

The Reader's Helper: A Personalized Document Reading Environment

Jamey Graham

California Research Center

Ricoh Silicon Valley, Inc.

2882 Sand Hill Road, Suite 115

Menlo Park, CA 94025, USA

jamey@rsv.ricoh.com

ABSTRACT

Over the last two centuries, reading styles have shifted away from the reading of documents from beginning to end and toward the skimming of documents in search of relevant information. This trend continues today where readers, often confronted with an insurmountable amount of text, seek more efficient methods of extracting relevant information from documents. In this paper, a new document reading environment is introduced called the Reader's Helper™, which supports the reading of electronic and paper documents. The Reader's Helper analyzes documents and produces a relevance score for each of the reader's topics of interest, thereby helping the reader decide whether the document is actually worth skimming or reading. Moreover, during the analysis process, topic of interest phrases are automatically annotated to help the reader quickly locate relevant information. A new information visualization tool, called the Thumbbar™, is used in conjunction with relevancy scoring and automatic annotation to portray a continuous, dynamic thumb-nail representation of the document. This further supports rapid navigation of the text.

Keywords

document annotation, information visualization, content recognition, intelligent agents, digital libraries, probabilistic reasoning, user interface design, reading online

INTRODUCTION

Around 1750AD there was a dramatic change in the way people read documents [8]. Before this time, readers consumed documents *intensively*, reading the document from start to finish, sometimes several times or even out loud to a group. By the early 1800's, however, readers tended to read *extensively*, reading documents only once or skimming the documents in search of relevant information to determine whether the document was worth reading in its entirety. Today, with the advent of the World Wide Web (WWW) and the growing collection of electronic documents, this style is likely to continue: there are simply too many potentially use-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI '99 Pittsburgh PA USA

Copyright ACM 1999 0-201-48559-1/99/05...\$5.00

ful documents and not enough time to read them all [14]. Office workers, in particular, are forced to optimize their daily reading by sifting through the vast amount of information, establishing a balance between in-depth understanding and expediency. Reading intensively versus extensively can be thought of as *vertical versus horizontal* reading [6]. That is, in the past, readers read the document from beginning to end (vertical); now, they scan and browse the text (horizontal).

Few applications available today fully support the reading process. There are, however, several applications which condense or locate documents for the user. Applications such as [13, 15] provide a synopsis of the text which can sometimes be used to determine the document's relevance. Other systems search for and retrieve documents relevant to an evolving user profile [3, 16]. The learning of user profiles over time provides an evolutionary process which enables the system to improve the quality of documents retrieved for the user. Another system supports users as they search digital libraries by showing query keywords in the context of the sentences they appear in a document [17]. Thus, users can quickly access the database based on the presence or absence of a particular context they are seeking. Another application inserts supplemental information in the form of an annotation into a news story if the story contains key phrases from the subject database [9]. This offers the reader additional information not necessarily provided by the author of the original text. The work by [7] supports the skimming of documents by representing the topics of a text as *content capsules*. Using a special visualization tool portraying the document as a thumb-nail image, topics are presented to the reader at the location in the text where they occur. This allows the reader to quickly view the highlights of the document in the context of the surrounding text structure.

Despite the growing number of applications used to locate and evaluate documents, there are few, if any, applications that focus on the actual *text reading* process. I believe that readers require a personalized environment that supports the skimming of documents and the extraction of information. I have created a new document reading environment called the Reader's Helper, to act as both the reader's document

browser and personal agent, advising the reader of relevant documents and of the relevant text within each document. The Reader's Helper is not a search engine; it does not search for or deliver documents to the user. Instead, it *helps readers help themselves* to be more productive in reading by evaluating documents the reader views and by providing visual tools for showing the locations of the relevant portions of the text. In the following sections the Reader's Helper system is described, both in terms of the user interface and the underlying content recognition subsystem. Future issues and potential research directions are also discussed.

THE READER'S HELPER

The Reader's Helper (RH) application integrates existing technology—a WWW browser, highlighting key words, and probabilistic reasoning—with a unique information visualization tool in support of readers who read both online and paper documents. The RH uses information about each reader in evaluating the content of a document. It calculates a *relevance* value to determine if a document is applicable to a reader. The current prototype is composed of a specialized WWW browser and an annotation agent responsible for recognizing the reader's *topics of interest* in a document. The way the annotation agent understands what is important to the reader is through a *reader profile* which contains personal information about the reader. In the following sections the electronic document browsing environment is described in terms of the user interface and followed by a description of the *paper-based* version of the RH.

