M., Frøkjær, E.: Browsing and querying in online documentation: A study of user interfaces and the interaction process. In ACM Transactions on Computer-Human Interaction, 3(1996) 136-161
[16] 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), Laguna De San Rafael, Chile (2001)
[17] Kazai, G., Lalmas, M., Reid, J.: Construction of a Test Collection for the Focussed Retrieval of Structured Documents. To appear in: 25th European Conference on Information Retrieval (ECIR'03), Pisa (2003)
[18] Lalmas, M., Reid, J., Hertzum, M.: Information Seeking Behaviour in the Context of Structured Documents. In preparation
[19] Hertzum, M., Lalmas, M., Frøkjær, E.: How are Searching and Reading Intertwined during Retrieval from Hierarchically Structured Documents? In: INTERACT 2001, Japn (2001)
[20] Kazai, G., Lalmas, M., Roelleke, T.: Focussed Structured Document Retrieval. In: String Processing and Information Retrieval (SPIRE), Lisbon (2002)
[21] Chimera, R., Shneiderman, B.: An Exploratory Evaluation of Three Interfaces for Browsing Large Hierarchical Tables of Contents. In: ACM Transactions on Information Systems, 12(1994) 383-406
[22] Tague-Sutcliffe, J.: Measuring Information: An Information Services Perspective. ASIS (1995)