



ScentIndex and ScentHighlights: productive reading techniques for conceptually reorganizing subject indexes and highlighting passages*

Ed H. Chi¹
Lichan Hong¹
Julie Heiser^{1,2}
Stuart K. Card¹
Michelle Gumbrecht^{1,3}

¹Palo Alto Research Center, User Interface Research, 3333 Coyote Hill Road, Palo Alto, CA 94304, U.S.A.; ²Work done while at PARC, current address: Adobe Systems, 321 Park Ave., San Jose, CA 95110, U.S.A.; ³Work done while at PARC, Current address: Department of Psychology, Jordan Hall, Bldg. 420, Room 316, Stanford University, Stanford, CA 94305-2130, U.S.A.

Correspondence:
Ed H. Chi, Palo Alto Research Center,
User Interface Research, 3333 Coyote
Hill Road, Palo Alto, CA 94304, U.S.A.
Tel: +1 650 812 4312;
fax: +1 650 812 4258;
E-mails: echi@parc.com,
hong@parc.com,
julie.heiser@adobe.com,
card@parc.com,
mgumbrec@psych.stanford.edu

*This paper (or a similar version) is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. Based on 'ScentIndex: Conceptually Reorganizing Subject Indexes for Reading', by Ed H Chi, Lichan Hong, Julie Heiser, Stuart K Card which appeared in Proc. of the IEEE 2006 Visual Analytics Science and Technology (VAST2006) Symposium. © 2006 IEEE

Received: 23 June 2006
Revised: 31 July 2006
Accepted: 19 September 2006
Online publication date: 11 January 2007

Abstract

Agreat deal of analytical work has been carried out in the context of reading, in digesting the semantics of the material, the identification of important entities, and capturing the relationship between entities. Visual analytic environments, therefore, must encompass reading tools that enable the rapid digestion of large amounts of reading material. Other than plain text search, subject indexes, and basic highlighting, tools are needed for rapid foraging of the text. In this paper, we describe a technique that presents an enhanced subject index for a book by conceptually reorganizing it to suit particular expressed user information needs. Users first enter information needs via keywords, describing the concepts they are trying to retrieve and comprehend. Then our system, called ScentIndex, computes what index entries are conceptually related, and reorganizes and displays these index entries on a single page. We provide a number of navigational cues to help users peruse over this list of index entries and find relevant passages quickly. We report some initial results in a new technique called ScentHighlights that enhances skimming activity by conceptually highlighting sentences. Both use similar techniques by computing what conceptual keywords are related to each other via word co-occurrence and spreading activation. Compared to regular reading of a paper book, our study showed that users are more efficient and more accurate in finding, comparing, and comprehending material in our system.

Information Visualization (2007) 6, 32–47. doi:10.1057/palgrave.ivs.9500140

Keywords: Index; highlighting; annotations; information scent; contextualization; personalization

Introduction

'The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present-day interests, but rather that publication has been extended far beyond our present ability to make real use of the record.' – Bush.¹

Intelligence analysts spend a large amount of their time reading various articles and reports.² In fact, there is evidence that expert analysts set their filters lower (thereby accepting more irrelevant information) because they want to make sure that they do not miss information.² This increases the amount of reading, skimming, and text-searching by expert analysts compared to non-experts; they are just able to do it more quickly. Figure 1 shows a metric for describing this phenomenon as a graph. The figure illustrates that an expert can obtain more information in the same amount of time (or the same information in less time).

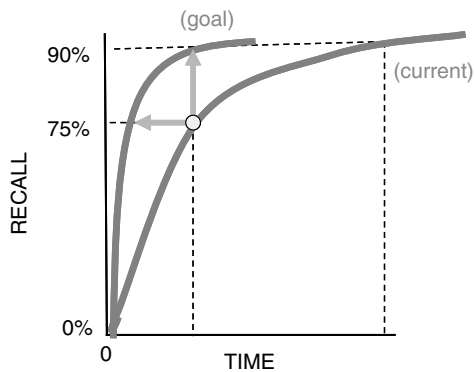


Figure 1 A cost structure metric for analyst information access. Improvements of system, method, or skill are reflected in raising the curve. Improvements may be harvested as greater information recall or as same recall within a given time.²

A goal for the system described in this paper is to enable novice analysts to raise their own curves by the use of intelligent highlighting and anticipatory semantic selection. To achieve this goal, we applied visual analytic methods in which a visual front-end helps the analyst direct her attention and see patterns, and an analytical semantic-processing back-end computes information scent relative to the analyst's changing interests. We are interested in ways in which user attention can be warped by the structure of the information environment, so that users can become more productive.

As the analyst of the large amount of time in reading and as the role this reading plays a role in expertise, visual analytic environments must encompass reading as an essential activity of the visual analytic cycle.^{3,p.45-47} While paper often remains as the preferred medium for reading,⁴ analysts are increasing reading and analyzing reports directly on the computer screen. One can trace back the history of various devices invented for reading and see a trend of ever increasing sophistication, such as the switch from linear scrolls to pages in books, or the utilization of library catalogs, table of contents (TOC), and indexes.⁵ In many ways, the giant leaps forward each time have been marked by new and better ways to find, correlate, and comprehend information. The subject index is an exemplar invention that furthers our ability to process information contained in documents.

Here, we explore the enhancement of subject indexes and the skimming of large amount of text by automatic highlighting. Latest efforts in digital libraries, such as the Million Book Project,⁶ have focused on the preservation and digitization of the vast archive of paper documents that human efforts have accumulated. Many, if not nearly all, of these books contain subject indexes that have been painstakingly generated manually. These subject indexes represent the authors' and editors' meticulous care in organizing the conceptual ideas in each and every one of these documents. Instead of seeking to automatically gen-

erate new indexes for texts,^{3,7-9} we seek to understand the problem of how to utilize these existing subject indexes and infuse them with new capabilities. We are interested in ways of enhancing these indexes so that they reorganize themselves to better suit the information needs of the analyst. We accomplish this chiefly in two ways:

- We have invented a new method to narrow down the index entries that are conceptually related to a query. To obtain this, we computed word co-occurrences within the entire book, which forms a conceptual word association matrix of the relationships between the words. Using this conceptual matrix, we were able to extract index entries that are conceptually relevant to the keywords the users entered. In this way, a large index containing thousands of entries can be quickly narrowed down to extremely relevant entries and displayed on a single page for the user to peruse.
- We have invented a method to intelligently extract summary passages that are relevant to the user's information need. We propose to direct the reader's attention by automatically highlighting relevant text. We highlight sentences by first computing the related conceptual keywords by using the word association matrix above. A sentence is highlighted if it contains conceptual keywords that are highly relevant to the user's interests.

There are often large lists of pages in the index to check for relevant passages. We help users in checking these pages by enhancing the navigation and browsing interactions between the index and the document. We use highlighting techniques to give navigational cues to the users. These cues tell users: (1) which index entries are likely to satisfy their query, and (2) which passages on the pages are likely to contain the relevant information pieces.

The third contribution is that we conducted a user study to understand the performance of ScentIndex (SI). Before the study, we were not sure if our entire interaction model based on the SI would hinder or speedup users. Would they get confused by the conceptual reorganization? Would they find the interaction cumbersome? Are users faster in locating information using the SI as compared to the standard practice of reading the paper book? We studied tasks for retrieving, comparing, and comprehending information, and measured the performance and accuracy of both content experts and novices. We found that users were faster in finishing tasks and more accurate in their answers using the SI, regardless of their expertise level in the textual content.

Previously, we described the interaction scenario of the SI and ScentHighlights ideas in non-archival short conference notes.^{12,13} In both cases, the computational algorithms have never been presented before and are presented here. In the case of SI, we also additionally present a detailed user study.

The rest of this paper is organized as follows. We present related work next. We then demonstrate the interaction model using a realistic scenario, specifically focusing on

the user's interactions with the SI. We present the computational method of how we reorganized the subject indexes and highlighted relevant passages. We also present details of the user study and analyze its results.

Background and related work

Researchers have focused on the possibility of utilizing computing devices for reading.¹⁴ The devices proposed range from the Memex¹ to mass-marketed devices such as Rocket eBook and SoftBook.¹⁵ There are also considerable efforts in the software-based document readers, including the representation of page content, distributing, displaying, reading, and searching over documents (e.g. DigiPaper,¹⁶ DjVu,¹⁷ Portable Document Format (PDF),¹⁸ Microsoft Reader¹⁹). Researchers have also been interested in using computer graphics to provide the look and feel of a real physical document on the computer screen. Early efforts have included the SGI Demo Book²⁰ and WebBook,²¹ and recent efforts have included the British Library's Turning the Pages,²² and 3Book.¹³ Typically, enhancement technologies for reading are studied in the context of an electronic book reader.^{23,24} These efforts laid the foundation for our work.

We are interested in ways in which the reading experience can be improved in digital environments. First, we studied whether one traditional component of a paper document – the subject index – could be improved upon by its digital counterpart. Second, we studied whether we could effectively compute and highlight relevant passages to draw users' attention and increase their skimming abilities.

Past work in navigating in documents using indexes

Currently, there are three basic ways to locate information relevant to a concept: (1) keyword search engines, (2) cross-referencing table like a subject index, and (3) browsing with a dynamically generated keyword or phrase hierarchy. We will discuss each of these methods in turn.