The User Interface

One of the most important aspects of an electronic document reading environment is the user interface. The difficulty of reading electronic documents is well known [18]. Readers therefore often rely on the printed document because of the high resolution and flexibility offered by paper. A design goal for the RH was to provide the reader with an easy to use interface that emphasizes the relevant content of the document, and, as a consequence, *personalizes* the document. This should not only increase the appeal of reading electronic documents but also the efficiency with which they can be consumed. There are three main methods that work in concert to improve the readability of an electronic document. First, a relevance score is computed for each topic that is important to the reader with respect to a document. Each topic score is an estimation of the relevancy of the document to the reader, offering a *first appraisal* of the document. This is directly associated with the horizontal reading trend mentioned earlier: most readers do not have time to completely analyze all of the documents they must process and so it is desirable to have a quantitative measure of each document's relevancy. Second, a new information visualization tool, called the Thumbar, shows an overview of both the document and the results of annotating the document. Hence, the reader can quickly navigate to the more relevant parts of a document based on the

visual information present in a thumb-nail. Thirdly, after the RH has completed the analysis of a document, it is automatically annotated to depict the most relevant portions of the document. As the reader reads the document, key phrases are pointed out by way of highlighting to help guide the reader through the text.

The Thumbar

Figure 1 shows the RH document browser displaying a HTML source document (from the CHI '97 Online Proceedings [19]). On the left hand side of the display is the Thumbar (11 and 12). The Thumbar is a unique information visualization tool whose name is derived from the concept of a thumb-nail image combined with the functionality of a scrollbar. The reader may drag a lens (12) up and down to reposition the view of the document in the document area (13) in the manner of a scrollbar. The Thumbar allows readers to *look ahead* in the text to see more of the document's structure and content. By using a Thumbar, a reader can quickly scan the document for a desired image, chart or even text structure and scroll accordingly. The work by [22] supports the notion that humans are very good at recognizing small images. Recent work by [20] further exploits this concept in a system where icons of documents are used as the retrieval cues to a large document database. It was found that users could recognize the thumb-nail images of documents (based on the text structure and the images in the text) to retrieve several documents which "looked" like the document they were seeking.

The Thumbar is a dynamic representation of the document contained in the browser. Its contents and shape change as the document itself is personalized. For instance, after annotating a document, the Thumbar presents relevant keyword phrases as red lines (instead of the typical gray or black lines). This clearly indicates to the reader where the relevant information is located in the document, similar to the way attribute-mapped scroll-bars [24], TileBars [10] and the Mural [12] depict relevant information in a document. This visual information, however, can be changed. For example, by "turning off" a concept that has scored well in the document, the red lines in the Thumbar change back to the normal gray or black lines as if there were no annotation at that location at all. Thus, the reader can create a representation of the document based on a combination of concepts (14) either turned on or off. Another way of altering the annotation is by using the *sensitivity threshold meter* (15) which can be used to manipulate the concepts that are active in the document. This is done by setting a threshold and only allowing concepts whose similarity scores equal or surpass this threshold to be visible as annotations.

As with any HTML document, there is no formal pagination as long as the document is contained in the browser (the pagination is set when the document is sent to the printer). Instead, there is simply the concept of "a screenful" of the document which can change if the user resizes their browser

