ScentIndex: Conceptually Reorganizing Subject Indexes for Reading

Ed H. Chi, Lichan Hong, Julie Heiser[+], Stuart K. Card

Palo Alto Research Center

ABSTRACT

A great deal of analytical work is done 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 amount of reading material. Other than plain text search, subject indexes, and basic highlighting, tools are needed for rapid foraging of 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 also provide a number of navigational cues to help users peruse over this list of index entries and find relevant passages quickly. 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.

CR Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interfaces— Graphical User Interfaces; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia—Navigation; User Issues H.5.m. [Information interfaces and presentation (e.g., HCI)]: Miscellaneous **General Terms:** Design, Human Factors.

Additional Keywords: Book Index, eBooks, Information Scent, contextualization, personalized information access.

1   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 [8].
Intelligence analysts spend a large amount of their time reading various articles and reports [26]. 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 [26]. This increases the amount of reading, skimming, and text-searching expert analysts do relative 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 the expert can obtain more information in the same amount of time (or the same information in less time). 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 apply 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.
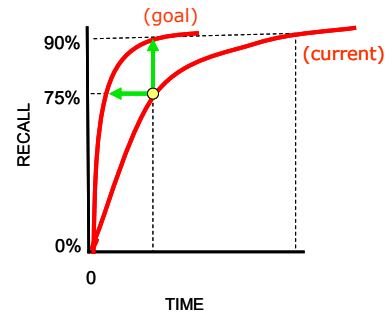

Figure 1: A cost structure metric for analyst information access. Improvements of system, method, or skill are reflected in raising the curve. Improvements may be harvested as greater information recall or as same recall within a given time [26],

Because of the large amount of time analysts spend in reading and because the role this reading plays in expertise, visual analytic environments must encompass reading as an essential activity of the visual analytic cycle [36, p. 45-47]. While paper often remains as the preferred medium for reading [34], increasingly analysts read and analyze 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, and indexes [20]. 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. Latest efforts in digital libraries such as the Million Book Project [12] 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 generate new indexes for texts [13, 22, 24, 37], 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 do 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 tens of extremely relevant entries and displayed on a single page for the user to peruse.

- 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 eBook. We use a wide variety of highlighting techniques to give navigational cues to the users. These cues tell users: (1) what index entries are likely to satisfy their query, and (2) what 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. Before the study, we were not sure if our entire interaction model based on the ScentIndex 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 ScentIndex 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 ScentIndex, regardless of their expertise level in the textual content.

Previously, we described the interaction scenario of the ScentIndex idea in a short conference note [10]. Here we describe the computational method as well as present the 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 ScentIndex. We present the computational method of how we reorganize the subject indexes. We also present details of the user study and analyze its results.

2   RELATED WORK

Researchers have focused on the possibility of utilizing computing devices for reading [18]. The devices proposed have ranged from the Memex [8] to mass-marketed devices such as Rocket eBook and SoftBook [30]. 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 [19], DjVu [14], Portable Document Format (PDF) [1], Microsoft Reader [21]). 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 Book [35] and WebBook [9], and recent efforts have included the British Library's Turning the Pages [7], and 3Book [10]. These efforts lay the foundation for our work. Here we study whether one traditional component of a paper document – the subject index – can be improved upon by its digital counterpart.

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 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™" [3].

Sophisticated search engines based on Information Retrieval (IR) techniques such as Google and AltaVista effectively provide indexes to large textual pages.

Fundamentally, our work is motivated differently from keyword search engines because we're interested in how to enhance 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, subject indexes, and visual interfaces. Keyword search methods do not typically integrate subject indexes and visual interfaces with searching simultaneously. (2) Second, we use some of the same conceptual IR techniques as search engines do, such as vector-space model and ranking based on cosine similarity, except we apply these techniques to a subject index. (3) Third, conceptually, our technique relates 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 creates a searching interface.

