無意識で文書管理を可能にする新しいパラダイム

Towards Zero-Effort Personal Document Management: A New Paradigm for Users

ジョナサン ハル[*]　ピーター イー ハート[**]
Jonathan J. HULL　Peter E. HART

要　旨

最近の研究によれば，3%から5%の文書が紛失する恐れが常にあり，Fortune 500企業における文書紛失に伴う平均コストは300万ドルから500万ドルに達している．また，作成した文書のうち，15%程度しか後で参照されないとの報告もある．他方，磁気ディスクのコストは近年目覚しく低下し続けており，1999年末には100キロバイト(KB)あたり0.1セントにまで低下した．他方，オフィスにおいて，文書はパーソナルコンピュータで作成されプリンターで出力されるか，あるいは，紙文書は複写機やファックスで配布される．もし，ここでの紙情報や電子情報すべてを，安価なメモリーに収納すると，「ほとんど無意識に，かつ，努力なしに，情報蓄積ができる」ことになる．最近の文字認識技術のおかげで，これら「無意識で蓄積された」文書画像はOCRによってテキスト変換することができるから，全文テキスト検索によって検索可能である．もちろん，文書を縮小表示したサムネイルや，収納した日付で検索することも可能である．CRC(20〜30名)での実験によれば，10ヶ月間で約30万ページの文書が蓄積され，これに費やしたメモリーが18GBであった．蓄積された文書は，83%がプリント文書，17%が複写文書であった．検索方法は，66%が最新文書からの遡及検索，23%が日付けによる検索，そして11%がテキスト検索であった．面白いことにテキスト化されているはずのプリント文書の検索もかなり多い．理由はファイルから検索するより容易だからである．このように，ほとんど無意識で蓄積された文書ファイル(本文中では，Shoeboxと表現されている．これは，家庭で靴の空き箱が，レシートなどの保管箱として再利用されており，年度末の税金払い戻し申請時に利用される)から，有効に文書検索ができることが判明した．

ABSTRACT

We introduce “Save everything!” as a radically simple paradigm for helping individual users manage their personal documents. For this approach to succeed the effort to capture and file documents must be very nearly zero, yet document retrieval must be at least as convenient and effective as it would be using conventional retrieval methods.

To achieve zero-effort document capture, we have modified conventional printers, digital copiers and fax machines to automatically save, as a side effect, representations of every document processed. To enable effective document retrieval in the absence of user-assigned filenames, paths, or keywords, we have developed a new set of retrieval methods that exploit the user’s familiarity with a personal document collection.

This perhaps counter-intuitive approach has been implemented and placed in daily use for three years in a workgroup of some 20-30 people. We describe here the system components and report quantitative experimental results.

[*] MDA Group Leader of Ricoh Silicon Valley,Inc.

[**] Senior Vice President of Ricoh Company,Ltd.
President and Chairman of Ricoh Silicon Valley,Inc.

1. Introduction

1-1 An Age-Old Problem

A problem that everyone has seems to be as old as civilization itself: What shall I keep, and what shall I throw away? In everyday life we face this problem every time we think about cleaning out a closet or wonder if we'll ever again use that guidebook to the Paris sewers. In the office the recurring theme is information, most especially documents: Which documents shall I save, and which shall I discard?

If I choose to save a document I have to think about how or where to file it and how much space it consumes. For most of us, there's the additional nagging thought in the back of our minds that regardless of how we file the document we might have trouble locating it again. This concern is well-founded. A recent study estimates that between 3% and 5% of documents are lost at any one time, and the average cost of lost documents to a Fortune 500 company is in the range of $3 million to $5 million [10]. On the other hand, if I choose to discard a document it might prove 6 months or 5 years from now to contain the very information that I need. That thought leads me to save the document—but then another study has shown that a mere 15% of documents are ever referred to again [3], so perhaps I can safely throw it away after all.

We propose a policy for resolving this very old dilemma that is as radical as it is simple: Save everything!

For this policy to succeed, a number of fundamental issues must be addressed:

1. The personal effort required to file a document has to be small to zero.
2. The cost of storage has to be economically reasonable.
3. Document retrieval has to be at least as efficient and effective as today's conventional methods, even though the filing effort is far lower than today's methods demand.
4. Users must be confident that security and privacy concerns have been adequately addressed.

For the past three years we have conducted an experiment to evaluate this radical "Save everything!" approach to personal document management, using as our experimental domain the daily work of a research group ranging between 20 and 30 members [4]. We describe here the capabilities that were built to support the workgroup and the results that were recorded using a comprehensive monitoring system.

