Information Retrieval, 7, 265–290, 2004

Within-Document Retrieval: A User-Centred Evaluation of Relevance Profiling

DAVID J. HARPER djh@comp.rgu.ac.uk
IVAN KOYCHEV ik@comp.rgu.ac.uk
YIXING SUN sy@comp.rgu.ac.uk
IAIN PIRIE ip@comp.rgu.ac.uk
Smart Web Technologies Centre, School of Computing, The Robert Gordon University, St. Andrew Street, Aberdeen AB25 1HG, UK

Received May 19, 2003; Revised July 24, 2003; Accepted September 8, 2003

Abstract. We present a user-centred, task-oriented, comparative evaluation of two within-document retrieval tools. ProfileSkim computes a relevance profile for a document with respect to a query, and presents the profile as an interactive bar graph. FindSkim provides similar functionality to the web browser "Find" command. A novel simulated work task was devised, where participants are asked to identify (index) relevant pages of an electronic book, given topics from the existing book index. The original book index provides the ground truth, against which the indexing results of the participants can be compared. We confirmed a major hypothesis, namely ProfileSkim proved significantly more efficient than Find-Skim, as measured by time for task. The study indicates that ProfileSkim was as least as effective as FindSkim in identifying relevant pages, as measured by traditional information retrieval measures, and there is some evidence that ProfileSkim is a precision-enhancing tool. Based on qualitative data from questionnaires, we also provide strong evidence to support our conjecture that the participants would be more satisfied when using ProfileSkim than FindSkim. The experimental study confirmed the potential of relevance profiling for improving within-document retrieval. Relevance profiling should prove highly beneficial for users trying to identify relevant information within long documents.

Keywords: within-document retrieval, relevance profiling, interactive information retrieval, task-oriented evaluation, language models

1. Introduction

A user faced with finding textual information on the Web, or within a digital library, is presented with three challenges. First, the user must identify relevant repositories of digital text, usually in the form of document collections. In the context of the Web, this might be by identifying appropriate content portals, or by selecting an appropriate search engine(s). Second, the user must find potentially relevant documents within the repository, usually through a combination of searching, navigating inter-document links, and browsing. Third, the user must locate relevant information *within* these documents. This paper is concerned with the latter challenge, which is becoming increasingly important as longer documents are published, and distributed, using Web and other technologies. Various approaches have been proposed for within-document retrieval, including passage retrieval (Kaszkiel and Zobel 1997), and user interfaces supporting content-based browsing of documents (Hearst 1995). We have proposed a tool, ProfileSkim (previously known as SmartSkim),

for within-document retrieval based on the concept of relevance profiling (Harper et al. 2002). In that paper we provided a comprehensive overview of approaches and interfaces for within-document retrieval. Subsequently, we reported on a user-centred evaluation of the ProfileSkim tool, in which a preliminary analysis of quantitative performance data was presented (Harper et al. 2003). In this paper, we extend that quantitative data analysis, and include a comprehensive analysis of qualitative data derived from questionnaires filled in by the participants during the evaluation study.

We have designed, developed and implemented a tool called ProfileSkim, which enables users to identify, efficiently and effectively, *relevant passages* of text within *long* documents. The tool integrates passage retrieval and content-based document browsing. The key concept underpinning the tool is relevance profiling, in which a profile of retrieval status values is computed across a document in response to a query. Within the user interface, an interactive bar graph provides an overview of this profile, and through interaction with the graph the user can select and browse *in situ* potentially relevant passages within the document.

The evaluation study reported herein was devised to test key assumptions underlying the design of the ProfileSkim tool, namely:

- That relevance profiling, as implemented and presented by the tool, is effective in assisting users in identifying relevant passages of a document;
- That by using the tool, users will be able to select and browse relevant passages more efficiently, because only the best matching passages need be explored;
- That users will find the tool satisfying to use for within-document retrieval, because of the overview provided by relevance profiling.

We report experimental results in support of these three assumptions, based on quantitative and qualitative data collected in the user study. The user study involved a comparative evaluation of ProfileSkim with another within-document retrieval tool, FindSkim, which provides similar functionality to the well-known Find-command delivered with most text processing and browsing applications. We investigated the tools within a simulated work task situation (Borlund and Ingwersen 1997), in which the participants in the study were asked to compile (part of) a subject index for a book. Within this setting, we evaluated the comparative effectiveness and efficiency of the tools, and assessed user satisfaction.

This study methodology is based on an evaluation approach that is beginning to emerge through the efforts of those involved in the 'interactive track' of TREC (Beaulieu et al. 1997), through end user experiments in the Information Retrieval community (Borlund and Ingwersen 1997, Hersh et al. 1996, Jose et al. 1998), and through the effort of groups such as the EC Working Group on the evaluation of Multimedia Information Retrieval Applications (Mira) (Dunlop 1997). Major elements of the approach are:

- The observation of 'real' users engaged in the performance of 'real-life' tasks (or, at least, convincing simulations of such tasks);
- A range of performance criteria are used, pertaining both to quantitative aspects of task performance (efficiency and effectiveness), and qualitative aspects of the user experience;
- A range of methods for acquiring and analysis of data are used, which can be quantitative in nature (e.g., time for task), and qualitative in nature (e.g. attitudes and reactions to the system, the task, etc.).

The paper is structured as follows. In Section 2, we provide an overview of relevance profiling, and describe how language modelling can be used as a basis for this. An overview is provided in Section 3 of the salient features of the two within-document retrieval tools used in the study. The research questions are presented in Section 4, and the experimental methods in Section 5. In Section 6, we present the results of the experimental study, and these are discussed in Section 7. Finally, we offer some concluding remarks concerning the efficacy of relevance profiling as a basis for within-document retrieval, highlight the advantages of our particular approach for evaluating this type of retrieval tool, and suggest some future directions for research.

2. Overview of relevance profiling based on language modelling

Relevance profiling using language modelling was introduced in Harper et al. (2002), and we provide a brief overview here. Based on a query, we want to compute a relevance profile across the document, and presented this profile to the user in the form of a bar graph. By interacting with this bar graph, the user can identify, and navigate to, relevant sections of a document. Effectively, a retrieval status value (RSV) is computed for each word position in the document. This RSV will be based on a text window (fixed number of consecutive words) associated with each word position. Language modelling is used to construct a statistical model for a text window, and based on this model we compute the window RSV as the probability of generating a query.

We employ the language modelling approach proposed for document retrieval (Ponte and Croft 1998, Song and Croft 1999), and adapt it for relevance profiling. We model the distribution of terms (actually stemmed words) over a text window, as a mixture of the text window and document term distributions as follows:

$$P(query \mid window) = \prod_{t_i \in \text{query}} p_{\text{mix}}(t_i \mid win) \tag{1}$$

where

$$p_{\text{mix}}(t_i \mid win) = w_{\text{win}}\, p_{\text{win}}(t_i \mid win) + (1 - w_{\text{win}}) p_{\text{doc}}(t_i \mid doc) \tag{2}$$

Thus, the probability of generating words is determined in part by the text window, and in part by the document in which the window is located. The estimates are smoothed by the document word statistics using the mixing parameter, w_{win}. The best value for this parameter needs to be determined empirically, and we have used 0.8 in our system. The individual word probabilities are estimated in the obvious way using maximum likelihood estimators:

$$p_{\text{win}}(t_i \mid win) = n_{iW}/n_W \quad p_{\text{doc}}(t_i \mid doc) = n_{iD}/n_D \tag{3}$$

where $n_{iW}(n_{iD})$ and $n_W(n_D)$ are the number of word occurrences of word i in the window (document), and total word occurrences in the window (document) respectively.