One prominent way is to use a keyword search engine. Here, the user enters keywords to retrieve a set of pages that contain those keywords. Indeed, the digitization of books has recently excited the possibility of searching over a large set of book pages. For example, our work has been inspired by Amazon.com's effort to digitize some 120,000 books and enable users to search for words and phrases in a feature called 'Search Inside the Book™'.²⁵ Sophisticated search engines based on Information Retrieval (IR) techniques such as Google and AltaVista effectively provide indexes to large textual pages.

Fundamentally, our work was differently motivated from keyword search engines because we were interested in how to enhancing the use of subject index in a reading activity. The relationship of our work to search engines can be summarized in three points: (1) First, we study the integration of three different components: searching,



Figure 2 Microsoft Help Index. Microsoft Help application's search over index simply scrolls the subject index.

subject indexes, and visual interfaces. Keyword search methods do not typically integrate subject indexes and visual interfaces simultaneously with searching. (2) Second, we used some of the same conceptual IR techniques as search engines do, such as vector-space model and ranking based on cosine similarity, except that we applied these techniques to a subject index. (3) Third, conceptually, our technique related to the *indexing structure* differently. A search engine takes a text collection and generates a structured index *dynamically*. In contrast, our technique takes a text collection and an *existing structure* (namely, the valuable subject index) and created a searching interface.

Another prominent way of locating information in documents is to use tables of contents, subject indexes, tables of names mentioned, and other cross-referencing tables. There are some computerized subject indexes out in the literature, suggesting its importance.²⁶ For example, computer-based help-applications as implemented by operating system vendors such as Apple and Microsoft are relevant here, even though they are not made to feel like a real physical user manual. These help-applications often have a subject index as well as a searching function. The searching function typically searches through a knowledge base of frequently asked questions, but the search is often integrated poorly with subject indexes. For instance, Microsoft's Help and Support Center application offers an alphabetical subject index. However, typing in a keyword simply scrolls the subject index to the first keyword or phrase match (Figure 2).

SuperBook²⁶ is probably the most well-known work related to the SI that uses computerized cross referencing, because it provides a TOC that is hierarchically and dynamically presented (Figure 3). Based on a query, a fish-eye function computes which part of the TOC is open.²⁷ However, the fisheye function uses only matching term frequency; thus, the TOC is not reorganized according to the concepts expressed by the keywords. For locating concepts, SuperBook uses an automated keyword search, like many current book readers.

In the arena of systems using cross-referencing techniques, the SI technique shares some similarities to systems that use topical hierarchies or network of related

Table of Contents	
31	*COMMON FUNCTIONAL SECTIONS
13	*TANDEM SUPPLEMENT
128	*FEATURE SPECIFIC DOCUMENTS
	*81 RESIDENCE AND BUSINESS CUSTOMER FEATURES
184	*82 PRIVATE FACILITY FEATURES
	*84 CUSTOMER SWITCHING SYSTEM FEATURES
3	*18 COIN AND CHARGE-A-CALL
	*15 PUBLIC SAFETY
17	*28 MISCELLANEOUS
	*25 INTEROFFICE
	*38 CALL PROCESSING
	*31 SERVICE SWITCHING
	*35 SYSTEM MAINTENANCE
	*48 TRUNK, LINE AND SPECIAL SERVICE CIRCUIT TEST
4	*45 ADMINISTRATION

Figure 3 SuperBook Index. A fisheye function based on term frequency computes which portion of SuperBook's Table of Content is displayed. (Term frequency on the left of each entry).

concepts to retrieve documents in a collection. These systems include term suggestion systems such as Schatz *et al.*,²⁸ and a bibliographical system called BoW.²⁹ Schatz *et al.*²⁸ suggested the simultaneous use of subject index and word co-occurrences to make term suggestions for retrieval. BoW enables users to search and insert entries into a hierarchical bibliographic system. BoW also uses a manually generated hierarchy, but it is used to index a collection of articles instead of concepts in a book. Moreover, BoW's searching algorithm is based simply on term frequency, not on conceptual word relationships. Also closely related is Rajashekar and Croft's^{30,31} examination of the use of thesaurus and keywords (but not subject indexes), to enhance query specification and retrieval.

These techniques are all related to the last prominent way of locating information in a book, which is browsing with a dynamically generated hierarchy. Natural Language Processing (NLP) researchers have looked into the automatic indexing of unstructured text for the purpose of browsing.^{8-11,32} For example, Cutting *et al.*^{11,33} discuss the idea of automatically generating hierarchical browsing structures for a collection of texts.³³ Often this is carried out by syntactically parsing the text for noun-phrases and other grammatical structures,⁸ or sometimes an existing taxonomy is used.³² Nevill-Manning *et al.*⁷ discuss a hierarchical phrase browsing system where the phrases are discovered using lexical parsers. Typically, an entity identification parsing algorithm is used to tag the text. For example, Wacholder *et al.*⁹ discuss how to increase the precision of noun-phrase identification for the purpose of discovering potential index entries.

There are two key differences between our work and these NLP efforts: (1) First, our idea is that much effort has been spent on the manual generation of the indexes of many books, so we should take advantage of this effort. (2) Second, the indexes generated by previous NLP techniques are usually not organized conceptually for a given task. With the exception of the dynamic clustering used in Cutting D. Scatter/Gather,^{11,12,33} these systems have

little semantic understanding of how keywords are related conceptually. Instead, our focus is to reorganize the existing manually generated index so that the conceptually relevant entries are presented together.

Past work in locating conceptually relevant passages

There are current deficiencies in reading/browsing interfaces.³⁴ For example, current search technology typically allows only exact keyword matches. Once the search is performed, a list of search results is displayed to the users, and they are then allowed to select from this list. As only exact keyword matches are given, users searching for the keyword 'Tennis' will only get articles that mention 'Tennis' explicitly. Articles that are related but do not contain many mentions of 'Tennis' will be ranked low or missed completely.

One related area of active research is conceptual search or associative search. That is, finding documents that refer to 'concepts' described by a set of terms. Typically, this is done by first applying keyword expansion techniques from information retrieval to find related conceptual keywords, and then use these conceptual keywords to actually do the search. In this way, related keywords will also be included in the search process.

For example, in the above search, the system does not know that 'Sharapova' is a famous tennis player. What is needed is a way to conduct conceptual search, where the specification of a keyword like 'Tennis' will incur the possibility of searching for other relevant keywords like 'Sharapova' and 'McEnroe', who are both famous tennis players.

For example, Google has a '~' syntax for a very limited keyword expansion capability. As another example, Autonomy has a conceptual search product available.³⁴ To our knowledge, this system is not based on spreading activation with word co-occurrence, but is based on the related technique of Bayesian networks. Researchers have also examined the possibility of using Latent Semantic Analysis (LSA) to perform associative search. Russell and Osborne³² described in their patent a method using LSA and a neural network for the purpose of performing a conceptual search.

While these methods have some conceptual similarities to our algorithm, they produce a result list of documents instead of directly highlighting passages in text. Unlike our system, these methods do not highlight the resulting search within the text *in situ*, but instead shows the results in a result page list. For example, in most search engines, once the user is brought to the result document, the relevant passages are not highlighted. Although sometimes the interface highlights the exact keyword matches on the document (e.g. Google Toolbar³⁵), what is needed is a way to direct user attention to the sentences or portions of the document that is most relevant to the concepts described by the user keywords, even when they do not mention the user-specified keyword directly. In

contrast, our highlighting method is designed to help user's skim for relevant passages directly within the text.

Perhaps, the closest work to ours is Hypertext Concordance explored by Schütze,³⁶ which lists the exact-matched terms *in situ* with the surrounding context in which these terms occur in. Again, while this is useful, what is needed is a way to highlight the relevant conceptual passages even when they do not refer to the search terms directly.

We want to construct an automatic system that highlights relevant passages. Whichever page the user is currently perusing, the system automatically highlights the relevant passages according to the current user interest.

Past work in conceptual association networks

The application of spreading activation to word co-occurrence is used in the construction of both the SI and ScentHighlights:

First, our system is based on a statistical NLP technique called Word Co-occurrence.³⁷ This technique has been used in noun-phrase identification and conceptual semantic mapping. Studies have shown that word co-occurrence patterns over a large corpus can be used to identify groups of keywords that are related conceptually to each other.³⁷ We use these word co-occurrence patterns in conjunction with algorithmic methods of Information Scents.³⁸

Second, SI and ScentHighlights make use of semantic and association networks to do conceptual search, which have been utilized in the past in both the AI and IR literature. Indeed, spreading activation has been studied extensively for the purpose of both information retrieval³³ and modelling human semantic memory.³¹ Spreading activation is a cognitive model developed in psychology that simulate how memory chunks and conceptual items are retrieved in the brain.³⁹ This model is suitable for our purpose of identifying related keywords and sentences. In our own work, spreading activation has also been shown to intelligently model user behavior in browsing a web site. We constructed a system to simulate the behavior of a group of users seeking to a piece of information in a web site.³⁸

SI and ScentHighlights are implemented in an eBook system called 3Book.⁴⁰ The SI displays only conceptually relevant entries to the user-keywords by using spreading activation. ScentHighlights is complimentary to SI, because it can be used in conjunction with SI. Once the user clicks on an index entry to go to a page of the book, relevant passages to that index entry will be highlighted in the book.

