Toward Zero-Effort Personal Document Management

Integrating document capture into daily work practices and exploiting familiar retrieval methods are the keys to Save everything!—a radically new approach to personal document management.

Jonathan J. Hull

Peter E. Hart
Ricoh
California
Research
Center

Deciding what to keep and discard is a problem as old as civilization itself. When we clean drawers or closets we decide whether saving something is worth the space, time, and effort required for storage and retrieval. In the office, the problem is how to store information efficiently and economically, especially documents.

The first decision is whether to save a document. According to one study,[1] only 15 percent of filed documents are ever referred to again. However, if we discard a document, it might contain valuable information we need in the future. If we elect to save a document, we must consider how and where to file it for easy retrieval. It's commonly estimated that 3 to 5 percent of documents are lost at any one time; to a Fortune 1000 company, this translates into an annual loss ranging from $3 million to $5 million.

An effective solution to the problem of personal document management in the Information Age must satisfy several criteria:

- The personal effort required to capture and file a document must be nearly zero.
- The cost of document storage must be economically reasonable.
- Document retrieval must be at least as efficient as conventional methods.
- Users must have confidence in the security and privacy of stored documents.

During the past three years, the California Research Center (CRC) has evaluated a radically simple approach to personal document management called Save everything![2] To achieve zero-effort document capture, we modified conventional printers, digital copiers, and fax machines to automatically save representations of every document processed. Our approach exploits users' familiarity with their personal document collection to retrieve documents without using filenames, paths, or keywords.

A NEW PARADIGM

A zero-effort document management system should enable users to acquire and file documents automatically without even thinking about saving this information. Digital documents can, of course, be automatically captured by straightforward software modifications to users' personal computers; hence, we did not focus on capturing e-mail, spreadsheets, Web pages, and other documents from computer sources.

Capturing paper documents with minimal effort, however, is a challenge. Although scanners, which have been available at reasonable prices for many years, would seem to offer an obvious solution for capturing page images, they are not commonly used in offices. Even many of the attendees at a seminar[3] on document analysis and retrieval did not routinely scan documents. Thus, scanners are not the route to zero-effort document management.

However, using a family of paper-handling office machines—printers, copiers, and fax machines—is part of nearly everyone's daily work practice. We can modify these devices to support automatic document capture and equip them with a large memory and a virtually infinite backing store to create an *infinite memory multifunction machine* (IM^3).

Storage cost

An obvious practical obstacle to the IM^3 is storage cost. To address this concern, we compared the cost of storing a digital page image with the cost of paper. We assumed that each sheet of paper costs about a penny and that the image of a typical bilevel office document compresses to around 100 Kbytes.

Whereas the cost of paper has been stable, the price performance of conventional, rotating magnetic storage media has plummeted in the past three decades.[4,5] Our informal analysis indicates that disk price-performance has doubled every year during this 30-year period. This record handily beats Moore's law, which states that semiconductor-price performance doubles every 18 to 24 months. By the end of 2000, the retail price of disk drives was approaching 0.4 cent per Mbyte, or 0.04 cent per 100 Kbytes, which means that storing a piece of paper costs 25 times more than storing a page image.

Obviously, this analysis ignores other factors such as the costs for file cabinets, office space, and maintaining computer systems. Our principal concern, however, is scaling—what happens when the demand for "infinite" memory becomes finitely very large—and, on this point, we are satisfied that IM^3 storage costs are economically reasonable.

Document retrieval

The most challenging research issue facing the IM^3 is how to easily retrieve automatically captured documents without filenames, a visible path or file directory, or assigned keywords. We can use an optical character recognition (OCR) front end for document bitmaps captured from copiers and fax machines to augment the well-developed, conventional indexing and retrieval technology typically used for searching a large full-text database. However, this method is too limiting for a personal document management solution. Information about a user's acquisition method or relevant activities at the time of data acquisition provides personal metadata that creates opportunities for inventing novel retrieval methods.