The relevance profile is given by the retrieval status value at each word position i:

$$RSV_{\text{window}}(i) = P(query \mid window_i) \tag{4}$$

where text window i is the sequence of words $[w_i \ldots w_i + L_W - 1]$, and L_W is the fixed length of each text window.

In order to provide a plot of the relevance profile, and to support direct navigation to relevant parts of a document, retrieval status values are aggregated over fixed size, non-overlapping sections of text we call text tiles. We assume that the document text is divided into fixed length, non-overlapping text tiles. Let us assume that each tile is L_T words long. The aggregate RSV for a given tile j is given by:

$$RSV_{\text{tile}}(j) = aggfun(\{RSV_{\text{window}}(i),\ i = (j-1)L_T + 1 \ldots jL_T\}) \tag{5}$$

Examples of aggregate functions (*aggfun*) include average, minimum and maximum, and we opt for the maximum as this corresponds to the best text window starting within the tile. Note that some text windows will extend beyond the end of a tile.

Text windows and text tiles, although related, serve two different purposes. A text window is used to compute an RSV at each word position in the document. The fixed size of a text window is set to the "typical" size of a meaningful chunk of text, such as the average size of a paragraph (or possibly section). The average size of a paragraph can be determined empirically, and in our system we have set it to 200 words. A text tile is used to aggregate or combine the RSVs of all text windows that start within the given tile, and tiles are used for summarizing (and thence displaying) relevance profiles. The size of a fixed tile is computed based on the length of the document, and depends on the number of tiles, and hence bars, we wish to display in the relevance profile meter. The heights of the bars in the profile meter are proportional to the tile RSV, and are based on the logarithm of the tile RSV (Harper et al. 2002).

3. The document skimming tools

Two within-document retrieval tools are used in the comparative user evaluation. One, ProfileSkim, is based on relevance profiling, and the other, FindSkim, is based on the ubiquitous Find-Command provided within most word processing and web browser applications. FindSkim will be described first, as much of its functionality is common to both tools. Then, ProfileSkim is described.

3.1. The FindSkim tool

The FindSkim tool is based on the Find-command, although in many respects it provides additional functionality. A screenshot of the tool is illustrated in figure 1.

A user selects a file to skim, using the file chooser, and the file is displayed in a scrollable panel. Given a query, the tool highlights all query word variants that appear in the document in cyan. The document is positioned in the display panel at the first word occurrence, which

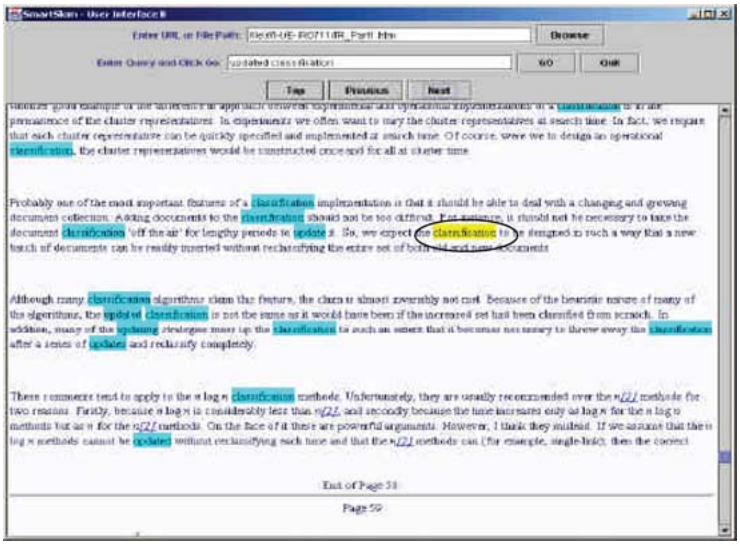

Figure 1. Screen shot for FindSkim tool.

becomes the current word. The current word is always highlighted in yellow (circled in figure 1). The user can navigate from the current word to the next (or previous) query word occurrence in the document using the Next/Find buttons. Query words which are not present in the document are flagged as possible misspellings, and the user may choose to edit the query, if appropriate.

Note, that the query is treated as a "bag of words". Hence, no phrase matching is performed based on query word order.

3.2. *The ProfileSkim tool*

The ProfileSkim tool is based on relevance profiling, and displays an interactive analogue of the relevance profile for a given query, in the form on a bar graph. A screenshot of the tool is illustrated in figure 2.

File selection and query input are identical to the FindSkim tool. Query term variants are also highlighted in cyan, and the document is displayed in a scrollable panel.

Based on a query input by the user, a relevance profile is computed over the document (see Section 2), and presented in the form of an interactive bar graph. Each bar corresponds to a fixed length section (tile) in the text of the document, with the leftmost bar corresponding to the start of the document, and the rightmost bar to the end of the document. The height of a bar corresponds to the computed retrieval status value of the corresponding tile. By clicking on a bar, the corresponding tile within the document is centred in the document viewer. Effectively, the bars of the relevance profile meter act as "hyperlinks" into the body of the document.

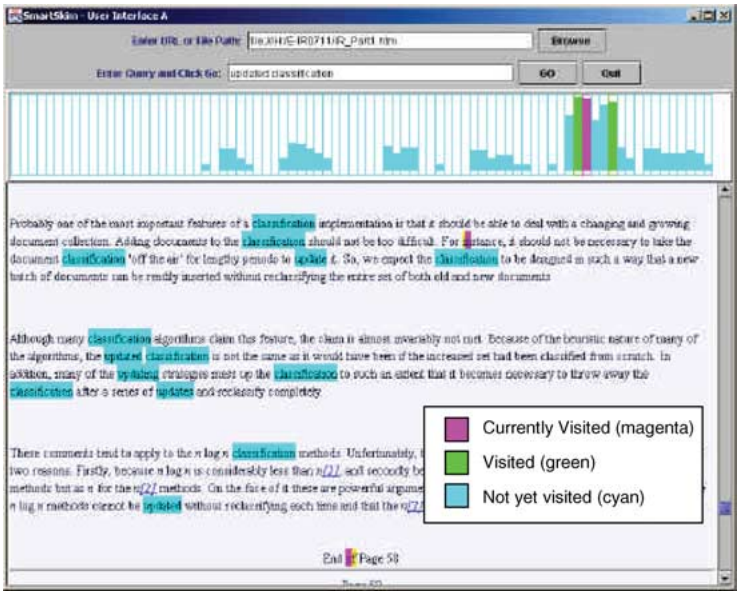

Figure 2. Screen shot for ProfileSkim tool.

To assist the user in browsing the document using the relevance profile meter, feedback is provided as to which bars (and corresponding tiles) had been visited. Colour coding of the bars indicates which bar/tile has: yet to be visited (cyan), currently being visited (magenta) and visited (green). This colour-coding scheme reinforces the view that the bars acts as hyperlinks, and the colours used correspond broadly to those used typically when browsing web pages. The currently visited tile is also indicated with a pair of yellow/magenta and magenta/yellow "brackets" on the document display.

A critique of the ProfileSkim interface using the Cognitive Dimensions Framework (Green 1991) is provided in Harper et al. (2002).

3.3. Choice of skimming tools

In setting up the comparative user evaluation of ProfileSkim, we gave careful thought to the choice of the other skimming tool.

