

Stuff I've Seen: A System for Personal Information Retrieval and Re-Use

Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, Daniel C. Robbins

Microsoft Research

One Microsoft Way

Redmond, WA 98052 USA

[sdumais; cutrell; jjcadiz; gavin; a-rsarin; dcr]@microsoft.com

ABSTRACT

Most information management technologies are designed to facilitate information discovery. However, much knowledge work involves information re-use and integration. We describe the design and evaluation of a system, called *Stuff I've Seen (SIS)*, that is designed to facilitate information re-use. This is accomplished in two ways. First, the system provides a unified index of information that a person has seen, whether it be email, web pages, office documents, calendar appointments, etc. Second, because the information has been seen before, rich contextual cues such as date, author, thumbnails and previews should be especially helpful in finding items. The system has been used internally by more than 200 employees. We report on both qualitative and quantitative aspects of system use. Initial findings show that people find information more easily using Stuff I've Seen, and use other search tools less frequently after installation.

Keywords

Information retrieval, personal information management, user studies, field studies.

INTRODUCTION

Most information retrieval tools, like popular web or intranet search engines, are designed to facilitate information discovery. Given a short query, they do a remarkable job of finding relevant materials using a variety of content, anchor text, link and popularity cues. However, much knowledge work involves integration and re-use of information that has previously been created, found or received. For example, creating a presentation or paper may involve some web searching, but it also involves pulling together information from existing information sources like presentations, documents, spreadsheets, data analyses, email messages, etc. Studies have shown that 58-81% of web pages accessed were re-visits to pages previously seen [9,23,29]. Similar re-access patterns have been observed in retrieving information from technical manuals, Unix

commands [14], library book borrowing [6], and human memory [2].

We developed a system, called *Stuff I've Seen (SIS)* that makes it easy for people to find information that they have seen before. There are two key aspects to the design that enable us to effectively support personal information management. First, the system provides a unified index of information that a person has seen, whether the information was received as an email, web page, document, calendar appointment, etc. Indexing happens automatically without any need for user interaction. Today, people have to manage several different organizations of information – e.g., the file system hierarchy for files, email folders for email, favorites or history for web pages. With SIS, all of these sources are integrated into a single index regardless of what form the information originated in. Second, because a person has seen the information before, rich contextual cues such as time, author, thumbnails and previews can be provided in the search results interface. Because they lack personal context, web search results must be organized by rank, showing titles, URLs and abstracts. With personal information, a much richer variety of cues can be used to search for and present information.

RELATED WORK

Vannevar Bush's vision of a memex [7,8], "*a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility*" captures the essence of SIS. The details of implementation are obviously very different, but the functionality is much the same as what we have implemented.

Jones and Thomas [18] surveyed people's use of new computer-based personal information management technologies and found rather low adoption rates at the time of their survey. They speculate that the limited applicability of new technologies, each focusing on a limited conception of personal information, is the main reason for slow adoption. Erickson [12] and Bellotti and Smith [4] describe field observations of new personal information management systems. Each built a new information management artifact (a personal notebook in [12] and a new email application in [4]) and observed usage patterns as well as the co-evolution of the systems and

LEAVE BLANK THE LAST 2.5 cm (1") OF THE LEFT
COLUMN ON THE FIRST PAGE FOR THE
COPYRIGHT NOTICE.

work practices. Our focus is not to create a new information management application per se but rather to develop a unified search interface to existing content regardless of how it was created initially.

Jones et al. [16,17] observed several methods that people use to organize web information for re-use. In addition to the book marking and history list tools provided by web browsers, people used a wide variety of other techniques including emailing URLs to themselves and others, printing paper copies of web pages, saving pages to the hard drive, pasting URLs into documents, and adding URLs to personal web pages. A functional analysis provides interesting insights into techniques that people use for managing web pages for subsequent re-access. However, this work is limited to web pages and does not develop tools, as we have, to address the problem.

Several groups have looked at tools for improving access to subsets of the information people receive. Malone [21] looked at how people organize paper materials and suggested how computer systems could better support these activities. Barreau and Nardi [3] examined how people manage files on their computer. Whittaker and Sidner [30] described the email overload problem and linked it to difficulties that people have with deferred processing and classification of items. Nardi et al. [24] developed a system that integrates email with Instant Messaging designed to overcome these difficulties. Several groups have developed systems to improve re-access to web pages, including the use of rich graphical representations (see [10] for a review), integration of back, history and favorites [19], predictive models of information needs [25], and full text indexing of pages [22]. All of these systems have focused on improving access to information within a single content type.