Security and privacy

We conducted a study of CRC's user population to evaluate the combined conventional and novel retrieval methods of the IM^3. Prior to the study, we discussed privacy and security issues with the CRC users, who expressed concern about the controllability of the document-capturing process, whether password-controlled access would provide sufficient protection, and the availability of permanent file deletion. We also considered the problem of whether users could manually override the default capture of all documents.

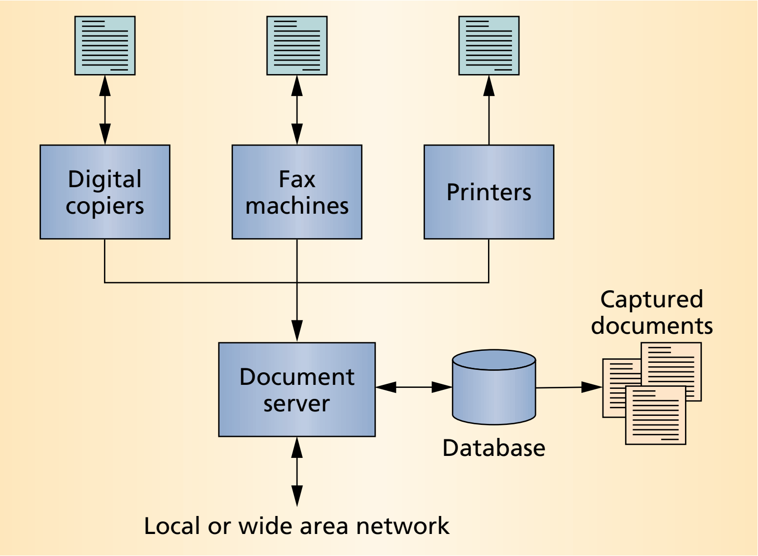


Figure 1. The infinite memory multifunction machine (IM^3) system. Digital copiers, fax machines, and printers automatically save a copy of every document they receive on a server that indexes them for later retrieval.

The IM^3 gives users control over what documents they capture, whether their documents are encrypted, and who can access them. The default print spoolers capture every job. A one-touch selection from the copier touch pad activates or deactivates the document-capturing feature. Users can also select password protection only or password protection plus encryption using PGP. Finally, a user can open an IM^3 database to other workgroup members.

IM^3 SYSTEM DESIGN

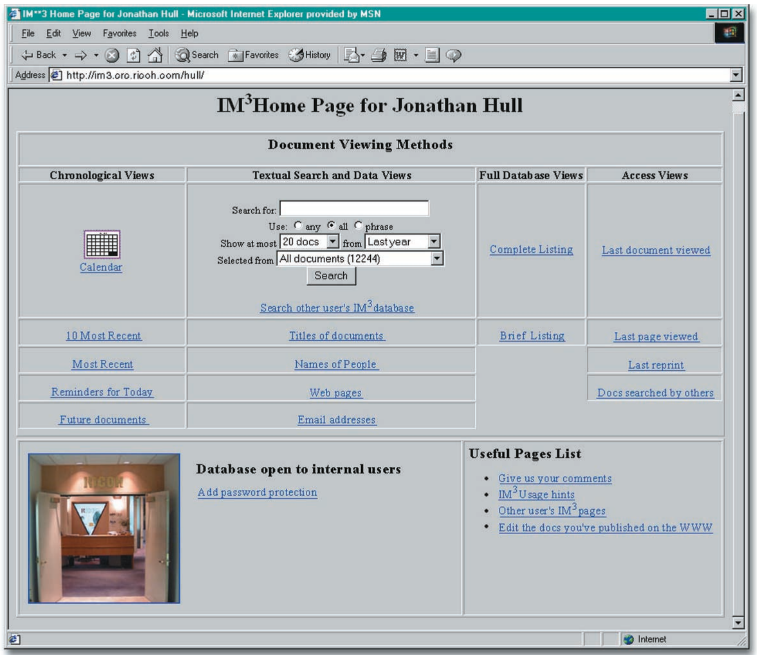
For our study, we modified conventional printers, digital copiers, and fax machines to automatically capture an image of every processed document. Aside from users of copiers identifying themselves by pressing a button on a touch screen, this process is effortless and hidden from view. The IM^3 transfers captured images to the document server for permanent storage and indexing for later retrieval. Figure 1 illustrates the IM^3 system's basic design.

