Describing Documents: What Can Users Tell Us?

Daniel Gonçalves, Joaquim A. Jorge
Computer Science Department
Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
+351 2184 1772, +351 2131 00363
djvg@gia.ist.utl.pt, jorgej@acm.org

ABSTRACT

With the increasing number of computers per user, it has become common for most users to deal with growing numbers of electronic documents. Those documents are usually stored in hierarchic file systems, requiring them to be classified into the hierarchy, a difficult task. Such organization schemes do not provide adequate support for the efficient and effortless retrieval of documents at a later time, since their position in the hierarchy is one of the only clues to a document's whereabouts. However, humans are natural-born storytellers, and stories help relate and remember important pieces of information. Hence, the usage of narratives where a user "tells a story" about the document will be a valuable tool towards simplifying the retrieval task.

To find out if there are common patterns in stories about documents, we performed a study where 60 such stories were collected and analyzed. We identified the most common story elements (time, storage and purpose) and how they are likely to relate in typical stories. This preliminary study suggests that it is possible to infer archetypical stories. Further, we present a set of guidelines for the design of narrative-based document retrieval interfaces.

Categories and Subject Descriptors

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

General Terms

Design, Human Factors.

Keywords

Narratives, document retrieval, Personal Document Spaces.

1. INTRODUCTION

Computers are increasingly common, both in the workplace and at home, and many everyday tasks (from ordering products to e-government initiatives) now require the production of electronic documents. However, the ways of storing and retrieving those documents remain largely unchanged. They involve classifying documents into a hierarchy (a *polyarchy*, if several machines are considered). This forces the users to categorize all the documents, even when no existing category (or more than one) seems to apply. This causes an undue cognitive load. Later retrieving the documents is even more difficult, since one of the only hints to a document's whereabouts is its position in the hierarchy and how a document was classified at storage time isn't necessarily how it will be remembered at the time of retrieval. As a result, finding documents is often a painstakingly time-consuming task. These problems will grow worse as ubiquitous computing becomes a reality and the numbers, types, and locations of documents increase [1]. New tools that allow users to more easily find a specific piece of information (regardless of location), or to visualize their Personal Document Space (PDS) as a whole will soon become an imperative necessity.

The problems inherent to hierarchic document organization have been studied for a long time. The work of Thomas Malone [7] showed that users tend to avoid hierarchies when organizing office documents. Similar results have been found for email messages [2]. However, even unclassified messages can be easily found since they are associated to useful autobiographical information elements (sender, date or reception, other messages received at the same time, etc.). Such information is of capital importance for the retrieval of documents. Works such as Dourish et al's *Placeless Documents* [3] and Freeman and Gelernter's *Lifestreams* [4], where properties associated to documents can be used to find them, try to take advantage of that information. However, they often require users to handle (and remember) arbitrary sets of properties, each of them an isolated piece of information with no apparent relation to the others. Some way to organize those properties into a coherent whole is needed and the most natural way to do it is in the form of stories or narratives, to which users are accustomed to. A narrative-based document retrieval interface would make that task easier and more natural.

To better design such interfaces, it is important to find how document-describing stories are structured. Hence, we performed a study where those stories were analyzed allowing design guidelines to be extracted.

We'll start by describing how the study was conducted. Next, we'll analyze the results thus obtained and present the design guidelines. Finally, we'll discuss the main conclusions and refer possible future work on the area.

IUI'04, Jan. 13–16, 2004, Madeira, Funchal, Portugal.
ACM 1-58113-815-6/04/0001.

2. PROCEDURE

Stories are important not only because of the information therein, but also of how that information inter-relates. We performed a set of 20 semi-structured interviews where we tried to identify not only the most common story elements, but also in which ways they are connected to each other. The participants were chosen from several professional and academic backgrounds, with ages ranging from 24 to 56, to prevent biasing the results.