Another prominent way of locating information in documents is to use tables of contents, subject indexes, and other cross-referencing tables. SuperBook [29] is probably the most well-known related work to the ScentIndex that uses computerized cross referencing, because it provides a table of content (TOC) that is hierarchically and dynamically presented. Based on a query, a fisheye function computes which part of the TOC is open [15]. 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 ScentIndex technique shares some similarities to systems that use topical hierarchies or network of related concepts to retrieve documents in a collection. These systems include term suggestion systems such as Schatz et al. [32], and a bibliographical system called BoW [16]. Schatz et al. suggested the simultaneous use of a subject index and word co-occurrences to make term suggestions for retrieval [32]. 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 examination of the use of thesaurus and keywords (but not subject indexes) to enhance query specification and retrieval [28].

These techniques are all related to the last prominent way for locating information in a book, which is browsing with a dynamically generated hierarchy. Natural Language Processing (NLP) researchers have looked into automatic indexing of unstructured text for the purpose of browsing [13, 22, 24, 31, 37]. For example, Cutting et al. discuss the idea of automatically generating hierarchical browsing structures for a collection of texts [13]. Often this is done by parsing the text syntactically for noun-phrases and other grammatical structures [24], or sometimes an existing taxonomy is used [31]. Typically an entity identification parsing algorithm is used to tag the text. For example, Wacholder et al. discuss how to increase the precision of noun-phrase identification for the purpose of discovering potential index entries [37]. Nevill-Manning et al. discuss a hierarchical phrase browsing system where the phrases are discovered using lexical parsers [22].

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 Scatter/Gather [13], 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.

Our system is based on a statistical NLP technique called Word Co-occurrence [33]. 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 [33]. We use these word co-occurrence patterns in conjunction with Information Scent algorithms of Chi et al. [11]. We discuss the details of our algorithm later in this paper.

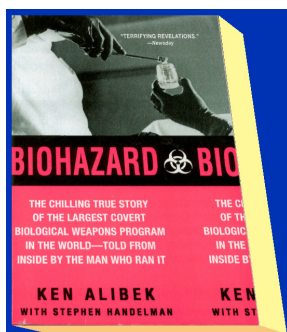### 3 USAGE SCENARIO AND USER INTERACTION

In this section, we will describe how ScentIndex works. We have produced 3Books of various types, but the user study and the scenarios below are based on *Biohazard* by Ken Alibek [2] (Figure 2), which is a non-fiction retelling of his experiences working on biological weapons in the former Soviet Union. There were 13 index pages in two columns, consisting of 829 entries.

Figure 2: Digitized version of the book used.

In the following description of the system, we use one of the comparison tasks in our user study to demonstrate the user interaction with the ScentIndex. 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 3 shows the ScentIndex after the index entries have 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 are important and included them in the single-page index view. After browsing several index entries, the user decides that the "Iraq" entry is the most relevant, and skims the page entries under "Iraq".

Figure 4 shows the book turned to a new page describing the 6th 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 helpful in tasks that require skimming for related facts through a large list of pages. The answer is 1995.
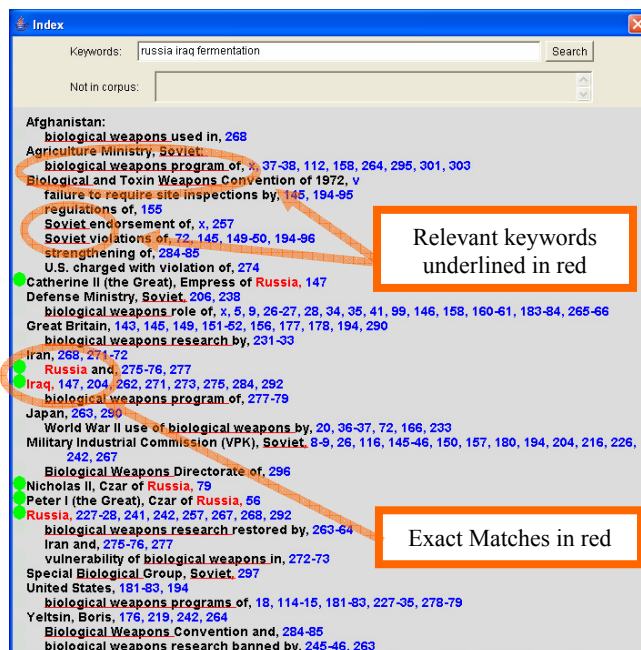

Figure 3: ScentIndex showing the reorganization after "russia iraq fermentation" was entered as the information need.