Software

Software running on a Unix print server automatically captures PC, Apple computer, and Unix workstation print jobs. A filter in the printer's spooling system transfers a copy of every printed document to the document server independent of any application software. To index saved documents, software applies OCR to photocopier images and extracts text from the Postscript files of printed documents. The software chooses keywords for each document and builds data structures for full-text retrieval. The software calculates thumbnail images at 4 dpi, 8 dpi, and 72 dpi for use in various browsable interfaces.

Retrieval interfaces

A Web server that runs on the document server provides platform-independent document retrieval.


IM**3 Home Page for Jonathan Hull - Microsoft Internet Explorer provided by MSN
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://im3.crc.ricoh.com/hull/
Internet

IM^{3}Home Page for Jonathan Hull

Document Viewing Methods			
Chronological Views	Textual Search and Data Views	Full Database Views	Access Views
Calendar	Search for: Use: ○ any ● all ○ phrase Show at most 20 docs from Last year Selected from All documents (12244) Search Search other user's IM^{3}database	Complete Listing	Last document viewed
10 Most Recent	Titles of documents	Brief Listing	Last page viewed
Most Recent	Names of People		Last reprint
Reminders for Today	Web pages		Docs searched by others
Future documents	Email addresses		

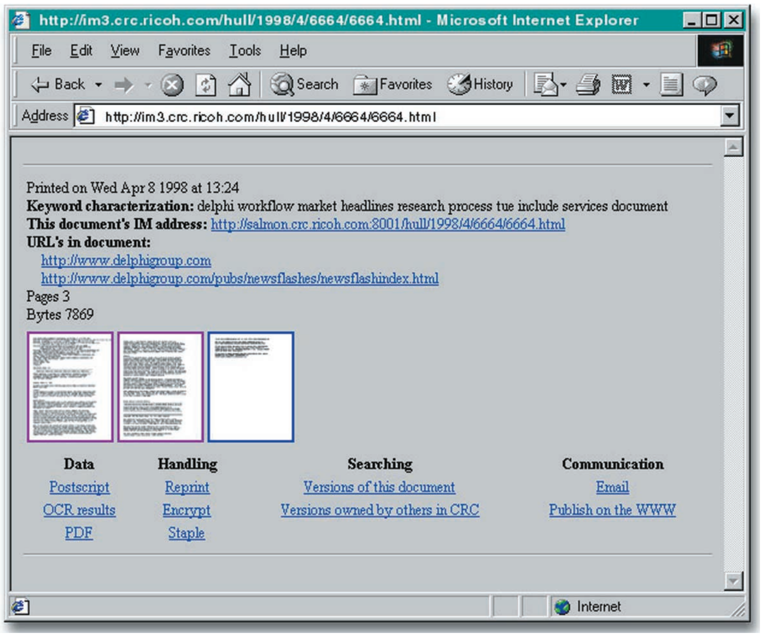
Database open to internal users

Add password protection

Useful Pages List

- Give us your comments
- IM^{3}Usage hints
- Other user's IM^{3}pages
- Edit the docs you've published on the WWW

Figure 2. IM^3 home page. The search and retrieval interface provides several chronological views of the user's documents, conventional full text search, and lists of extracted data.


http://im3.crc.ricoh.com/hull/1998/4/6664/6664.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://im3.crc.ricoh.com/hull/1998/4/6664/6664.html
Internet

Printed on Wed Apr 8 1998 at 13:24
Keyword characterization: delphi workflow market headlines research process tue include services document
This document's IM address: http://salmon.crc.ricoh.com:8001/hull/1998/4/6664/6664.html
URL's in document:
http://www.delphigroup.com
http://www.delphigroup.com/pubs/newsflashes/newsflashindex.html
Pages 3
Bytes 7869

Data	Handling	Searching	Communication
Postscript	Reprint	Versions of this document	Email
OCR results	Encrypt	Versions owned by others in CRC	Publish on the WWW
PDF	Staple		

Figure 3. Example of captured document representation, with various data formats and operations enabled. The 8-dpi thumbnails are hotlinked to 72-dpi thumbnails on the corresponding page.

