Evaluating the Accuracy of Document-Describing Stories

Daniel Gonçalves Joaquim A Jorge
Dep. Engª. Informática, IST
Av. Rovisco Pais, 1000 Lisboa
daniel.goncalves@inesc-id.pt, jaj@inesc-id.pt

Abstract

An increasingly difficult task found by most computer users is searching for specific documents. Traditional document organization forms are more and more ineffective, due to the growing amount of information that even a common user now has to deal with on a daily basis. Additionally, most of those organization schemes are based on artificially imposed conventions, such as the need to classify every document into a hierarchy: the file system. This leads to undue cognitive loads both when storing and retrieving documents. Furthermore, little support exists for non-textual documents. It is urgent to develop new document retrieval mechanisms that reflect the ways in which users naturally remember and refer to their documents, by taking advantage of a wealth of autobiographic information related to those documents.

Our research has shown that narrative-based interfaces can be a natural and effective alternative to facilitate document retrieval. A set of interviews allowed us to identify what shape document-describing stories take, and what contents to expect in those stories. Based on those results, two low-fidelity prototypes were produced and evaluated. The most promising one, dubbed Quill, was then implemented. It includes a knowledge-based infrastructure, used to understand the stories captured by the interface.

A crucial question remained unanswered: are stories sufficiently accurate? With the prototype's help, we collected thirty stories whose contents were then compared to the documents they portrayed, allowing us to conclude that, for the most part, document-describing stories are trustworthy enough to allow the retrieval of documents (81%-91% of all information is correct). We also confirmed that stories told to the computer are similar to those told to human interviewers.

Keywords

Narrative-based document retrieval, Personal Information Management, Knowledge-based interfaces

1. INTRODUCTION

Organizing and retrieving documents have been important tasks since the inception of computing. For some time, the numbers of documents users had to deal with were limited, as were their types. Nowadays, those numbers have grown larger. Not only must the average user deal with thousands of documents, but many of those documents are no longer text-based. Despite that, most tools to organize and retrieve documents remain largely unchanged, based on the document's location in hierarchical file systems. Organizing documents that way has never been easy. It is not unusual for a document to seemingly belong to more than one category. It can also appear not to belong to any of the existing ones. The decisions this forces upon the users give rise to classification problems and undue cognitive loads.

Thomas Malone was one of the first to study the ways in which users organize their documents [Malone83]. The study made evident that users have a hard time classifying all of their documents. Most simply store their documents in unstructured piles, resorting to their visual and spatial memories to later find them. Nowadays the problem still exists.

Recent, popular systems, as Google Desktop (http://desktop.google.com), for Windows, and Spotlight (http://www.apple.com/macosx/features/spotlight) for Macs, automatically index the users' documents, collecting relevant information about them. This allows the file system hierarchy to be, to some extent, circumvented while searching for documents. Those tools' major limitation resides on their interface, centered on keyword search. This might not be enough for non-textual documents, and not be expressive enough to allow users to mention all they remember about their documents. Unlike the more general case of Internet search, it is common for users to remember additional information about their documents other than names, locations or text keywords. Using that information might be advantageous when retrieving them.

Some proposed solutions try to use a wider range of information to facilitate the documents' retrieval. Temporal-based approaches, such as Freeman's

Lifestreams [Freeman96] and Rekimoto's Timescape [Rekimoto99] recognize the importance of time in the way the users' memories are organized. Lifestreams presents the users with all their documents in an ordered temporal stream. Substreams can be created by filtering the main stream with elements such as keywords or the documents' sender. In Timescape, the desktop is a window over a given time period of the users' document collections that can be moved back and forth in time.

Other solutions are property-based. The first such approach was Gifford's Semantic File System [Gifford91]. There, the users were presented with a hierarchy of virtual directories, whose contents were the results of queries based on keywords previously associated to the documents, either automatically or manually, such as their authors or subjects. More recently, we find Dourish's Placeless Documents System [Dourish00], Baeza-Yates' PACO [Baeza-Yates96] and the Haystack system, by Karger et al. [Karger02]. In the Placeless Documents system the organization and retrieval of documents is made by creating document collections, in practice the result of queries on the properties. PACO is similar in its organizational approach. Haystack builds a semantic interface, relating all kinds of personal and web-based information.

Properties often relate only to the users' interactions with their documents, rather than to a wider context. Some studies have shown that autobiographical information to be invaluable when organizing documents [Whittaker96]. Much of that contextual information can be gleaned from the different applications ran by the users. The Stuff-I've-Seen system [Dumais03] tries to integrate information about documents from several desktop applications to organize the documents. So does Ariel Shamir's system [Shamir04].

While property-based systems are promising, in terms of interface they often resort to querying the user about the values of properties. This can lead to problems of its own, since it might be necessary to remember the available properties and possible values. Our research shows that narratives are good alternative to traditional query formulation interfaces for document search. Storytelling is a natural way for humans to communicate. In stories, the several information elements are related as a coherent whole, appearing in a context that facilitates their recall. Stories about documents can, thus, be the means to extract large amounts of relevant information about documents from the users in a natural way. From the analysis of document-describing stories and the evaluation of low-fidelity prototypes, we designed an interface to collect the users' stories. Also, in order to correctly understand them, world and domain knowledge are necessary. Hence, we've studied how a knowledge base can be used to better comprehend the users' narratives.