For the second part of the question, we need to find out what year *Vladimir Kryuchkov became chairman of the KGB*. Figure 5 shows the ScentIndex 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 6 shows the results after clicking on the 2nd page entry. We see the answer is 1988.

After completing these two parts to the question, the user can then compare the two facts and find the 2nd 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.
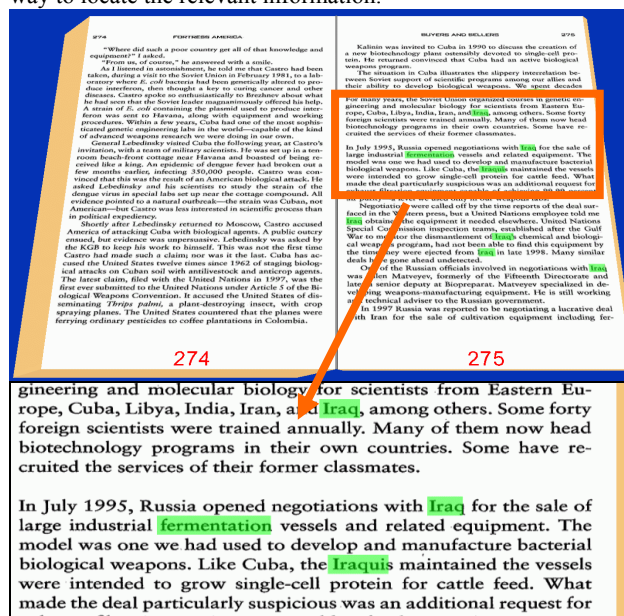

Figure 4: Clicking on the 6th "Iraq" page number 275, the book opens up to that page (top) and highlights the relevant keywords (detailed bottom).
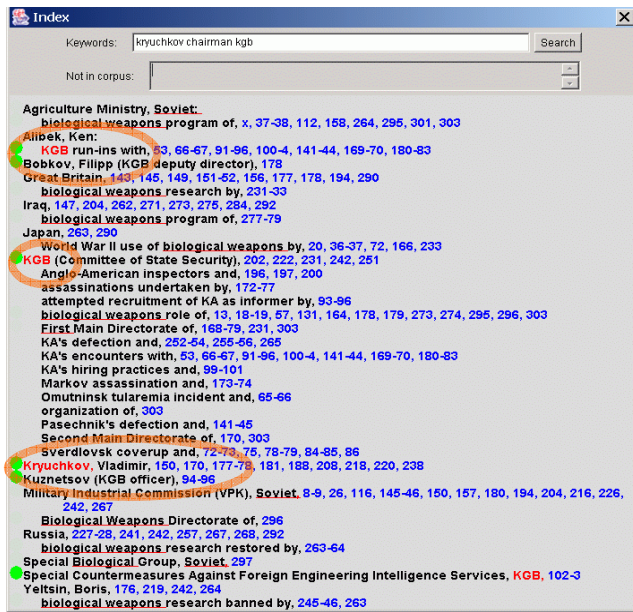
Figure 5: ScentIndex showing the reorganization after "kryuchkov chairman kgb" was entered.
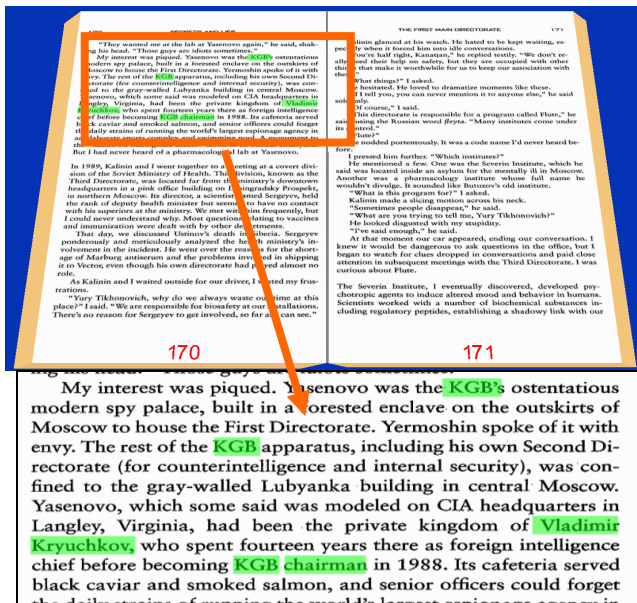