As Figure 2 shows, the IM^3 home page provides information about saved documents. The home page serves as a search and retrieval interface that provides two chronological views of each user's documents: a calendar view and a list of the 10 most recently saved documents. It also provides conventional full-text search and lists of various data extracted from the documents. Users can browse a list of all saved documents and directly access either the most recently viewed document or the one most recently processed.

Figure 3 illustrates how stored documents appear to users. The interface indicates whether the document was printed or copied and specifies the date and time when the information was captured. It also displays 10 automatically extracted keywords so users can quickly determine whether a document is relevant.

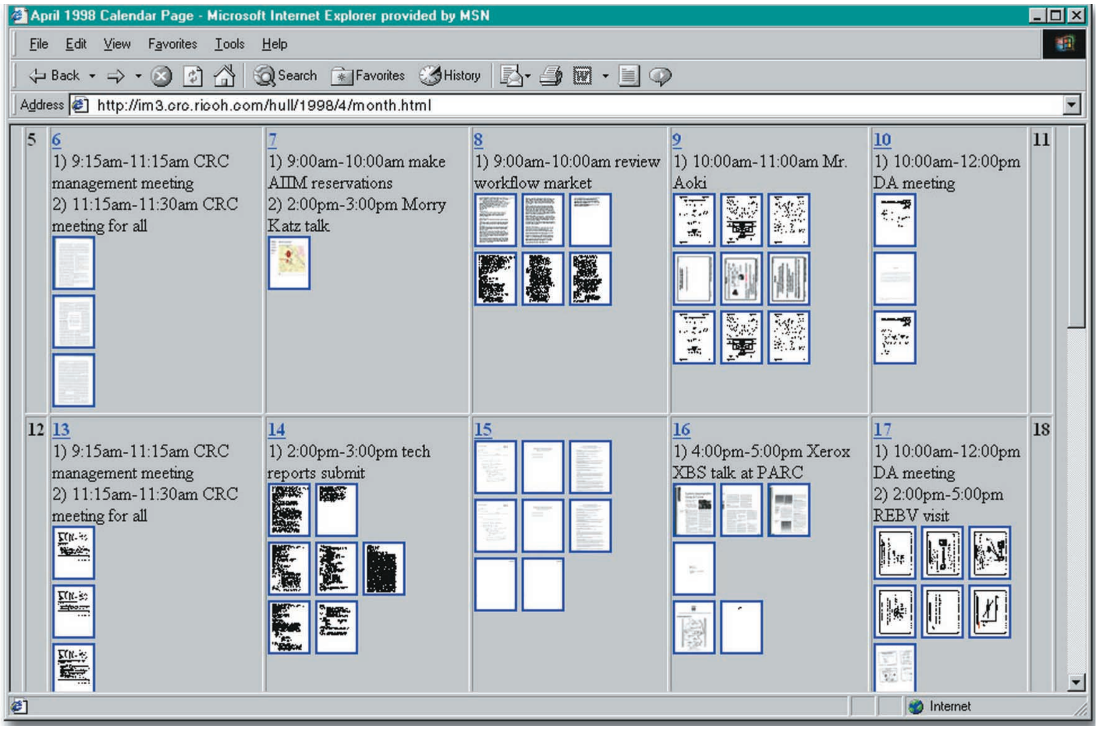
Documents are available in various formats. Users can access the original Postscript file, compressed with gzip, of printed documents as well as extracted ASCII text. Copied documents are also available in Postscript form as a 400-dpi binary image compressed with CCITT group 4 and with the appropriate Postscript header. Thumbnails are generated in GIF format for every page at 4 dpi, 8 dpi, and 72 dpi. OCR results are also saved for copied documents.