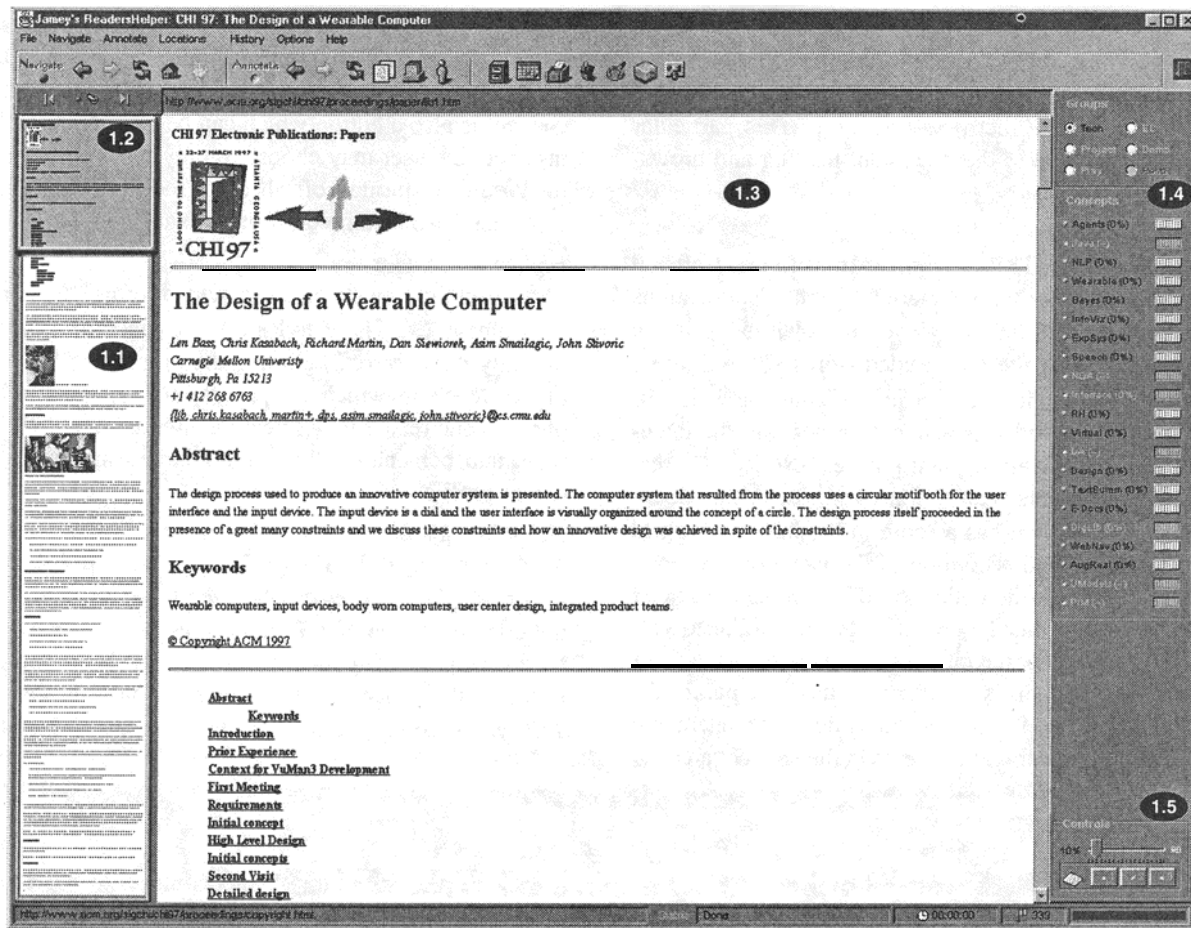


Figure 1 - The Reader's Helper WWW Browser

window. The Thumbbar (1.2) represents a reduced version of the original document based on a user defined reduction ratio (e.g. in figure 1, the reduction ratio is 6 which means that the Thumbbar is 1/6th the size of the document area (1.3)). The entire representation of text in the Thumbbar is presented using a proportionally reduced line for each word. This method of portraying a navigable thumb-nail of text is similar to [4], used for software visualization.

If the document is resized, the Thumbbar is resized to represent the same screenful depicted in the document area. The use of a thumb-nail icon to represent a page in a document is of course not a new idea. Adobe Systems, for instance, uses a similar method of displaying icons for the individual pages in a document [1]. The Adobe method, however, portrays a static representation of the document (i.e., the thumb-nail image does not automatically change when the document is edited) and does not support using a lens for navigation throughout the document. The Adobe method is also based on a formally paginated document with distinct thumb-nails for each page. In the Thumbbar, there is no pagination and therefore, the document is viewed as a continuous stream of text and images by which one can navigate to any location in the document.

If the document cannot be fully contained in the Thumbbar

(because the document is too long), a green line is drawn across the bottom of the Thumbbar. Dragging the lens down to the green line causes the entire thumb-nail image to scroll upwards, thereby revealing more of the document. Once the document has been scrolled in this way, a green line will also appear at the top of the Thumbbar indicating that more of the document is available above. To go back to the beginning of the document, the reader may again drag the lens upward to the green line which, thereby causing the thumb-nail image to scroll back down. Alternatively, there are buttons at the top of the Thumbbar for repositioning the lens at either the top (1.4) or the bottom (1.5) of the document. The user may also choose to view the entire document in the Thumbbar. This has the disadvantage, however, of reducing the clarity of thumb-nail information.

With the help of a Thumbbar attached to the document browser, the reader can look ahead in the document to evaluate its content and structure. This is not possible in most document browsing systems available today where readers are restricted to one page or screenful at a time. This ability to look ahead in a document is quite useful for reading documents as it leverages off of the design principle of providing a global (macro) and local (micro) representation of information [23]. The global representation portrays coarse infor-

mation while the local representation portrays fine grained information. In this instance, the Thumbbar presents an overview (macro) of the document and the document area presents the detailed (micro) information. This particular innovation adds value to the document reading and browsing experience.

Document Annotation

Figure 2 shows the HTML document from figure 1 after it has been annotated by the annotation agent. The locations of relevant phrases in the document are shown in red in the Thumbbar. Also, the pattern of bolded words (22) in figure 2 matches the highlighting pattern in the Thumbbar lens (21). Using the lens the reader can quickly reposition the document to areas containing relevant phrases. Notice also that there are four concepts with non-zero scores in this document (23). Each concept has a score in the label and a grid meter that is populated according to the value of the score. By simply looking at the collection of concepts the reader can quickly see that this document is indeed similar to several topics of interest based on which grid meters have color and to what extent. This is an important but subtle part of the interface design. This is based on the belief that readers will quickly learn the location of concepts in the concept area (23)

and as a consequence of highlighting in the grid meter, quickly know if the document discusses the desired topics of interest and which topics are covered. Note that each concept is also a button which can be turned on or off. For instance, the user may choose to only view the annotations for Wearable, turning off all other scoring concepts. Concepts may also be turned off prior to analyzing a document.

Highlighting Styles

Three highlighting styles are provided for use while reading a document (24). These styles currently include 1) highlighting only the phrase, 2) bolding the phrase and highlighting the sentence in which it appears, and 3) underlining the phrase. The reader may select the highlighting style at any time, and personalize the color schemes and styles in the reader profile.