Figure 6: Clicking on the 2nd entry of "Kryuchkov, Vladimir" from Figure 5, the book opens up to page 170 (top) and highlights any words in the list of "kryuchkov chairman kgb Vladimir". The automatic highlighting helped in locating the relevant passages. We automatically highlight not just the user search keywords, but also the words that were in the index entry, such as "Vladimir".

3.1 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 specified [27]. 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 are 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 6. 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 of 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.

4 METHOD AND ALGORITHM

As a summary, Figure 7 depicts the process that produces our ScentIndex. 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 use the matrix to compute the Degree-Of-Interest (DOI) function for the ScentIndex, ultimately producing a single page of conceptually relevant index entries.

The ScentIndex algorithm is based on the theoretical notion of *information scent* [11] developed in the context of *information foraging theory* [25]. Information Foraging is related to other research, such as Berrypicking [5] and ASK [6], on how users optimize behavior to seek information both in directed structured and opportunistic unstructured ways. 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 to how much the content reachable via that link coincides with the information goal. Olston and Chi applied this notion to obtain an algorithm called ScentTrails [23] that predicts the paths of users given some goal.
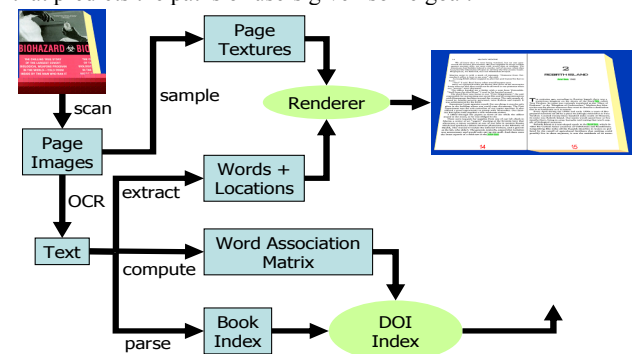

Figure 7: Overall flowchart of steps used to produce an ScentIndex in our system.

Here we adapt this method to the problem of predicting which index entries are most relevant to the information goal given by the user. 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.

In the following methods, we employ the standard vector space model to represent keyword vectors. Entries in keyword vectors are numbers that describe the importance of a word.

Figure 8 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 obtain a word list L, and a word co-occurrence matrix M. The word co-occurrence matrix is computed using a 40-word window. 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 x 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 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 find 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 [4]. We set the initial activation vector, $A(1) = Q$. The algorithm goes through $t=1 \ldots n$ number of iterations: $A(t) = \alpha M A(t-1) + Q$. The parameter α modulates the process, avoiding the values from increasing exponentially. The number of iterations is typically from 1 to 4, depending on designer's preference. The resulting activation vector $Q'=A(n)$ gives us a set of relevant keywords to the original query Q.

Taking the index entries in the book, we obtain the keyword vector for each entry $E(k)$, where $k=1 \ldots m$, and m is the number of index entries for the book. We use the same spreading activation method described above to expand the index entry keyword vectors $E(k)$. For each book, we can pre-compute the $E(k)'$ vectors and cache the results.