Haystack and Lifestreams provide access to a variety of personal information. In Haystack [1,15], users create “haystacks” that form an integrated personal information repositories to archive content and user-specific metadata. The initial version worked with input from the browser and editor, although extensions to email and files were planned. Lifestreams was described as a replacement for the desktop metaphor [13]. It provides a single time-ordered stream of electronic information, and allows for applications like searching, filtering and summarization.

Several commercial products have functionality to index personal information as well. Some software enables users to index information on their PCs (e.g., Microsoft’s Indexing Services, Apple’s Sherlock). These systems work on files, but not email stores. Other desktop search applications like Enfish Personal, PC Data Finder, 80-20 Retriever, and Scopeware index both files and email. Some of these also index web pages, and others extend indexing to enterprise content as well. However, despite the existence of these commercial applications, we are not aware of published research on how using these products, or how they affects people’s work patterns.

SIS is built on an extensible platform and we believe it covers a wider range of information sources and file types than the prototypes and products mentioned above. More importantly, our focus is on exploring novel information retrieval algorithms and user interfaces that capitalize on the user’s familiarity with their own content, and on studying user’s experiences in using SIS with log analysis and questionnaire techniques.

STUFF I’VE SEEN (SIS)

Today it is often easier to find information on the web than on your own desktop, email store, or intranet. This is due to both the multiplicity of independent applications used to manage information and to the limited search capabilities in many of them. The goal of SIS research and development was to remedy this problem by providing a unified index of all the information a person has seen, regardless of whether the information was seen as email, web pages, documents, calendar appointments, etc. If a user wants to restrict search to a particular medium they can, but this is not a prerequisite. Indexing happens automatically without any need for user interaction.

In addition to the core indexing capabilities, we can also explore new ranking and presentation ideas. Because the information is personal and has been seen before, we believe that rich contextual cues such as time, author, thumbnails and previews will be especially useful. Moreover, the index allows for very fast searching. This enables interesting interactions for filtering, sorting and browsing of the search results.

We report on the design and functionality of the system as well as our experiences in deploying it.

System Architecture

The SIS application is built on top of a modular indexing architecture. There are five main components – Gatherer, Filter, Tokenizer, Indexer, and Retriever. The Gatherer specifies the interface to different content sources in their native format. Files, http, and MAPI are examples of gatherers that are supported in the current prototype. The Filter decodes individual file formats (e.g., .doc, .pdf, .ps, .html) and emits a character stream for further processing. The Tokenizer breaks the stream of characters into ‘words’ and can also handle additional linguistic processing such as date normalization, stemming, etc. The Gatherer, Filter and Tokenizer components are extensible to handle new data sources, file types and languages. The Indexer builds a standard index structure to support quick retrieval. The Retriever is the query language for accessing stored information. It supports Boolean as well as best match retrieval on the full text and metadata properties and allows phrase, wildcard and proximity searches. All of these components run on the client machine.

By default, the users’ mail profile, web cache, and personal files are indexed. No additional work is required on the part of the user, although other sources are easily added. As new mail is received, web pages viewed or content is created or modified, the index is automatically updated.