Sensitivity Meter

Also represented in figure 2 is the *sensitivity threshold meter* (25). This meter can be used to set the threshold for the document area and Thumbbar, restricting the annotation to concepts scoring above the specified threshold. For instance, the concept Design currently has a score of 76%. The reader could use the sensitivity meter to increase the

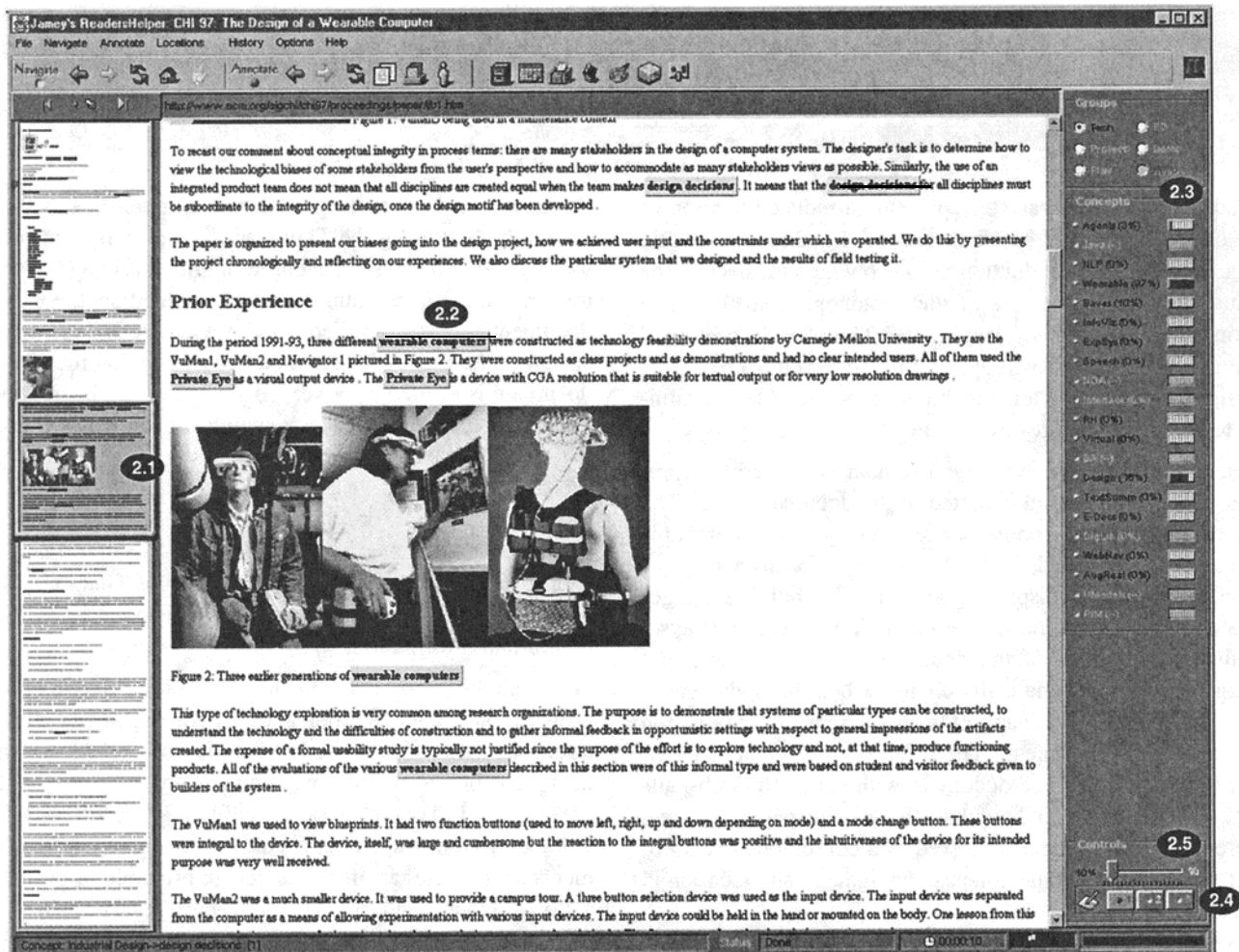


Figure 2 - An Annotated HTML Document

threshold to a value greater than 76% and all annotations associated with the Design concept would disappear from the display. This gives the user control of showing only the higher scoring concepts when reading a document, further supporting the concept of personalizing the document.

Summaries

The annotation agent automatically generates a summary of the document based on the concept hits found in the text. For instance, given the annotated document in figure 2, a user might want to read only the sentences that have hits in them for the concept Design. This document is generated automatically by the RH after it is done analyzing the document. Figure 3 shows a summary of the document annotated in figure 2. The toolbar button 3.1 invokes the summary function to present the reader with a list of available summaries. In figure 3 each sentence containing a keyword phrase for the concept Design is listed in the order of appearance in the document 3.2. This interface provides a way for the reader to view all relevant sentences at once. Associated with each sentence is a hypertext link 3.3 back into the same sentence in the annotated document. Another summary generated by the RH lists all key phrases for each concept and the number of hits each phrase has in the document. Each hit is represented as a hypertext link which points back to the location of the hit in the annotated document. The value of this summary is to show how broad the coverage is with respect to the concept's key phrases.