Two crucial questions remained unanswered. Firstly, we needed to confirm if stories told with no human intervention were similar to those told to the interviewers. In short, is it possible to tell stories to computers in the same effective and natural way than when telling them to humans? Secondly, in order for stories to be useful as a means of document-retrieval, the information in them should be accurate, at least to some extent. If the users' memories betray them and the information in the stories is false, it cannot be used as a criterion for choosing promising documents. To answer those questions we collected thirty new stories told without any human intervention using our prototype. All information in those stories was then compared with actual data about the documents described. The stories themselves were compared with those collected in previous studies. We verified that, indeed, stories told to the computer are identical to those told to humans, both in terms of structure and content. Furthermore, we were able to show that 81% to 91% of all information in stories is accurate, validating stories as a good vehicle for extracting information for document retrieval.

In the next section we will describe the relevant aspects of the prototype we used. The experimental methodology will be mentioned next, following which the results will be described. After a discussion of those results and their implications for interface design we will conclude, pointing to relevant future work.

2. THE PROTOTYPE

To understand what to expect from stories, we conducted a series of semi-structured interviews in which 30 users were interviewed and 60 stories collected. Those stories described three different document types, in search of eventual disparities: Recent Documents, created by the users up to a week ago; Old Documents, created by the users at least a year ago; and Other Documents, from other authors.

After performing a contents and a relational analysis on the stories, we were able to identify their most likely contents, grouped into 17 different categories: Time, Place, Co-Authors, Purpose, Author, Subject, Other Documents, Personal Life, World, Exchanged, Type, Task, Storage, Version, Contents, Event, and Name. Furthermore, we discovered the order in which those categories can be expected to appear in stories. From that data we were able to create several guidelines for the design of narrative-based personal document retrieval interfaces [Gonçalves04].

We then produced two low-fidelity prototypes of two possible interfaces created using the guidelines. One of the interfaces was based on the direct manipulation of graphically represented story elements and the other on the textual representation of the story. The elements were entered with the help of special-purpose dialogues. From the start we rejected the possibility of allowing a freeform text entry interface. When queried about that possibility, most users stated they would not have the patience to write down entire stories. Also, from our data, we know the users tend to digress when telling stories. A completely unconstrained environment would allow this, making it difficult for relevant information to be

