Using Autobiographic Information to Retrieve Real and Electronic Documents

Daniel Gonçalves, Tiago Guerreiro, Renata Marin, Joaquim A. Jorge

Dep. Engª. Informática, IST
Av. Rovisco Pais, 1000 Lisboa, Portugal
{daniel.goncalves,tjvg,jaj}@inesc-id.pt, renata.marin@gmail.com

Abstract. Current systems force users to store their documents in hierarchic filesystems. However, users remember their documents not in terms of ad-hoc categories, but of their contents, why they were written, etc. Describing documents using such autobiographic information would be a simpler and more effective way for users to retrieve them. Furthermore, we found that users mention printed versions of documents when retrieving them. However, current systems don't regard the printed and electronic versions of a document as facets of the same entity. To solve these problems we developed an interface that allows users to browse their autobiographic information, gathered by a special-purpose monitoring subsystem, to retrieve their documents. We bridged the gap between the real and electronic worlds by maintaining the association between paper and electronic documents resorting to RFID technology. Tests show our interface allows users to efficiently retrieve their documents.

Keywords: personal information management, document retrieval, RFID

1 Introduction

It is easy for users that have tried to find an electronic document stored somewhere in their computers to remember how ineffective and frustrating traditional retrieval mechanisms can be. All files containing documents are stored somewhere in the hierarchic file system and have a file name. Unfortunately, apart from that information, there is little else that can be used to find the correct document. The file system's hierarchy could help but it is often hard for users to classify all their documents into categories in that hierarchy. A document might seem to belong to more than a folder, or none of the existing ones might apply. These and other problems make it difficult to properly classify documents. Indeed, Thomas Malone found that many users would rather not classify their documents at all, preferring to store them in roughly unsorted piles [5]. But, as we have already discussed, a good classification into the hierarchy is crucial, as it is one of the few cues users have available at a later time to help them find their documents.

Compounding on the classification problem are the large numbers of documents that users must deal with nowadays, on a daily basis. Not only have computers left schools, labs and the enterprise to become commonplace on most citizen's homes, but

also each user now performs a large and growing number of tasks with them. From writing letters to paying taxes, computers are increasingly common. For each task one or more documents are often produced. Thus, and considering that most users are now non computer-savvy, better and easier ways to organize and retrieve personal documents must be developed.

While research on this area has been taking place for decades, it concerns itself mostly with the retrieval of text-based documents. Keyword search is common, even in recent systems such as Google Desktop. This reflects the reality of some years ago. Today, with the lowering costs of storage and bandwidth, multimedia formats are gaining more and more importance, as attested by popularity of web sites as Flickr and YouTube. Many users now store hundreds or thousands of digital photos, music files, and video clips. These should be amenable to retrieval in a way similar to that of text-based documents.

Finally, many works in the field of Information Retrieval try to help users to retrieve any kind of text-based documents, from news articles to scientific papers in digital libraries. Such works don't give enough relevance to *personal* documents. The retrieval of a user's own documents, unlike the general case of document searching on the web or libraries, has a different but very specific nature. Users looking for personal documents are not tying to find any documents on a given subject, or some document they have some reference to. Instead, they are trying to find *their own documents*, with which they have intimately interacted in the past, having written or read them for a reason, at some point in time. Some particular tasks might have been performed to complete that document, or they had to send it to a friend or colleague for revision. In short, users remember about their documents a wealth of autobiographic information they cannot associate to documents in the general case. Remembering this autobiographic information is easier than recalling arbitrary values imposed upon them by the computer, such as the location in the filesystem. Thus, using this information would undoubtedly be more natural and easier. Furthermore, users are able to recall a large amount of such information and consequently can provide the retrieval system with a large number of hints of where to find a document they seek [3].

It is this ease of using autobiographic information to store and retrieve documents that makes users sometimes resort to their email tools to perform those tasks, even if they are not directly supported by the tools [8]. This is due to the fact that email messages are perhaps the type of document the users most frequently deal with that have autobiographic information associated to them. The message's sender or receiver, its date, subject, etc. can be used to finding them more easily than it would be otherwise possible.