Usage scenario and user interaction

ScentIndex

In this section, we will describe how the SI works. We have produced 3Books of various types, but the user study

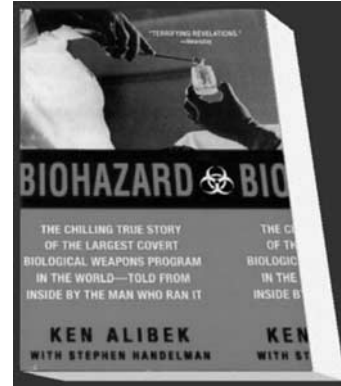


Figure 4 Digitized version of the book used.

and the scenarios below are based on *Biohazard* by Ken Alibek⁴¹ (Figure 4), which is a non-fiction retelling of his experiences while working on biological weapons in the former Soviet Union. There were 13 index pages in two columns, consisting of 829 entries.

In the following description of the system, we used one of the comparison tasks in our user study to demonstrate user interaction with the SI. The task is 'What year did Russia open negotiations with Iraq for large fermentation vessels? What year did Vladimir Kryuchkov become chairman of the KGB? Which occurred first?'

Figure 5 shows the SI after the index entries had been reorganized according to the information need of 'russia iraq fermentation'. We see keywords highlighted in red showing exact keyword matches. Relevant words such as 'biological', 'weapons' are highlighted by red underlining.

There are many entries that are relevant to the query. For example, for the keywords 'russia iraq fermentation', the system determined 'biological weapon' and 'Soviet' related entries to be important and included them in the single-page index view. After browsing several index entries, the user decided that the 'Iraq' entry is the most relevant, and skims the page entries under 'Iraq'.

Figure 6 describes how the user interacts with this index view. After the user has specified the concepts in the keyword box, the method computes a new single-page index view. The user can then click on a page number associated with an index entry, which opens the book to that particular page. User query keywords and the words in that index entry are used for keyword highlighting on the book page.

Figure 7 shows the book turned to a new page describing the sixth page entry of 'Iraq'. The words 'russia iraq fermentation' were automatically highlighted on this page. The bottom figure shows the specific paragraph that gives the first answer. The key to finding the answer quickly is looking for a page that contains all three of these words in a single passage. As shown, the automatic highlighting enables the user to quickly find the passage that might be relevant to the user query. This highlighting is extremely

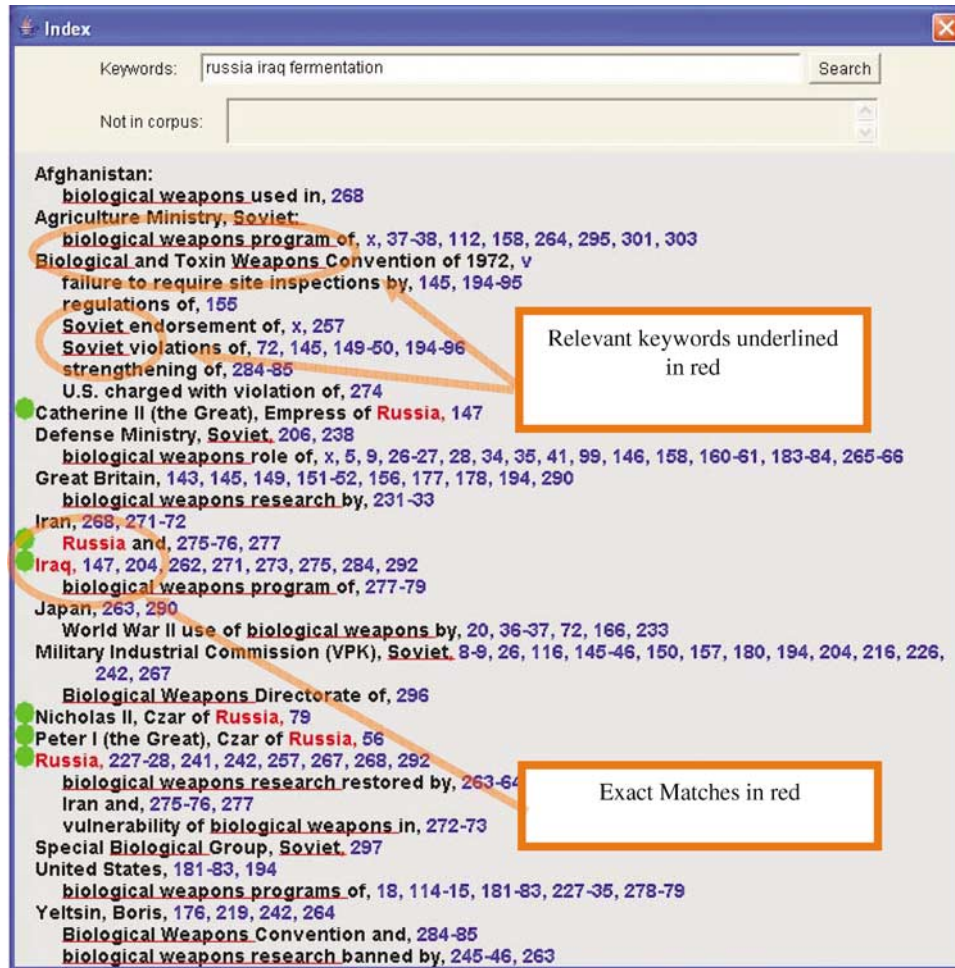


Figure 5 An example SI screenshot. SI showing the reorganization after 'Russia Iraq fermentation' was entered as the information need.

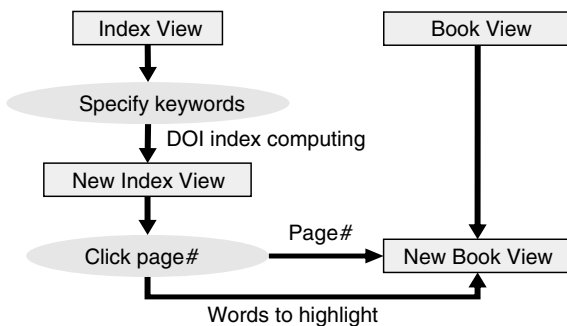


Figure 6 SI interaction flowchart. Flow chart describing the user interaction for SI.

helpful in tasks that require skimming for related facts through a large list of pages. The answer is 1995.

For the second part of the question, we need to find out what year *Vladimir Kryuchkov became chairman of the KGB*. Figure 8 shows the SI after the query 'kryuchkov

chairman kgb' is entered, thus eliminating hundreds of index entries. There are still many potentially relevant entries, but 'Kryuchkov, Vladimir' is probably the most relevant. Figure 9 shows the results after clicking on the second page entry. We see the answer is 1988.

After completing these two parts of the question, the user can then compare the two facts and find that the second fact occurred first in 1988. There are other ways to navigate through the index to find the solution to this question. We have merely shown one possible way to locate the relevant information.

Discussion The tasks here seem easy for a number of reasons: (1) First, by using conceptual reorganization, users have a high confidence that relevant entries are not omitted, because we do not rely solely on exact keyword matches. It is known that users generally have a hard time formulating a good set of search keywords, as there are large subject variations in formulating search queries in search engines, even when the task is explicitly

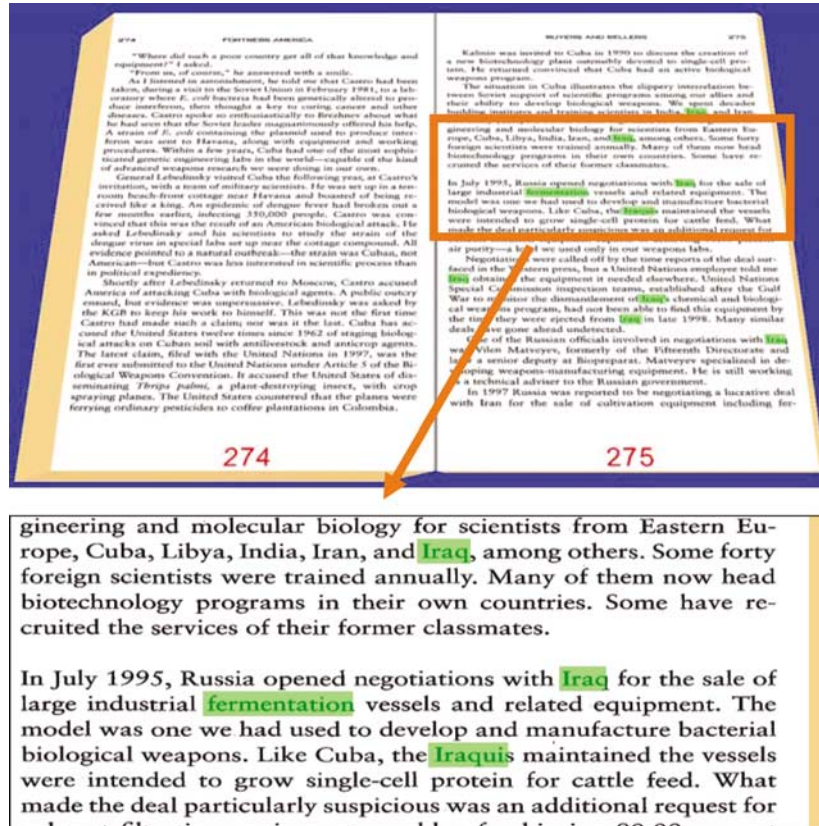


Figure 7 Keyword highlighting after clicking on an index entry. Clicking on the 6th 'Iraq' page number 275, the book opens up to that page (top) and highlights the relevant keywords (detailed bottom).

specified.⁴² For the first question above, for example, without the reorganization and fitting the index on a single page, a user might have first looked under other potential entries such as 'Russia' and 'fermentation' without success. Instead, users were able to decide 'Iraq' is the most promising entry.