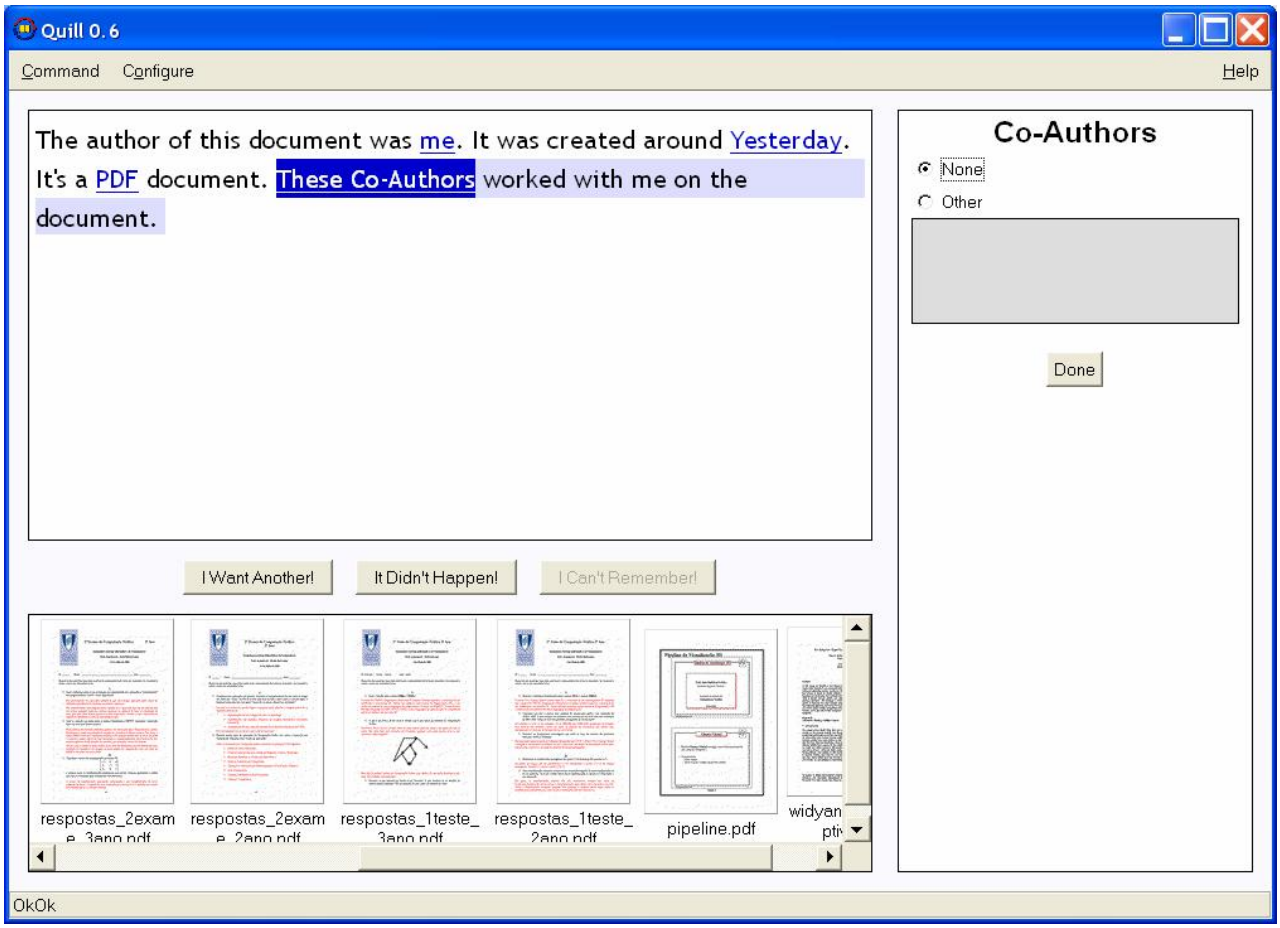

Figure 1– The Quill Inteface

separated from the rest. A new set of 30 stories was collected, and their comparison to those previously gathered showed that, unlike the direct-manipulation interface, the text-based one was able to elicit stories similar, both in terms of structure and contents, to those told to humans. Also, the users were questioned regarding their subjective opinion of the prototypes. It was undoubtedly clear that the text-based prototype was far better understood and liked than the other [Gonçalves04].

A prototype of this interface was implemented using the Python programming language, chosen for its power and flexibility. The resulting application was dubbed Quill.

2.1 The Interface

Figure 1 depicts the overall look of the interface at the time when the study was conducted. The application window is divided into three main areas. On the top left, the stories that describe the documents being sought are incrementally created as the users tell them. In turn, each of the possible story elements is suggested to the user in the form of an incomplete sentence. The missing information is entered into the story with the help of specialized dialogues, which appear to the right of the story area. Those dialogues reflect the different element types that might be mentioned. A degree of flexibility was implemented into the dialogues. For instance, the Contents dialogue allows the users to specify the overall visual aspect of the document. If the document type, text, for instance, had been previously mentioned, only the appearances textual documents can take are shown.

The evaluation of the low fidelity prototypes made clear that it is important for the story to be presented to the users as a coherent whole if storytelling illusion is be maintained. Hence, when the sentences are completed, they undergo some changes to better reflect the new information. This includes number agreement, among others, taking care to keep the sentences as similar to the original as possible to prevent undue confusion.

Quill presents the story elements to users in the order inferred from stories told to humans. Few users ever deviated from that order in our studies. Nevertheless, they can control the flow of the story with the help of three buttons under the story area. The leftmost, “I Want Another” lets them choose what element to mention next from a list. The “It Didn’t Happen” button can be used to state that something didn’t take place. Finally, the “I Can’t Remember” button should be pressed when the users cannot remember some element (a new one will be suggested by the interface). The latter two buttons allow the distinction between not knowing something and knowing something not to have occurred, often detected in the users’ stories.

As the story grows, Quill continuously looks for possible matches. These are displayed in the document suggestion area at the bottom of the interface. Their name is

displayed together with a small thumbnail image of their overall look, whenever possible. Presenting the users with the thumbnails will minimize the cognitive load of scanning the suggestion list for a match, and distract them from the storytelling process as little as possible.

All of Quill's interface elements can be displayed in different languages (at this time, English and Portuguese). This is important for user testing, since English is not the native language of many users of our prototype.

2.2 Infrastructure

In order for the stories told to the prototype to be understood, it is necessary for both domain and world knowledge to be present in the system. Underlying the interface, a knowledge base (KB) is the basis for the narrative-based retrieval system. After a survey of possible formalisms for knowledge representation [Gonçalves04b], we decided to use RDF, from the Semantic Web initiative. It not only has enough expressive power (in several complexity levels to which we can upgrade if necessary: RDF, RDFS, and OWL), but also because it promises to become standard way to represent semantic information on the web. This will eventually make it possible to collect and interpret meaningful information about the users' online actions. We created a semantic network-like KB by defining several RDF case-frames in a RDF Schema we called iQuill. This schema provided an expressivity equivalent to that of first order logic, with the exception of existential quantification and negation (omitted for efficiency reasons). A Python library, Scroll, was implemented creating an abstraction layer over the RDF formalism and allowing the use of the iQuill schema in an easier, straightforward way. In another schema, Quill, we defined all concepts required to store relevant information about the users' documents and actions.

The information in stories is, mainly, autobiographical. Thus, it is important to collect as much data regarding the users and their actions as possible. Most of it can be found on the users' computers. From the documents themselves to the users' actions, a careful monitoring allows all relevant information to be stored in the KB. Also, events away from the computer can be reflected in information therein, such as the users' agendas.