Finally, we take the expanded query vector Q' and compute its cosine similarity with each $E(k)'$. We rank these similarity computation results in descending order. Since the subject indexes are hierarchically organized, we use a Degree-Of-Interest (DOI) function [15] to compute what entries to 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 can ensure that a word is always highly associated with itself by simply inserting arbitrarily large values onto the diagonal of the M matrix.
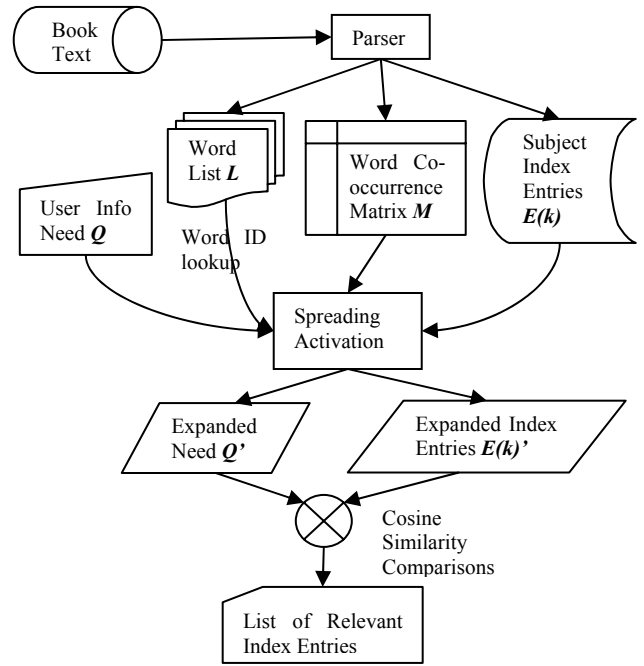

Figure 8: Flow chart of the ScentIndex algorithm describing how the word semantic association matrix is used.

5   USER STUDY

5.1   Possible Evaluation Studies

There are several possible ways to compare ScentIndex 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 many people still print out their emails to read.

The paper document has a number of advantages: (a) The readability of the paper is highly superior to the digital screen, and many reading tasks require large amounts of reading, skimming, and scanning [18]; (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 ScentIndex might be too confusing to the user.

(2)   an existing digital document reader such as Adobe Acrobat [1], MS Reader [21], or Rocket eBook [30]. 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 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 first compare with the paper book (experiment 1 above), because we wanted to compare the entire system with the existing practice of reading on paper.

5.2 Experimental Design

We conducted a user study to find out if ScentIndex 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 (ScentIndex [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: 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 ScentIndex (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 ScentIndex interface. For the Paper Subject Index (PSI) condition, subjects used a paperback copy of the book.

Tasks: An experimenter without prior knowledge of how the ScentIndex 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 Comparision:

- What is the death rate of smallpox and tularemia? Which virus has a higher death rate?
- 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?

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 Paper Subject Index interface first, and the other

eight used the ScentIndex first. Subjects were trained to use the ScentIndex 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 is randomized. Between the two sets of questions, half of 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=2min, Comparison=4min, Comprehension= 6min). 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.

5.3 Completion Time Analysis

The first question was whether ScentIndex is faster for users. We first observed that there were more tasks that users could not complete in the time allotted using the Paper Subject Index. Of the simple retrieval tasks, 6 out of 7 incomplete tasks were using the Paper Subject Index, and 7 out of 8 for comparison, and 3 out of 5 for comprehension tasks. Thus, the average completion time data presented below is actually biased *against* the ScentIndex interface.

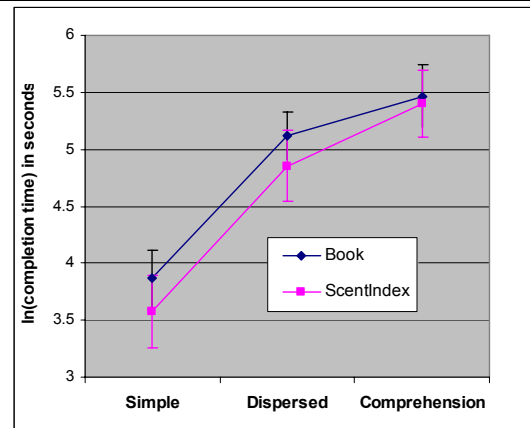| (secs)	S1	S2	S3	S4	D1	D2	D3	D4	C1	C2	C3	C4
SI	25.9	75.4	23.4	69.4	148	82.1	177	165	274	217	201	285
PSI	25.1	79.8	29.1	113	157	188	151	206	265	169	240	317

Figure 9: (top) Raw time data in seconds. (bottom) Plot of log transformed completion time for the two interface conditions over the three question types. (Error bars represent standard deviation)

On average, SI (M=145) is about 20 seconds 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 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 normality of within-cell distributions. Figure 9 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, $F_{(1,12)}=12.96$, $p<.01$.