Figure 4 illustrates the calendar retrieval interface. When a user inserts a new document, the system generates a calendar for the current month. This process automatically extracts the appointments users have recorded with their Unix calendar manager software and places them in the appropriate cell of the display. The display cell also contains 4-dpi thumbnails for the first three pages of the last three documents the document server received on that day. Users can click on one of the thumbnails to navigate directly to a particular document. Alternatively, users can click on a day to view a list of all documents processed on that day.

The following example illustrates the use of the calendar retrieval interface. Suppose a user is looking for an e-mail message about workflow systems. He remembers printing the document when he attended a meeting on this topic in early April. Browsing through his calendar page for April, he sees an appointment to "review workflow market" on 8 April. The image thumbnails shown in this calendar's cell reveal two images with the texture pattern characteristic of an e-mail message. Clicking on one of these thumbnails yields the desired document.

Security

Users want to share documents with one another and, at the same time, guarantee that fully searchable and retrievable confidential documents remain secret. To satisfy these requirements, we adopted a two-tier security model. The simplest method allows users to designate their document collection as either open or closed. The Web server allows access if a collection is open, but a user must enter the appropriate password to gain access to a closed collection. Pressing the "Encrypt" link, as Figure 3 shows, permits users to encrypt individual documents with the PGP algorithm. By not modifying the full-text index, the system retains the capability to locate encrypted documents using full-text search.


April 1998 Calendar Page - Microsoft Internet Explorer provided by MSN
File Edit View Favorites Tools Help
Back
Search
Favorites
History
Address
Internet

http://im3.crc.ricoh.com/hull/1998/4/month.html

5	6 1) 9:15am-11:15am CRC management meeting 2) 11:15am-11:30am CRC meeting for all	7 1) 9:00am-10:00am make AIIM reservations 2) 2:00pm-3:00pm Morry Katz talk	8 1) 9:00am-10:00am review workflow market	9 1) 10:00am-11:00am Mr. Aoki	10 1) 10:00am-12:00pm DA meeting	11
12	13 1) 9:15am-11:15am CRC management meeting 2) 11:15am-11:30am CRC meeting for all	14 1) 2:00pm-3:00pm tech reports submit	15	16 1) 4:00pm-5:00pm Xerox XBS talk at PARC	17 1) 10:00am-12:00pm DA meeting 2) 2:00pm-5:00pm REBV visit	18

Figure 4. The calendar retrieval interface displays captured document thumbnails in context with events that occurred when the documents were saved.

CASE-STUDY METHODOLOGY

Save everything! raises several obvious questions about storage requirements, the usefulness of different document sources, and the relative popularity of various retrieval interfaces. For example, will users already familiar with full-text search from their interaction with Web search engines readily adopt the calendar retrieval technique?

Any study to answer these and similar questions must first identify suitable workgroup populations. From a document-handling viewpoint, we can generally classify workgroups as either *structured* or *unstructured*. A structured workgroup deals with relatively few document types, regardless of volume. A human resources activity, which handles many resumés and personnel files, is an example of a structured workgroup. An unstructured workgroup, such as an academic or corporate research center like CRC, deals with many types of documents.

Throughout the study, we installed a series of prototype machines in our CRC workgroup until the entire system became "IM^3-enabled." In addition to maintaining detailed usage statistics for the past three years, we conducted user interviews during the first year to gather feedback.

CASE-STUDY RESULTS

From March 1996 to November 1999, the CRC workgroup's IM^3 captured 72,492 documents totaling 295,981 pages. On average, printed documents contained 4.3 pages, and copied documents contained 3.2 pages. The complete data set includes 83 percent printed documents and 17 percent copied documents.

Storage space

We reviewed data saved from May 1998 to November 1999 to compare the storage required for printed and copied documents. During that time, the system saved every printed document and two IM^3 copiers were available. Overall, we captured 38,384 documents comprising 152,251 pages and saved them in 12 Gbytes of disk space. Figure 5 illustrates how much storage space the various formats consumed. The Postscript files used 61 percent of the space for printed documents and consumed 76 percent of the storage used for copied documents. These high percentages were not surprising given that Postscript files allow perfect reprinting of the original document.