Relying on keywords or annotations provided explicitly by the users was doomed to fail, both because the users won't consistently provide them, and because they would not, in any case, be sufficient. Hence, a monitoring system that takes notice of the users' actions at the computer and enters all relevant data into the KB was implemented. At the time of the experiment only the users' documents were indexed, but currently also are their emails, web pages visited and applications used. Other sources are planned (using RFID tags to bridge the gab between real and virtual documents, for instance), and will be easily implemented into the system given its modular nature.

The knowledge gleaned by the monitoring system is structured, in the KB, mainly as instances of two difference classes defined in the Quill schema: Document and Event. Document instances represent documents on the users' computers. Provisions were made to record a document's evolution (different versions, when it was modified, etc.), as well as keywords or other relevant metadata (the contents of ID3 tags of mp3 files, for instance). The keywords are extracted from text documents with the help of the tfidf algorithm [Salton88]. In order to use it, the text is first tokenized, and all words stemmed, using the Porter stemmer [Porter80]. This process is undertaken in a modular way, allowing for the easy adaptation to different languages.

Event instances represent actions undertaken by the users, visiting a web page or sending an email, for instance. Events are related to the documents involved in them. Apart from this automatically gathered knowledge, more has been stored in the KB. Namely, world- and common-sense related knowledge is used, to help relate the different story elements and understand them. For instance, some knowledge explaining when certain holidays occur is present in the KB, in case they are mentioned by the users.

Each time a new element is entered by the user several custom-designed inference rules are created and evaluated in the KB. The documents that match those rules are given a score by the system. Those with the highest scores are suggested to the user as possible matches.

In some element-entering dialogues, freeform text entry is allowed. It is the case of the Time element, relative to when a document was created or handled. In this case, it is simpler to understand what might be entered by the users, since we're dealing with a limited and well defined domain. As such, we used Context Free Grammars and a chart parser to parse the sentences entered by the users. Furthermore, those grammars were augmented with lambda-calculus formulae that are able to compositionally derive the semantics of phrases during the parsing process itself. Parsing a sentence describing a time instant ("before Christmas a couple of years ago", for instance), will automatically yield a timestamp for the corresponding moment. Coupling those semantics to the knowledge in the KB, it is possible to better comprehend what information the user is referring to.

3. METHODOLOGY

To evaluate the accuracy of stories told to the computer and their similarity to those told to humans, we performed a study in which a set of stories was collected with the help of the prototype we just described.

For this study, access to the users' computers was required, allowing us to assess the accuracy of stories by comparing them with the actual documents they describe. Consequently, we conducted the interviews either at the users' homes or at their workplace. This placed some limitations on the number of users that could be

interviewed, especially for those that handle most of their documents at work, as our presence in their workplace was disruptive of both their and their colleagues work. Overall, we interviewed ten users, six of which at the workplace. We tried to interview not only colleagues and students, but a wider range of users, to prevent biasing the results. The background of the users ranged from a Computer Science consultant to a lawyer. Their ages were between from 26 and 56 years old. Six were male and four female. Each interview took from 45 to 60 minutes.

After meeting the users where their computer is located, we explained the interview's goal and how it was going to be conducted. Then, the prototype was installed in the users' machines. Ensuring that all the files required for the prototype to run were placed in a well identified directory and guaranteeing that they could be deleted without a trace was very important in securing the users' collaboration.

While the program indexed the users' documents, a quick cursory tutorial on how the interface works was provided. We also filled in the interview forms during that time, gathering information about the users (age, profession, etc.). The indexing process would then be interrupted, if it hadn't still finished. This led most interviews to be conducted with only a partial index of the user's documents. Since we were not trying to actually retrieve documents this was not problematic.

The users were then asked to tell three stories about three different documents: a Recent document, an Old document and a document of Other authors. These are the same document kinds for which stories had been previously collected in other studies, allowing a direct comparison between the two. To prevent a bias due to the users' increasing familiarity with the interface, the order in which stories about the different document were requested varied from user to user (Table 1).

	1st place	2nd place	3rd place
Recent	5	3	2
Old	3	7	0
Other	2	0	8

Table 1 – Position of stories in the interviews

The time it took to tell the stories was registered. After each story was told, the users were requested to actually find the document they had just described. After finding the document, actual facts concerning it would be compared to those in the story. Both the data in the stories and the actual facts were saved for future analysis.

3.1 Assessing the Accuracy of Story Elements

Not all elements are amenable to the same degree of verification. For instance, a document's filename can be easily checked, making its confirmation a trivial matter. Verifying if a document was somehow given to someone is not as easy. It would entail checking every email message and every file-transportation medium. Even if this was possible we could never be 100% certain. The users were questioned about those elements and had to make a case for their choices. If it seemed reasonable enough, given other hints gathered from the users' computers and the documents themselves, we considered the information to be accurate. The elements were thoroughly explained, examples of meanings that might have eluded the users were given, and "no stone was left unturned" when questioning them.