We opted for a tool based on the Find-command for three reasons. First, the Find-command is the de facto standard for document skimming, albeit in a number of guises in word processing applications and web browsers. Relevance profiling is a possible alternative to the Find-function, and it is therefore useful to provide comparative performance data. Second, we wanted to measure the relative performance of ProfileSkim against FindSkim to provide a benchmark for future developments of ProfileSkim itself. Third, developing our own Find-command variation might suggest ways of improving the Find-command itself.

We accept that the functionality of the tools is different, and in particular that additional information is made available to the users through the relevance profiling tool. However, we thought is best to establish the comparative performance of ProfileSkim against a de facto standard in the first instance, and investigate possible variants of relevance profiling tools at a later stage.

4. Research questions and hypotheses

In general terms, we wanted to investigate whether within-document retrieval based on relevance profiling was more efficient in user time, and more effective in identifying relevant sections of long documents, than the competing tool based on functionality similar to the Find-command. Beyond that, we wanted to measure user satisfaction in relation to using the competing within-document retrieval tools. The interactive experiment was designed to test user efficiency, user effectiveness and user satisfaction in performing the book indexing task. The effectiveness measures are described in Section 5.4, as are questionnaires used to capture user satisfaction.

More formally, a number of hypotheses and conjectures were formulated, based on the expected performance of, and user satisfaction with, ProfileSkim and FindSkim. These are, with justifications:

Hypothesis HT: The 'time to complete' an indexing task is less using ProfileSkim compared with FindSkim.

We expected that the relevance profile meter would enable the user to readily identify relevant sections of the text, and not spend time browsing less relevant sections.

Hypothesis HP: ProfileSkim is more effective than FindSkim as measured by Precision.

Hypothesis **HP** is based on the observation that ProfileSkim encourages a user to explore the highest peaks of the relevance profile (potential relevance hotspots), and thus we might expect a user to achieve higher precision when using ProfileSkim.

Hypothesis HR: FindSkim is more effective than ProfileSkim as measured by Recall.

Hypothesis **HR** is based on the observation that FindSkim encourages a user to visit all query word occurrences in the text and thus we might expect a user to achieve higher recall, and this possibly at the expense of precision. However, it is possible that ProfileSkim might achieve comparable levels of recall, depending on the extent to which a user is prepared to explore comprehensively the relevance profile.

Conjecture CF: Supposing that hypotheses **HP** and **HR** hold, then we conjecture that effectiveness, as measured by the combined F-measure, will be comparable.

This conjecture is simply a consequence of the fact that the F-measure "trades off" precision against recall.

Conjecture CU: Users will be more satisfied when using ProfileSkim compared with FindSkim.

This conjecture is based on the hypothesised efficiency of ProfileSkim compared with FindSkim, and to a lesser extent the hypothesised Precision differential. Further, we believe that the visual "model of relevance" presented by ProfileSkim will provide a better starting point for the book indexing task.

As a side effect of the study, we are also interested in establishing whether the book indexing experiment is suitable for evaluating within-document retrieval, and specifically whether users were indeed able to perform the indexing tasks satisfactorily. We will investigate a number of sources of evidence including the quantitative measures of task efficiency and effectiveness, and user reactions to the tasks.

5. Methods

In this evaluation of within-document retrieval using relevance profiling, and specifically the comparative evaluation of ProfileSkim and FindSkim, we wanted to address the following issues:

- the participants in the experiment should be placed in a simulated work task situation (Borlund and Ingwersen 1997), such that document skimming is central in performing the task;
- the focus of the task should be document skimming, and not document retrieval;
- the documents used in the study should be long, in order to provide a realistic assessment of the tools being studied;
- the tasks should be realistic, understandable to the participants, and able to be completed in a reasonable time; and
- task performance can be measured against some ground truth established for the task.

A novel work task situation was devised that satisfied our requirements, namely creating a subject index for an electronic book.

5.1. Participants

The participants for the study were all graduate students drawn from various places in our University. We would have preferred to select from a homogeneous group, but this was not possible given that the experiment was performed with 24 participants (plus 6 additional participants for the pilot). Instead, we selected from a number of programmes, namely students in: MSc Information and Library Studies (10), MSc Knowledge Management (7), MSc Electronic Information Management (2), PhD in Business Studies (1) and PhD in Computing (4). Summary statistics about the users, based on an entry questionnaire, are presented in Section 6.1.

5.2. *Instruments*

Collection. An electronic version of van Rijsbergen's classic information retrieval text was obtained, and we added page numbers which are necessary in creating a subject index. The book was divided into four sections, two sections for training and two for the main experiment (see Table 1).

Topics. Eight topics[1] were selected at random from the subject index provided with the original textbook (see Table 2). The selected topics met the following criteria:

- between 4 and 7 pages indexed for the topic;
- at least two distinct ranges of page numbers;
- two or more words for the topic;
- (preferably) indexed pages present in both Part 1 and Part 2 of the text; and
- (as far as possible) no overlap between the sets of relevant pages for the different topics.

Table 1. Collection details.

Filename	Content	No of pages	Word count
Training1	Chapter 4	29	9526
Training2	Chapter 7	40	13181
Part1	Chapter 2, 3	52	18087
Part2	Chapter 5, 6	49	17296

Table 2. Indexing task groups.

Task group	Order	Topic/Subject	File to skim	Indexed pages
1	Training	Expected search length	Training1	none
			Training2	160–163
	First	Loss (or cost) function	Part1	29
			Part2	116–117, 126
	Second	Boolean search	Part1	none
			Part2	95–97, 109
	Third	Information measure	Part1	41–42, 57
			Part2	123, 136, 138
2	Training	Relational data model	Training1	67, 90
			Training2	none
	First	Maximum spanning tree (MST)	Part1	56, 57
			Part2	123, 132, 139
	Second	Relevance feedback	Part1	none
			Part2	105–108, 112
	Third	Cluster based retrieval	Part1	47, 56
			Part2	103–105

These criteria ensured that the corresponding indexing tasks could be performed in a reasonable time, and that the participants would be required to browse comprehensively both parts of the book. We opted for multi-word topics for two reasons. First, it would have proved more difficult for the participants to judge page relevance for one-word topics. Second, we were interested in exploring the effect of relevance profiling on multi-word topics. The final criterion was included to try and minimize the learning effect of viewing many times the same, albeit, long document.

5.3. Procedures

Scenario for simulated work task. The participants were asked to imagine they were graduate students, who had been asked by their tutor to assist him/her in creating a subject index for a book he/she has written. For a given topic they were asked to locate pages that should appear under that topic, using one of the skimming tools. The criteria for including a page, i.e. assessing the page relevant for the topic, were:

- the page must be topically relevant, i.e. about the subject;
- the page must be substantially relevant, i.e. the page would add to a potential reader's understanding of the topic;
- all pages in a set of contiguous relevant pages should be included; and
- pages in the bibliographies at the ends of chapter were not to be indexed.

These instructions accorded in general with the way the book was originally indexed by the author (Private communication with C.J. van Rijsbergen).

Tasks and task groups. Each topic was the basis for an indexing task, and to assist the participants, a short definition was provided for each topic. This provided some context for evaluating the relevance of page to a topic, and plays a similar role to the extended topic descriptions in TREC-1 (Harman 1992). The topics were divided into two groups for the experiment, and we refer to these as Task Groups (see Table 2). Within each task group, the first task was used as a training task, and the other three tasks were arranged in increasing order of difficulty. This ordering was established based on a pilot study we performed.

Experiment design. The design is summarised in Table 3.

Table 3. Experiment design. ProfileSkim is 'A', and FindSkim is 'B'.