1.1 The Gap Between Real and Electronic Documents

One particular aspect of autobiographic information has been for the most part neglected in previous solutions: real documents. It often occurs that users trying to find a specific electronic document remember they printed it to give to some one or to review it. They might even have a printed copy of the document with them. However, this information is for the most part useless. The reason for this is that once a

document has been printed, nothing connects it to its electronic version. The computer is incapable of regarding the two different versions of the document, the printed and the electronic one, as different facets of a same entity. This would make possible two different scenarios: retrieving a paper document from its electronic version and retrieving an electronic document from its print version. A way to bridge the gap between the real and electronic worlds was, thus, sorely needed.

Some previous works try to accomplish this in different ways. Video-Based Document Tracking [4] tries to solve the physical/virtual document association problem with the help of an overhead camera, used to track physical documents on a desktop and link them to the corresponding virtual documents. The system detects changes in a document stack, and uses this information to establish the desk's contents. The movement of documents is tracked with a video camera. The video is analyzed with computer vision techniques to connect the document with his virtual copy on the disk. It relies on the visual pattern of the first page of a document to identify it, which is clearly not enough when several documents with similar first pages are present at the same time. Also this technique can only recognize documents if their first page is facing upwards and not occluded by some other document or object. As such, this and other similar approaches (such as using barcodes) are ill suited for document retrieval, not only because of their inaccuracy, but also because they don't support the users' work practices: users don't store their documents side-by-side, face up, placing them in piles instead.

Other solutions, aiming at a more natural and accurate linking between virtual and real documents are based on RFID (Radio Frequency Identification) technology. Historically, the roots of the RFID technology can be traced back to the World War II. It was used by the British to distinguish their aircrafts from the enemy's. RFID is a generic term for technologies that use waves to automatically identify objects. Each tag is identified with a unique serial number. A microchip is attached to an antenna that enables the RFID to transmit the identification to a reader. This reader converts the radio waves reflected back from the RFID tag into digital information that can be passed on to computers [6]. Recently, the growing popularity of RFID technology has made readers and tags less expensive, to the point where tagging each document is feasible.

One of the earliest works that tried to bridge physical and virtual worlds through the use of RFID technology was Want et al. [7]. It tries to connect the physical objects with its virtual representation using various types of tags.

More recently, the association between virtual representations and real objects has been extended to encompass personal items that accompany the users on their everyday tasks [2]. This work tries to help people not to forget important objects. The different relevant objects are tagged. RFID readers are present at each of the locations where the user usually spends some time at. Users are given a mini personal server that communicates with their watches. Whenever a user passes close to a reader without an important object (that he could have forgotten) the watch reminds the user of if. This work shows how RFID technology can be employed not only to help users to initiate the retrieval of objects they remember, but also how the computer itself can identify a user's needs and act proactively to provide the required objects.

Abu Safiya et al. [1] developed a project that more directly addresses the information retrieval problem. The concept of Document Database (where all

electronic documents are indexed) is extended to allow the representation of printed documents and their physical location. With this, a company can manage all its documents, becoming possible to know at all times where a specific document is or which documents are in a specific room. The RFID readers must be located in strategic positions, places where documents usually accumulate. While interesting, this work is a large scale project, trying to deal with an organization's entire document collection. Thus, only important documents are tagged identified. It is mostly a localization-based project and the retrieved information is quite poor. Also, the retrieval process is unidirectional: a user can't retrieve a virtual document from its real replica. A proper way to handle *personal* documents was still to be developed.

1.2 Our Approach

To solve the aforementioned problems, we developed a system that automatically collects a wide range of autobiographical information. Provisions were made to allow it to semi-automatically associate printed and electronic documents, with the help of RFID tags. This information is stored interrelated with the remaining autobiographic data, forming a coherent whole.

A custom-designed interface can be used to browse that information in search of personal documents. User tests show this retrieval to be simple and effective.