As predicted, experts performed tasks faster than novices overall, Expert Mean=4.58, S.D.=.212, Novice Mean=4.85, S.D.=.212, $F_{(1,12)}=17.7$, $p<.01$. There were no interactions. This is surprising, because we had surmised that experts are less likely to find the ScentIndex 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 pleasing surprise, experts and novices alike were able to take advantage of the ScentIndex 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, $F_{(2,24)}=204$, $p<.01$. As shown in Figure 10, Mean Log Completion Times are: Simple=3.72 S.D.=.257, Dispersed=4.99 S.D.=.230, Comprehension=5.435 S.D.=.245.
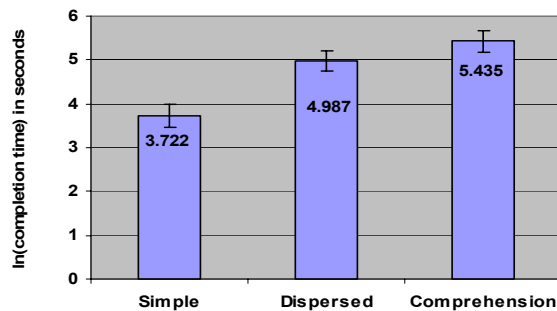

Figure 10: Average log transformed completion time for the three question types.

5.4 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 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.

(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	Comprehen -sion
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

Table 1: (top) Mean of points earned. (bottom) Mean and Standard Deviation on points scored for each question type and interface combination.

We found that users performed better with more points using SI, reaching marginal significance (see Table 1), $F_{(1,12)}=3.991$, $p=.06$. 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.

5.5 User Comments

The post-experiment survey showed that the participants overwhelmingly preferred ScentIndex (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 experiments, 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.

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

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

REFERENCES

1. Adobe. What is Adobe PDF? http://www.adobe.com/products/acrobat/adobepdf.html, Retrieved March, 2006.

2. Alibek, Ken, Handelman, Stephen. *Biohazard.* Delta Publishing, New York, NY, 1999.

3. Amazon.com. Search Inside the Book. http://www.amazon.com. Retrieved March, 2006.

4. Anderson, J. R., Pirolli, P. L. Spread of Activation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10 (1984): pp. 791-798.

5. Bates, M.J. The design of browsing and berrypicking techniques for the on-line search interface. *Online Review*, 13(5): pp. 407-431, 1989.

6. Belkin, N. J. Anomalous states of knowledge as the basis for information retrieval. *Canadian Journal of Information Science*, 5, pp. 133-143, May 1980.

7. British Library. Turning the Pages on the Web. http://www.bl.uk/collections/treasures/digitisation.html, 2006.

8. Bush, V. *As we may think*. The Atlantic Monthly 176, 1 (July 1945), 101 – 108.

9. Card, S. K., Robertson, G. G., & York, W. The Webbook and the Web Forager: An Information Workspace for the World Wide Web. In *Proc. of Human Factors in Computing Systems (CHI 96),* pp. 111-117. ACM Press, 1996.

10. Chi, E.H., Hong, L., Heiser, J., and Card, S.K. eBooks with Indexes that Reorganize Conceptually. In *Proc. of the Human Factors in Computing Systems Conference (CHI2004) Conference Companion*, pp. 1223-1226. ACM Press, 2004. Vienna, Austria.

11. Chi, E.H., Pirolli, P., Chen, K., and Pitkow, J. Using information scent to model user information needs and actions on the Web. *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI 2001)*, pp. 490-497. 2001.