The use of RH outside the target document

The RH also provides access to an archive of previously annotated documents. Each time a document is annotated in the RH, a local copy of the document is stored in the reader's private document archive so that it can be accessed again.

Calendar Interface

Readers can use a calendar interface to the archive for retrieving documents based on their date of annotation. The calendar keeps track of all documents annotated by record-

ing the date, time and the most relevant concept on the day the annotation occurred. There are two ways to view the calendar: as a standard calendar month (days listed in columns from left to right, etc.) with entries on each day an annotation occurred, or as a *timeline* so that more information about the document can be presented to the user.

Similar Documents Interface

Another way to access the document archive is by viewing documents that are similar in content. This is possible using the *similar documents* interface. Each time a document is annotated, the RH annotation agent records how each concept in the reader's profile performed (i.e. the concept score is logged with respect to the document's unique ID). When a reader requests to see all documents which are similar to the current document in the browser, the system uses the concept information for the current document to generate a list of documents from the archive that have scored similarly. For instance, in figure 4 we see the current state of the concepts after annotating a document (4.1 & 4.2). The ExpSys 4.1 concept (Expert Systems) has a score of 80% and the NLP 4.2 concept (Natural Language Processing) has a score of 63%. Now the system must generate a collection of documents which have scored similarly to the way this document scored. In other words, *what documents in my collection or the group's collection have an emphasis on both expert systems and natural language understanding?* Because the RH annotation agent records all concept scores for each document annotation, the answer can be easily generated and presented to the reader.

Figure 4 shows the title of the document 4.3 previously annotated which generated the scores in the concept area: Adding Intelligence to the Interface. The similar document subsystem creates a list for each concept 4.4 of the documents which have scored above a user defined threshold (e.g. 50%). The list is composed of the title of the document which is a hypertext link to the document in the archive. From this list we can see that one document scored

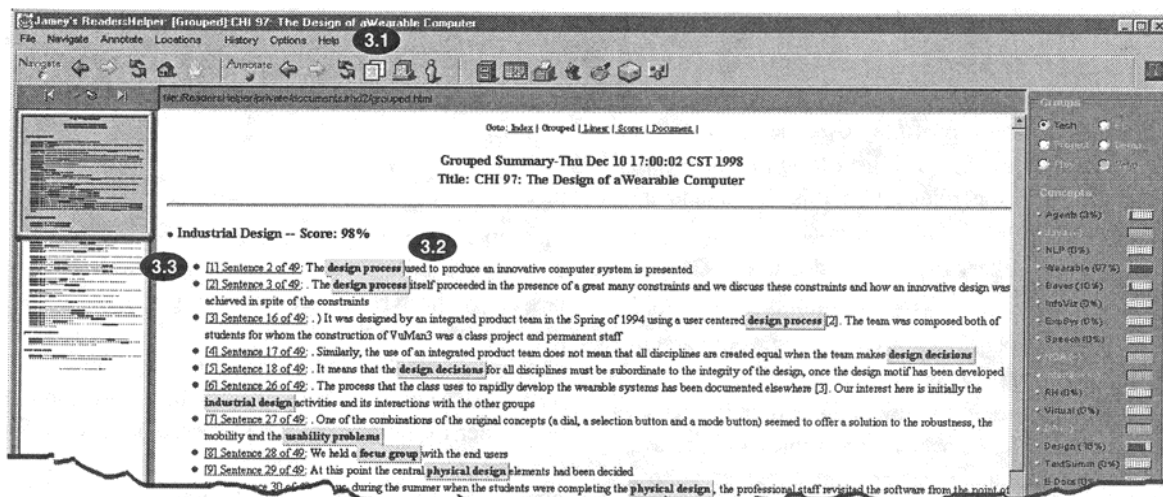


Figure 3 - Summary of sentences relating to the "Design" concept



Each page in the document contains six thumb-nail images of the surrounding five pages along with the current page in the left-hand margin (5.4). This allows the reader to look ahead in the paper document to view the surrounding pages and their relevance with respect to the topics of interest. The annotations in the document are presented as yellow boxes surrounding the key phrases (5.5). The page number

The RH uses reader *profiles* to define the topics of interest for each reader. The topics of interest are called *concepts* since they tend to represent a meaning which the reader would like the system to recognize when analyzing documents. Concepts are **defined** by a collection of keyword phrases which represent the overall topic of interest. For instance, if a reader is interested in the concept of Intelligent Agents, several keyword phrases can be used to define this topic: adaptive agents, personal assistants, Patti Maes, cooperative information agents. When a document is processed by the RH, a probability is generated for each concept representing the likelihood the document is about that concept. This likelihood is called *the concept similarity measure*. At present, concept definitions are hand-coded by the reader in the profile. The process is actually quite simple and only involves defining a name for a concept and populating it with the keyword phrases.