Even so, to ensure correctness, we distinguish between two different accuracy levels: **Correct elements**, that we managed to directly verify (filenames, for instance), and **Probable elements**, those we just had no way of verifying directly but that seemed to be correct from all the indicia collected. Some concrete strategies used to assess the truthfulness of some of the most problematic story elements (a full list is given in the technical report that describes the study [Gonçalves05]) were:

- **Purpose:** we questioned the users, in cases where purpose was not evident from the contents.
- **Other Documents:** we requested to see those documents whenever possible (often in the same folders), confronting the users with them. In cases where bits of documents were used in the target document, the verification was immediate.
- **Personal Life:** elements in the users' agendas or known by the researcher (some of the interviewed are old acquaintances), we considered it ok.
- **World Events:** the main way to verify this was to resort to the users' and the interviewer's own knowledge of world events (verifying it when necessary).
- **Exchanges** (sending a document to someone by email, using a CD, etc.)**:** we estimated the information's accuracy from the document's contents and apparent purpose.
- **Tasks:** if the tasks described by the users reflected on the document's contents (inserting images, preparing graphics, etc.), we considered the element as Correct..
- **Events** (occurring while interacting with the document, such as someone entering the room): there really was no way to verify this apart from dialoguing with the users.

4. RESULTS

In this section we will describe the study's main results. We'll start by verifying if stories told using the Quill interface are similar to those told to humans or not. Then, we'll evaluate the accuracy of the information contained in the stories, both globally and regarding each if the possible story elements.

4.1 Telling the Stories...

No user required more detailed explanations about how to use the interface apart from the initial tutorial. With a few notable exceptions (described below), all interface features were correctly used and understood. A learning curve was observed: the time spent on the third story was on average only 60% of the time spent on the first one (from 427 to 269 seconds), regardless of the type of document being described. A sample story, copied from the interface, is:

> The author of this document was me. It was created around 10 of May of 2004. I created it for PCM Report. I worked on the document while I was at home and the workplace and At my colleague's home, in college. André Martins worked with me on the document. The document is about CGEMS Advanced Search Engine. This document reminds me of no other. I sent it to André Martins using email and LAN (shared folders, etc.). It's a PDF document. The document contains the words or expressions "Search Engines, CGEMS, Java, SIGGRAPH" and looks like a two-column with lots of images and a little text. The document is stored in Laptop and Other computer. To write it, I had to developed a prototype for PCM, Search the Web, Read many related papers. It had different versions. Its filename was something like "pcm final".

The English in it isn't perfect, but the sentences' adaptations were enough to produce a human-readable text. Two users told their stories to a Portuguese language prototype, while the rest used it in English (Portuguese computer users are often used to English-based interfaces).

A learning curve was observed, reflecting on the time it took the users to tell their stories (Figure 2). On average, the time spent with the third story is only 60% of the one used for the first one, independently of the document type being described (from 427 to 269 seconds)

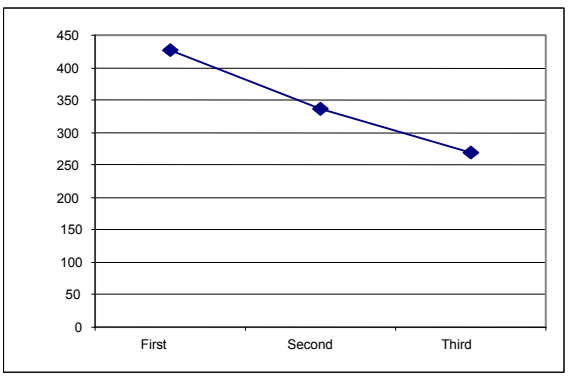

Figure 2 – Time spent telling stories

4.2 Comparing the Stories

To understand if stories told to a computer-based interface are similar to the ones told to humans and if, consequently, the results of previous studies still apply, we compared those told to an interviewer with those told to the prototype, regarding both their structure and contents.

4.2.1 Story Structure

Stories told using the prototype have lengths similar to the ones narrated to interviewers: around 14 elements (Table 2). The ratios between the lengths of current and previous stories are 98%, 101.5% and 100.7%, for Recent, Old and Other documents, respectively. All seems to indicate that, regarding the story lengths, the two sets of stories are, indeed, equal. T-tests confirmed this, with 95% confidence.

	Current		Previous	
	Avg	**StDev**	**Avg**	**StDev**
Recent	14	1.05	14.3	2.06
Old	13.5	1.08	13.3	1.25
Other	13.4	1.43	13.3	2.06

Table 2 – Story Length (in elements)

We evaluated the order in which the different elements occur in stories by taking into account that the order in which they were suggested to the users by the prototype is the one inferred from stories told to humans. Deviations from that order were possible only if the users so wished, simply by clicking on a button and choosing the next element form a list (the users were instructed about this feature beforehand). The number of such deviations is, thus, inversely proportional to how natural the users feel the order to be. Only one user ever chose to mention an element different than the one suggested at the time, once for all three of her stories. On average, such order changes occurred only 0.1 times per story, leading us to conclude that the order in which the elements were presented to the users was natural to them.

In short, the structure of stories told using the prototype is similar to that of stories told to human interviewers.

