THE INFINITE MEMORY MULTIFUNCTION MACHINE (IM^3)

JONATHAN J. HULL AND PETER HART
Ricoh California Research Center,, Menlo Park, CA
E-mail: hull@crc.ricoh.com

A complete, working document storage and retrieval system is described. Designed to help users solve the problem of lost documents, this system illustrates the concepts of automatic document capture and easy retrieval. *Every document* a user copies or prints is automatically indexed and saved for later retrieval. A prototype implementation of such a system was constructed and used daily by approximately 20 people for over two years. The design of this system is presented and experimental results are discussed.

1 Introduction

Lost documents are a significant problem for office workers. A recent study estimates that at any time 3% to 5% of documents are lost and the average cost of lost documents to a Fortune 500 company is in the range of $3 million to $5 million [10]. An obvious solution to this problem is for users to scan all their documents. However, for any number of reasons, not the least of which are users' natural reluctance to alter their work practices, such an approach has not been widely adopted [3]. In fact, recent results indicate that for new retrieval techniques to gain wide acceptance they should be easy to use, be familiar and require as little user effort as possible [6].

This paper proposes a system design called the Infinite Memory Multifunction Machine (IM^3) in which *every* document a user copies, prints, or faxes is automatically captured and indexed for later retrieval with a web browser. The automatic capture process is performed as a natural side-effect of copying, printing, or faxing and is almost completely transparent to the user. This removes any need for the user to decide at the time a document is processed whether it should be saved. By so doing, users are almost guaranteed that when they need to find a document, the system will contain a copy of it.

Economic considerations are always important factors when users decide whether to adopt new technologies. An important consideration in the design of the IM^3 system was the relative cost of printing a document on paper vs. storing an image of the same document on magnetic disk. Of course, there is a wide variation in the price of paper and toner needed to print the range of documents encountered in the typical office. For the purposes of this analysis, it was assumed that, on average, the cost of an 8.5x11 inch sheet of paper is one cent. It was also assumed that a 400 dpi binary image of the same document on average would require 100 KB. At the time the IM^3

Appeared in Proceedings of the Third IAPR Symposium on Document Analysis Systems, Nagano, Japan, Nov. 4-6, 1998.

project was started (late 1993) it was observed that it cost about 3 cents for 100 KB of magnetic disk storage. This was just for the disk space. It did not take into account the cost of the computer, etc. However, we projected that over time these costs would significantly decrease and eventually become less than the cost of a sheet of paper. That time arrived sometime in 1996. Today, it costs about 0.27 cents for 100 KB of disk space. This is significantly less than the cost of a sheet of paper. The 4:1 difference in cost of the two media now favors the adoption of a document storage and retrieval system like the IM^3.

Easy-to-use methods for retrieval of stored documents are also a key factor for success in user acceptance of such a system. This is especially the case since by reducing the filing-time effort almost to zero, users may need to perform additional tasks at retrieval time. Our objective is to reduce this overhead as much as possible.

Because of the relatively large volume of documents that are captured, it was deemed necessary to provide multiple, complementary retrieval techniques. Following the themes of ease-of-use and familiarity, retrieval interfaces were created that utilize paradigms users encounter daily. For example, one method displays thumbnails of document images in a calendar along with appointments. Users can retrieve documents by browsing the calendar for images that were saved at some time near that of an appointment. Other retrieval techniques that have been the subject of our research include image-based methods for document retrieval [1]. The detection of duplicate documents is another obvious concern, both for storage reduction and version detection in a user interface. Text-based and image-based techniques for duplicate detection have been developed [4, 5, 7].

Several document image storage and retrieval systems relevant to the IM^3 approach are described in the literature. An approach for maintaining a personal database of scanned scientific papers has been described [8]. An approach designed for storing and retrieving scientific papers in a client-server environment was discussed in [9] and the need for sophisticated retrieval techniques was presented in [2].

The rest of this paper describes a prototype implementation of the IM^3 system concept. The copiers and printing system in an office with about 20 users were modified so that every document they processed could be automatically saved. Captured documents were stored on a central server and made available for retrieval with a web browser. Provisions were made for selective encryption and security of users' databases in a way that promoted sharing information among users. A variety of retrieval and document communication interfaces were provided. User interactions with the retrieval interfaces were recorded and analyzed to determine the popularity of the different interfaces.

2 System Design

An outline for the design of the IM^3 system is presented in Figure 1. Specially modified digital photocopiers were developed that automatically capture an image of every copied document. Aside from a user identifying himself by pressing a button on a touchscreen, this process is completely transparent. The captured images are transferred to the document server where they are permanently stored and indexed for later retrieval.