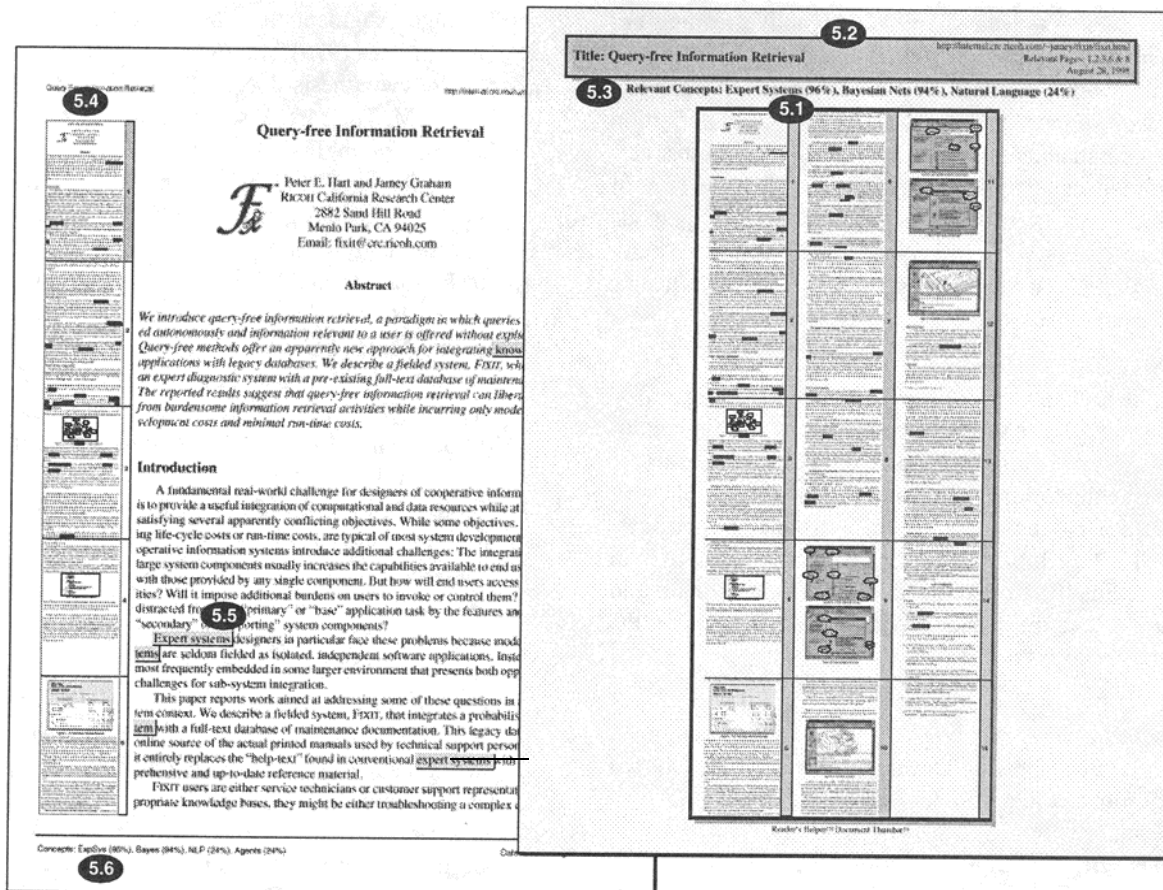


Figure 5 - The Paper-based RH showing the cover sheet and the first page from the document.

Probability-Based Content Recognition

The concept similarity measure is computed in a Bayesian belief network [11] based on keyword phrase frequency and the location and proximity of keywords in the document (figure 6). Most text summarization systems exploit the location of a sentence in a document [13], and for this reason the location of the **key-word phrase** similarly influences the probability of the concept it supports. The **phrase proximity** feature rewards hits occurring in close proximity to each other. This supposition is based on rewarding hits in areas of the document that might be emphasizing one of the reader's topics of interest. Even though the current three-feature concept belief network represents a somewhat simplified expression of content evaluation in a document, it has provided adequate results in our initial prototype. In the future we plan to experiment with more sophisticated content analysis methods by taking into consideration document structure and semantics and potentially the reader's browsing and reading history. The pattern matching is performed by an annotation agent using the key word phrases which define the reader's topics of interest. It

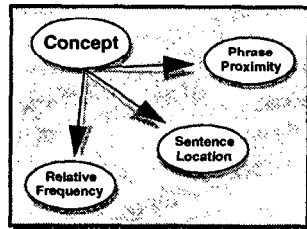


Figure 6 - A sample belief

is conducted before the document is parsed, as the stream of characters is received from the site hosting the document so that special tags can be added to the document to denote relevance and location in the text. The result is a version of the document automatically annotated to point out relevant areas of the document. A concept similarity value is computed at the end of this analysis and is immediately used to update the user interface so that the reader can see the concept scores while the document is being parsed and laid out in the browser.