4.2.2 Story Contents

Trying to establish to what extent were the contents of stories told to humans similar to those of stories told to the prototype, with no human intervention whatsoever, we compared the frequencies with which each of the story elements appeared in stories from both the current study and the interviews described in section 2. The graphic in Figure 3 allows us to compare those frequencies.

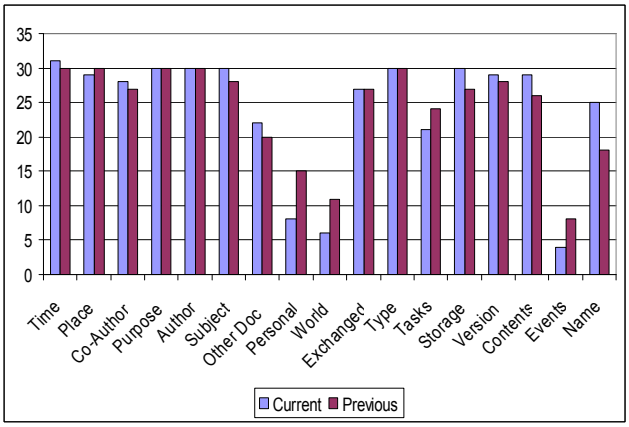

Figure 3 – Element Frequency Comparison

It is immediately apparent that, with few exceptions, the frequencies are very similar. The larger differences occur for four elements. Document Names were mentioned 39% more often in stories told to the prototype. This might be explainable by a conjunction of two factors. Firstly, all users were working on their computers prior to our arrival, which may have made them more conscious of their directory structures and naming conventions. In second place, the simple fact that they were sitting at a computer might have placed them in a more appropriate mindset than in a more informal environment, as was the case of the previous study. Regardless of why, more information is always welcome. The elements Personal Life, World Events and Events occur less frequently (47, 45 and 50 percent, respectively). Those are the three elements that have proven to be more unreliable and harder to remember in all of our studies. Most users don't associate them with documents. A high individual variability has been found and given that we are comparing fairly small numbers of stories, this is enough to account for the changes we found.

Looking at the relative importance of story elements, regardless of absolute value, directly comparing the position of the different story elements in an overall ordering would be inadequate since relatively small changes can lead to order swaps. Instead, we noticed that there are two different types of elements: those that are mentioned in nearly all stories and those far less important mentioned much rarely. We divided the element set according to the following criterion: all elements mentioned in at least 70% of stories went into the "Common" group, and the remaining went into the "Rare" group. That cutoff value was chosen as the value for which a large gap in the element frequency distributions occurs (at least 20%), clearly separating the two element sets.

For Recent documents, no changes were found. For Old and Other documents and overall, only the Name element that, as we have already seen, was mentioned more often, changes from Rare to Common. Personal Life, Events and World Events remain in the same group: Rare. They were always unimportant and remain so.

In conclusion, the contents of stories remain largely unchanged from those collected in the previous study, with the exception of the slightly more important Name element.

4.3 Story Accuracy

We'll now study the accuracy of stories, verifying to what extent we can trust the information in them. First, we'll look at the overall correction of stories, and then we will focus on the individual accuracies of the different story elements.

4.3.1 Overall Accuracy

The graphic in Figure 4 summarizes the percentages of accurate elements in stories for the different element types. As stated before, we considered two element kinds: Correct, those whose accuracy was verified without a reasonable doubt and Probable, those that while believed to be accurate, were not directly verified.

There doesn't seem to be relevant differences between stories about Recent or Old documents. It would seem that users are equally good remembering them. T-tests confirm this. Also, stories describing documents of Other authors seem more inaccurate than those describing the users'. Again, the t-tests confirm this, establishing they have, indeed, different accuracies (with 95% confidence).

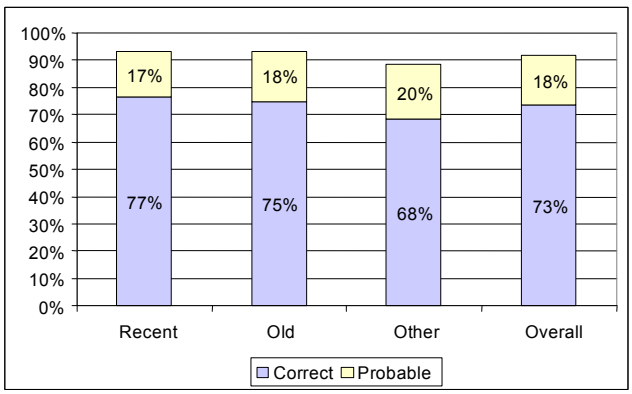

Figure 4 – Overall Story Accuracy Summary

The overall accuracy of stories, while not perfect, it is fairly good. Looking at all the elements, Correct and Probable, all values are around 90%. In the conservative worst-case scenario in which we consider only Correct elements, the numbers are around 70%. On average, between 73% and 92% of what users tell in their stories is accurate. Since there are 17 possible elements in a story, it means that between 5 and 1 will be wrong.