(2) Second, there are many page entries that are potentially relevant. Our keyword highlighting on the book pages helped in skimming for relevant passages. By highlighting the keywords 'KGB chairman' users could easily locate the year as shown in the zoom of Figure 9. In this case, 'KGB chairman' must be entered as query keywords for the highlighting to help in the skimming process.

As shown in this usage scenario, by reorganizing the index entries, the user can narrow down the number of entries that one must search through to find the correct answer. By entering all the relevant keywords, users can see in one single screen what might be relevant without having to consult multiple index entries dispersed through several different index pages.

ScentHighlights

ScentHighlights enhances skimming activity by conceptually highlighting keywords and sentences that relate to

current user interest. The topic profile can be specified on-the-fly by the user using search keywords or can be generated from the user's reading and browsing history. We perform the conceptual highlighting by computing what conceptual keywords are related to each other via word co-occurrence and spreading activation.

We illustrate here how ScentHighlights can help readers locate relevant passages with a realistic scenario. Suppose we are looking to find out the symptoms of anthrax. We first type the keywords 'anthrax symptoms' into the search box (Figure 10a). Searching forward from the beginning of the book produced the result shown in Figure 10b.

The system identified three profitable regions to examine. Zooming up to the relevant passages that were highlighted on the left page shows that Alibek had worked on creating an anthrax weapon (Figure 11a). The conceptual keywords that caused the sentences to be highlighted are highlighted in gray, distinguished from the exact keyword matches shown in pastel-like colors (using Google's highlight color scheme). The boundaries of the highlighted sections are defined by sentences, as the algorithm attempts to highlight the top 3-5 most relevant sentences.

The spreading activation process produced highlights that were relevant to the task at hand. Zooming up to the

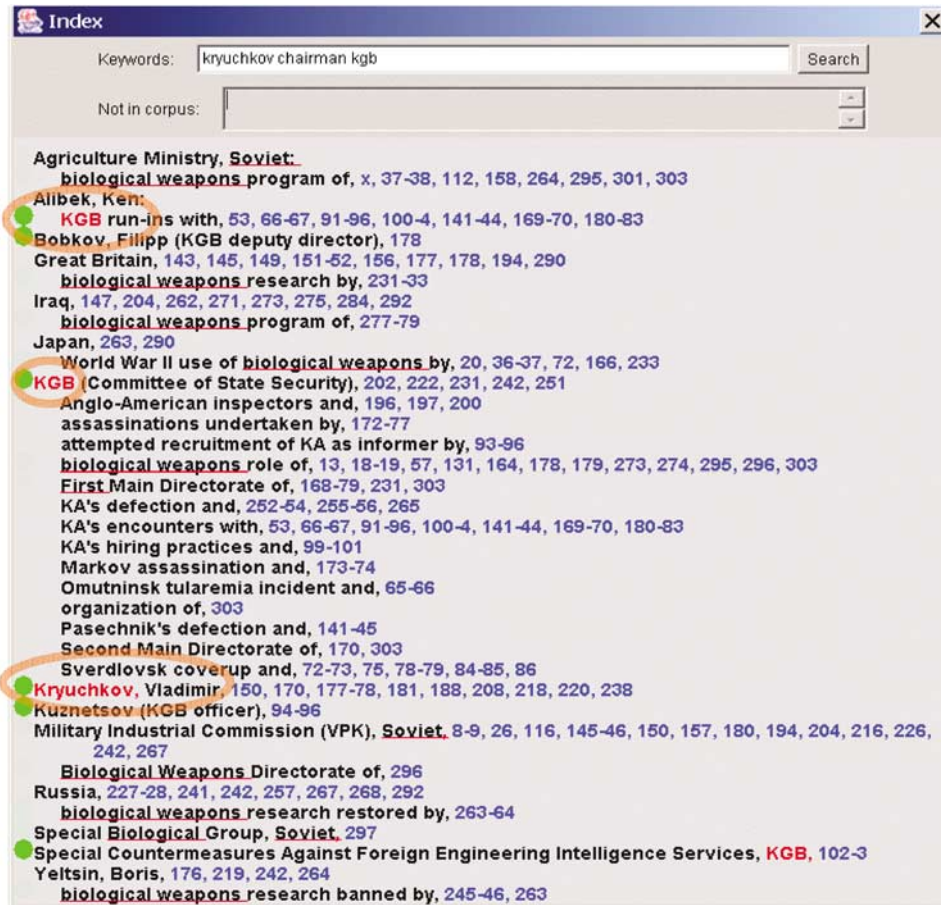


Figure 8 An example SI reorganization. SI showing the reorganization after 'kryuchkov chairman KGB' was entered.

relevant sections that are highlighted on the right side of the page gave us the information we were seeking (Figure 11b).

We see that the anthrax symptoms are *nasal stuffiness, twinges of pain in joints, fatigue, and a dry persistent cough*. Searching forward or turning to each new page will continue to produce highlights that are only relevant to the search keywords. As shown in this example, the ScentHighlights technique enables novel interactive browsing of electronic text in which users' attention is guided toward the most relevant sentences according to the user interest.

Method and algorithm

In this section, we describe the method used to create the SI and ScentHighlights.

As a summary, Figure 12 depicts the process that produces our SI. First, the paper document is scanned and OCR'ed, producing page images. The word locations are extracted to enable highlighting of the individual words. The recognized text is then used to compute the word association matrix. We used the matrix to compute the Degree-Of-Interest (DOI) function for the SI, ultimately

producing a single page of conceptually relevant index entries.

Figure 13 depicts the process that produces documents with ScentHighlights. Most of the process is exactly the same, except that the sentence structures are determined and the word association matrix is used in the spreading activation process to find other related conceptual words and sentences.

Method for computing related conceptual keywords

As the algorithms share the same first step, we will describe this step first. Given the keywords from a user profile (either specified directly with search keywords or implicitly generated from past user behaviors), we propose to find related conceptual keywords using word co-occurrence in conjunction with spreading activation.

The algorithms are based on the theoretical notion of *information scent*³⁴ developed in the context of *information foraging theory*.⁴³ Information Foraging is related to other research, such as Bates on Berrypicking^{34,45} and ASK,⁴⁴ on how users optimize behavior to seek information both in directed structured and opportunistic unstructured ways.

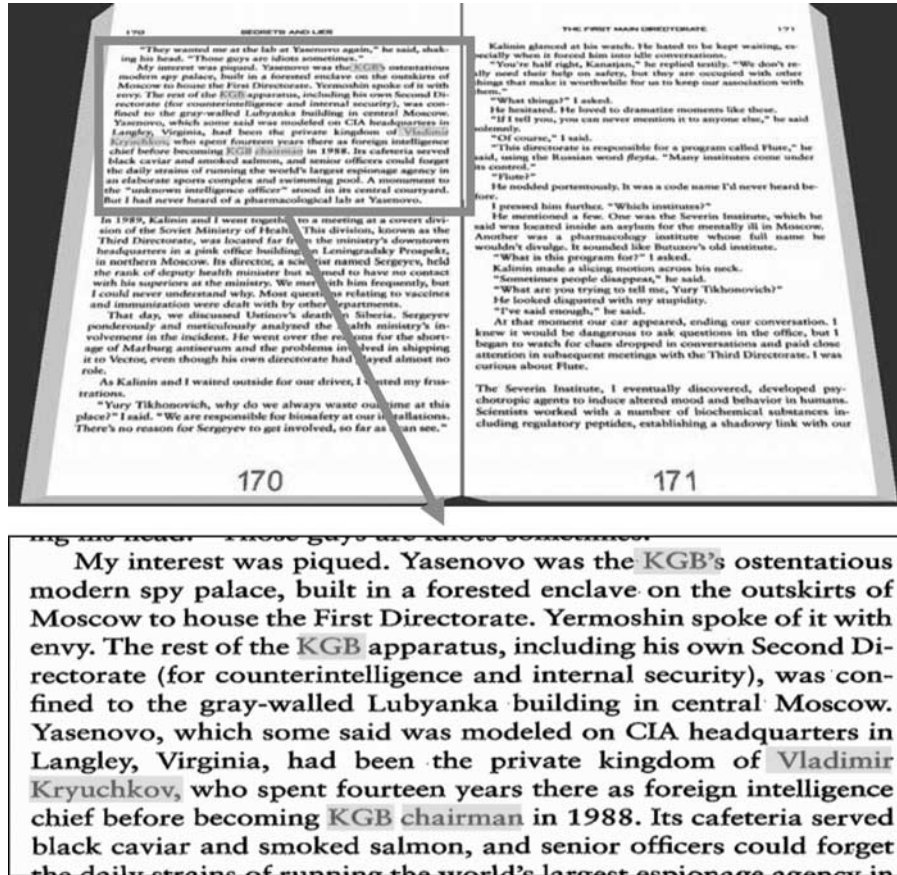


Figure 9 Keyword highlighting after clicking on an index entry. Clicking on the second entry of 'Kryuchkov, Vladimir' from Figure 8, the book opens up to page 170 (top) and highlights any words in the list of 'kryuchkov chairman KGB Vladimir'. The automatic highlighting helped in locating the relevant passages. We automatically highlight not just the user search keywords, but also the words that were in the index entry, such as 'Vladimir'.