Document sources

To determine the source of captured documents the workgroup later accessed, we analyzed the Web server logs from May to October 1999. On average, our users accessed their personal IM^3 document stores every other day. Of the 1,182 high-resolution page images collectively viewed during this time, 33 percent were originally copied documents, and the rest were printed. Thus, copied documents were accessed roughly twice as often as their frequency in the population would predict.

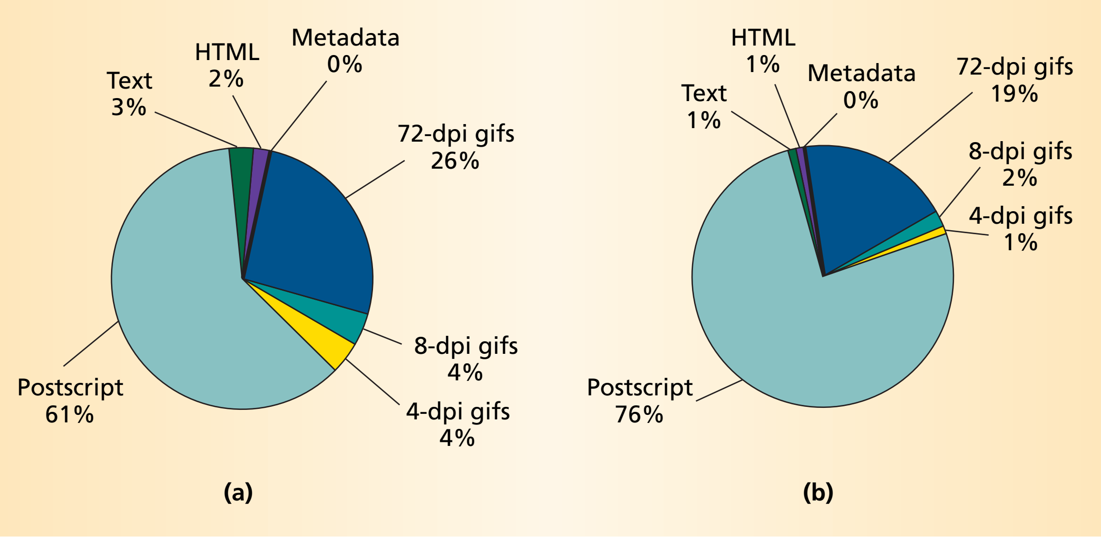


Figure 5. Storage required for the various formats of printed and copied documents captured between May 1998 and November 1999 at the California Research Center. (a) 25,507 printed documents—7 Gbytes, and (b) 8,877 copied documents—5 Gbytes.

The results confirmed our initial assumption that people use copied documents more frequently than printed ones. Because all print documents had a symbolic source file at one time somewhere in the workgroup, however, the relatively high utilization rate for printed documents was surprising. Users evidently could more easily retrieve a desired print document from the IM^3 database than from their own file directory.

The average age of page accesses—that is, the difference between creation and access time—was 44 days. Interestingly, 38 percent of accesses were to one-week-old documents, and 10 percent of accesses were to documents older than six months. These statistics suggest that old documents are indeed useful and that they should be accessible. Because automatic capture almost guarantees that a document will be present in the future when a user needs it, the IM^3 database became more useful to the CRC workgroup over time.

Retrieval interfaces

The CRC workgroup quickly embraced the time-based retrieval interfaces that exploited users' familiarity with their personal document metadata. Indeed, the group used these novel retrieval interfaces 10 times as often as the traditional text-based methods that were prominently available on each user's IM^3 home page. The size of this disparity surprised us and suggests that these new retrieval interfaces provide real value to users.

RELATED WORK

Personal document management systems that use multiple technologies provide fertile ground for exploring new methods for data capture and retrieval. Despite its similarity to other approaches, Save everything! is unique. For example, in one method users manually categorize documents before scanning.[6] Other work in document image storage and retrieval systems emphasizes improving the accuracy of image analysis, OCR, and indexing that occur after document scanning.[7] Both approaches contrast sharply with our emphasis on eliminating an explicit scanning step, which we believe actually inhibited common use of previous systems.