In the following section, we describe the monitoring system, responsible for collecting all relevant autobiographic information. Next, we will explain the browsing interface, to which a discussion of the user tests performed using it will ensue. Finally, we will conclude pointing to relevant directions for future work.

2. Collecting the Autobiographic Information

To be able to build an interface that allows users to describe the autobiographic information they remember about their documents to retrieve them, it is necessary to somehow collect all that information beforehand so that the correct documents can be identified. There is a wide range of potentially relevant information [3]. As such, the users won't be willing to enter it all by hand, either using a special purpose interface or annotating their documents. Thus, it is necessary to automatically and continuously collect it, in an effective but unobtrusive way.

To that end, we created a monitoring system that monitors the users' actions at the computer and stores all relevant information. A problem we faced when designing this system is the large degree of variability that can be found in terms of software applications, configurations and services employed by users. It soon became apparent that a single monolithic system would never satisfy the needs of all users. We decided to build a plugin-based system instead. Each plug-in is responsible for collecting data from a different source and can be independently configured to match each users' particular details. Also, it will be possible to enable only those plugins that make sense in any given system.

The overall architecture of the system can be found in Fig. 1. There, we can see that the different monitoring plugins are managed by the monitoring system,

configurable with the help of its graphic user interface. This monitoring system stores all data in a special-purpose knowledge-base. The data therein can then be used by the retrieval interface to allow users to quickly and effectively find their documents.

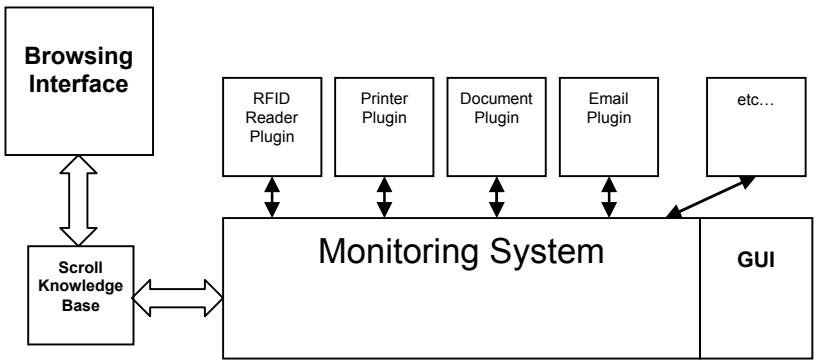

Fig. 1. Overall architecture of the monitoring system.

Overall, we created plugins to collect data from the following sources:

- All of the users' documents stored in their hard drives
- All changes performed on those documents, in real time
- All emails already present in the users' email clients (currently, Mozilla Thunderbird, Microsoft Outlook and Microsoft Outlook Express are supported)
- All email messages sent and received by the user
- All programs ran by the users
- The users' agenda (Microsoft Outlook and Palm Desktop)
- The web pages visited by the users
- All documents printed by the users
- All paper documents located by RFID readers

The challenges were many, concerning mainly poorly documented formats, improperly followed standards and the wide range of encodings, formats and software versions that can be found. In the end, we were able to obtain adequate behavior for all plugins.

Each plugin collects as much information from the relevant source as possible. For instance, the email plugins record not only an email's subject, but also when it was received, to whom it was sent, whether it was a reply of another message, any documents attached to it, etc. All meta-information that can be gleaned from the documents themselves is extracted. For instance, all data in an mp3 file's ID3 tags is collected. Also, text-based documents are analyzed, tokenized, stemmed, and relevant keywords selected using the tf-idf algorithm.

All this data is stored in the *Scroll Knowledge Base*. This is a special-purpose knowledge base developed especially to store autobiographic data. It is based on the RDF standard. It abstracts from it by providing facilities to handle high-level concepts, such as classes and their properties, in easy ways. An RDF Schema dubbed QUILL was created to provide enough expressiveness to allow the representation of all relevant information pertaining a given document (author, dates, title, keywords, etc.), and all events that influence it (being sent by email as an attachment, for instance).