Applied to the Web, users typically forage for information by navigating from page to page along hyperlinks. The content of pages associated with these links is presented to the user by some snippets of text or graphics called *proximal cues*. Foragers use these browsing proximal cues to access the *distal content*: the page at the other end of the link. *Information Scent* is the imperfect, subjective perception of the value, cost, or access path of information sources obtained from proximal cues. During information seeking, when choosing from a set of outgoing links on a page, the user examines some of the links and compares the cue (i.e., link anchor and/or surrounding text) with her information goal. The user takes the degree of similarity as an approximation of how much the content reachable via that link coincides with the information goal. Olston and Chi applied this notion to obtain an algorithm called ScentTrails⁴⁶ that predicts the paths of users, given some goal.

Applied to the subject index, the proximal cues are the words in the index entries. The distal content is the pages pointed to by these entries. Foragers use the proximal cues

(the words of the index entries) to find relevant pages to the concept that they are seeking.

Here, we adapt this method to the problem of predicting which index entries and passages are most relevant to the information goal given by the user. In the following methods, we employed the standard vector space model to represent keyword vectors. Entries in keyword vectors are numbers that describe the importance of a word.

Figure 14 describes the flow chart of the method. First, the book is funnelled through a parser that cleans and tokenizes the text. From this parser, we obtained a word list L , and a word co-occurrence matrix M . In our case, the word co-occurrence matrix is computed using a 40-word window, which seems to work well in practice. Specifically, for each word i , $M(i, j)$ is the number of times the word j has been co-mentioned within a ± 20 word span around each instance of i . The word \times word co-occurrence matrix M gives us the association strength between the words in the text. Conceptually, words that are co-mentioned together in the text should have a high degree of relevancy with each other. We experimented with a Porter stemmer

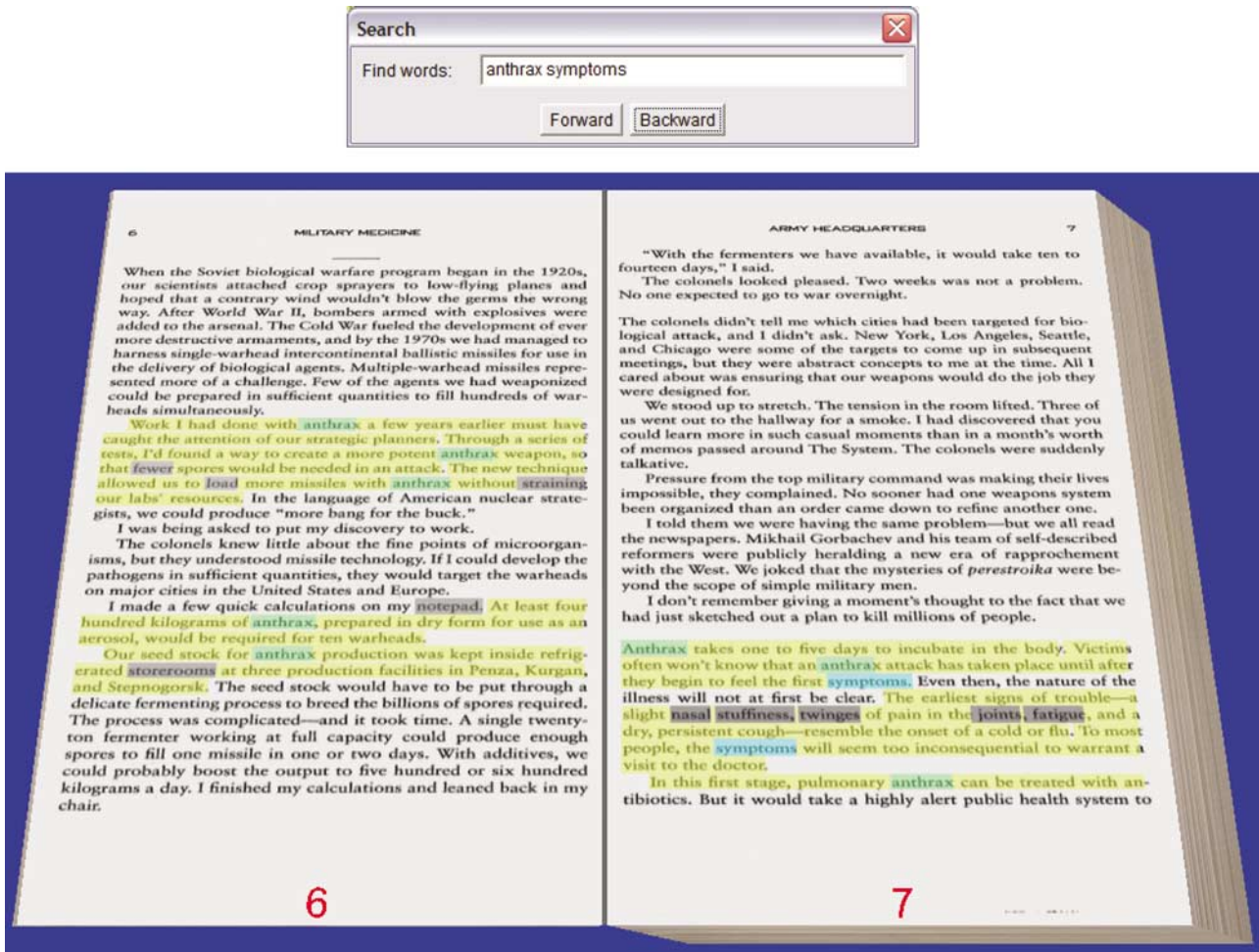


Figure 10 Example of highlights made by ScentHighlights. (a: top) keyword search box; (b: bottom) highlights obtained for ‘anthrax symptoms’.

and found that the results were not as good. Note that M is a symmetrical matrix.

A user’s information need is expressed as a keyword query vector Q . Given a query Q , we found other words that are relevant to the words on the list Q . Using spreading activation, we can compute words that are closely associated with the words in Q . There are various reasons to use spreading activation. The best explanation is that spreading activation has been shown to mimic the retrieval of relevant items in human memory.³⁹ We set the initial activation vector, $A(\mathbf{1}) = Q$. The algorithm goes through $t = \mathbf{1} \dots n$ number of iterations: $A(t) = \alpha MA(t - \mathbf{1}) + Q$. The parameter α modulates the process, avoiding the values from increasing exponentially. The number of iterations is typically from 1 to 4, depending on the designer’s preference.

We have observed in practice that words occurring often, in general, have a high probability of showing up in $A(n)$. That is intuitive, especially because we know word frequency follows a well-known power law curve called

Zipf’s Law. Therefore, to weigh down those keywords, we compute the term frequency of every word in the corpus, obtaining a vector TF which specifies the term frequency of each keyword. We then modify the weights of $A(n)$ to obtain Q' :

$$Q' = A(n)/(c * TF), \text{ where } c \text{ is a constant.}$$

The resulting activation vector Q' gives us a set of relevant keywords to the original query Q .

Therefore, the vector Q' specifies the conceptual words that are relevant to the original specified query words. Q' also specifies the relative weight (or importance) of the conceptual keywords. Q' can be viewed as the basis for a DOI function.²⁷

Method for ScentIndex (SI)

Taking the index entries in the book, we obtained the keyword vector for each entry $E(k)$, where $k = \mathbf{1} \dots m$, and

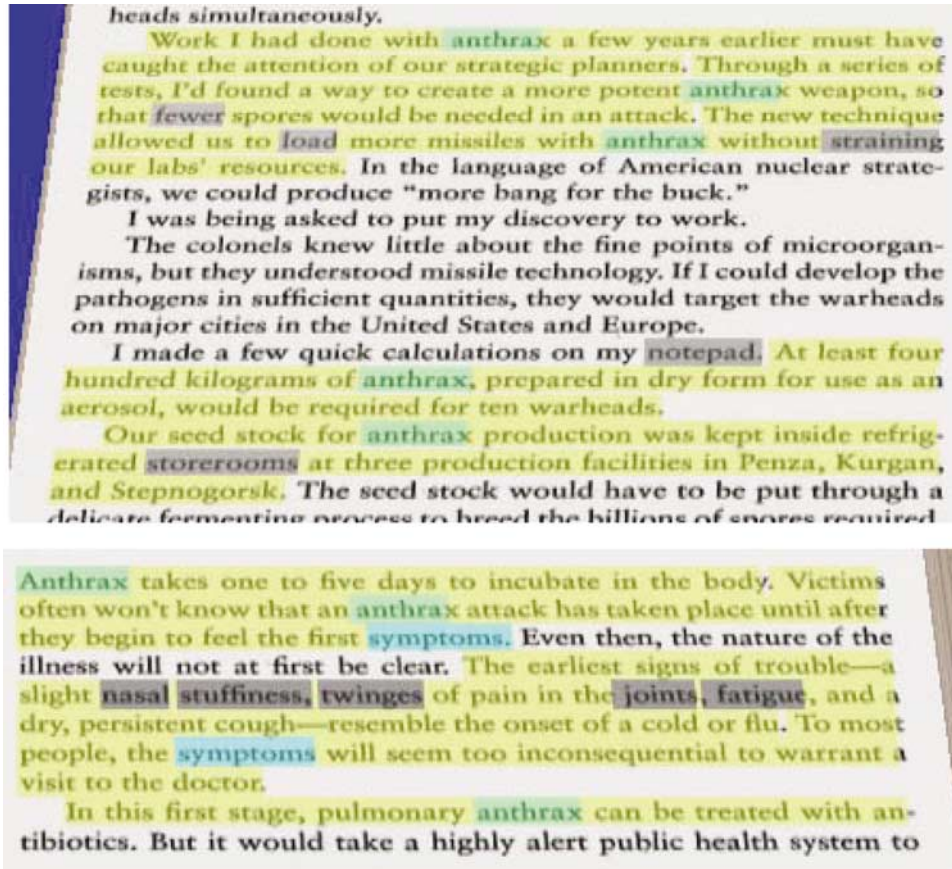


Figure 11 Example of keyword and sentence highlights. (a: top) Zoomed detail of the highlights on the left page; (b: bottom) Zoomed detail of the highlights on the right page.