Print jobs are automatically captured by software running on a Unix print server. A copy of every printed document is transferred to the document server as it is sent to a printer. This is done by a filter in the spooling system that is applicable to jobs printed on PC's, Apple computers, and Unix workstations. In this way the capture of printed documents is completely transparent to the user and is independent of any application software. Every document sent to printers serviced by the Unix server is saved.

An indexing process is applied to every saved document. The images from the photocopiers are OCR'd. Text is extracted from the postscript files for printed documents. This is used to choose keywords for each document and build data structures for full text retrieval. Thumbnail images are also calculated at several resolutions (4 dpi, 8dpi, and 72 dpi) for use in various browsable interfaces.

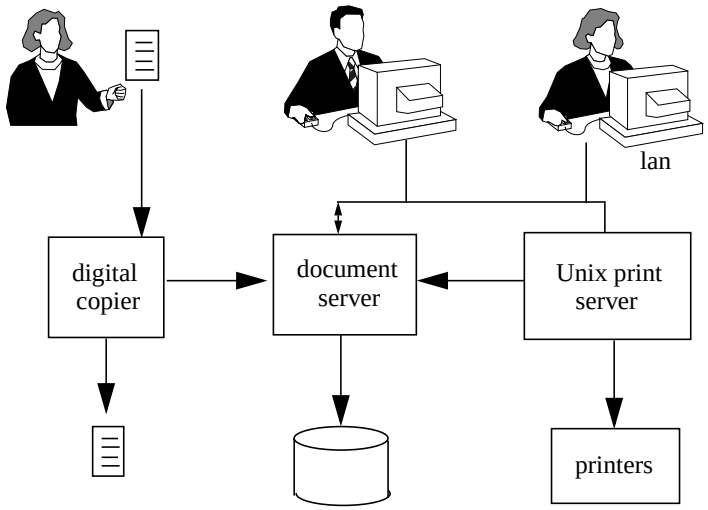

Figure 1. IM^3 System Design

Access to stored documents is through a web server running on the document server. This provides a platform-independent technique for document retrieval in a format (the web browser) that is already familiar to most users. Each user has their own home page which provides an entry point into their collection of saved documents.

The retrieval interfaces available to each user are illustrated by Figure 2 which shows an example of one user's home page. Two chronological views of each user's

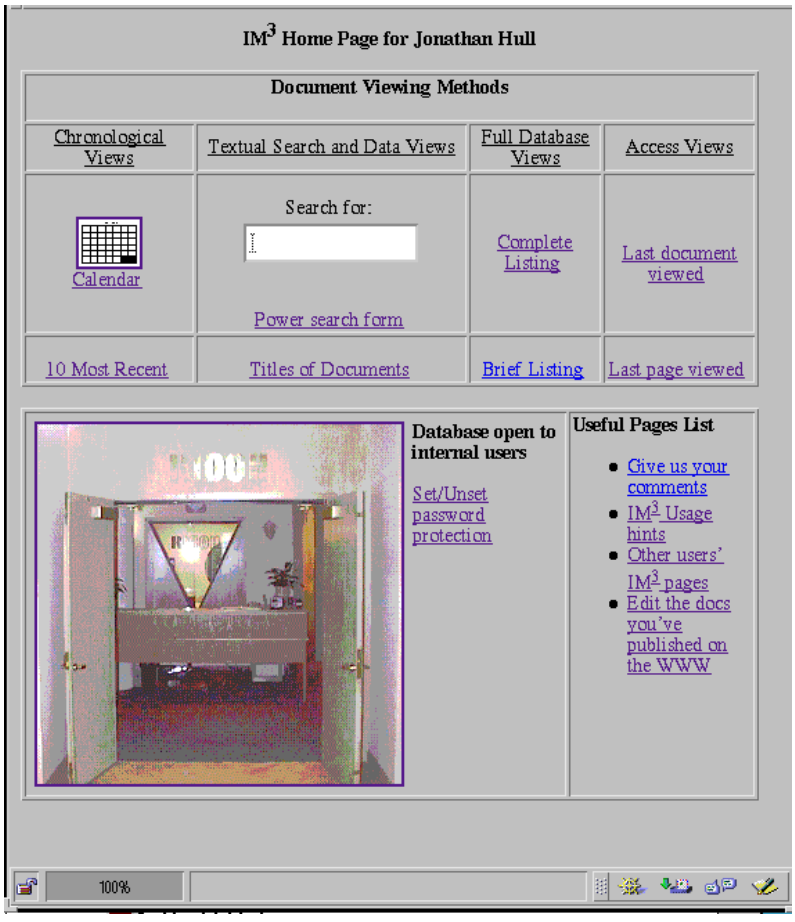

Figure 2. One user's IM^3 home page

documents are provided. These include a calendar view as well as a listing of the ten most recently saved documents. Conventional text search is also provided as well as listings of various data extracted from the documents. Users can also browse a listing of all the documents that have been saved for them. They can also directly access the document they most recently viewed or the document they most recently processed (e.g., re-printed).