Although each plugin works separately from the others, an effort was made to ensure all the information stored in the knowledge base is tightly coherent. For instance, if a document that had previously been indexed by the Document Plugin is found to be sent as an email attachment by the Email Plugin, instead of creating a duplicated entry for that document in the knowledge base the existing one is annotated with the new information. The same happens for other similar cases.

2.1 Monitoring Printed Documents

Of the different plugins that were developed, one that merits special attention is the printer monitor plugin. This plugin is responsible for creating the association between the digital and paper versions of a document.

To accomplish this, we rely on an RFID infrastructure composed by a fixed RFID reader (with a 30-40 cm range) capable of reading multiple tags simultaneously, connected to a PC. The relatively short range is important, as we want to know precisely where our documents are (a desk or a shelf, for instance). Our RFID tags are passive (with no power source), as we only need to receive the tag ID. Each printed document will have its own tag, which identifies it. The printing plugin associates the tag's ID with the corresponding virtual document.

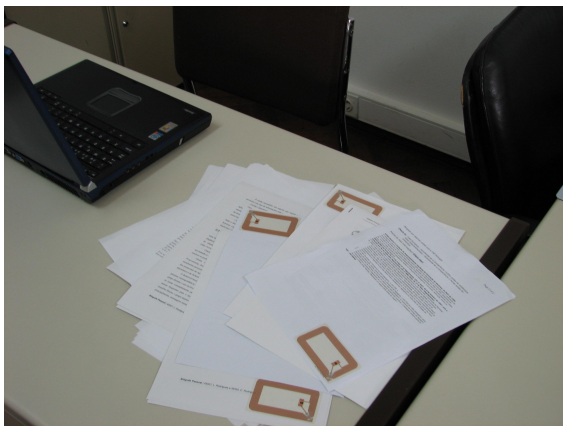

Fig. 2. Tagged documents on the user's desk

Whenever a document is printed, the operating system generates an event. Our plugin intercepts that event to perform the association. From the event, we are able to get the name of the file being printed. The user is supposed to fetch the printed document from the printer and stick an adhesive RFID tag to it. Then, the document can simply be placed in the user's desk, where the RFID reader is placed. The plugin will detect the new tag and automatically associate it to the electronic document that was printed (Fig. 2). From that moment onwards, both versions will be treated by the system as the same document.

The only problem that sometimes hinders this process is that we can get the file name from the print job event, but not the entire path. Thus, it is necessary to inspect the Scroll knowledge base, where all the users' documents are indexed, to identify all files with that name. If there is only one such file, the association is automatic. If the query returns more than one result, several different files with the same file-name, the users are asked to choose between them. All different possibilities will be displayed in a list, sorted by modification date, since it is likely that a file being printed has been recently modified. However, a study we performed shows that over 81% of names identify a single document, with nearly 96% of documents sharing their name with, at the most, two others. Thus, in the general case, the need to choose the correct document is not a problem.

3. Browsing Interface

To allow the users to take advantage of all the autobiographic information stored in the knowledge base, a browsing application was created. This application, depicted in Fig. 3, allows the users to see and inter-relate all different information factoids, regardless of their source. It is based on the concept of *views*. Just as there were different plugins in the monitoring system to monitor different data sources, there are different views in the interface to display information about different kinds of entity in the knowledge base. These views are not fixed or hard-coded. Rather, they are automatically built from the data in the knowledge base, allowing the system to be quickly and easily extended, if necessary. At this moment, the interface supports the following views:

- **Document View**, displaying all information regarding a document;
- **Email View**, that shows all pertinent data about an email message;
- **Person View**, in which everything about a person is presented;
- **Date View**, where everything that took place around a certain moment in time can be seen.

Each view displays all information to which an object of the appropriate type is connected. For instance, the Document View shows all properties of a document, such as its title, authors, creation and modification date, keywords, etc. All views also show other elements of other views with which the displayed element is connected (all emails a certain document was attached to, for instance).