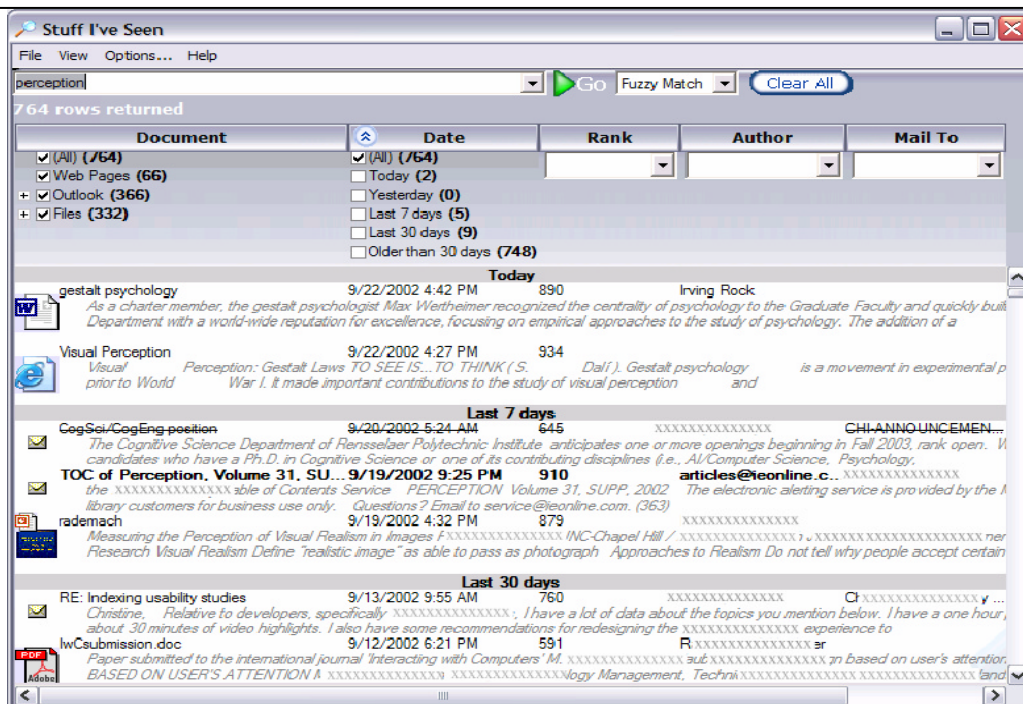


Figure 1. Screen shot of SIS interface, with standard Top View.

User Interface

The SIS interface allows users to specify queries and to view and manipulate results. Contrary to many query interfaces where users specify several properties and then press a button to launch a query, SIS launches its queries whenever any of the filtering widgets in the UI are manipulated or when the user presses return. Because SIS works from a local index of all content, query results can be returned very quickly, allowing users to search their content using a different interaction strategy than is used in other search interfaces.

Figure 1 shows the first interface we developed, called the Top View. The query text box is in the upper left hand corner. By default, query words are combined using an AND operator. Users can specify other Boolean operators, a fuzzy matching alternative in which morphological variants are also used (e.g., *car* matches *cars* as well), or fielded search in which matches are restricted to certain fields (e.g., *author="Jane Doe"*). We have worked to produce understandable and accurate default performance, and in practice few users modify the default operation.

The details of search results are shown in the lower portion of the display. In Figure 1, each result includes a preview showing the first 256 bytes of a message as well as thumbnails for images and PowerPoint files. The previews can be turned off, increasing the number of results displayed. Five fields are present in the default view: Document Title, Date, Rank, Author and MailTo. Additional fields are available through the options menu: File Type, Mail CC, Mail HasAttachment, Message Type, Message Read, Path, Size, Title. The results are sorted by either Date or Rank by default (different versions were deployed). Clicking on any column header sorts the results

by that column. A double arrow is shown in the column by which results are currently sorted (Date in Figure 1). When Date is the sort field, markers showing the main date groupings (today, yesterday, etc.) are displayed to help group the results visually. The scroll bar on the right allows users to quickly move through the results.

Results lists can be further refined by selecting filters. In the Top View, filters for each column are located at the top of each column just below the column header. Checkboxes are shown when there are only a few alternatives (e.g., Document Type and Date), and text boxes with drop down lists are used when there are many possible alternatives (e.g., author). Filters can be applied even when there is no query. Users can click the filters to return all items from a certain date range, or all items from a specific person.

The user interface shown in Figure 1 is somewhat complex and the filters at the top reduce the number of results that can be displayed. An alternative, called the Side View, is shown in Figure 2. This interface has the main query box and list view of results, but the filters have been simplified and moved to the side. In this view, filters are revealed serially. Selecting a specific item type, like Outlook, filters the results by that type and exposes additional fields that the user can specify for fielded search. In the case of Outlook items, the user can specify From, To, Path or whether the item has an attachment.

The Side View has the advantage that it's somewhat easier to understand and is less cluttered. And, because the filters are moved to the side, more results can be displayed. In contrast, the Top View is considerably more flexible. For instance, in the Side View it is not possible to filter by multiple types of items (e.g., Outlook and Web Pages), or by a specific column across all document types (e.g.,