Figure 3 shows an example of how stored documents appear to users. The source of the document (printed or copied) is given as well as the date and time when it was captured. Ten keywords are shown that were automatically extracted from the document. They provide a way for users to quickly determine whether a given document is relevant to their interests. The 8 dpi thumbnails shown in Figure 3 are hot-linked to 72 dpi thumbnails of the corresponding page.

Various formats are stored for each document. These include the original postscript file for printed documents. Copied documents are also stored in postscript form as a 400 dpi binary image compressed with CCITT group 4 with the appropriate postscript header. The postscript file is compressed with gzip. Thumbnails are generated

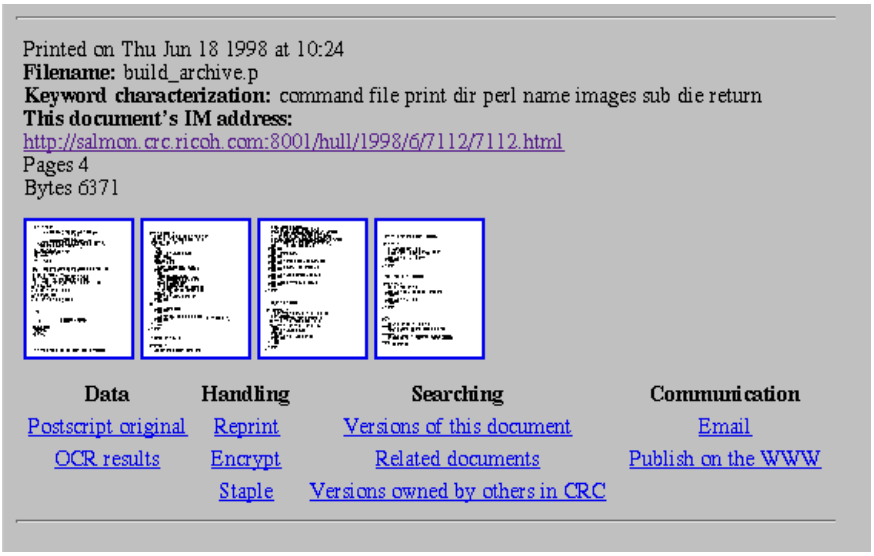

Figure 3. Example of a document showing the operations that can be applied to it.

in GIF format for every page at 4 dpi, 8 dpi, and 72 dpi. OCR results are also saved for copied documents. ASCII text is extracted from printed documents.

The calendar retrieval interface is illustrated in Figure 4. Every time a new document is inserted in the system, the calendar for the current month is generated. This process automatically extracts the appointments users have recorded with their Unix calendar manager software. These data are placed in the appropriate cell of the display