1-2 Paper vs. Digital Documents

Our general concept is that all user documents should be captured (or capturable) with no user effort. Metaphorically, we think of the proverbial giant shoebox into which all documents are 'unconsciously' thrown. (We have rather deliberately adopted the term 'unconscious'. Our experience has been that even thinking about whether or not to save a document places a cognitive burden on the user that we wish to remove.) In practice, this means capturing documents from both personal computer and paper sources. It seemed clear at the outset that digital documents could be unconsciously captured from personal computers, providing of course that privacy and security concerns were adequately addressed. Accordingly, no attention was given to this part of the problem in our study.

Paper-sourced documents, however, present a much more difficult problem. How can paper documents be captured with little or no user effort? A first thought is that scanners provide an obvious solution. But even though high-quality scanners have been available at very reasonable prices for many years, they have not been widely adopted in most offices and are not part of the work practices of most people. A kind of "scanning barrier" seems to prevent most people from using them routinely. Indeed, at a recent seminar on document analysis and retrieval it was found that only a small minority of attendees used scanners in their daily routine, even though everyone had scanners readily available for capturing page images for their experimental work [2]. Scanners, then, are not the route to unconscious, zero-effort document capture.

There are, however, a family of paper-handling office machines that are part of everyone's work practice: Copiers, printers and fax machines are a daily part of all of our lives.

The three functions are combined in today's so-called multi-function machines. This leads naturally to the thought that a multi-function machine could be modified to support automatic document capture, and could be equipped with a very large memory and backing store that from the user's perspective appears to be 'infinite'. Additionally, individual printers, digital copiers and fax machines could be equipped with the same capabilities and can be conveniently deployed around the workgroup. We think of this collection of capture devices and storage as the "infinite memory multi-function machine", or IM^3. The IM^3 is our proposed approach for resolving the first major issue noted earlier, viz., how to minimize the effort required to acquire and file a document.

1-3 Storage Costs

The most obvious economic concern regarding the IM^3 is storage costs. To address this concern, we compared the cost of storing a digital page image with the cost of the paper on which the image is printed. For this purpose, we made two assumptions: First, we assumed that paper costs around a penny a sheet. Second, we assumed that the image of a typical bi-level office document compresses to around 100 Kbytes.

While the cost of paper has been relatively stable, the price-performance of conventional, rotating magnetic storage media has plummeted for three decades [6,8]. Our informal analysis is that over this period disk price-performance has doubled roughly every year. This record handily beats the celebrated Moore's Law for semi-conductors, which is usually quoted as stating that semi-conductor price-performance doubles every 18 – 24 months. Indeed, we find it curious that a more authoritative version of this "law" has not received wider attention. In any event, at the end of 1999 the street price of disk drives was approaching 1 cent per Mbyte, or 0.1 cents/100 Kbyte, which means that the cost of storing a page image is one-tenth the cost of paper.

This quick analysis of course ignores many other factors, such as the cost of file cabinets and the cost of the office space for holding file cabinets, as well as the cost of maintaining computer systems. But our principal concern is with scaling—what happens when demands for "infinite" memory becomes finitely very large—and on this point we are satisfied that storage costs are economically reasonable.

1-4 Document Retrieval Issues

As the reader has most-likely already concluded, the most challenging research issue for the IM^3 approach is document retrieval: If every document is automatically captured as a side effect of copying, printing and faxing, how will it be retrieved in the future at low user effort? "Unconscious capture" means that no file name has been given to the document by the user, no path or file directory is visible, and no key words have been assigned. Indeed, the shoebox metaphor—"just throw the document over your shoulder into the shoebox and forget about it"—seems rather apt.

A first approach to retrieval relies on the well-developed, conventional indexing and retrieval technology that is typically used for searching a large full-text database. This search technology is readily augmented by an OCR front end for document bitmaps captured from copiers and fax machines.

While useful, these conventional methods seemed too limiting for the IM^3 application. But the IM^3 is first and foremost a personal document management solution, not a general-purpose full-text database solution. Frequently, the user has information about a sought-for document that might, for example, relate to how the document was acquired or to what other activities the user was engaged in at that time. The availability of this personal metadata creates opportunities for inventing novel retrieval methods that might not be suitable for more general full-text-search applications. A combination of conventional and novel retrieval methods are offered by the IM^3 system, and we have been able to measure their popularity in our user population.

1-5 Related Work

Personal document management systems utilizing multiple technologies can provide fertile ground for exploring new methods for data capture and retrieval. As with many new developments in this field, the work reported here bears a tantalizing similarity to several other previously-described ideas and yet appears to be different from any of them.

For example, work in document image storage and retrieval systems might appear to be closely related to ours, but the emphasis there is on improving the accuracy of the image analysis, OCR, and indexing that occur after a document has been scanned [1,9]. This contrasts sharply with the emphasis we place on eliminating an explicit scanning step. Indeed, we suppose it's precisely this step that has inhibited the common use of previous systems.