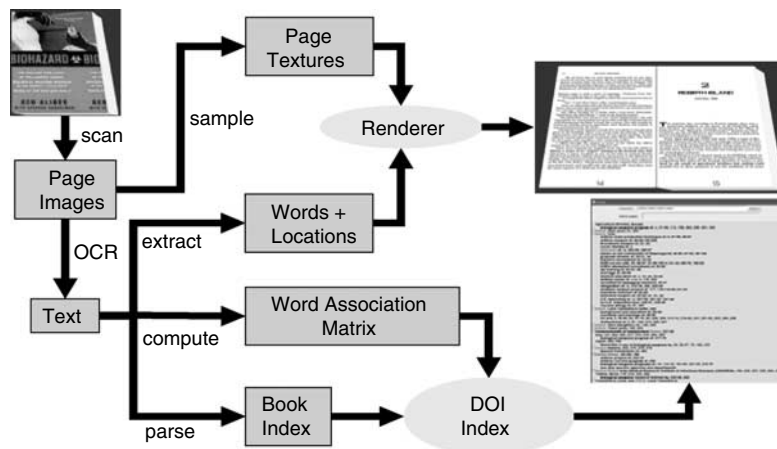


Figure 12 SI system flowchart. Overall flowchart of steps used to produce a SI in our system.

m is the number of index entries for the book. We used the same spreading activation method described above to expand the index entry keyword vectors $E(k)$. For each book, we could pre-compute the $E(k)$ vectors and cache the results.

Finally, we took the expanded query vector Q' and compute its cosine similarity with each $E(k)$. We ranked these similarity computation results in descending order. As the subject indexes are hierarchically organized, we used a DOI function²⁷ to compute on what entries to

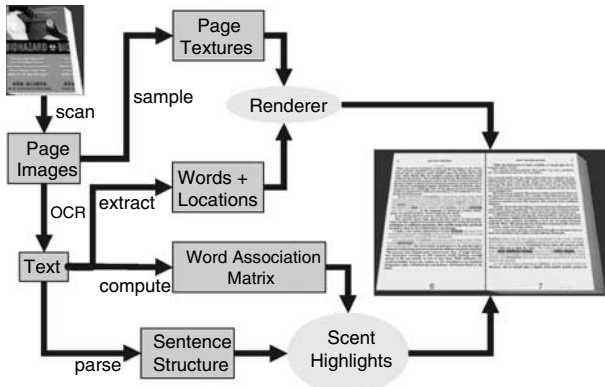


Figure 13 ScentHighlights system flowchart. Overall flowchart of steps used to produce an eBook with ScentHighlights.

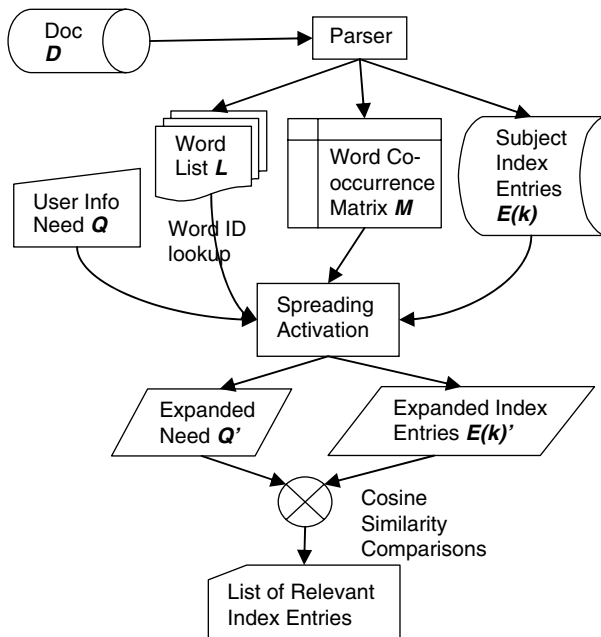


Figure 14 SI algorithm flowchart. Flow chart of the SI algorithm describing how the word semantic association matrix is used.

show. Boiled down simply, if a sub-entry is displayed, it would cause all of its parent entries (ascendants) to be displayed as well.

There is a caveat to the above algorithm. We found that an index entry $E(k)$ is not guaranteed to show up on the result list even if it contains a keyword i that is in the query vector Q . This is because keyword i might not have occurred with high frequency in the text, giving it a low magnitude in the word co-occurrence matrix M . There are two solutions to this problem. First, we can employ a keyword search algorithm to go through the index entries and make sure any entry that contains one of the query terms would show up on the final result list. Second, we

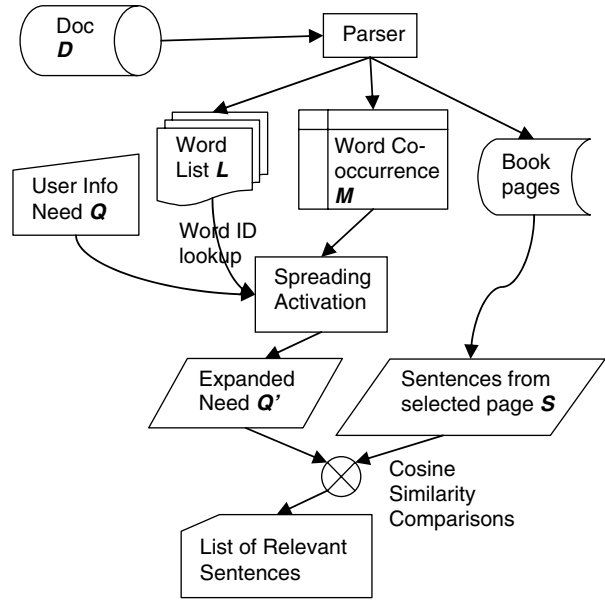


Figure 15 ScentHighlights algorithm flowchart. Flow Chart of the ScentHighlights algorithm describing how the word association matrix is used.

can ensure that a word is always highly associated with itself by simply inserting arbitrarily large values onto the diagonal of the M matrix.

Method for ScentHighlights

For ScentHighlights, we used these related keywords to figure out what sentences or portions of the documents are the most relevant conceptually to the profile keywords, and then actually highlight the relevant keywords and sentences in the text. Figure 15 depicts the algorithms used to calculate which sentence should be highlighted.

Given a document D , we now seek to highlight sentences that are conceptually related to the user specified query words. First, given a set of keywords Q , we obtain a new conceptual keyword vector Q' with the above process.

Now, we break the document D into sentences. For each sentence S , we compute an expanded keyword vector S' that represents that sentence. We then compute the dot product of S' with Q' . Essentially, we count how many times related conceptual keywords are shared between S and Q . This produces a score T for each sentence S . Given a page of text comprising a set of sentences, we then choose to highlight the top ranked sentences. For example, say the top five sentences that are the most related to the original query Q .

The above process is essentially a way to interactively generate dynamic summaries. The user is given a way to interactively specify an information need, and an interactive summary of the electronic document can be computed. We used this process to help readers be more

productive in choosing which sentence to read. Notice that the user can specify how much highlighting is needed by changing the number of sentences to be highlighted.

User study

Having described the methods for SI and ScentHighlights, we turn our attention toward evaluation of these techniques. We are interested in understanding the effectiveness of our techniques for directing users' attention to the profitable areas of the document. For example, we are conducting an on-going user study on the ScentHighlighting technique using eye-tracking technology to understand if users are paying attention to highlighted areas of the text.

For this paper, due to the enormous amount of different evaluation studies that could be done on this technology, we have chosen to report on the first user study on the SI and its effectiveness.

Possible evaluation studies

There are several possible ways to compare SI to existing methods of accessing a subject index:

- (1) The paper document and its paper-based subject index, which is the most common and familiar way currently used by intelligence analysts. The popularity of paper-based reading can be observed by noting that to read, many people still print out their emails. The paper document has a number of advantages: (a) The readability of paper is highly superior to the digital screen, and many reading tasks require large amounts of reading, skimming, and scanning;¹⁴ (b) Users might be familiar and fast with the paper-based subject index, and no technology might be able to improve upon that performance; (c) The digital interface might be unfamiliar to users to easily perform simple tasks such as turning the pages to the correct spot or complex tasks such as formulating search keywords and typing them into the search box; (d) The familiarity of the subject indexes might be so ingrained in the user that they are unable to use a new subject indexing technique. The dynamic reorganization of the SI might be too confusing to the user.
- (2) An existing digital document reader such as Adobe Acrobat,¹⁸ MS Reader,¹⁹ or Rocket eBook.¹⁵ Comparing our system with the best existing eBook systems would tell us how existing eBook systems could be improved.
- (3) A scrollable hypertext version of the original subject index in the 3Book, enabled with keyword search. The idea here is to compare the system with or without (a) the dynamic reorganization of the subject index and (b) the keyword highlighting navigational cues.

We chose to compare first with the paper book (experiment 1 above), because we wanted to compare the entire system with the existing practice of reading on paper.