12. CMU. The Million Book Project. http://www.library.cmu.edu/Libraries/MBP_FAQ.html, 2006.

13. Cutting, Douglass, David Karger, Jan Pedersen, and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proc. of the 15th Annual International ACM/SIGIR Conference*, 1992. Copenhagen.

14. DjVu Zone. http://www.djvuzone.org/wid/. 2006.

15. Furnas, G. W. Generalized Fisheye Views. In *Proc. of Conference on Human Factors in Computing Systems (CHI'86)*, pp.16-23. ACM Press, 1986.

16. Geffet, Maayan and Dror G. Feitelson. Hierarchical indexing and document matching in BoW. In *Proc. of ACM/IEEE Joint Conference on Digital Libraries*, pp. 259-267. 2001.

17. Golovchinsky, Gene, Cathy Marshall and Bill Schilit. Designing Electronic Books. In *Conference Companion of the ACM CHI99 Conference.* ACM Press, 1999. Pittsburgh, PA.

18. Harrison, B. L. E-books and the future of reading. *IEEE Computer Graphics and Applications*, 20(3):32-39, May 2000.

19. Huttenlocher, D., Moll, A. On DigiPaper and the Dissemination of Electronic Documents, *D-Lib Magazine*, Vol. 6, No. 1, January 2000.

20. Fischer, S. R. *A History of Reading*. London: Reaktion Book. 2003.

21. Microsoft Reader. www.microsoft.com/reader/, 2006.

22. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. Lexically-generated subject hierarchies for browsing large collections. *Int. Journal on Digital Libraries*, Vol. 2, No. 2/3, pp. 111-123. 1999.

23. Olston, Chris, Ed H. Chi. ScentTrails: Integrating Browsing and Searching on the Web. *ACM Transaction on Computer-Human Interaction*. 10(3), pp. 177-197. Sept, 2003.

24. Paynter, Gordon W., Ian H. Witten, Sally Jo Cunningham, George Buchanan. Scalable browsing for large collections: a case study, *Proc. of the fifth ACM conference on Digital libraries*, pp.215-223, June 2000, San Antonio, Texas.

25. Pirolli, P. and S.K. Card. Information foraging. *Psychological Review*. 106: pp. 643-675. 1999.

26. Pirolli, P., Lee, T., and Card, S. K. (in press). Leverage points for analyst technology identified through cognitive task analysis. Next Wave.

27. Pollock, Annabel and Andrew Hockley. What's Wrong with Internet Searching. *D-Lib Magazine*, March 1997. http://www.dlib.org/dlib/march97/bt/03pollock.html

28. Rajashekar, T. B., Croft, W.B. Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of the American Society for Information Science*, 46(4): 272 – 283. 1995.

29. Remde, J.R., Gomez, L.M., and Landauer, T.K. SuperBook: An automatic tool for information exploration – hypertext? In *Proc of Hypertext '87*, pp. 175-188. ACM Press, 1987.

30. Rocket eBook. www.rocket-ebook.com, Retrieved Mar. 2006.

31. Sacco, G. Dynamic taxonomies: a model for large information bases. *IEEE Transaction on Knowledge and Data Engineering,* 12 (3), pp. 468 – 479. IEEE Press, May/Jun 2000.

32. Schatz, B.R., Johnson, E.H., Cochrane, P.A. Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval. In *Proc. of the first ACM international conference on Digital libraries,* pp. 126 – 133 ACM Press, 1996.

33. Schuetze, H., Manning, C. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

34. Sellen, Abigail J. and Richard H. R. Harper. *The Myth of the Paperless Office*. Cambridge, Mass. and London: The MIT Press, 2001.

35. Silicon Graphics, "Demo Book", Silicon Graphics, Mountain View, California, Computer program 1993.

36. Thomas, J. J. and Cook, K. A., ed. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE CS Press, 2005.

37. Wacholder, Nina, David K. Evans and Judith L. Klavans. Automatic identification and organization of index terms for interactive browsing. In *Proc. of the first ACM/IEEE-CS joint conference on Digital libraries,* pp. 126-134. 2001.