Some of the novel retrieval methods provided in the IM^3 resemble innovations in user interface design [12] and information visualization (e.g. [5,7]). There, however, the focus is on representing document content. Instead, we represent information *about* documents (personal metadata) in ways that make it easy for users to find what they want. In another related project, users wore special badges so that personal metadata could be gathered [11]. In contrast, our approach requires almost no change in user behavior.

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 users identifying themselves 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.

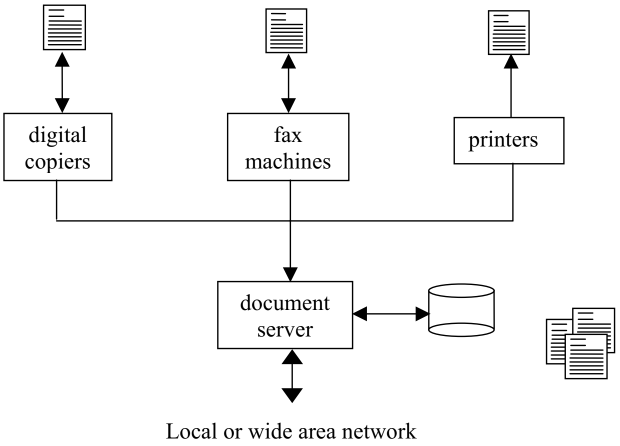
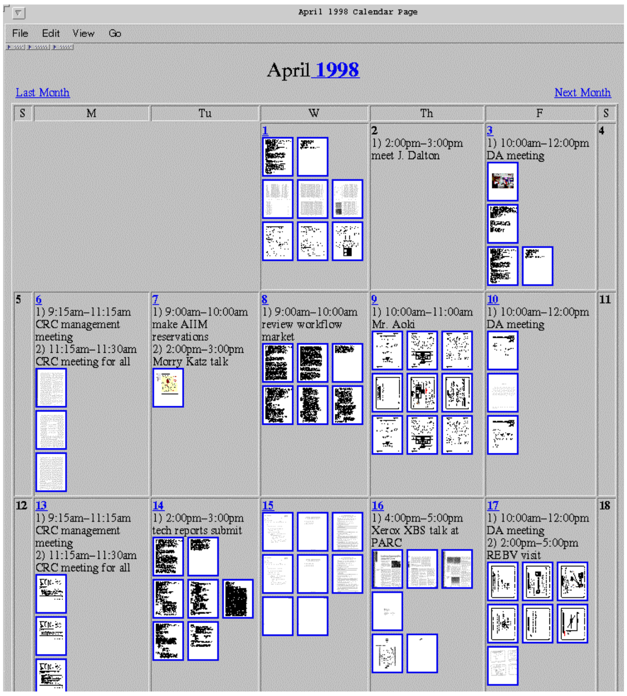


Fig.1 IM^3 system design. Copied, faxed, and printed documents are automatically saved on a document server.

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.

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 include several chronological views. These include a calendar view (see Fig.2) as well as a listing of the ten most recently saved documents. The calendar displays thumbnail images and apppointments (extracted from a calendar manager). Users access documents by clicking on the thumbnails. Conventional text search is also provided. Users can also browse a listing of all the documents that have been saved for them.



April 1998

S	M	Tu	W	Th	F	S
			1	2 1) 2:00pm–3:00pm meet J. Dalton	3 1) 10:00am–12:00pm DA meeting	4
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

Fig.2 Calendar retrieval interface. Document thumbnails are displayed in context with events that occurred when the documents were saved.

3. System Usage

The automatic document storage and retrieval system described in this paper has been in constant use for over three years. There are several obvious questions: how much storage is required; how useful are various document sources (printers vs. copiers); which of the various retrieval interfaces is most popular. 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.

An analysis of the storage for documents saved between March, 1996 and November 1999 shows that 72,492 documents with a total of 295,981 pages were captured. A total of 18 GB were needed to store all the information needed by the system. Printed documents contain 4.3 pages on average, while copied documents contain 3.2 pages. The complete data set contains 83% printed documents and 17% copied documents.

A comparison of the storage required for printed and copied documents considered the data saved from May, 1998 to November, 1999. During this time every printed document was saved and two copiers were constantly available that could perform unconscious capture. Overall, 38,384 documents comprised of 152,251 pages were stored in 12 GB of disk space.

The age (difference between their creation time and their access time) of the 1182 pages accessed between May and October, 1999 was also analyzed. The average is 44 days and the median is 3 days. The difference between these statistics is reflected in the long tail on the distribution. It's interesting to note that 38% of the accesses were to documents more than one week old and over 10% of the accesses were to documents over 6 months old. This suggests that users find old documents useful and that they should be made available. It also points out the value of automatically capturing documents without asking since the effort required to locate an old document is often enough for users to simply give up or to re-create the document from scratch.