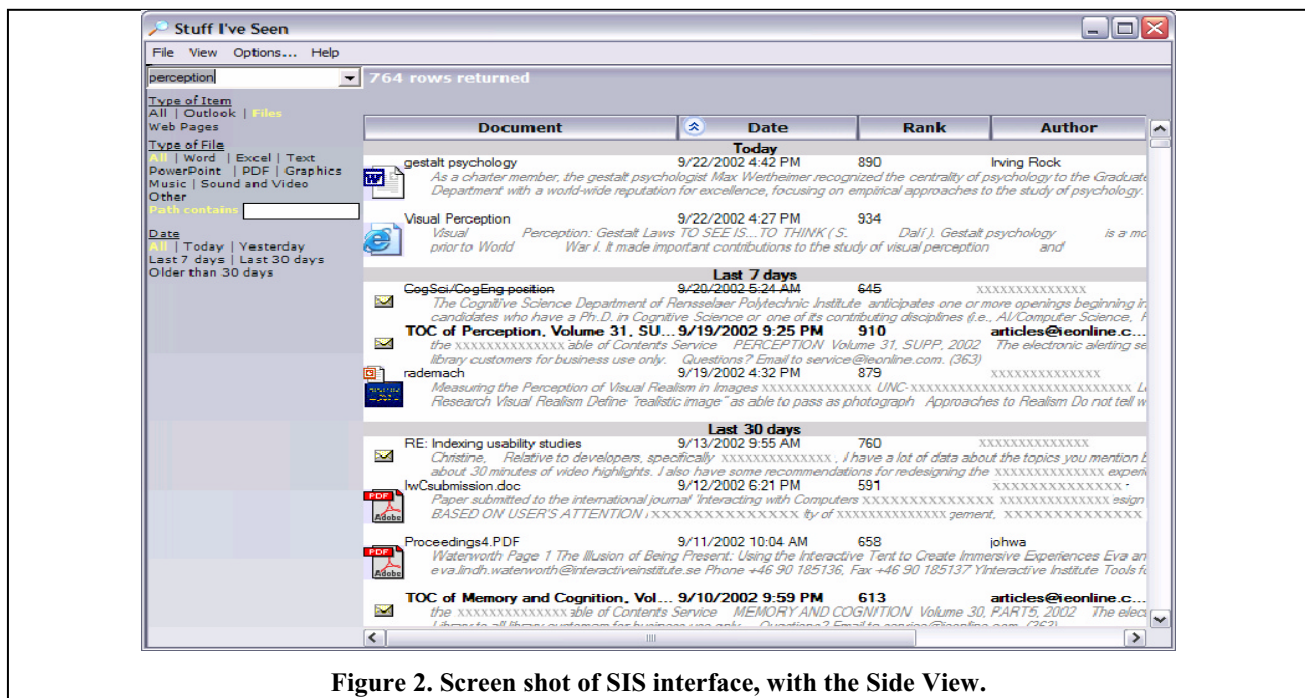


Figure 2. Screen shot of SIS interface, with the Side View.

author). In addition, in the Top View, the filters are associated with the columns they affect to create a more 'direct' filtering experience.

In both interfaces, double clicking on an item opens it up in the appropriate native application. Right-clicking an item brings up a context menu that allows the user to go to the folder containing the item (for mail and files), or to the cached version (for web pages).

Deployment

The SIS prototype has been available for internal use at our institution since June 2002. We maintain an internal web page with general information about the system, a FAQ, and installation instructions, as well as an email distribution list for announcements and user feedback. Since the first deployment, six updates to SIS have been released. These updates have added new features, refined existing features, and fixed a variety of bugs.

As of September 16th, 2002, 310 users had installed SIS. Although our users are arguably more computer savvy than the average person, they come from variety of backgrounds, including consulting, legal, product support, administration and sales. Our user group also includes individual contributors, managers, and even executives.

The following sections describe how people used SIS, and how we studied the effects of their usage.

Usage and Analysis

We studied SIS using two main techniques: surveys and log file analysis. First, before people installed SIS, we asked them to take a brief survey about their current behavior in searching for e-mail, files, and web pages. We also asked people to categorize themselves as people who didn't file e-mail, filed e-mail in bursts, or who filed e-mail regularly (the same categories as described in [30]). 317 responses to this survey were received. (This number is greater than the

number of installations, presumably because a few people could not successfully install SIS.)

Second, SIS was instrumented such that all user actions were uploaded to a database. Examples of actions logged include: query text, use of filters, and the number of results returned by each query. For privacy reasons, we do not log any information about what content is indexed, or the content of search results.

Third, in early September, a longer usage survey was distributed to all SIS users, and 57 responses were received.

We report usage data from 234 people. Bugs, and logging and robustness issues with early prototypes meant that we were unable to use data from everyone who downloaded. These 234 people used SIS during the time period studied (August 1st, 2002 through September 16th, 2002).

Usage Log Data

During the time period studied, SIS users executed more than 8000 queries. Some general summary statistics are presented in Table 1. Users took advantage of the system on 84% of the work days they had it installed. They issued an average of 4.4 queries per day of use, but there is high variance with some users issuing no queries and one user issuing 45 queries in one day. On an average work day, 61 people used SIS at least once.

| N=234 | Mean | Std | Med | Min | Max |
|-----------------------------|------|------|------|-----|------|
| Calendar Days Since Install | 20.1 | 14.3 | 19.5 | 1 | 47 |
| Work Days Since Install | 14.5 | 10.1 | 14 | 1 | 33 |
| % Cal Days Used | 59% | 28% | 55% | 11% | 100% |
| % Work Days Used | 84% | 27% | 70% | 13% | 100% |
| Queries Per Day of Use | 4.4 | 6.1 | 2.5 | 0 | 45 |
| Users Per Day | 46.9 | 27.1 | 56 | 0 | 98 |

Table 1. Summary SIS usage statistics

Query Characteristics

Although our users are more computer savvy than the typical web user, many characteristics of their queries are similar to those reported in analyses of web query logs [27,28]. Their queries are short, averaging 1.6 words. Only 7.5% of the queries involved explicit Boolean operators (AND, OR, NOT, +, -), phrases, or field restrictions specified in the main query box (e.g., from="Jane Doe"). Although field restrictions are seldom used in the main query box, they are used frequently through direct manipulation of filters in the interface. More than half of the queries (4112 of 8067) involved a filter specified using the checkboxes (selection in the Side view). Since personal stores are much smaller than the web (an average of 740 results), simple queries along with the sorting and filtering mechanisms work well.