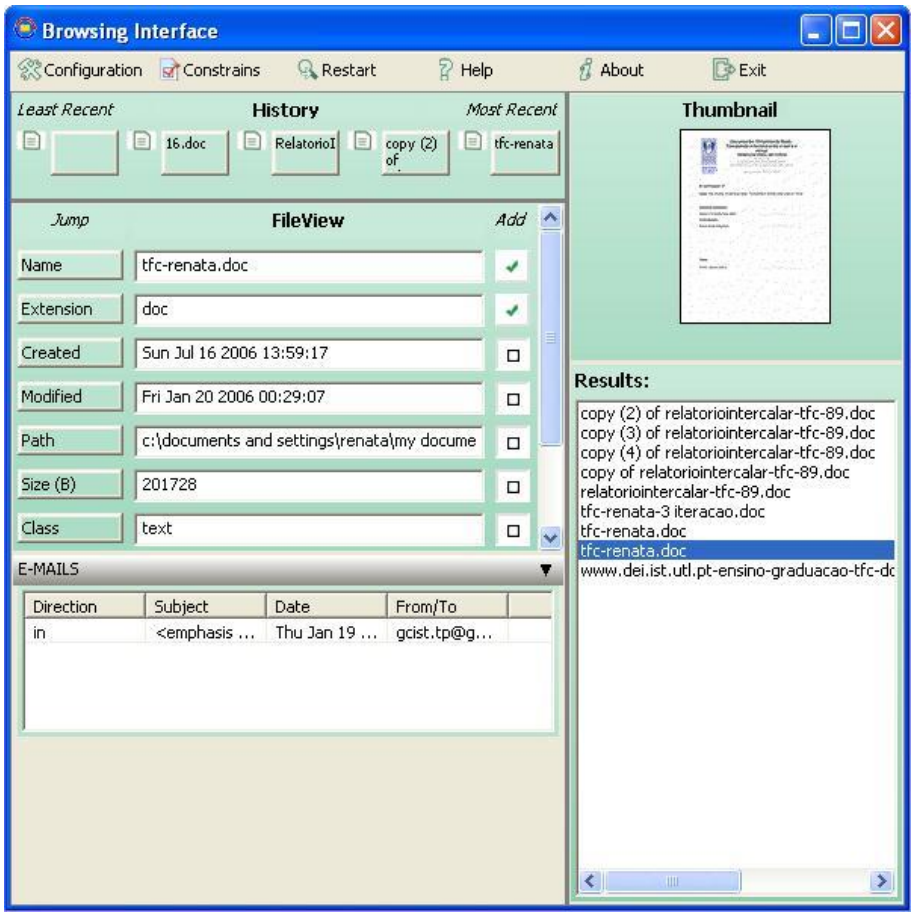

Fig. 3. Browsing Interface

It is possible to navigate between the views by double-clicking on a specific element in the interface. For instance, clicking on the text field with a document's creation date will transfer the user to the View Date for the appropriate time period. It is thus possible to navigate throughout the different information in the knowledge base in search of the appropriate document. Furthermore, it is possible to enter constraints in the interface. For instance, by entering the "doc" string in the field corresponding to the document's extension, in the document view, and then pressing the "Add" button, we will constrain the system to display information only about "doc" files. At all times, every document that satisfies the active constraints is displayed in the list at the right side of the interface. The user can click on each of those to see their information in the Document View. Finally, a thumbnail of the document being displayed in the Document View is shown in the top right corner of the interface. In this way, we take advantage of the users' associative memories and help them recognize the documents they seek.

4 Evaluation

After explaining to the users how the system works, they were asked to print out two documents and associate their printed and electronic versions. Both the time it took to do so and the steps they had to were recorded. Six users participated in this part of the study. We found that it is fairly easy for users to establish the connection between both versions of a document. It took them, on average, 3.7 steps to do so. There was statistically significant difference between the documents regarding the number of steps. However, performing those steps was faster for the second document (49.6s vs. 97.3s for the first one), showing that there is a learning effect and that proficient users might get even lower times.

The times it took to retrieve an electronic document from its printed counterpart and vice-versa, using the browsing interface, were also measured. Finding a real document from its electronic counterpart took, on average, 36.3 seconds. This time includes finding the electronic document using the browsing interface and shows this process to be efficient. Retrieving an electronic document from its printed counterpart was faster, taking only 23s on average.