Participant group	First task set (System/task group)	Second task set (System/task group)
1	A/TG1	B/TG2
2	A/TG2	B/TG1
3	B/TG1	A/TG2
4	B/TG2	A/TG1

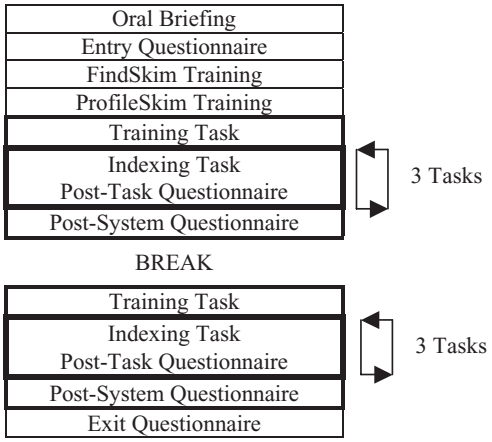

Figure 3. Procedure for experiment.

Experiment procedure. The procedure is summarised in figure 3.

The participants were asked to complete the indexing tasks as quickly as possible, while at the same time achieving good levels of indexing specificity and exhaustivity. The pilot study established that most tasks could be completed in 6–10 minutes, and thus we allocated 40 minutes for each task group. However, the participants were asked to complete all tasks in a group, even if they over-ran the allocated time. The majority of participants completed each task group within the 40 minutes.

A few observations are necessary regarding this procedure. We would have preferred to run the experiment with each participant individually. This was not possible due to timetabling and resource constraints. However, we minimised as far as possible interaction between the participants. We would have preferred to do the system training just prior to use of each system. This was not possible given the experiment was performed with participants from all participant groups (see Table 3). In mitigation, the training was mostly concerned with task training, as the systems were relatively easy to learn and use. Moreover, prior to using each system, there was a specific training task.

5.4. Measures

For each indexing task, allocated one at a time, the user was asked to record the page numbers of relevant pages they would include in the topic (subject) index. Using this information, we were able to assess the specificity and exhaustivity of the indexing, using traditional precision and recall measures (see below). The time for each task was recorded in minutes and seconds. Using this information, we were able to assess the user efficiency of the indexing process.

Precision, recall and the F-measure were computed as follows. The original subject index of the book provides the ground truth for the indexing tasks. That is, the pages indexed originally by the author of the book, are effectively the pages deemed relevant.

Hence for a given subject, if A is the set of pages indexed by the author and B is the set of pages indexed by a participant in the study, then precision (P) and recall (R) can be computed in the obvious way:

$$P = |A \cap B|/|B| \quad R = |A \cap B|/|A| \tag{6}$$

The F-measure (F), which is a single measure of performance, is simply the harmonic mean of precision and recall, namely:

$$F = 2PR/(P + R) \tag{7}$$

This measure effectively "values" precision and recall equally, and thus it enables us to trade off precision and recall.

Questionnaires. As indicated in figure 3, the participants completed a number of questionnaires throughout the study, and these are detailed in Appendix A. In summary, they are: entry questionnaire (1 per participant), post-task questionnaire (6 per participant, 1 for each indexing task), post-system questionnaire (2 per participant, 1 for each system), and exit questionnaire (1 per participant). These questionnaires were based on those of the TREC 2002 Interactive Track (TREC 2002). The questions, together with unique question identifiers we use throughout the paper, are given in Appendix A. Many of the questions require a response on a 7-point scale of attitude measurement. In the questionnaires, we elicit responses to these questions, and we note here the standard labels used for the scale points: 1 (not at all), 4 (somewhat), and 7 (extremely).

For each participant and system pair, the mean values of Time, Precision, Recall and F-measure were calculated by averaging over the three tasks within the associated task group. The responses for the post-task questions were averaged in the same way. These averaged values were used in the formal statistical comparison of the two systems reported below.

6. Experimental results

In this section, we present and analyse both quantitative data, derived from measuring indexing task efficiency and effectiveness, and qualitative data derived from the questionnaires filled in by the participants. In Section 6.1, we present some summary data about the users derived from the entry questionnaire, and provide some observations thereon. Quantitative data relating to task efficiency and effectiveness is presented in Section 6.2, and qualitative data from the questionnaires in Section 6.3.

6.1. Summary participant data

We present summary data on the participants, and offer some observations on this data to set the scene for reporting the main experimental results. This data is based on the entry questionnaire (see Appendix A and Section 5.4). The demographics of our group of participants are an equal number of males and females, and an approximately equal split

between those aged 25 and under, and those older than 25. This is broadly representative of the postgraduate population in our University.

The entry questionnaire captured data on the participants' familiarity with aspects of the book indexing task. In the following, all responses are on a 7-point scale and questions are uniquely identified using '**En**' labels as in Appendix A. The participants assessed themselves via the entry questionnaire as:

En4, En6 Highly familiar with MS-Word or similar (average: 6.29, median: 7.0), and slightly less familiar with Information Explorer (IE) or similar (avg: 5.71, med: 6.50);
En5, En7 Somewhat familiar with using the MS-Word Find/Edit command (avg: 4.63, med: 4.50), and less familiar with using the corresponding IE command (avg: 3.79, med: 4.0);
En12, En13 Somewhat experienced in using a book index (avg: 4.17, med: 4.0), and less experienced in manual indexing (avg: 3.22, med: 3.0);
En14 Medium level of expertise in the subject of the book, namely Information Retrieval (avg: 3.39, med: 4.0).

These results indicate that the book indexing task may have been quite challenging for the participants given their lack of expertise in manual indexing and their level of expertise on the subject matter of the book. In respect of reading and extracting information from electronic documents, the participants are quite familiar with typical "reading" tools (e.g. MS-Word, Information Explorer, etc.), and, on average, have 3.8 years of experience of reading electronic documents.

6.2. Quantitative analysis: Indexing task efficiency and effectiveness

In this section, we focus on the quantitative data, which are used to test the major hypotheses of the experimental study. Thus, we concentrate on the presentation and analysis of data relating to task efficiency, as measured by 'time for task', and task effectiveness, as measured by precision, recall, and F-measure. Note that, for each participant/system combination, the average value of each of these four variables was computed over the given task group prior to the analyses described below.

We note that the analyses which follow are fuller than those presented in Harper et al. (2003) insofar as the possible effects of factors such as the order in which the two sets of tasks were attempted are discussed.

For each of the variables, 'time for task', precision, recall, and F-measure, an analysis of variance (ANOVA) was performed to assess the significance of the four factors: 'system' (ProfileSkim or FindSkim), 'task group' (1 or 2), 'order' (first or second), and 'participants'. Summary statistics are presented in Table 4. The results from the ANOVAs are summarised in Table 5. Due to the nature of the experimental design, these ANOVAs had to be conducted under the assumption that there are no interactions between the factors. However, further analyses, including two-way ANOVAs, were conducted on subgroups of the data to establish whether relationships existed between factors, and any significant findings from these analyses are reported where relevant.

Table 4. Summary statistics for comparisons between ProfileSkim and FindSkim.

	Mean (standard deviation)	
	ProfileSkim	FindSkim
Time	5.808 (1.567)	7.744 (1.941)
Precision	0.622 (0.154)	0.550 (0.112)
Recall	0.739 (0.170)	0.687 (0.232)
F-measure	0.635 (0.133)	0.582 (0.150)

Table 5. P-values from ANOVAs.