Experimental design

We conducted a user study to find out if SI helps users to find, compare, and comprehend information in the Alibek book more quickly and more accurately than the subject index in the Paper Book. The user study was a within-subjects design with factors being interface condition (SI vs Paper Subject Index (PSI)) and task type (retrieval, compare, and comprehend), with the order of the interface used and the expertise level as the between-subjects variables.

Subjects: In all, 16 subjects participated, and were recruited from the authors' workplace, consisting of researchers, interns, and junior employees. Educational levels ranged from college grad to post-graduate. Eight subjects were content experts (read the book at least once). The other eight were novices (never read the book).

Materials: For the SI condition, subjects used a standard PC desktop machine with two LCD monitors. The left monitor displayed the Alibek eBook, and the right monitor displayed the SI interface. For the PSI condition, subjects used a paperback copy of the book.

Tasks: An experimenter without prior knowledge of how the SI system works, devised a total of 12 tasks. The tasks were divided into two groups of six tasks each. Tasks from one group were designed to be one-to-one equivalents of the other group. Of these six tasks, two were Simple Fact Retrieval questions, two were Dispersed Comparison questions, and two were Comprehension questions. Here is a sample of the questions:

Simple Fact Retrieval:

- The last natural-occurring case of WHICH virus occurred in Somalia in 1977?
- Who received a state award for developing a Q fever weapon?

Dispersed Comparison:

- What is the death rate of smallpox and tularemia? Which virus has a higher death rate?
- Which year did Russia open negotiations with Iraq for large fermentation vessels? Which year did Vladimir Kryuchkov become chairman of the KGB? Which occurred first?

Comprehension:

- Pasechnik's defection to the West had grave implications for the Soviet biowarfare program. Match the person with the fact that describes how they're involved:

Persons: Frolov, Chernyayev, Karpov, Vinogradov

Facts: (A) First told Alibek about Pasechnik's defection. (B) Deputy minister who refused to sign formal diplomatic reply. (C) Given demarche that said US have 'new information', presumably given by Pasechnik. (D) Told American visitors that Pasechnik's jetstream milling machine was for 'salt'.

- Diseases caused by different agents have different symptoms. Connect items on the agent list to the symptoms on the right.

Agents: Smallpox, Marburg, Tularemia

Symptoms: Chills, Nausea, Tiny bruises on the body, Toxic shock, Headache, Painful blisters, Stiffness, Fever, Unable to communicate.

Procedure: Each subject was first briefed on the experiment and filled out an initial survey on computing and search experiences.

All subjects used both interfaces. Four expert and four novice subjects used the PSI interface first, and the other eight used the SI first. Subjects were trained to use the SI right before they needed to use it.

All subjects also completed all twelve questions. For each interface, subjects performed the simple fact retrievals first, the dispersed comparisons second, and the comprehension questions last. Within each question type, the presentation order of the questions was randomized. Between the two sets of questions, half the subjects received one set first; the other received the other set first.

Each task was given on a separate sheet of paper, and subjects read and understood each question completely before they started the task. Subjects were told that they were being timed for each task after they finished reading each question and to do the best they could, but that there was a time limit for each task also (Simple retrieval=2 min, Comparison = 4 min, Comprehension = 6 min). Subjects were given one minute time warnings for each task. Incomplete tasks were recorded as the maximum allotted time. The time limits enabled us to keep each run of the experiment down to about an hour. After each run, subjects were asked to fill out a questionnaire on their preferences between the two interfaces, and any comments they might have.

Completion time analysis

The first question was whether SI is faster for users. We first observed that there were more tasks that users could not complete in the allotted time using the PSI. Of the simple retrieval tasks, six out of seven incomplete tasks were using the PSI, and seven out of eight for comparison, and three out of five for comprehension tasks. Thus, the average completion time data presented below is actually biased *against* the SI interface.

On average, SI ($M = 145$) is about 20s faster than PSI ($M = 162$). We performed a two-way within-subjects ANOVA with factors being interface (SI vs PSI) and task type (Simple Retrieval, Dispersed Comparison, and Comprehension), with between-subjects variables of order of interface used and expertise level, and the dependent measure being time to complete the task. The completion times were on the order of minutes, attesting to the difficulty of the tasks. We used a natural log transformation on the completion time for the analysis, which is a standard procedure in statistics to obtain the normality of

(secs)	S1	S2	S3	S4	D1	D2	D3	D4	C1	C2	C3	C4
SI	25.9	75.4	23.4	69.4	148	82.1	177	165	274	217	201	285
PSI	25.1	79.8	29.1	113	157	188	151	206	265	169	240	317

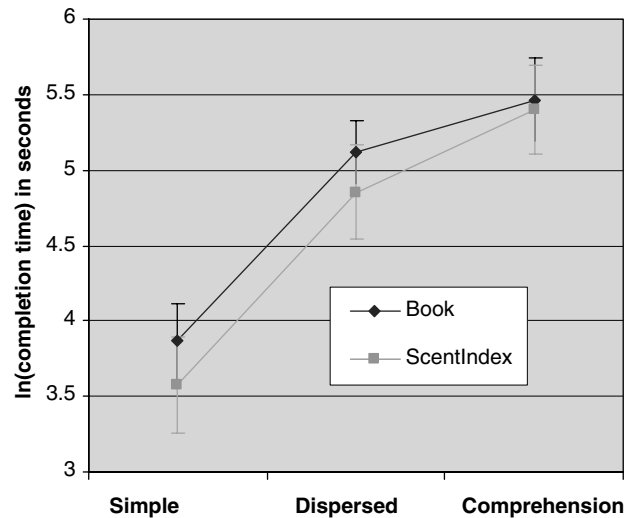


Figure 16 Comparing the two interface performances. (a: top) Raw time data in seconds. (b: bottom) Plot of log transformed completion time for the two interface conditions over the three question types (Error bars represent SD).

within-cell distributions. Figure 16 shows the summary plot of the data for the two different interface conditions over the three question types. We found that the participants using SI interface performed tasks faster than those using PSI, $F(1,12) = 12.96$, $P < 0.01$.

As predicted, experts performed tasks faster than novices overall, Expert Mean = 4.58, $SD = 0.212$, Novice Mean = 4.85, $SD = 0.212$, $F(1,12) = 17.7$, $P < 0.01$. There were no interactions. This is surprising, because we had surmised that experts are less likely to find the SI helpful in locating relevant content, because they should be able to navigate within the book effectively, due to their existing knowledge of the book. To our pleasant surprise, experts and novices alike were able to take advantage of the SI interface and complete the tasks faster.

Also, as predicted, Dispersed Comparison tasks took longer than Simple Retrieval tasks, and Comprehension tasks took longer than Dispersed Comparison tasks, $F(2,24) = 204$, $P < 0.01$. As shown in Figure 17, Mean Log Completion Times are: Simple = 3.72, $SD = 0.257$, Dispersed = 4.99, $SD = 0.230$, Comprehension = 5.435, $SD = 0.245$.

Accuracy analysis

The second question we wanted to answer was, whether users using SI produced answers that were better or on par with users using the PSI. We compared the answers given

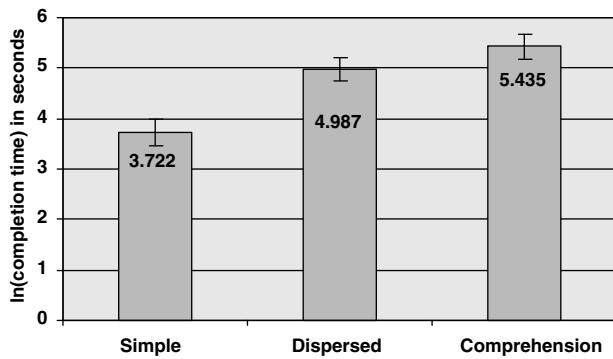


Figure 17 Comparing question type performance. Average log-transformed completion time for the three question types.

(pts)	S1	S2	S3	S4	D1	D2	D3	D4	C1	C2	C3	C4
SI	1.00	1.00	1.00	0.89	2.50	2.00	2.75	2.00	11.6	3.67	3.63	11.3
PSI	1.00	1.00	1.00	0.43	2.88	1.75	2.50	1.00	12.5	4.00	3.63	9.16

(measured score)	Simple Retrieval	Dispersed Comparison	Comprehension
ScentIndex (SI)	M=1.88 S.D.=.342	M=1.88 S.D.=.269	M=1.77 S.D.=.284
Paper Subject Index (PSI)	M=1.75 SD=.447	M=1.58 S.D.=.516	M=1.84 S.D.=.259

Figure 18 Comparing accuracy performance for both interfaces. (top) Mean of points earned. (bottom) Mean and SD on points scored for each question type and interface combination.

by the subjects with an answer key for each task to assess the accuracy. Some tasks have a single point score; others have as many as 13 points. So we converted the point scores to percentage scores. For each subject, there were two questions for each question type and interface combination. Therefore, we added the scores for these two questions together to get a combined score, giving a maximum of 200% = 2.0 for each combination.

We found that users performed better with more points using SI, reaching marginal significance (see Figure 18), $F(1,12) = 3.991, P=0.06$. We found no difference between experts and novices on points. We had surmised that experts might be more accurate in their answers, but instead novices were just as accurate in their answers as experts, using both interfaces. Again, there were no interactions.

User comments

The post-experiment survey showed that the participants overwhelmingly preferred SI (15 out of 16 subjects). The reasons given for this preference include ‘can search using keyword combinations’, ‘clicking on page number to navigate’, ‘highlighting enables faster scanning and skimming’, and ‘easier to compare index entries because it’s all on one page.’ In free-form discussion after the exper-

iments, some subjects mentioned that they would prefer the paper book version for extensive reading. Several users suggested that the index should not be organized alphabetically like a real index, but more like a search engine with the entries listed in decreasing relevance.