IMPLEMENTATION

The RH system was implemented in Java on a PC running Windows NT 4.0. It also runs under the **Solaris** and Linux operating systems as well. It is HTML 3.2 compatible and can be used for regular browsing of the WWW.

CONCLUSION

With the emergence of the WWW, it is now quite clear that sophisticated online reading environments are not only useful but required for keeping pace with the increasing volume of text available in electronic form. The goal of this project was to design a system that makes reading electronic documents more efficient. Based on our experience using the RH, we feel that automatic document annotation, used in conjunction with the Thumbar, provides many benefits over the current state of the art. There is, of course, a lot of work

yet to be done. As discussed in [2,21], active reading not only requires identification of relevant text but writing on the text, i.e. manually annotating the document. Future RH systems will support manual annotations. The definition and automatic maintenance of reader profiles will most certainly provide challenges for researchers [3,16]. This of course is directly tied to the content recognition methods which also presents a challenge because of the complexity of understanding natural language and presenting a quantitative measure of relevancy.

ACKNOWLEDGEMENTS

I would like to thank my colleague Dr. David Stork for first proposing the Reader's Helper concept and for his continued brain storming support, especially in the areas of learning based methods for profile maintenance. I would also like to thank Dr. Ziming Liu for his assistance in discovering and distilling information regarding how people read documents. Finally, I'd like to thank Dr. Jonathan Hull and Dr. Peter Hart for their leadership and guidance during the course of this project.

REFERENCES

1. Adobe Acrobat, Adobe Acrobat 3.0 Technology: A Mini-White Paper for Developers. Available at <http://www.adobe.com/prodindex/acrobat/devwhitepaper.html>.
2. Andler, A., Gujar, A., Harrison, B., O'Hara, K. and Sellen, A. A Diary Study of Work-related Reading: Design Implications for Digital Reading Devices. *Proceedings of CHI '98* (Los Angeles, CA, April 1998), ACM Press, 241-248.
3. Balabanovic, M. and Shoham, Y. Learning Information Retrieval Agents: Experiments with Automated Web Browsing. *AAAI-95 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.
4. Ball, T. and Eick, S.G. Software Visualization in the Large. *IEEE Computer*, Vol. 29, No. 4, April 1996, pp. 33-43.
5. Bass, L., Kasabach, C., Martin, R., Siewiorek, D., Smailagic, A., and Stivoric, J. The Design of a Wearable Computer. *Online Proceedings of CHI '97* (Atlanta, GA, March 1997), <http://www.acm.org/sigchi/chi97/proceedings/paper/ljb1.htm>.
6. Birkerts, S., *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. Fawcett Books, November, 1995.
7. Boguraev, B., Kennedy, C. Bellamy, R., Brawer, S. Wong, Y.Y., Swartz, J. Dynamic Presentation of Document Content for Rapid On-Line Skimming., *AAAI Spring 1998 Symposium on Intelligent Text Summarization*, Stanford University, 23-25 March, 1998.
8. Darnton, R. Toward a history of reading. *Wilson Quarterly* 13(4): 87-102, 1989.
9. Elo, S. PLUM: Contextualizing News for Communities Through Augmentation. Master's Thesis, MIT Media Laboratory, (1995) <http://mu.www.media.mit.edu/people/elow/www/thesis-doc.html>.
10. Hearst, M. TileBars: Visualization of Term Distribution Information in Full Text Information Access. *Online Proceedings of CHI'95* (Denver, CO, May 1995), http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/mah_bdy.htm.
11. Heckerman, D. and Wellman, M.P. Bayesian Networks. *Comm. ACM*, Vol. 38, No. 3, March, 1995, pgs. 27-30.
12. Spring, M. B., Morse, E. and Heo, M. Multi-level Navigation of a Document Space. *Leveraging Cyberspace Conference* (Xerox PARC, Palo Alto, CA, October 1996), <http://www.lis.pitt.edu/~spring/mlnds/mlnds/mlnds.html>.
13. Kupiec, J., Pedersen, J., and Chen, F. A trainable document summarizer. *Proceedings of the 18th ACM-SIGIR Conference*, 1995.
14. Levy, D. I Read the News Today, Oh Boy: Reading and Attention, *ACM Digital Libraries*. Philadelphia, PA, USA, pgs. 202-211, 1997.
15. Mahesh, K. Hypertext Summary Extraction for Fast Document Browsing. *AAAI-97 Symposium on Natural Language Processing for the World Wide Web*, Stanford University, March 24-26, 1997.
16. Moukas, A., and Maes, P. Amalthaea: An Evolving Multiagent Information Filtering and Discovery System for the WWW. *First issue of the Journal of Autonomous Agents and Multi-Agent Systems*, 1998.
17. Nevill-Manning C.G., Witten I.H. and Paynter G.W. Browsing in digital libraries: a phrase-based approach. *Proc. Digital Libraries '97*, Philadelphia (1997), 230-236.
18. O'Hara, K. and Sellen, A.J. A Comparison of Reading Paper and On-line Documents. *Online Proceedings of CHI '97*, (Atlanta, GA, March 1997), <http://www.acm.org/sigchi/chi97/proceedings/paper/koh.htm>.
19. *Online Proceedings of CHI '97*, available at <http://www.acm.org/sigchi/chi97>.
20. Peairs, M. Iconic Paper. *Proceedings of the Third International Conference on Document Analysis and Recognition*, Montreal, Canada, August 14-16, 1995, 1174-1179.
21. Schilit, B. Golovchinsky, G. and Price, M. Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations. *Proceedings of CHI '98* (Los Angeles, CA, April 1998), 249-256.
22. Standing, L., Conezio, J. and Haber, R.N. Perception and memory for pictures: *Single-trial learning of 2500 visual stimuli*, *Psychon. Sci.*, 1970 Vol. 19 (2).
23. Tufte, E.R. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, 1990.
24. Wroblewski, D. and Hill, W.C. Attribute-mapped Scroll Bars. U.S. Patent Number 5,479,600, December 26, 1995.