	Time	Precision	Recall	F-measure
System	0.000	0.061	0.392	0.219
Task group	0.484	0.004	0.082	0.004
Order	0.003	0.760	0.507	0.689

The boxplots in figure 4 show the distributions of the measures for 'time for task', precision, recall and F-Measure, for the ProfileSkim tool and the FindSkim tool.

For the efficiency measure 'time for task', there is very strong evidence ($p = 0.000$) of a difference between the two systems. With the participants taking, on average, 1.94 minutes less to complete a set of tasks with ProfileSkim than with FindSkim, there is very strong evidence to support the hypothesis HT and conclude that:

The 'time to complete' an indexing task is less using ProfileSkim compared with FindSkim.

This difference is shown clearly on the boxplot for 'Time'.

With respect to 'time for task' there is also an interesting 'order' effect with participants taking on average, 7.54 minutes to complete the first set of tasks, and 6.01 minutes to complete the second set. This difference is significant at level $p = 0.003$ (Table 5).

In terms of precision, the ProfileSkim mean (0.622) is higher than the FindSkim mean (0.550) with $p = 0.061$ suggesting that there is weak evidence that ProfileSkim is more effective than FindSkim. The results shown in Table 5 also indicate a significant difference between the average levels of precision achieved for the two task groups. A more in depth analysis resulted in the means displayed in Table 6 and suggests that when ProfileSkim was

Table 6. Mean precision and F-measure by 'system' and 'task group'.

		Task Group 1	Task Group 2
Precision	ProfileSkim	0.712	0.533
	FindSkim	0.578	0.523
F-measure	ProfileSkim	0.737	0.534
	FindSkim	0.615	0.549

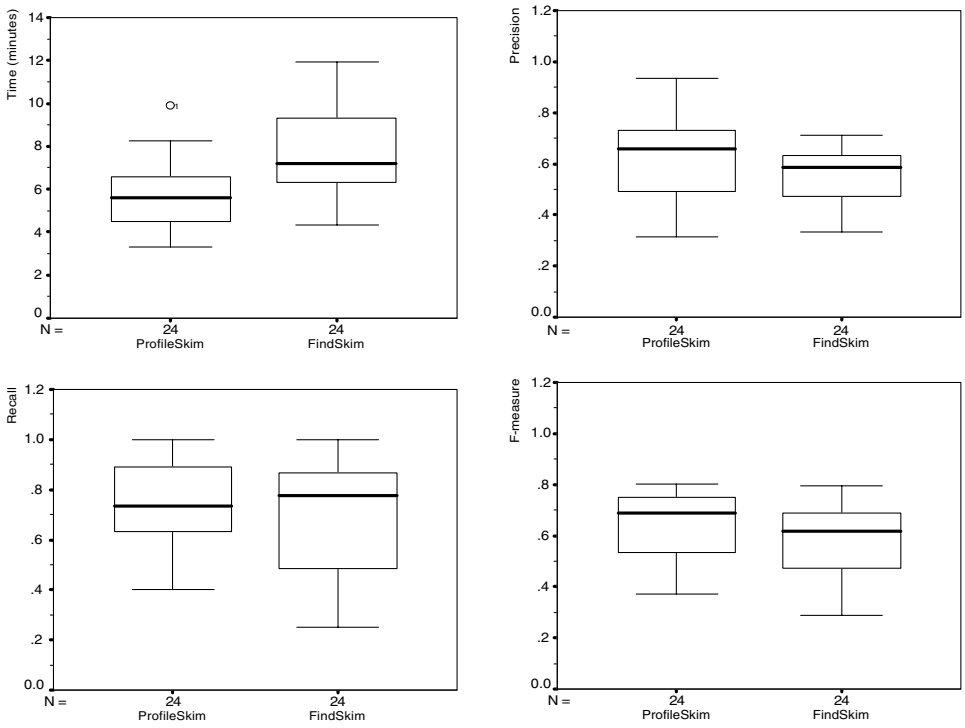

Figure 4. Boxplots for time, precision, recall and F-measure for ProfileSkim and FindSkim.

applied to Task Group 1 the average level of precision was significantly higher than with any other combination of 'system' and 'task group'. However, on the basis of our results, it would be unwise to infer that, in general, ProfileSkim is more effective than FindSkim in terms of mean precision.

With respect to recall, there is no evidence of a significant difference between the performance of the two systems.

The results presented in Table 5 suggest that the average values of the F-measure differ only with respect to the two task groups. However, the results of further analysis, together with the relevant means reported in Table 6, suggest that when ProfileSkim was applied to Task Group 1 the average F-measure achieved was considerably higher than with any other combination of 'system' and 'task group'. As the precision is used in the computation of the F-measure, this is not entirely unexpected. In keeping with our findings with respect to precision, we cannot infer that, in general, ProfileSkim performs better than FindSkim in terms of the F-measure.

We note for each of the three measures of effectiveness, the ProfileSkim sample mean is greater than the FindSkim sample mean. It seems reasonable to infer that ProfileSkim is at least as effective as FindSkim.

Table 7. Result from statistical analysis of post-task questionnaires for ProfileSkim (A) and FindSkim (B).

	Mean		Median		Min		Max		
Question abstract	A	B	A	B	A	B	A	B	*P*-value
PT1. Easy to index the topic?	4.78	4.03	5.00	3.67	2.33	2.33	7.00	6.00	0.003
PT2. Satisfied with results?	4.78	4.04	5.00	3.83	1.33	1.67	7.00	6.67	0.002
PT3. Previous knowledge?	3.55	3.51	3.67	2.83	1.00	1.00	6.50	6.67	0.911

6.3. Qualitative analysis: Participant satisfaction from questionnaire data

Here we present qualitative results derived from the questionnaires filled in by the participants. See Appendix A for details of all questionnaires.

In Table 7, the average and median of the responses (7-point scale) are given for each system for the questions from the post-task (**PT**) questionnaire. Recall that the responses for an individual question for the set of three indexing tasks are averaged for a participant/system combination. The systems are then compared on the basis of these averaged figures per participant using the non-parametric Wilcoxon Signed Ranks test. The results show that from the participants' perspective:

PT1 It is easier to perform topic indexing using ProfileSkim (avg: 4.78, med: 5.00) than using FindSkim (avg: 4.03, med: 3.67), and this is significant at the level $p = 0.003$;

PT2 Users are more satisfied with the indexing results when using ProfileSkim (avg: 4.78, med: 5.00) than when using FindSkim (avg: 4.04, med: 3.84), and this is significant at $p = 0.002$; and

PT3 Users rated previous knowledge of the topic as not that helpful in performing the indexing task for both ProfileSkim (avg: 3.55, med: 3.67) and FindSkim (avg: 3.51, med: 2.84) which is not significant.

In Table 8, the average and median of the responses (7-point scale) are given for each system for five questions of the post-system (**PS**) questionnaire. The systems are compared for

Table 8. Result from statistical analysis of post-system questionnaires for ProfileSkim (A) and FindSkim (B).

	Mean		Median		Min		Max		
Question abstract	A	B	A	B	A	B	A	B	*P*-value
PS1. Easy to learn?	5.70	5.71	6	6	2	3	7	7	0.926
PS2. Easy to use to index?	5.57	4.13	5	3.5	3	2	7	7	0.001
PS3. Understand to use?	5.30	5.46	6	6	2	3	7	7	0.517
PS4. Accuracy of Index?	4.83	4.21	5	4	2	2	7	6	0.010
PS5. Completeness of Index?	4.48	3.96	5	4	2	1	7	6	0.041

each question on the basis of the individual participant responses using the non-parametric Wilcoxon Signed Ranks test.