Furthermore, we notice that the fairly large amount of unverified elements (17, 18 and 20 percent for each of the three document types) is due mostly to three elements: Personal Life, World Events and Events. Those elements account for 59, 41, and 47 percent of all unverified information for Recent, Old, and Other documents, respectively. This was due to the verification method used: since most users would just tell that "nothing happened" concerning those elements, it was impossible to verify them. Ignoring those elements when computing the stories' overall accuracy, we find that the numbers of

unverified element percentages decrease dramatically, as can be seen in the graphic depicted in Figure 5. Considering these new values, between 81% and 91% of elements can be expected to be accurate. This will correspond to 1 to 3 untrustworthy elements per story.

Ignoring the three disruptive elements is not problematic. Firstly, they are rarely mentioned in stories, so their overall influence is low. Secondly, we intend to use stories to get information remembered in association with a document. If the users can't remember anything it is nearly the same as if it hadn't, indeed, happened. Rather than providing incorrect information (that *would* be problematic when looking for documents) the users are providing none at all.

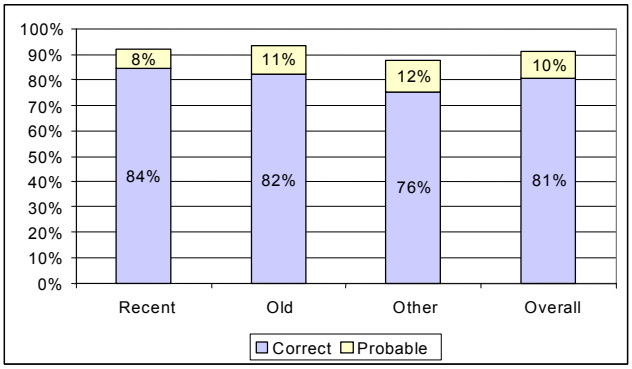

Figure 5: Corrected Overall Story Accuracy

4.3.2 Story Element Accuracy

It is important to know what elements are more often wrong, to better cope with their imprecision. The graphic on Figure 6 shows the accuracy of each separate element, for all document types.

Some elements were notoriously difficult to verify. It was the case of the three least accurate elements, Personal Life, World Events and Events discussed in the previous section. Also in this situation are Exchanges and Tasks. We were convinced they were correct in most cases, but unable to get hard data to verify them.

Name is well remembered least frequently. We witnessed cases that were altogether wrong, but also some in which the users had some idea of the real name but swapped parts of it (*"janeiro2005"* for *"2005janeiro"*, User 7), abbreviated it, or where part of the name suggested by the users was part of the real name.

Next, we find Time. For the most part, the wrong elements were "near misses", falling just outside the predefined tolerance intervals, indicating they should be adjusted.

The third less accurate element is Other Documents. More often than not, the users got something right, but part of the information would be wrong. For instance, User 9 correctly stated the document had the same subject but not its name. The same elements cause problems for either the target document or the Other Documents.

Type mix-ups were due, mainly, to confusions between formats of the same kind: plain text or PDF for Word (Users 1 and 7), for instance. Online documents also caused some trouble: User 6 was unsure of whether a Microsoft Access database file he made available at his personal web site was a "Web Document" or a "database". This might motivate some interface changes.

The other elements all have accuracies above 90%. No further relevant error trends could be identified. Regarding eventual differences in accuracy for the different document types, the only noteworthy aspect is that Author is far less well remembered for Other Documents than for those of the user (80% vs. 100%).

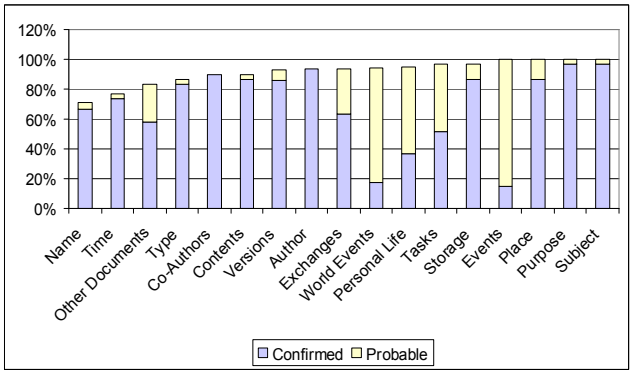

Figure 6: Element Accuracy

5. DISCUSSION

A comparison of stories told to humans and those told using the narrative-based prototype shows that (given the right interface), *no relevant differences occur*. The stories' structures remain the same, both in terms of length and element order. Their content is also similar, except for the Name element that appears somewhat more frequently in stories told to the computer. This is actually positive, since it means the interface has more information with which to work to retrieve documents. It is possible to maintain the feeling of "telling a story" even when the receptor is not human, showing that the design of narrative-based interfaces is possible, without resorting to completely unconstrained text entry, that the users would not be willing to do.

About the trustworthiness of stories, we found that, *for the most part, we can believe the users***.** We strove to verify the accuracy of story elements beyond any reasonable doubt. We managed to do so for 81% of them, and verified that at least between 73% and 92% of story elements can be trusted to be accurate. If some adjustments are made taking into account the nature of some story elements, we can expect accuracy rates ranging from 81% to 91% (1 to 3 inaccurate elements per story). This implies that narrative-based systems must not blindly restrict the documents to be suggested to the users based on the stories, but, instead, implement an approach where each element's influence is limited, in case it is wrong. This can be easily accomplished by weighing each element and establishing a rank order of all documents. Even if some elements cause the document to unduly move in that ordering, their influence will be limited.