Some novel IM^3 retrieval methods resemble innovations in user-interface design[8] and information visualization.[9,10] Our focus, however, is on presenting information about documents in ways that make it easy for users to find what they want rather than on representing document content. Another related project gathered personal metadata through special badges worn by users,[11] while our approach requires little change in user behavior.

The IM^3 system in our study captured only documents with a paper source (copier, fax machine), or paper destination (printer). Ricoh's eCabinet, which implements Save everything!, also captures computer documents (see http://www.rsv.ricoh.com). There are many opportunities to improve and extend the first-generation document-capture capabilities we have developed thus far. For example, as a document corpus becomes very large, we will need a means for automatically identifying multiple versions of the same document.

Given the extended time and large volume of documents captured in our study, we are confident that our conclusions about using this system are equally applicable to other unstructured workgroups. In addition, we speculate that by augmenting the IM^3 with structured filing capabilities, Save everything! also offers an effective paradigm for structured workgroups. ✷

Acknowledgments

We thank many individuals who contributed to developing the IM^{3}system, including Michael Baxter, John Cullen, Jamey Graham, Toshio Kanoh, Dar-Shyang Lee, Ziming Liu, Alex Ono, Mark Peairs, Robert Runyon, and Kiyoshi Suzuki.

References

1. M.D. Gordon, "It's 10 A.M.: Do You Know Where Your Documents Are? The Nature of Information Retrieval

Problems in Business," *Information Processing and Management,* vol. 33, no. 1, 1997, pp. 107-121.

2. J.J. Hull and P. Hart, "The Infinite Memory Multifunction Machine," *Proc. 3rd IAPR Workshop Document Analysis Systems,* IAPR, Nagano, Japan, pp. 49-58.
3 H. Fujisawa and H. Stabler, "Needs of the Market and User Requirements," *Document Analysis Systems,* A.L. Spitz and A. Dengel, eds., World Scientific, River Edge, N.J., 1995, pp. 452-454.
4. M. Lesk, *Practical Digital Libraries,* Morgan Kaufmann, San Francisco, 1997.
5. H. Paulapuro, "The Future of Paper in the Information Society," *The Electronic Library,* June 1991, pp. 135-143.
6. R. Rao et al., "Protofoil: Storing and Finding the Information Worker's Paper Documents in an Electronic File Cabinet," *Proc. Conf. Computer-Human Interaction* (CHI 94), ACM Press, New York, 1994, pp. 180-185.
7. G.A. Story et al., "The RightPages Image-Based Electronic Library for Alerting and Browsing," *Computer,* Sept. 1992, pp. 17-26.
8. R. Wilensky, "Toward Work-Centered Digital Information Services," *Computer,* May 1996, pp. 37-44.
9. J. Lamping, R. Rao, and P. Pirolli, "A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies," *ACM Conf. Human Factors Software,* ACM Press, New York, 1995, pp. 401-408.
10. J.D. Mackinlay, G.G. Robertson, and S.K. Card, "The Perspective Wall: Detail and Context Smoothly Integrated," *Proc. Conf. Human-Computer Interaction* (CHI 91), ACM Press, New York, 1991, pp. 173-179.
11. R. Want et al., "The Active Badge Location System," *ACM Trans. Information Systems,* Jan. 1992, pp. 91-102.

Jonathan J. Hull *is the leader of the Multimedia Document Analysis Group at the Ricoh California Research Center. His research interests include document analysis, computer vision, and information retrieval. He received a PhD in computer science from the State University of New York at Buffalo. Hull is a Fellow of the IAPR and a member of the ACM and the IEEE Computer Society. Contact him at hull@crc.ricoh.com.*

Peter E. Hart *is Senior Vice President of Ricoh Company Ltd. and Chair and President of Ricoh Innovations Inc. His research interests include information appliances and multimedia document analysis. He received a PhD in electrical engineering from Stanford University. Hart is a Fellow of the IEEE and the AAAI and a member of the ACM and the IEEE Computer Society. Contact him at hart@rii.ricoh.com.*