In relation to questions on "ease of learning" and "understanding how to use" (**PS1** and **PS3**), both systems were assessed at the median value of 6.0, and there was no discernible difference between the systems.

From the participants' perspective, ProfileSkim was better than FindSkim as follows:

PS2 For the indexing task, it was easier to use ProfileSkim (avg: 5.57, med: 5.00) than FindSkim (avg: 4.13, med: 3.50), and this is significant at the level $p = 0.001$;

PS4 The accuracy[2] of the index entries was assessed by the participants as more accurate when using ProfileSkim (avg: 4.83, med: 5.00) than FindSkim (avg: 4.21, med: 4.00), and this is significant at $p = 0.01$; and

PS5 The completeness[2] of the index entries was assessed by the participants as more complete when using ProfileSkim (avg: 4.48, med: 5.00) than FindSkim (avg: 3.96, med: 4.00), and this is significant at $p = 0.041$.

Here we present the summary results from the exit questionnaire that is completed after both systems have been used. Note, one participant did not fill in an exit questionnaire, and another omitted question **EX5**. The participants judged (on a 7-point scale) that they:

EX1 Understood the nature of the indexing task at a very high level (avg: 5.65, med: 6.00); and

EX3 Found the systems were different at a high level (avg: 5.05, med: 5.00).

In respect of the perceived difference between the systems, further questions were asked comparing the systems (**EX4**, **EX5** and **EX6**), and summary results are given in Table 9. Responses to these questions could be: preferred ProfileSkim (response A), preferred FindSkim (response B), or neither preferred (response NoDiff). For each question, the Sign Test, ignoring NoDiffs, was used to determine whether the number of users who preferred ProfileSkim is significantly different from the number who preferred FindSkim. For these three questions, ProfileSkim was adjudged better than FindSkim as follows:

EX4 ProfileSkim easier to learn to use than FindSkim (A: 13, B: 4, NoDiff: 6), and this is significant at $p = 0.049$ and not significant at $p = 0.093$ (using a more conservative Sign Test that distributes NoDiffs equally between A and B);

EX5 ProfileSkim easier to use than FindSkim (A: 17, B: 4, NoDiff: 1), and this is significant at $p = 0.007$; and

Table 9. Result from statistical analysis of exit questionnaire.

Questions abstract	ProfileSkim	FindSkim	No difference	*P*-value (Sign test)
Ex4. Easier to learn to use?	13	4	6	0.0490 (ignore NoDiff) 0.0931 (distrib. NoDiff)
Ex5. Easier to use to index?	17	4	1	0.0072
Ex6. Best overall?	20	3	0	0.0005

EX6 ProfileSkim was preferred overall to FindSkim (A: 20, B: 3, NoDiff: 0), and this is significant at $p = 0.0005$.

Given the significance tests for **EX4**, it would not be safe to conclude that ProfileSkim was easier to learn to use than FindSkim. Indeed, we anticipated that ProfileSkim would be slightly more difficult to learn, and now speculate whether the participants conflated 'ease of learning' and 'ease of using' in their response.

7. Discussion of results

In this section, we discuss the results presented in Section 6. We discuss the quantitative results in Section 7.1, and the qualitative results in Section 7.2. We consider evidence on the suitability of the book experiment itself in Section 7.3. In Section 7.4, we consider whether our experimental results can be generalised and thus apply to other within-document retrieval task settings.

7.1. Quantitative data: Indexing task efficiency and effectiveness

Our results provide very strong evidence that relevance profiling, as presented and implemented in ProfileSkim, is more efficient than the FindSkim for the book indexing task. In relation to hypothesis **HT**, there is very strong evidence that:

The 'time to complete' an indexing task is less using ProfileSkim compared with FindSkim.

The ProfileSkim mean 'time for task' of 5.81 minutes is 25% lower than the FindSkim mean of 7.74 minutes.

The analysis of 'time for task' also shows a significant 'order' effect with the second set of tasks being completed in, on average, 1.53 minutes less time than the first set of tasks. This 'order' effect is similar for both systems and is not accompanied by any reduction in the accuracy of the indexing performed by the participants. We believe that these results indicate the presence of a learning effect as participants became more familiar with the type of indexing task to be performed. We note that a balanced experimental design is important for highlighting such (possible) effects.

In respect of effectiveness as measured by precision, and the hypothesis **HP**, although there is weak evidence for the statement "*ProfileSkim is more effective than FindSkim as measured by Precision*" we are not able to firmly conclude that this hypothesis is true.

In relation to the hypothesis **HR** concerning recall, there is no evidence for a difference between FindSkim and ProfileSkim, and we cannot conclude that the statement "*FindSkim is more effective than ProfileSkim as measured by Recall*" is true.

Technically, the conjecture **CF** is not supported, given that hypotheses **HP** and **HR** are not supported. However, on the basis of the results presented it seems entirely reasonable to conclude, without caveat, that:

ProfileSkim is at least as effective as FindSkim as measured by the F-measure.

We are greatly encouraged by the fact that for all three measures of effectiveness, the sample mean precision, recall and F-measure, are all higher for ProfileSkim than for FindSkim.

Returning to the precision results, there is weak evidence that ProfileSkim may be more effective than FindSkim. Interestingly, for the particular combination of ProfileSkim and Task Group 1, the average level of precision is significantly higher than for any other combination of 'system' and 'task group', and in the case of the F-measure considerably higher. We posit that the topics in Task Group 1 are more specific than those of Task Group 2, and particularly for the third topics (see Table 2). It may simply be easier to perform the indexing for the topics in Task Group 1. Nevertheless, for these topics, ProfileSkim is better than FindSkim in respect of precision and the F-measure. Tentatively, we might conclude that, for specific-type topics, ProfileSkim is more effective than FindSkim. Given this discussion, there are grounds for suggesting that ProfileSkim may be a precision-enhancing device, although further experimentation would be required to provide clear evidence for this conclusion.

In summary, ProfileSkim has been shown to be more efficient than FindSkim for the book indexing task, and moreover that ProfileSkim was at least as effective as FindSkim. Given the central role of within-document retrieval in the task, we have demonstrated the efficacy of relevance profiling for within-document retrieval.

7.2. Qualitative data: Participant satisfaction from questionnaire data

We conjectured that the participants would be more satisfied when using ProfileSkim than FindSkim for the book indexing task. The results presented in Section 6.3 show there is considerable evidence in support of this conjecture. ProfileSkim was judged by the participants to be better than FindSkim in relation to the following questions pertaining to ease of use and learning, perceived task performance and overall satisfaction, and in all cases significantly so. We simply list the questions (see Section 6.3 for details):

Post-task PT1	Ease of indexing on the topic
Post-task PT2	Satisfied with indexing results
Post-system PS2	Ease of indexing using the system
Post-system PS4	(Perceived) accuracy of indexing
Post-system PS5	(Perceived) completeness of indexing
Exit Ex5	Easier to use
Exit Ex6	Best overall

Perhaps most interestingly, topics were perceived to be easier to index when using ProfileSkim compared with FindSkim (**Post-task PT1**), strongly suggesting the suitability of ProfileSkim for the indexing task, and this was reinforced by responses to **Post-system PS2**.

There was no significant difference between the systems in relation to: ease of learning to use; and understanding how to use the system. In absolute terms, both systems achieved very high ratings on these aspects (**PS1** and **PS3**).

On most measures of user satisfaction, and arguably the most critical ones for task execution, there is considerable evidence for conjecture **CU**:

Users will be more satisfied when using ProfileSkim compared with FindSkim.