Some query characteristics of interest (e.g., use of names) were coded by hand. Names that were ambiguous (e.g., bill, grant) were not counted unless they were paired with a last name (e.g., bill smith) or other disambiguating context (grant cell number). Many name queries involved just a first name or email alias. Queries involving a name accounted for 25% of the queries.

Opening Items From the Search Results

There were over 8000 searches performed and nearly 2500 files opened using SIS (several files could be opened after one search, and not all searches led to files being opened). The failure to open items after a search is difficult to interpret. It could mean that the search was a failure, or that the search results were used in other ways. For example, the preview and metadata shown in the interface could provide the needed information. For items that were opened, we recorded the type, date and position in the list of results. Email is by far the most common type opened (76%), followed by web pages (14%) and files (10%). The most common file types were Microsoft Word (14%), plain text files (11%), and Microsoft PowerPoint (11%), with the remaining types accounting for less than 10% each.

We also looked at the time distribution of opened items. Figure 3 shows the number of items opened as a function of time. Not surprisingly, recent items are accessed frequently, but there is also a long tail with items up to eight years old being opened. These long-tailed distributions have been reported for a variety of information access activities, but to our knowledge never before for personal items [5,31,32]. Overall, 6.6% of the items opened were first seen that day, 21.9% within the last week, 45.9% within the last month, and 89.4% during the last year. Because the majority of items accessed are more than a month old, it is important to provide rich content and contextual cues to aid retrieval.

Figure 4 plots the access patterns on a log-log scale, and focuses on just recent items. The linear fit in the log scale is quite good ($r^2 = .62$, $p < .0001$). The fitted function is $\log(\text{Frequency}) = -0.68 \cdot \log(\text{DaysSinceItemFirstSeen}) + 2.02$. Others have reported similar power functions for re-access to web pages [9] and human memory [2].

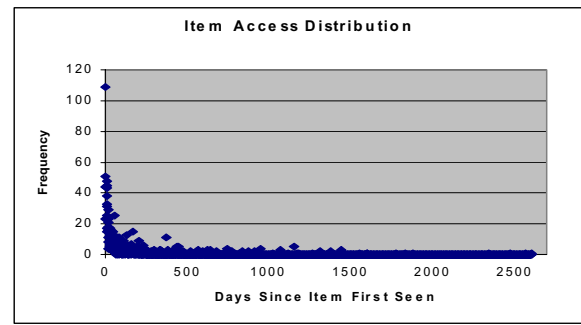


Figure 3. Frequency of access for items over time.

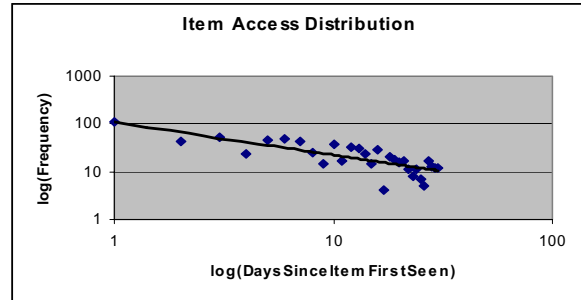


Figure 4. Frequency of access for items (log-log scale for 1 month).

Interface Experiments

We randomly deployed different versions of the search interface to different users. Our logging data allow us to look at how many users changed these default settings. In general users stay with what they are assigned by default, but there are some interesting details.

Top vs. Side Layout

Half of the users started with the Top view and half with the Side view. Table 3 shows the total number of queries whose results were shown using the Top or Side view, broken down by which condition the user was initially assigned. People who started with the Top view issued more queries than those who started with the Side view, but this difference is not reliable statistically ($t(228)=1.32$, $p=0.19$). People who started with the Top view were less likely to switch to the Side view (34%) than vice versa (44%) and this effect was reliable ($t(228)=2.04$, $p<0.05$).

| START: | Top | Side | Total |
|--------|------|------|-------|
| N | 115 | 115 | 230 |
| USE: | | | |
| Top | 3043 | 1567 | 4610 |
| Side | 1621 | 1986 | 3607 |
| Total | 4664 | 3553 | 8217 |

Table 2. Use of layouts for results presentation.