Summary and discussion

Experts and novices were equally accurate using either interface. The advantage of the prior knowledge in experts only showed when we compared their completion times. Experts were faster in completing their tasks with both interfaces. More importantly, the analysis results show that the interface condition did not have any interactions with the expertise level for both experimental measures. This means that expertise level affected the experimental measures independently of the interface used.

Overall, the SI performed better than the PSI. Subjects using the SI were faster in completing their tasks no matter whether they were experts or novices. Moreover, the answers they provided while using the SI were more accurate than the answers given when they used the PSI. Users also overwhelmingly preferred the SI interface for these tasks.

Conclusion

Reading, skimming, and text searching are essential activities in the visual analytic cycle, and rapid examination of an increased number of sources is associated with expert analyst behavior. Reading occupies a significant amount of the analyst’s time. Improving analyst reading and reading-like activities is therefore a place where computer-enhancement has real leverage. Our method of attacking this problem has been to bring visual analytic methods to bear.

To do this, we coupled intelligent visual highlightings of text (ScentHighlights) that helps direct the analysts attention, with analytic semantic background processing that filters a book’s index down to the most relevant entries (SI), including those semantically but not textually related. In this way, we amplified the role that subject indexes have had for books since they were invented in the 15th century. SI conceptually reorganizes large subject indexes according to the information need. Our user study suggests that this works. Both expert and novice users are quicker in completing fact-finding, comparison, and comprehension tasks using the SI, and the answers produced by the users are more accurate. We hope this will inspire a new line of research in augmenting reading with new innovations.

Acknowledgements

The user study portion of this research has been funded in part by ARDA NIMD/ARIVA program MDA904-03-C-0404 to Stuart Card and Peter Pirolli. We thank Jock Mackinlay,

Michelle Gumbrecht, Tan Lee, Michael Nguyen, Haixia Zhao, Pam Desmond, and Brian Tramontana for their help.

References

- 1 Bush V. As we may think. *The Atlantic Monthly* 1945; **176**: 101–108.
- 2 Pirolli P, Lee T, Card SK. Leverage points for analyst technology identified through cognitive task analysis. *Next Wave* (in press).
- 3 Thomas JJ Cook KA (Ed). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE CS Press: Silver Spring, MD, 2005.
- 4 Sellen AJ, Richard HRH. *The Myth of the Paperless Office*. The MIT Press: Cambridge MA, London, 2001.
- 5 Fischer SR. *A History of Reading*. London: Reaktion Book, 2003.
- 6 CMU. The Million Book Project. (WWW document) http://www.library.cmu.edu/Libraries/MBP_FAQ.html, Retrieved September 18, 2006.
- 7 Nevill-Manning CG, Witten IH, Paynter GW. Lexically-generated subject hierarchies for browsing large collections. *International Journal on Digital Libraries* 1999; **2**: 111–123.
- 8 Paynter GW, Ian HW, Cunningham SJ, Buchanan G. Scalable browsing for large collections: a case study. In: *Proceedings of the Fifth ACM Conference on Digital Libraries (San Antonio, Texas), 2000*; 215–223.
- 9 Wacholder N, David KE, Judith LK. Automatic identification and organization of index terms for interactive browsing. In: *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, Springer: Berlin. 2001; 126–134.
- 10 Douglass C, Cutting D, Karger D, Pedersen J, Tukey JW. Scatter/gather: a cluster-based approach to browsing large document collections. In: *Proceedings of the 15th Annual International ACM/SIGIR Conference*, pp. 318–329, 1992, ACM Press: Copenhagen, Denmark.
- 11 Sacco G. Dynamic taxonomies: a model for large information bases. *IEEE Transaction on Knowledge and Data Engineering* IEEE Press, 2000; **12**: 468–479.
- 12 Chi EH, Hong L, Heiser J, Card SK. eBooks with indexes that reorganize conceptually. In: *Proceedings of the Human Factors in Computing Systems Conference (CHI2004) Conference Companion (Vienna, Austria)*, ACM Press: New York. 2004; 1223–1226.
- 13 Chi EH, Hong L, Gumbrecht M, Card SK. ScentHighlights: highlighting conceptually-related sentences during reading. In: *Proceedings of the 10th International Conference on Intelligent User Interfaces*, ACM Press: New York. 2005; 272–274.
- 14 Harrison BL. E-books and the future of reading. *IEEE Computer Graphics and Applications* 2000; **20**: 32–39.
- 15 Rocket eBook. (WWW document) www.rocket-ebook.com, Retrieved March 2006.
- 16 Huttenlocher D, Moll A. On DigiPaper and the dissemination of electronic documents. *D-Lib Magazine* January 2000; **vol. 6**(1).
- 17 DjVu Zone. (WWW document) <http://www.djvuzone.org/wid/>. Retrieved September 18, 2006.
- 18 Adobe. What is Adobe PDF? [WWW document] <http://www.adobe.com/products/acrobat/adobepdf.html>, Retrieved March 2006.
- 19 Microsoft Reader. (WWW document) <http://www.microsoft.com/reader/>. Retrieved September 18, 2006.
- 20 Silicon Graphics, 'Demo Book', Silicon Graphics, Mountain View, California, Computer program 1993.
- 21 Card SK, Robertson GG, York W. The webbook and the web forager: an information workspace for the world wide web. In: Michael J. Tauber, (Ed.), *Proceedings of Human Factors in Computing Systems (CHI 96)*, ACM Press: New York. 1996; 111–117.
- 22 British Library. Turning the Pages on the Web. (WWW document) <http://www.bl.uk/collections/treasures/digitisation.html>, 2006.
- 23 Golovchinsky G, Marshall C, Schilit B. Designing electronic books. In: *Conference Companion of the ACM CHI99 Conference (Pittsburgh, PA)*. ACM Press: New York. 1999;
- 24 Remde JR, Gomez LM, Landauer TK. SuperBook: An automatic tool for information exploration – hypertext?. In: *Proceedings of Hypertext '87*, ACM Press: New York. 1987; 175–188.
- 25 Amazon.com. Search Inside the Book. (WWW document) <http://www.amazon.com>. Retrieved March 2006.
- 26 Wilson R, Landoni M, Gibb F. Guidelines for Designing Electronic Books. In: *Proceedings of the Sixth European Conference on Research and Advanced Technology for Digital Libraries*, Springer-Verlag: Berlin. 2002; 47–60.
- 27 Furnas GW. Generalized Fisheye views. In: *Proceedings of Conference on Human Factors in Computing Systems (CHI'86)*, ACM Press: New York. 1986; 16–23.
- 28 Schatz BR, Johnson EH, Cochrane PA. Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval. In: *Proceedings of the First ACM International Conference on Digital Libraries*, ACM Press: New York. 1996; 126–133.
- 29 Geffert M, Feitelson DG. Hierarchical indexing and document matching in BoW. In: *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, 2001*; 259–267.
- 30 Rajashekar TB, Croft WB. Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society for Information Science* 1995; **46**: 272–283.
- 31 Quillian MR. Semantic memory. In: Minsky M (Ed). *Semantic Information Processing*. MIT Press: Cambridge, MA. 1968; 216–270.
- 32 Russell DW, Osborne J. Method and system for searching text. US Patent 6,598,047. 2003.
- 33 Cohen PR, Kjeldsen R. Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management* 1987; **23**: 255–268.
- 34 Autonomy. Conceptual Search. (WWW document) http://www.autonomy.com/c/content/Products/IDOL/f/Conceptual_Search, 2004.
- 35 Google. Google Toolbar. (WWW document) <http://toolbar.google.com/>. Retrieved September 18, 2006.
- 36 Schütze H. The hypertext concordance: a better back-of-the-book index. *Proceedings of Computerm '98* 1998; 101–104.
- 37 Schütze H, Manning C. *Foundations of Statistical Natural Language Processing*. MIT Press: Cambridge, MA, 1999.
- 38 Chi EH, Pirolli P, Chen K, Pitkow J. Using information scent to model user information needs and actions on the Web. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2001)*, 2001; 490–497.
- 39 Anderson JR, Pirolli PL. Spread of Activation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 1984; **10**: 791–798.
- 40 Card SK, Hong L, Mackinlay JD, Chi EH. 3Book: a 3D electronic smart book. In: Maria Francesca Costabile (Ed.), *Proceedings of the Advanced Visual Interface (AVI2004)*, Italy, 303–307.
- 41 Alibek K, Handelman S. *Biohazard*. Delta Publishing: New York, NY, 1999.
- 42 Pollock A, Andrew H. What's Wrong with Internet Searching. *D-Lib Magazine*, March 1997. <http://www.dlib.org/dlib/march97/bt/03pollock.html>
- 43 Pirolli P, Card SK. Information foraging. *Psychological Review* 1999; **106**: 643–675.
- 44 Belkin NJ. Anomalous states of knowledge as the basis for information retrieval. *Canadian Journal of Information Science* 1980; **5**: 133–143.
- 45 Bates MJ. The design of browsing and berrypicking techniques for the on-line search interface. *Online Review* 1989; **13**: 407–431.
- 46 Olston C, Chi EH. ScentTrails: Integrating Browsing and Searching on the Web. *ACM Transaction on Computer-Human Interaction* 2003; **10**: 177–197.