The popularity of several retrieval interfaces (last 10 documents captured, calendar, and full text search) was also investigated (see Fig.3). The web server access logs for May to October, 1999 showed that the last 10 documents were used 3 times more often than the calendar. This is not surprising since the last-10 is an easy-to-use browsable listing. Also, of the three, it requires the least cognitive effort by the user. The calendar, on the other hand, is also a browsable interface but is a new metaphor for document retrieval. However, the novelty apparently did not affect its adoption since it was used more than twice as often as the retrieval method most people use on a regular basis for searching the web – full text search. These results suggest that users can readily adapt to new (ten most recent and calendar) retrieval interfaces if they are intuitive and provide utility for the effort that must be invested in learning how to used them.

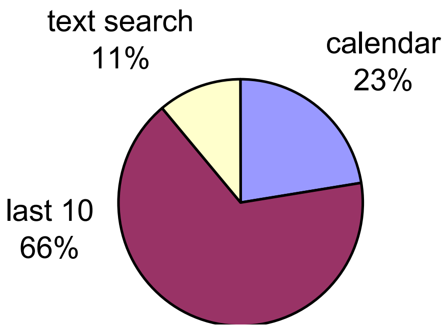


Fig.3 Popularity of retrieval methods.

4. Conclusions

We have presented the results of a multi-year experiment based on what seems to be an inherently self-contradictory approach: Personal document management headaches can be reduced not by minimizing your personal document store but instead by maximizing it. The success of this approach was shown to depend on two key points: First, the effort to acquire and file documents was reduced to nearly zero by making document capture a side effect of everyday work practices. Second, new document retrieval methods were developed that exploit the user's familiarity with their own personal document corpus.

The implemented system accumulated over 70,000 documents totalling over 300,000 pages in its three years of experimental use. The incremental disk space needed to hold this document corpus costs only a few hundred dollars today and prices continue to drop rapidly.

During a recent 6 month period, our users accessed their personal IM^3 document stores on average about every other day. More than 10% of these accesses were to documents more than 6 months old. This statistic supports comments made in user interviews that the IM^3 database becomes more useful over time. We attribute this to the fact that unconscious capture almost guarantees that a document will be present at some point in the future when a user needs it.

The new retrieval interfaces were quickly embraced by our test population: They were used ten times more often than the traditional text-based methods that were prominently available on the user's IM^3 home page.

Interestingly, users frequently retrieved printed documents from the IM^3. Since all printed documents had, at one time at least, a symbolic source file somewhere in the workgroup, we had not anticipated this observation. Evidently, it was easier for users to find and retrieve a desired printed document from the IM^3 database than from their own file directory (if indeed the document still existed there).

The IM^3 system described here captures only documents that have a paper source (i.e., copiers or fax machines), a paper destination (i.e., printers), or a Web page source. Our general concept is that personal computers should be document sources for the IM^3 as well, but that was not implemented in the study[1]. However, even with this restriction on document sources we believe that the "Save everything!" approach is a useful step along the way to zero-effort personal document management.

Acknowledgments

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

References

1) 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.

2) H. Fujisawa and H. Stabler, "Needs of the market and user

[1] Ricoh has introduced the eCabinet™ that adds the capability to capture computer documents as well as those described here.

requirements," in Document Analysis Systems, World Scientific, 1995, 452-454.

3) 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, v. 33, no. 1, 1997, 107-121.

4) J.J. Hull and P. Hart, "The infinite memory multifunction machine," Pre-proceedings of the Third IAPR Workshop on Document Analysis Systems, Nagano, Japan, November 4-6, 1998, 49-58.

5) J. Lamping, R. Rao, and P. Pirolli, "A focus+context technique based on hyperbolic geometry for visualizing large hierarchies," ACM Conference on Human Factors in Software (CHI '95), Denver, Colorado, 1995.

6) M. Lesk, "Practical Digital Libraries," Morgan Kaufmann, San Francisco, 1997.

7) J.D. Mackinlay, G. G. Robertson, and S. K. Card, "The perspective wall: Detail and context smoothly integrated," ACM Conference on Human Factors in Computing Systems (CHI '91), 173-179, 1991.

8) H. Paulapuro, "The future of paper in the information society," The Electronic Library, v. 9, no. 3, June, 1991, 135-143.

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) Wang Laboratories, Inc., "Introduction to document imaging: solving the paper dilemma," White Paper, 1996.

11) R. Want, A. Hopper, V. Falcao, and J. Gibbons, "The active badge location system," ACM Transactions on Information Systems, v. 10, no. 1, January, 1992, 91-102.

12) R. Wilensky, "Toward work-centered digital information services," IEEE Computer, v. 29, no. 5, May, 1996, 37-44.