Evaluating Stories in Narrative-Based Interfaces

Daniel Gonçalves, Joaquim A. Jorge
Instituto Superior Técnico
Universidade Técnica de Lisboa
Av. Rovisco Pais, 49
1050-001 Lisboa, Portugal
Tel: +351-213100289
{daniel.goncalves,jaj}@inesc-id.pt

ABSTRACT
Traditional ways to help users organize and retrieve their documents don't scale well, nor do they properly handle non-textual documents. This paper evaluates narrative-based interfaces as a natural and effective alternative for document retrieval. We have identified what shape document-describing stories take, and what contents to expect. This led to an interface that is able to capture stories, and a knowledge-based infrastructure to understand them. A prototype of the interface was used to validate narrative-based interfaces, with emphasis on story accuracy. To this end, we collected thirty stories whose contents were then compared to the documents they portrayed. Results allow us to conclude that, for the most part, such stories are trustworthy enough to allow humans to retrieve documents reliably (81%-91% of all information is correct). We also confirmed that stories told to a computer are similar to those told to human interviewers.

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

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces – *evaluation/methodology, interaction styles, user-centered design*. H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *query formulation*.

General terms: Design, Experimentation, Human Factors

INTRODUCTION
Organizing and retrieving documents are important tasks made difficult by the growing numbers of documents average users must deal with. Also, 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, resulting in misclassifications and undue cognitive loads in users. Recent systems, such as Google Desktop, automatically index the users' documents, collecting relevant information about them. Those tools' major limitation resides on their interface, centered on keyword search. This doesn't help to retrieve non-textual documents, and might not be expressive enough to allow users to mention all they remember about their documents. Other solutions try to use a wider range of information to facilitate the documents' retrieval. Temporal-based approaches, such as Lifestreams [2] recognize the importance of time in the way the users' memories are organized. Others are property-based. It is the case of Placeless Documents [1] and Haystack [5]. While promising, in terms of interface those systems often resort to querying the user property values. This can lead to problems of its own, since it might be necessary to remember the available properties and respective values.

IUI'06, January 29–February 1, 2006, Sydney, Australia.
ACM 1-59593-287-9/06/0001.

Our research shows that narratives are good alternative to traditional query formulation interfaces for document search. 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. Previous studies [3] have shown what shapes can narrative-based document retrieval interfaces take. However, two questions remained unanswered: are stories told to the interface similar to those told to humans? And what is the accuracy of the information in stories? The study described in this paper allowed us give positive answers to both questions,

THE QUILL INTERFACE
After interviewing 20 users and collecting and analyzing 60 document-describing stories, we identified what elements are part those stories and defined a set of interface design guidelines for narrative-based interfaces. Two low-fidelity prototypes embodying those guidelines were built and evaluated. The most promising interface, dubbed Quill, was implemented (Figure 1).

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

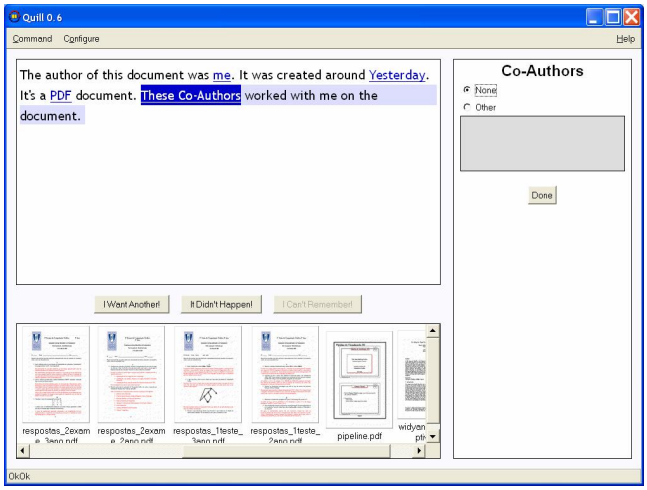

Figure 1: The Quill Interface Prototype

information is entered into the story with the help of specialized dialogues, one for each possible element type. The flow of the story can be controlled with the help of three buttons under the story area: "I Want Another" (lets them choose what element to mention next from a list), "It Didn't Happen" (used to state that something didn't take place), and "I Can't Remember" (to be pressed when the users cannot remember some element). As the story grows, Quill continuously looks for probable matches. These are displayed in the document suggestion area at the bottom of the interface. Presenting the users with document thumbnails minimizes the cognitive load of scanning the suggestion list for a match, and distracts the users from the storytelling process as little as possible.

METHODOLOGY

Ten users were interviewed and 30 stories collected. The background of the users ranged from a Computer Science consultant to a lawyer and their ages from 26 to 56 years. 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 and allowed to index their documents. The privacy of their personal data was ensured. After a short tutorial on how to use Quill, 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. 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. 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. Actual facts concerning that document were then compared to those in the story.

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. To ensure correctness, we considered two different accuracy levels: **Correct elements**, that we managed to directly verify, and **Probable elements**, those we just had no way of verifying directly but that seemed to be correct from all the indicia collected. More details can be found in the technical report describing this study [4].