The Reader's Helper: A Personalized Document Reading Environment (Page 481)

Jamey Graham

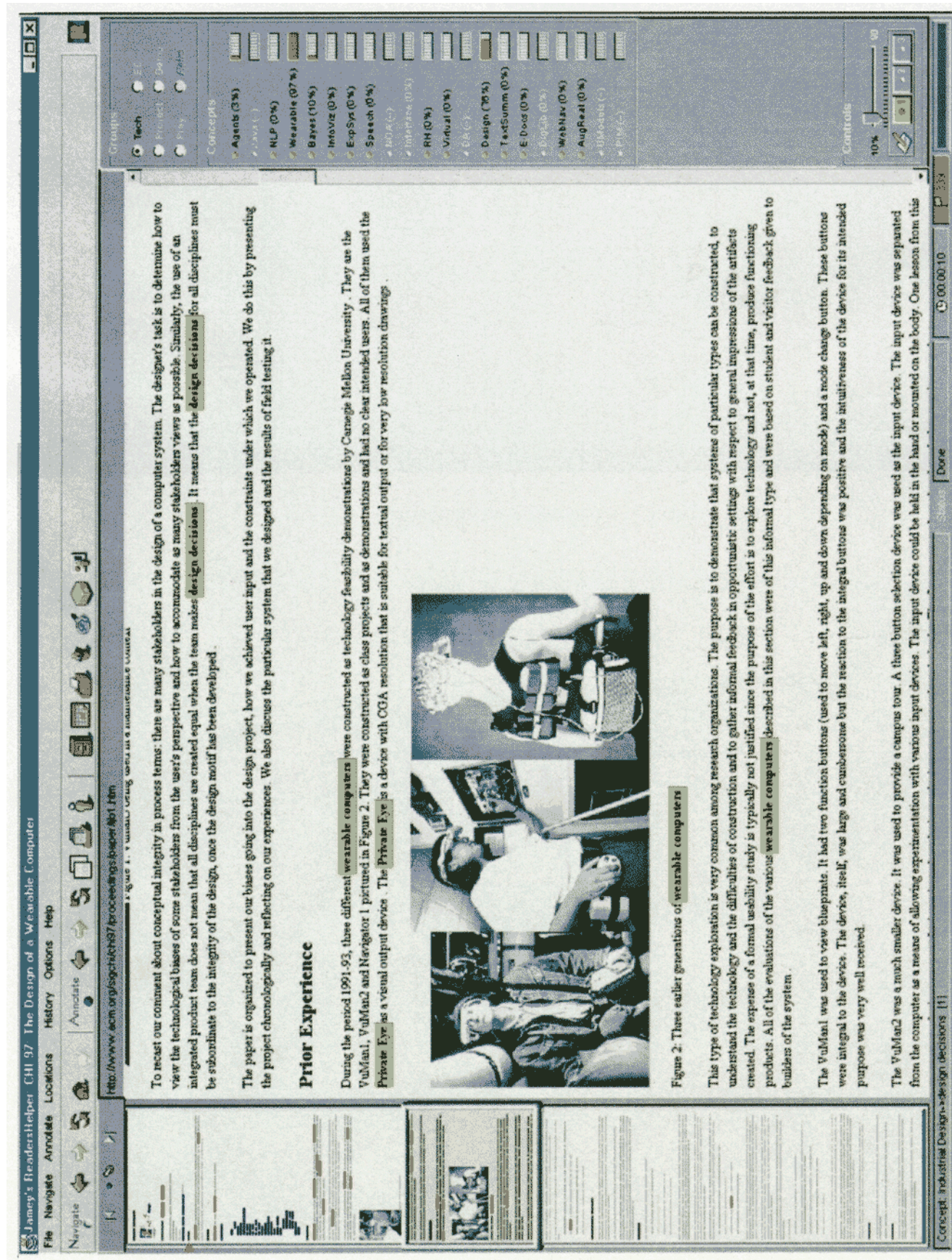


Figure 2 - An Annotated HTML Document