The narrative data collected via the questionnaires (e.g. **PS6**, **PS7**, **EX7**, **EX8** and **EX9**) provides a rich source of data on participants' reactions to both systems, and we can only highlight a few points here. Of the 24 participants, 10 stated explicitly in response to **EX7** that they liked the relevance profile meter (referred to almost universally as 'the bars'). Two representative comments by the participants illustrate the perceived value of the relevance profile meter: "gave a good picture of document content in terms, of keywords" (**PS7**) and "could see clusters of potentially relevant pages" (**EX7**). We may conclude from these, and many similar comments, that users liked the overview provided through relevance profiling, in that it provides a visual model of relevance across a document for a query.

7.3. Evidence on suitability of book indexing task for experiment

Here we consider whether the book indexing experiment is suitable for evaluating within-document retrieval, and specifically whether users were indeed able to perform the indexing tasks satisfactorily. We draw on both the quantitative and qualitative data analyses presented in Sections 6.2 and 6.3. This includes quantitative measures of task efficiency and effectiveness, and user perceptions on task difficulty and task achievement.

In absolute terms, the average time taken to complete the indexing task was less than the ten minutes allocated to each task, with sample means of 5.81 and 7.74 for ProfileSkim and FindSkim respectively. The average levels of Precision (P) and Recall (R) achieved were relatively high; for ProfileSkim ($P = 0.62$, $R = 0.74$) and FindSkim ($P = 0.55$, $R = 0.69$). This was achieved despite the relative lack of experience of the participants in book indexing (manual indexing), and the relative lack of knowledge of subject area of the book, namely Information Retrieval (see Section 6.1 for details).

In terms of participant satisfaction, the following results indicate user satisfaction with the book indexing task:

PT1 Ease of indexing was judged 'high' for ProfileSkim (avg: 4.78, med: 5.00), and somewhat (easy) for FindSkim (avg: 4.03, med: 3.67);
PT2 Satisfaction with the indexing was judged 'high' for ProfileSkim (avg: 4.78, med: 5.00), and somewhat (less satisfied) for FindSkim (avg: 4.04, med: 3.84);
EX1 Extent of understanding of the indexing task was judged as 'very high' (avg: 5.65, med: 6).

Taken together, there is strong evidence that the participants understood the book indexing task, and were able to achieve satisfactory results in the time available. Task achievability is clearly important when we are trying to assess comparative performance of competing systems, and provides a firm foundation in this case for comparing ProfileSkim and FindSkim.

7.4. Overall discussion of results

Given these results, what can we conclude about the efficacy of ProfileSkim, and by implication relevance profiling, for more general within-document retrieval tasks. That is, to what extent will these results on efficiency, effectiveness and user satisfaction carry over

into other task settings and situations? The experimental task required the participants to locate relevant pages of long documents given a topic. Within-document retrieval is clearly central to this task.

ProfileSkim proved significantly more efficient compared with FindSkim for the experimental tasks, and we believe that it is likely to be equally efficient in more general within-document retrieval settings. Relevance profiling could be usefully provided within word processing applications and document reading/browsing tools as a replacement for the commonly provided "Find" functionality.

ProfileSkim was shown to be at least as effective as FindSkim in our experiment. Suppose ProfileSkim were to be used in a general web search setting (say), and that long documents were being retrieved. That is, ProfileSkim was being used to browse (long) documents returned by a search engine, based initially at least on the submitted query. The within-document retrieval effectiveness will depend largely on the nature of the queries (topics), and we must therefore ask how typical of web queries are the topics used in the experiment? The topics were short (2–4) words and this is broadly typical of web queries (Jansen et al. 2000). However, the experiment topics were almost all quite specific, and generally in the form of phrases. Web queries are certainly more varied that, and many have no phrasal structure. We might therefore expect ProfileSkim to perform well with specific-type queries (e.g. phrasal queries) in a web setting. Equally, ProfileSkim would be useful for exploring long documents for specific information after retrieval. Earlier, we presented some evidence that suggested ProfileSkim may be a precision-enhancing device. Relevance profiling may therefore be valuable in within-document retrieval tasks that require high precision, such as question-answering. ProfileSkim is able to accurately pinpoint relevant sections of large text documents, and to do so using relatively short queries. These are characteristic of many question-answering tasks. For more general queries, it may be that ProfileSkim would at least match the effectiveness of FindSkim (and Find-like commands), but this would have to be confirmed in the setting of a web retrieval experiment.

In terms of user satisfaction, and given the positive reactions of the participants to ProfileSkim, we believe the conclusions are likely to hold more generally. Although some questions were indexing task-specific, the centrality of within-document retrieval to this task means the conclusions have wider applicability. It is clearly important for user satisfaction that comparable levels of within-document efficiency and effectiveness are attained, and this is dealt with above. The overall strong preference for the participants in favour of ProfileSkim is highly likely to hold for similar relevance profiling tools in more general within-document retrievals settings.

The simulated work task situation we used in our experiment, namely the book indexing task, proved highly successful in many respects. Analysis of the questionnaire data shows that the scenario and tasks were understood by the participants, although admittedly the participants were all postgraduates. The participants were able to perform the tasks both efficiently and effectively, as evidenced by the performance analysis.

The book indexing task provides a ready-made ground truth, namely the original subject index. However, it would not always be straightforward to ascertain the original indexing policy, and incorporate this within the experiment setting. The subject matter of the book is critical, and we were fortunate that our participants were able to comprehend the relatively

technical material we used. The provision of both the topic and a longer definition proved important in enabling these participants to make the necessary relevance assessments. It may be that using more assessable materials, such as general-interest reference books, e.g., an encyclopaedia, would make the task simpler for participants drawn from a wider population.

8. Conclusions and future work

In this paper, we have reported the results of a user-centred evaluation of within-document retrieval tools, in the simulated task of providing (part of) the subject index of an electronic book. Two tools were compared, one based on relevance profiling (ProfileSkim), and one based on a sequential search (FindSkim).

The major findings of our investigation are that, for the book indexing task:

- The 'time to complete' the task was significantly less with ProfileSkim than with FindSkim;
- ProfileSkim was at least as effective as FindSkim, as measured using precision, recall and the F-measure; and
- The users (participants) were significantly more satisfied when using ProfileSkim compared with FindSkim, based on a wide range of measures of user satisfaction.

We argued that there is some justification for believing that these findings will hold in more general task settings, in which within-document retrieval may be useful. Further, relevance profiling should prove a worthy replacement for the familiar Find-Command implemented in most text processing and/or browsing applications.

Tentatively, we also conclude from our study that:

- The average effectiveness measures suggest that ProfileSkim may prove more effective than FindSkim, at least for within-document retrieval tasks involving specific-type topics; and
- ProfileSkim may be a precision-enhancing tool based on weak evidence provided by the experiment.

Based on the latter finding, FindSkim may be suited to high precision within-document retrieval tasks, and specifically might find a role in question-answering systems.

The book indexing task proved highly satisfactory for evaluating the comparative performance of within-document retrieval tools, and based on our experiences, we would advocate its use for this kind of study. Arguably, an experimenter might need to choose the subject matter of the books carefully, depending on the background of the study participants, and indeed the indexing task may prove too taxing for some.

For the future, we plan further user experiments in the more general setting of an interactive IR experiment, in which both document retrieval and within-document retrieval are critical. We are considering using the experimental framework provided by the TREC Interactive Track, and specifically the TREC-9 Interactive question answering task (Hersh and Over 2001). This would provide evidence as to the general applicability of the relevance profiling concept.