Some of the more frequent use of and switching to the Top view is likely due to the richer search interface which encouraged frequent filtering and allowed for fielded search. In addition, the Top view required an additional query to select a single item class (e.g., just Outlook). With just the log data it is difficult to fully understand why more

queries were issued. This could be the result of poor results presentation (and thus more query refinements), or they could be the results of better results (thus more searches with a useful tool). We are currently examining the details of query patterns to better understand this effect.

Rank vs. Date Sort

About half of the users started with results sorted by Date and the other half with results sorted by Rank. Rank is the kind of match score that is used to order results from web or intranet searches, where date is typically unknown to the person issuing the query. For personal content, Date is a very reasonable alternative since people often remember roughly when something happened. In addition to these two fields, users can sort on any of the other fields such as Author, Size, Path, Title, etc. Table 3 shows the number of queries sorted by various fields, broken down by which sort condition the user was assigned by default.

| START: | Date | Rank | Total |
|--------|------|------|-------|
| N | 111 | 122 | 233 |
| USE: | | | |
| Date | 3062 | 1975 | 5037 |
| Rank | 508 | 1530 | 2038 |
| Author | 340 | 83 | 423 |
| Title | 250 | 186 | 436 |
| To | 57 | 52 | 109 |
| Path | 73 | 93 | 166 |
| Other | 26 | 19 | 45 |
| Total | 4316 | 3938 | 8254 |

Table 3 Use of sorting options for results presentation.

People who started with the Date sort issued somewhat more queries than those who started Rank, but the difference is not reliable ($t(231)=0.90$, NS). Regardless of which sort order people started with, they issued more queries in which they sorted the results by Date, suggesting that Date is a useful attribute for finding personal items. The difference was reliable for those who started with Date ($t(110)=4.90$, $p<0.01$) but not for those who started with Rank ($t(121)=1.15$, NS). The next most popular sort attributes after Date and Rank were Author and Title.

Item Previews

We recently added item previews to the interface. The previews show the first 256 bytes of a message as well as thumbnails for images and PowerPoint slides. The default setting for all users was to show the previews. Thirty nine of 234 users (16.7%) modified this default setting at some time. Some turned it off, and others turned it off then back on. The final distribution of interfaces for these users was 19 turned off the previews and 20 returned to the previews, so there does not seem to be much of a usage preference.

Questionnaire Data

In our questionnaire, we assessed how frequently and easily people searched for information both before and after using SIS. Differences in the pre- and post- measures give us an indication of overall effect of the SIS system on user's

searching behaviors. Forty five people responded to both the questionnaires. The questionnaires were completed about a month apart. Participants were asked to report on recent searching behavior using both Likert scale questions (e.g., "When I need to search for a web page that I have seen before, it is easy for me to find it quickly.") and frequency estimates (e.g., "Yesterday, approximately how many times did you use a search engine to find a web site that you had previously visited?"). We asked similar questions for search over email stores and the local file system as well. Figures 5 and 6 show the results of the ease of finding as measured by the Likert scale (Figure 5, top), and frequency of use estimates (Figure 5, bottom).

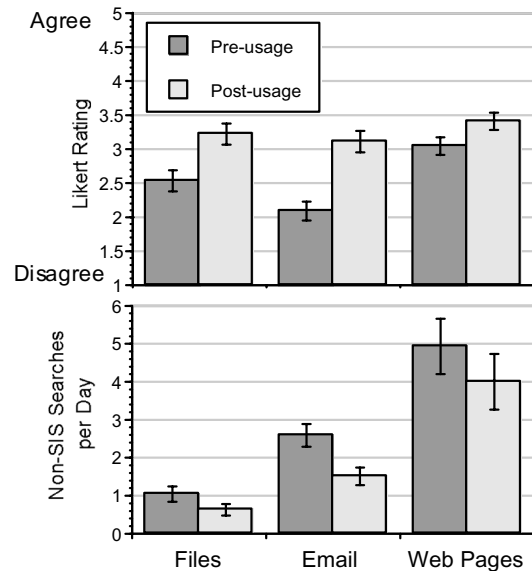


Figure 5. Ease of finding information, and frequency of use estimates, before and after SIS.

For all types of information seeking (files, email, web), users were more likely to say that they could find information quickly after SIS usage (paired t-tests: $t(43)=2.39$, $p<0.01$ for files; $t(44)=6.64$, $p<0.01$ for email; $t(44)=4.35$, $p<0.01$ for web pages). The difference is smallest for web pages, which isn't surprising since good web search tools exist. Larger differences are more evident for email and files. After SIS, people find it equally easy to find all three kinds of information (light bars in Figure 5).