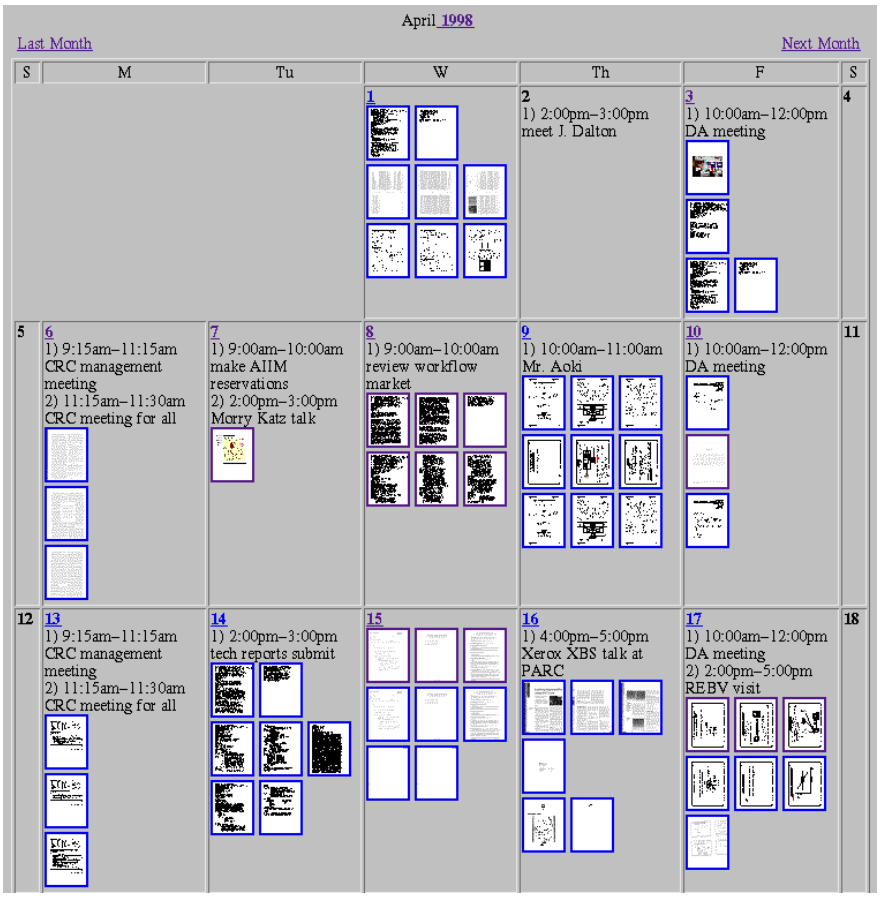

Figure 4. Portion of one user's calendar for April, 1998.

along with 4 dpi thumbnails for the first three pages of the last three documents received by the document server on that day. Users can navigate directly to a particular document by clicking on one of the thumbnails. Doing so displays the description shown in Figure 3. Alternatively, users can click on a day. This displays a list of all the documents processed on that day.

It's easy to imagine how the calendar is used. A user might be looking for an email message about workflow systems that he remembers printing around the time of a meeting he attended on this topic in early April. Browsing through his calendar page for this month he sees he had an appointment to "review workflow market" on April 8th. Inspection of the image thumbnails shown in this cell of the calendar reveals two images with the texture pattern characteristic of an email message. Clicking on one of these thumbnails yields the desired document.

Security is another important aspect of a system like the one described in this paper. Users would like the opportunity to share documents with one another. However, at the same time, they would like to guarantee that their confidential documents remain secret. Furthermore, they would also like to retain the full capability to search and retrieve such confidential documents.

A two-tier security model was adopted to satisfy these requirements. The simplest method allows a user to designate their document collection as either open or closed. If a collection is open, the web server allows access to any host. If a collection is closed, a user must enter the appropriate password to gain access.

Users can also encrypt individual documents by pressing the "Encrypt" link (see Figure 3 for an example). This uses the PGP algorithm to encrypt the data files for a document. Some rudimentary searching capability is retained by not modifying the full text index. Thus, full text search can still be used to find an encrypted document.

3 System Usage

The automatic document storage and retrieval system described in this paper has been in use at CRC for over two years. There are several obvious questions about such a technique. One concerns the amount of storage required. Other questions concern the usefulness of the various document sources (printers vs. copiers). Another question concerns the popularity of the various retrieval interfaces. That is, given that users are already familiar with full text search from their interaction with web search engines, will they readily adopt the additional techniques used here (10 most recent and calendar retrieval). The answers to these questions were investigated by examining the storage and access logs kept by the web server for CRC's IM^3 automatic document storage and retrieval system.

Table 1 presents an analysis of the storage used for all captured documents. From March, 1996 to June 1998, 41,174 documents with a total of 169,643 pages were captured by the system. Of these, 2468 originated on the copier. A total of 9 GB were needed to store all the information needed by the system. Analysis of printed vs. copied documents shows that on average a printed document contains 4.7 pages. A copied document contains an average of 3.5 pages. The complete data set contains 94% printed documents and 6% copied documents.

Table 1: Storage used

time	source [1]	number of docs	number of pages	total storage	storage per page
March 1996 to June 1998	printers + copier	41,174	169,643	9.0 GB	56 KB

[1] Not all document sources were operational throughout the test period

Table 2 presents an analysis of the source of accessed documents. These data were produced by analyzing the web server logs from a 3 month period (April - June, 1996). Every 72 dpi page image viewed by one of the users had been recorded along with the time of that access. The date when the image had been created was also available. A differentiation was also made between documents that had originally been printed and those that had been copied. The results in Table 2 show that of the 1261 pages viewed, 67% of them had originated as printed documents, 28% had been copied. This is much higher than the proportion of copied documents in the population (6%) and suggests, not surprisingly, that copied documents are used more frequently (relative to their sample size) than printed ones.