Since most elements are accurate, this will be enough to keep the search on track. Furthermore, we've seen that for some elements (Time, for instance) it will be possible to deal with their inaccuracy by considering better tolerance margins, that can be inferred from the data we collected.

Another important result is that *stories about the users' own documents, either Recent or Old, share the same properties*. Some previous results already pointed in that direction but it is, nevertheless, a surprising result. It was to be expected that the accuracy of the information in stories would decrease for stories about older documents. However such a decrease was not observed. In terms of interface design, this means that interfaces for narrative-based document retrieval can be simpler than expected, treating differently only two kinds of stories, rather than adapting to three, and having to cope with very inaccurate stories for old documents. In terms of document retrieval, it reinforces narratives as an effective way to elicit useful and valid information about documents from the users.

Some lessons about the interface were also learned. We've seen that *the interface is easy to understand and learn*. A single short tutorial was enough to teach its use, and after just three stories the time spent to search for a document had fallen to 60% of that of the first interaction. This shows the choices made based on the evaluation of low-fidelity prototypes to have been correct.

6. CONCLUSIONS

Retrieving documents in today's systems is a painstaking task, due both to the growing number of documents and their different types, for which traditional retrieval approaches don't apply well. Narratives about documents seem a good approach to allow users to naturally convey to the computer a wealth of autobiographical information useful to retrieve those documents. To validate that approach, two research questions were posed: can stories about documents be told to computers as they are to humans? Is the information in those stories accurate enough to allow it to be used for document retrieval?

We were able to satisfactorily answer both questions. Stories told to the computer are similar to those told to humans, containing enough information to find documents. In addition, that information is, for the most part, accurate. The adequacy of narratives for document retrieval was, thus, confirmed.

In the near future, and taking advantage of the lessons learned, the interface prototype will undergo some changes to better suit the users' needs and cope with their stories. We will fully integrate the context-monitoring system with the interface, to collect more information that might be used when making sense of stories. New inference rules will make use of that information.

Once the interface reaches a more mature stage, some extended user testing will be conducted. In those tests we will be able to better establish the interface's learning curve and to answer a third pertinent research question: are stories discriminative enough to distinguish between similar but different documents?

7. ACKNOWLEDGEMENTS

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.

8. REFERENCES

[Baeza-Yates96] Baeza-Yates, R., Jones, T. and Rawlins, G. A New Data Model: Persistent Attribute-Centric Objects, Technical Report, University of Chile, 1996

[Dourish00] Dourish, P. *et al*. Extending Document Management Systems with User-Specific Active Properties. *ACM Transactions on Information Syst.*,18(2), pp 140-170,ACM Press 2000.

[Dumais03] Dumais, S. T. et al. Stuff I´ve Seen: A system for personal information retrieval and re-use. In *Proceedings of SIGIR 2003*.

[Freeman96] Freeman, E. and Gelernter, D. Lifestreams: A Storage Model for Personal Data, *ACM SIGMOD Record,*25(1), pp 80-86, ACM Press 1996

[Gifford91] Gifford, D., Jouvelot, P., Sheldon, M. and O'Toole, J. Semantic File Systems. *13th ACM Symposium on Principles of Programming Languages*, October 1991.

[Gonçalves04] 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.

[Gonçalves04a] Gonçalves, D. and Jorge, J. Telling Stories to Computers. *In Proceedings CHI2004*, ACM Press, 27-29 April 2004, Vienna, Austria.

[Gonçalves04b] Gonçalves, D. and Jorge, J, Why RDF? Considerations on a Language for the Representation of Document-Describing Stories. *Technical Report*, *IST/UTL*. June 2004, (available *online* at http://immi.inesc.pt/~djvg/phd/files/why_rdf.pdf)

[Gonçalves05] Gonçalves, D. Real Stories about Real Documents: Evaluating the Trustworthiness of Document-Describing Stories. *Technical Report*, *IST/UTL*. March 2005, (available *online* at http://immi.inesc.pt/~djvg/phd/files/real_stories.zip)

[Karger02] Karger, D. et al. Haystack: A platform for creating, organizing and visualizing information using RDF. In Proceedings *Semantic Web Workshop, The Eleventh World Wide Web Conference 2002 (WWW2002)*. 2002.

[Malone83] 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.

[Porter80] Porter, M. F. An algorithm for suffix stripping. *Program* 14, 130-137, 1980.

[Rekimoto99] Rekimoto, J. Time-machine computing: a time-centric approach for the information environment. In *Proceedings of the 12th annual ACM symposium on User interface software and technology*, pages 45-54, ACM Press, 1999.

[Salton88] Salton, G. Automatic text processing, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1988.

[Shamir04] Shamir, A.. A View on Views. In *Proceedings SmartGraphics04*, Banff Center, Canada, May 2004.

[Whittaker96] Whittaker, S., Sidner, C. Email overload exploring personal information management of email. In *Conference proceedings on Human factors in computing systems*, pages 276-283, ACM Press, 1996