The browsing interface was also evaluated. Twenty users were asked to retrieve two documents: a recent one (created up to two weeks ago), an old one (created over six months ago). The order in which the documents were considered varied from user to user, to prevent unduly biases to the results. The users' own documents (and other autobiographic information sources such as emails) were indexed and used in the tests, as only for those documents would the users recall relevant autobiographic information.

The time to retrieve the recent document was, on average, 77s, while for the old document it was of 92.6s. This was to be expected, as users had a worse time remembering relevant information about older documents. Nevertheless, all documents were found which in itself is encouraging. More interestingly, 89% of recent documents and 72% of old ones were found in less than two minutes.

The number of steps it took to find a document was also greater for older documents than for more recent ones (6.2 and 5.5, respectively). Again, this reflects the need to perform a more extensive search when looking for older documents.

While the times we presented might seem a little high at first, if we consider that the retrieval was performed by novice users, and that the times include all stages of retrieval, including opening the target document to verify it is the correct one, it becomes clear that our approach produces results which are better than many traditional approaches.

5 Conclusions

While users remember a wealth of autobiographic information about their personal documents, current document retrieval tools don't make it possible to use that information. Furthermore, users often associate to their electronic documents printed versions of those documents. Sometimes they can even have the printed document in front of them and still be no closer to find its electronic counterpart.

To solve these problems, we developed a system that continuously monitors the user's actions at the computer. All documents created and modified are indexed. A plugin-based architecture allows the system to collect data from different sources in an integrated fashion, from the emails sent and received by the user to the web pages visited. This allows a knowledge base of the autobiographic information to be constructed without the explicit need for user intervention, which few would be willing to give. In particular, the monitoring system is able to bridge the gap between real and electronic documents using RFID technology. To navigate the information in the knowledge base, a special-purpose browsing tool was created. It allows users to explore all relevant information until a document is found.

User tests show that both the association between real and electronic documents and the retrieval of a document of any of those kinds can be done quickly and with little effort.

In the future, it would be interesting to extend this approach to other real-world entities besides documents. Also, while we have brought together electronic and real documents when the former are printed, scanned documents should also be handled. Finally, as technology prices continue to drop, an extended usage scenario with more readers should be considered.

Acknowledgments. The authors would like to thank all users that participated in the studies described in this paper. This work was funded by Project BIRD, FCT POSC/EIA/59022/2004.

References

1. AbuSafiya, M. and Mazumdar, S.: Accommodating Paper in Document Databases, DocEng'04.
2. Borriello, G., Brunette, B., Hall, M., Hartung, C., Tangney, C.: Reminding about Tagged Objects using Passive RFIDs, Sixth International Conference on Ubiquitous Computing, UbiComp 2004.
3. Gonçalves, D. and Jorge, J., "Tell Me a Story": Issues on the Design of Document Retrieval Systems. In Proceedings DSV-IS'04, Lecture Notes on Computer Science, Springer-Verlag, July 2004, Hamburg, Germany.
4. Kim, J., M.Seitz, S., Agrawala, M.: Video-Based Document Tracking: Unifying Your Physical and Electronic Desktops, UIST '04.
5. Malone, T. How do People Organize their Desks? Implications for the Design of Office Information Systems, ACM Transactions on Office Information Systems, 1(1), pp 99-112, ACM Press 1983.
6. RFID Journal – The History of RFID Technology http://www.rfidjournal.com/article/articleview/1338/1/129.
7. Want, R., Fishkin, K., Gujar, A., Harrison, B.: Bridging Physical and Virtual Worlds with Electronic Tags. ACM Conference on Human Factors in Computing Systems, Pittsburgh, PA, May 1999, 370-377.
8. Whittaker, S and Sidner, C. Email over-load exploring personal information management of email. In Conference proceedings on Human factors in computing systems, pages 276-283. ACM Press, 1996. ISBN 0-89791-777-4.