Table 2: Sources of pages viewed

printer	copier	misc.
848 (67%)	354 (28%)	59 (5%)

The age of accessed documents is presented in Table 3. It is seen that 19% of the pages accessed between April and June of 1998 were originally created in 1997. 12%

were created in 1996. That is, a significant percentage were captured about two years before they were accessed. This suggests that users find old documents useful.

Table 3: Capture date of pages accessed between April and June, 1998

1998	**1997**	**1996**
864 (69%)	244 (19%)	149 (12%)

The popularity of several of the retrieval interfaces (10 most recent, calendar, and full text search) was also investigated. The web server access logs for April to June, 1998 were analyzed and the frequency of use of each of the three interfaces was recorded. This analysis showed that the ten most recently captured documents, the calendar, and full text search were the three most popular interfaces. These results suggest that users can readily adapt to new (ten most recent and calendar) retrieval interfaces. One potential reason for this is that the new interfaces presented here are intuitive, easy to use, and fast

4 Conclusions

A design was presented for a document storage and retrieval system known as the infinite memory multifunction machine (IM^3). The concept for this approach includes capture of all printed, copied, or faxed documents and easy retrieval with a web browser. The capture of all documents processed by users, as a side-effect of their normal processing, eliminates the need for them to decide whether they should save the documents. This almost guarantees that a document will be present at some point in the future when a user needs it. This helps solve the problem of lost documents since when a user determines a document is lost, he or she can be assured the IM^3 has a copy of it.

An implementation of the IM^3 in a research lab with about 20 users was described. Over the course of more than two years, over 40,000 documents have been captured. A total of only 9 GB of disk space has been used. Today, this much disk space can be accommodated on a single 3.5 inch device costing just a few hundred dollars. The utility of this prototype implementation was determined by analyzing the web server access logs. It was shown that copied documents were accessed more frequently than printed documents. However, users still accessed old documents (more than two years old) that had originally been printed. This illustrates the utility of capturing printed documents without asking since it is unlikely that a user would have manually entered them in an image retrieval system when they were created.

The usefulness of several retrieval interfaces was also examined. It was shown that users readily adapted to two new interfaces (listing of the 10 documents most recently printed or copied and a calendar listing) that were different from one they were already familiar with (full text search). This could be attributed to the simplicity of design of these interfaces and their relative ease of use.

Acknowledgments

Many people in Ricoh contributed in various ways to the development of the system presented in this paper. They include Michael Baxter, John Cullen, Toshio Kanoh, Dar-Shyang Lee, Alex Ono, Mark Peairs, Robert Runyon, and Kiyoshi Suzuki.

References

1. J. Cullen, J.J. Hull, and P.E. Hart,''Document image database retrieval and browsing using texture analysis,'' Fourth International Conference on Document Analysis and Recognition,'' Ulm, Germany, Aug. 19-21, 1997, 718-721.
2. D. Doermann, J. Sauvola, H. Kauniskangas, C. Shin, M. Pietikainen, and A. Rosenfeld, ''The development of a general framework for intelligent document retrieval,'' in Document Analysis Systems II, World Scientific, 1998, 433-460.
3. H. Fujisawa and H. Stabler, ''Needs of the market and user requirements,'' in Document Analysis Systems, World Scientific, 1995, 452-454.
4. J.J. Hull, ''Document image matching and retrieval with multiple distortion-invariant descriptors,'' in Document Analysis Systems, World Scientific, 1995, 379-396.
5. J.J. Hull, ''Document image similarity and equivalence detection,'' International Journal on Document Analysis and Recognition, v. 1, no. 1, February, 1998, 37-42.
6. B.J. Jansen, A. Spink, and T. Saracevic, ''Failure analysis in query construction: Data and analysis from a large sample of web queries,'' The Third ACM Conference on Digital Libraries, Pittsburgh, PA, June 23-26, 1998, 289-290.
7. D.S. Lee and J.J. Hull, "Group 4 compressed document matching," Document Analysis Systems, Nagano, Japan, November, 1998.
8. A.L. Spitz, ''SPAM: A scientific paper access method,'' in Document Analysis Systems II, World Scientific, 1998, 242-255.
9. G.A. Story, L. O'Gorman, D. Fox, L.L. Schaper, H.V. Jagadish, ''The RightPages image-based electronic library for alerting and browsing,'' IEEE Computer, v. 25, no. 9, September 1992, 17-26.
10. ''White Paper: Introduction to Document Imaging,'' Wang Corporation, technical report, 1996.