RESULTS

To understand if stories told to Quill are similar to the ones told to humans, we compared them to those told to an interviewer, collected earlier in our research. Stories in both sets have the same lengths, around 14 elements. T-tests confirmed this similarity, with 95% confidence. The order in which they occur is also similar. Given that the order in which the elements were suggested to the users by Quill was inferred from stories told to humans, deviations from that order show it not to be felt as natural by the users. This happened only 0.1 times per story (on average). Thus, the order in which the elements were presented to the users was natural to them.

Regarding the stories' contents, we looked at the relative importance of story elements. Directly comparing the position of each element in an overall ordering would be inadequate since small changes can lead to order swaps. Instead, 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. The cutoff value corresponds to a large gap (at least 20%) in the element frequency distributions, clearly separating the two element categories. The only difference we found between both sets of stories was that for Old and Other documents and overall, only the Name element changes from Rare to Common in stories told using Quill. In short, both the stories' structure and contents remain largely unchanged from those collected in the previous study: stories are similar to those told to humans.

Figure 2 summarizes the percentages of accurate elements in stories for the different element types. There are no statistically significant differences between Recent and Old

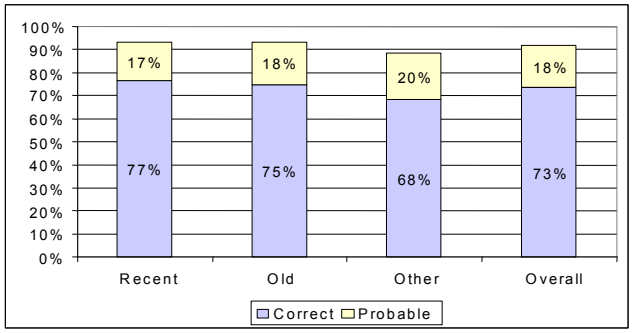

Figure 2:Overall Story Accuracy

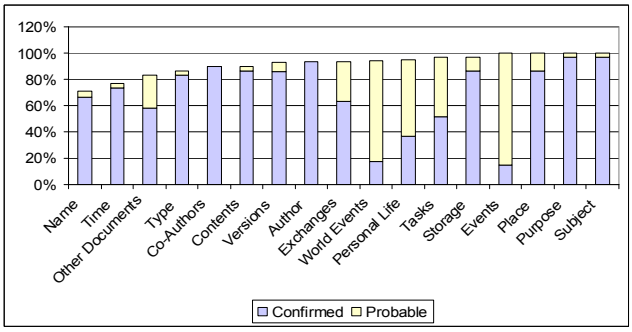

Figure 3: Element Accuracy

documents. On the other hand, stories describing documents of Other authors are more inaccurate than those describing the users'. The overall accuracy of stories is fairly good. On average, between 73% and 92% of what users tell in their stories is accurate, depending of whether only Correct elements are considered or if Probable elements are too. Furthermore, 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. Those elements are rarely mentioned. At best, the users would just tell that "nothing happened". Thus it was impossible to verify them. Ignoring those elements, the stories' overall accuracy rises to between 81% and 91% of accurate elements. This will correspond to 1 to 3 untrustworthy elements per story.

The graphic on Figure 3 shows the accuracy of each separate element, for all document types. There we see that Personal Life, World Events and Events were notably difficult to verify, as already mentioned. Also in this situation are Exchanges and Tasks. While convinced they were often correct, we were unable to get hard data to verify them.

Name is well remembered least frequently. Often, the users had some idea of the correct name, but were unable to recall it properly. Next, we find Time, where most of the wrong elements were "near misses", falling just outside the predefined tolerance intervals, that, thus, should be adjusted. The third less accurate element is Other Documents (suffering from the same problems as the target document). Type mix-ups were due, mainly, to confusions between formats of the same kind: plain text or PDF for Word, for instance. All other elements have accuracies above 90%. No further relevant error trends could be identified. Regarding eventual accuracy differences for the different document types, the only noteworthy aspect is that Author is not as well remembered for Other Documents than for documents of the user (80% vs. 100%).

CONCLUSIONS

Narratives about documents are a good approach to allow users to naturally convey to the computer a wealth of autobiographical information useful to retrieve those documents. We verified that *stories told to the computer are similar to those told to humans*. It is possible to maintain the feeling of "telling a story" even when the receptor is not a human, showing that the design of narrative-based interfaces is possible. In addition*, the information in stories is, for the most part, correct*. We can expect accuracy rates ranging from 81% to 91% (1 to 3 inaccurate elements per story). Thus, narrative-based systems must not blindly consider all information, but instead implement an approach where each element's influence is limited, in case it is wrong. Also, we found that *stories about the users' own documents, either Recent or Old, share the same properties*. This means that it won't be necessary to cope with very inaccurate stories for old documents, and reinforces narratives as an effective way to elicit useful and valid information about documents from the users.

In the near future we will conduct some extended user testing, to answer a third pertinent research question: are stories discriminative enough to distinguish between similar but different documents?

ACKNOWLEDGMENTS

This research was sponsored in part by Portuguese Science Foundation under grant BIRD - POSI/EIA/59022/2004.

REFERENCES

1. 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.

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

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

4. 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)

5. 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.