In addition, users estimated that they searched less frequently using native applications, again for all types of information (Wilcoxon signed ranks tests: $Z=2.25$, $p<0.05$ for files, $Z=3.76$, $p<0.05$ for email, and $Z=2.09$, $p<0.05$ for web pages). This is presumably because they now searched for this information using SIS instead.

The questionnaire asked about participants current filing strategies. Of the respondents, 39% were Frequent Filers, 52% were Spring Cleaners, and 9% were No Filers.

The questionnaire also asked for general information about SIS. Some interesting items are listed in Table 4. These questions were answered on a Likert scale, where

1=strongly disagree and 5=strongly agree. Users were overwhelmingly positive about SIS, claiming that a SIS-like service should be an essential functionality for any computer system. They also very much liked the item previews, but they had trouble understanding different kinds of searches such as fuzzy match and Boolean queries. Finally, note that they reported SIS was used much more for email than for web pages, mirroring the log findings.

| Question | Rating |
|--|--------|
| SIS-like search service should be an essential functionality in any computer. | 4.55 |
| The previews of items that SIS provides are useful | 4.06 |
| I typically find things very easily with SIS | 3.89 |
| I typically use SIS to look for e-mail messages | 3.87 |
| I would get more value out of SIS if it indexed documents over all of my machines rather than just one client. | 3.67 |
| I would likely put less effort into maintaining a detailed set of folders for my files if I could depend on SIS to find what I am looking for. | 3.45 |
| I typically use SIS to look for web pages | 3.00 |
| It is easy to understand the different kinds of searches (e.g., exact, fuzzy, and / or) that I can make with SIS. | 2.94 |
| I use advanced query syntax to help me find what I'm looking for. | 2.35 |

Table 4. Subset of post-installation questionnaire items, 1=strongly disagree, 5=strongly agree.

User Comments

When people filled out the SIS usage survey, we also gave them the opportunity to write general comments about SIS. People said that SIS was often quite useful when they could only remember one or two vague things about an item:

"I'd deleted an email that turned out to be vital for a customer engagement, only knew its general timeframe - and wouldn't have found it again without SIS." – A technology specialist in California

"SIS has been very helpful in finding older emails. It has especially been helpful in tracking down messages where 'I know I saw an email about something like that a while ago' but I couldn't remember any specifics." – A software tester in Washington

People also said that SIS was helpful for finding things that were 'buried' or filed in the wrong location:

"The ability to find misfiled documents and email has been extremely helpful." – A sales executive in Washington D.C.

"Without SIS I would sometimes never find the info I knew I had but was too buried to find." – A technology specialist in Washington D.C.

People have also said that SIS sometimes helps them find things that they don't remember doing at first:

"I was looking for contacts on .Net patterns, and didn't know where to begin. So I slammed '.Net patterns' into SIS and it found that I had browsed the MGB agenda (over a month ago!) and there were .Net patterns

sessions. I went back to the page and found that these sessions had contacts... who I mailed! Awesome!!" – A consultant in the United Kingdom

SUMMARY AND FUTURE WORK

We have designed, deployed, and evaluated a system that provides unified access to information a person has seen, regardless of where it came from. SIS continues to be installed and used by more and more people in our institution. Initial findings are quite positive. They show that people find information more easily using SIS, use other search tools less frequently after installation, and use dates and filters to hone in on relevant items. However, much work remains to be completed.

The first area for improvement is the overall performance of the search engine. Currently, the CPU resources required to maintain an index of an individual's rapidly changing content is still relatively high. The initial time required to index the content is understandable, and incremental costs are minimal. But, whenever SIS is stopped and restarted it has to check to see if anything changed. This time is noticeable especially on laptops which are turned on and off regularly. Second, on some systems query results sometimes take more than a second to retrieve which is longer than we would like. To address this in the short term we have added an explicit "go" button that is used to start all queries (instead of having queries start automatically anytime any filters are changed). These performance issues will be addressed in a future release.

We are exploring the possibility of providing users with the way to tag files with meta-data via SIS. We believe that fast and effective search coupled with simple tagging could greatly reduce the need to maintain separate organizational structures for files, email and web pages, as is also hypothesized by the Placeless System [11].

We are also developing very different visual presentations of results including timeline visualizations with personal landmarks to further tap individuals' memories for their own content. Finally, we would like to extend the prototype to cover information on the fringes of awareness (i.e., Stuff I Should See), and to collaborative settings.

ACKNOWLEDGMENTS

Anonymizing for now ...