In the course of the experiment, we logged all user interactions for both ProfileSkim and FindSkim. We believe this data will provide valuable feedback on how the participants used the tools, and provide insights on possible enhancements. Based on our observations of the participants, we believe that some users were employing an "information foraging" strategy (Pirolli and Card 1999). That is, they explore the highest bars and peaks first, exhaust the information found there, and then move on to the next highest bar/peak. We would like to test this conjecture. Further, we would like to ascertain what the optimal strategy is for using ProfileSkim based on the logging data about user interactions and task effectiveness measurements.

Relevance profiling in ProfileSkim is based on a relatively simple mixture language model. This model favours term frequency over term discrimination. We would like to investigate other possible formulations of relevance profiling, based on more advanced divergence models, which we believe would allow term frequency to be combined with term discrimination c.f. tf.idf weighting. We would expect to evaluate alternative relevance profiling approaches using the book indexing approach, albeit in a batch environment, i.e. without user involvement, at least initially. These experiments would simulate ideal users based on the results of the user interaction study sketched above.

Appendix A: Questionnaires for the experiment

The following types of questions have been used:

Scale: 7 point Scale of attitude measurement, where 1 (not at all), 4 (somewhat) and 7 (extremely).
A/B: A choice among '*system A preferred*', '*system B preferred*' and '*No preference*'.
Open: Any response acceptable e.g., free comments.
Closed: Only a particular answer is acceptable. e.g., age, gender.

Entry questionnaire

ID	Question	Type
En1–En3	Personal details. Include name, age, gender, subject of study, etc.	Closed
En4	Experience in using MS-Word or similar software?	Scale
En5	Using Edit/Find command in MS-Word?	Scale
En6	Using web browser (e.g. IE, Netscape)?	Scale
En7	Using Edit/Find command in web browser?	Scale
En8	Reading electronic documents?	Scale
En9	Reading E-document with other systems (e.g. Acrobat Reader), please specify.	Scale
En10	When I read an E-document, I can usually find what I am looking for.	Scale
En11	Overall, for how many years have you been doing online reading?	Closed
En12	Please indicate how much experience do you have with the "book index":	Scale
En13	Please indicate how much experience do you have with the manual indexing:	Scale
En14	Please indicate your level of expertise with the subject Information Retrieval:	Scale

Post-task questionnaire

ID	Question	Type
PT1	Was it easy to do the indexing on this topic?	Scale
PT2	Are you satisfied with your indexing results?	Scale
PT3	Did your previous knowledge help you with your indexing?	Scale

Post-system questionnaire

ID	Question	Type
PS1	How easy was it to learn to use this system?	Scale
PS2	How easy was it to use this system to index?	Scale
PS3	How well did you understand how to use the system?	Scale
PS4	How accurate do you think your Index is?	Scale
PS5	How complete do you think your Index is?	Scale
PS6	Did you adopt any strategies when you use system X?	Open
PS7	Please write down any other comments that you have about your indexing experience with this system X.	Open

Exit questionnaire

ID	Question	Type
Ex1	To what extent did you understand the nature of the indexing task?	Scale
Ex2	To what extent did you find this task similar to other searching task that you typically perform?	Scale
Ex3	How different did you find the systems from one another?	Scale
Ex4	Which of the two systems did you find easier to learn to use?	A/B
Ex5	Which of the two systems did you find easier to use?	A/B
Ex6	Which of the two systems did you like the best overall?	A/B
Ex7	Was there anything in particular you liked about system A and system B?	Open
Ex8	Was there anything in particular you disliked about system A and system B?	Open
Ex9	Please list any other comments that you have about your overall index experience.	Open

Acknowledgments

This work was undertaken in the Smart Web Technologies Centre, which was established using funding (Gant number: HR01007) from the Scottish Higher Education Funding Council (SHEFC) under the Research Development Grant (RDG) programme.

We would like to thank C.J. "Keith" van Rijsbergen for permission to use the electronic version of his textbook "Information Retrieval" in our experiment. We are indebted to the following people for advice on experimental design and analysis: Diane Kelly, Alex Wilson, Anna Conniff, Stuart Watt, Ayse Goker, and Peter Lowit. We would also like to thank Robert Newton and Alan Maclennan for volunteering their MSc students for the study. Finally, we would like to offer our heartfelt gratitude to the participants who gave freely of their time, and strived so hard to find "relevance feedback" in Part 1 of the book! Finally, we would like to thank the anonymous referees for their very helpful comments.

Notes

1. Although we normally refer to 'subject indexing' and 'subjects' for books, we will adopt the standard IR terminology of 'topic indexing' and 'topic' in this paper.
2. When describing the book indexing task to the participants, we explained what was meant by 'accuracy' (indexing specificity) and 'completeness' (indexing exhaustivity).

References

Beaulieu M, Robertson SE and Rasmussen E (1996) Evaluating interactive systems in TREC. Journal of the American Society for Information Science, 47(1):85–94.

Borlund P and Ingwersen P (1997) The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, 53(3):225–250.

Dunlop M (1997), Ed. Proceedings of the second mira workshop, Technical Report TR-1997-2. Department of Computing Science, University of Glasgow. Available online at URL: http://www.dcs.gla.ac.uk/mira/workshops/padua_procs/ (visited May 15 2003).

Green TRG (1991) Describing information artefacts with cognitive dimensions and structure maps. In: Diaper D and Hammond NV, Eds., Proceedings of the HCI'91 Conference on People and Computers VI. Cambridge University Press, pp. 297–316.

Harman D (1992) Overview of the First Text REtrieval Conference (TREC-1), National Institute of Standards and Technology, Gaithersburg, Maryland, pp. 309–318.

Harper DJ, Coulthard S and Sun Y (2002) A language modelling approach to relevance profiling for document browsing. In: Proceedings of the Joint Conference on Digital Libraries, Oregon, USA, pp. 76–83.

Harper DJ, Koychev I and Sun Y (2003) Query-based document skimming: A user-centred evaluation of relevance profiling. In: Proceedings of 25th European Conference on Information Retrieval. Lecture Notes in Computer Science, Springer-Verlag, Berlin, pp. 377–392.

Hearst MA (1995) TileBars: Visualization of term distribution information in full text information access. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Denver, CO, pp. 65–71.

Hersh W and Over P (2001) TREC-9 interactive track report. In: Proceedings of the Ninth Text Retrieval Conference (TREC- 9), Gaithersburg, MD: NIST, pp. 42–50.

Hersh W, Pentecost J and Hickam D (1996) A task-oriented approach to information retrieval evaluation. Journal of the American Society for Information Science, 47(1):50–56.

Jansen B, Spink A and Saracevic T (2000) Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management, 36:207–227.

Jose J, Furner J and Harper DJ (1998) Spatial querying for image retrieval: A user-oriented evaluation. In: Proceedings of the 21st International ACM-SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 232–240.

Kaszkiel M and Zobel J (1997) Passage retrieval revisited. In: Proceedings of the Twentieth International ACM-SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, ACM Press, pp. 178–185.

Pirolli P and Card SK (1999) Information foraging. Psychological Review, 106:643–675.

Ponte J and Croft WB (1998) A language modeling approach to information retrieval. In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, pp. 275–281.

Song F and Croft WB (1999) A general language model for information retrieval. In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 279–280.

TREC (2002) Interactive track home page. Available at http://www-nlpir.nist.gov/projects/t11i/ (visited May 15 2003).