REFERENCES

1. Adar, E., Karger, D. and Stein, L. A. (1999). Haystack: Per-user information environments. In *Proceedings of CIKM'99*, 413-422.
2. Anderson, J. R. and Schooler, L. A. (1991). Reflections of the environment in memory. *Psychological Science*, 10, 396-408.
3. Barreau, D. and Nardi, B. A. (1995). Finding and remembering: File organization from the desktop. *SIGCHI Bulletin*, 27(3), July 1995.
4. Bellotti, V. and Smith, I. (2000). Informing the design of an information management systems with iterative

- fieldwork. *Proceedings of DIS'00: Designing Interactive Systems*, 227-237.
5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. L. (2000). Graph structure in the Web, In *Proceedings of the 9th International WWW Conference*, 309-320.
6. Burrell, Q. L. (1980). A simple stochastic model for library loans. *Journal of Documentation*, 36(2), 115-132.
7. Bush, V. (1945). As we may think. *Atlantic Monthly*, 176, 101-108.
8. Bush, V. (1967). *Science Is Not Enough*. Chapter:: Memex revisited, 75-101. New York: William Morrow.
9. Catledge, L. and Pitkow, J. (1995). Characterizing browsing strategies in the World Wide Web. In *Proceedings of the 3rd International World Wide Web Conference*, 1995.
10. Cockburn, A. and Greenberg, S. (1999). Issues of page representation and organization in web browsers revisitation tools. In *Proceedings of OzCHI'99*, 7-14.
11. Dourish, P., Edwards, W. K., LaMarca, A. and Salisbury, M. (1999). Presto: An experimental architecture for fluid interactive document spaces. *ACM Transaction on Computer-Human Interaction*, 6(2), 133-161.
12. Erickson, T. (1996). The design and long-term use of a personal electronic notebook: A reflective analysis. *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, 11-18.
13. Fertig, S., Freeman, E. and Gelernter, D. (1996). Lifestreams: An alternative to the desktop metaphor. In *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, 410-411.
14. Greenberg, S. (1993). *The computer user as toolsmith: The use, reuse and organization of computer-based tools*. Cambridge, MA: Cambridge University Press.
15. Huynh, D., Karger, D. and Quan, D. (2002). Haystack: A platform for creating, organizing and visualizing information using RDF. Available at <http://haystack.lcs.mit.edu/papers/computer-network2002.pdf>.
16. Jones, W. P., Bruce, H. and Dumais, S. T. (2001). Keeping found things found on the web. In *Proceedings of CIKM'2001*, 119-126.
17. Jones, W. P., Dumais, S. T. and Bruce, H. (2002). Once found, what next? A study of 'keeping' behaviors in the personal use of web information. To appear in *Proceedings of ASIST 2002*.
18. Jones, S. R. and Thomas, P. J. (1997). Empirical assessment of individuals' 'personal information management systems'. *Behaviour and Information Technology*, 16(3), 158-160.
19. Kaasten, S. and Greenberg, S. (2001). Integrating back, history and bookmarks in web browsers. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI'02)*, 379-380.
20. Kaasten, S. and Greenberg, S. and Edwards, C. (2002) How people recognize previously seen WWW pages from titles, URLs and thumbnails. In *Proceedings of Human Computer Interaction 2002*, 247-265.
21. Malone T. (1983). How do people organize their desks? Implications for the design of office information systems. *ACM Transactions on Office Information Systems*, 1(1), 99-112.
22. Marais, H. and Bharat, K. (1997). Supporting cooperative and personal surfing with a desktop assistant. *Proceedings of UIST 1997*, 129-138.
23. McKenzie, B. and Cockburn, A. (2001). An empirical analysis of web page revisitation. In *Proceedings of the 34th International Conference on System Science (HICSS34)*, CD Rom.
24. Nardi, B., Whittaker, S., Isaacs, E., Creech, M., Johnson, J. and Hainsworth, J. (2002). Integrating communication and information through ContactMap. *Communications of the ACM*, 45, 89-95.
25. Pitkow, J. and Pirollo, P. (1997). Life, death, and lawfulness on the electronic frontier. In *Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems*, 383-390.
26. Recker, M. M. and Pitkow, J. (1994). Predicting document access in large, multimedia repositories. *Georgia Tech, Tech Report*, August 23, 1994.
27. Silverstein, C. Henzinger, M., Marais, H. and Moricz, M. (1998). Analysis of a very large Alta Vista query log. *SRC Technical Note 1998-014*, October 26, 1998.
28. Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 53(2), 226-234.
29. Tauscher, L. and Greenberg, S. (1997). How people revisit Web pages: Empirical findings and implications for the design of history systems. *International Journal of Human Computer Studies, Special issue on WWW Usability*, 47(1), 97-138.
30. Whitaker, S. and Sidner, C. (1996). Email overload: Exploring personal information management of email. *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, 276-283.
31. Yule, G. U. (1944). *Statistical Study of Literary Vocabulary*, Cambridge University Press.
32. Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*, Addison-Wesley.