The participants were asked, in turn, to remember specific documents of three different kinds: a Recent Document (worked on in the last couple of weeks), an Old Document, and a document not created by themselves. Then, they were asked to 'tell the document's story', describing the documents including all they could remember, not only regarding interactions with the computer but reporting to a wider, real-world, context. All interviews were subjected to a Contents Analysis [6]. We manually coded for several elements that, from preliminary studies, we expected to find in the stories (Table 1). No new elements were found in this study. Elements were identified semantically, and could span more than one sentence.

Table 1 – Story Elements

Time	Place	Co-Authors
Purpose	Author	Subject
Other Docs.	Personal Life	World Events
Doc. Exchanges	Doc. Format	Tasks
Storage	Versions	Contents
Events	Name	

We coded for frequency rather than for occurrence, to obtain an estimate of the relative importance of elements. Also, we took notice of what elements were *spontaneous* (proposed by the interviewees) and *induced* (promptly remembered by the interviewee after a question or suggestion from the interviewer). We also took into account that not knowing something is different from knowing something not to have happened. An element was recorded only in the latter case. We also performed a Relational Analysis to estimate how the elements relate in the story. No relation between two elements was considered when the destination element was induced, since in that case no real connection between them exists in the interviewee's mind.

3. RESULTS

Overall, we collected and analyzed 60 different stories, 20 for each document type. In what follows, all values are averages. The stories are 15.85 elements long (st.dev.=5.97). The fairly large standard deviation reflects the different sizes for the user's own documents (17.7) and those with other authors (12.15). Not surprisingly, users remember their own documents better. There is no relevant correlation between length and age, showing that narratives might help alleviate cognitive problems. Regarding gender, women tell slightly longer stories than men (16.81 vs. 14.67 elements). We also found by looking into the elements/transitions ratio that nearly half of each story was not induced by external influences.

The number and length of the uninterrupted element sequences (often spanning more than one sentence) gives us a measure of how the stories are structured. Most stories have three or less sequences. Even if the story is longer, three of them matter the most: the first two account for over 50% of the story, and the last for another 25%. Users tend to easily remember half the available information, and often add a 'burst' of information when they feel the story is coming to an end.

The most common story elements were **Time**, **Place**, **Co-Author**, **Purpose**, **Subject**, **Other Documents**, **Format**, **Exchanges, Tasks**, **Storage** and **Contents** (Figure 1). Some (notably Time) appear more than once in a story, showing that users sometimes provide additional information to reinforce or clarify some element. The most uncommon elements were **Authors**, **Personal Events**, **World Events**, **Versions**, **Events**, and **Names**. They are harder to remember or considered less important by users.

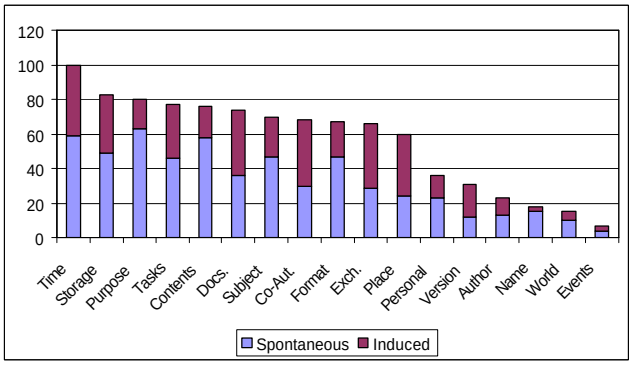

Figure 1 – Overall Element Frequencies

Recent and Old Documents seem to follow similar element distributions, with the exception of **Subject**, more common in Recent Documents. When comparing documents created by the user and those of others, we find differences for **Place**, **Co-Authors**, **Purpose, Author**, and **Version**.

Regarding the percentage of induced elements in stories, the less often induced elements are also the more infrequent, with the exception of **Purpose**. This shows that they are hard to remember (asking won't help), appearing spontaneously when important enough. **Purpose** is probably relevant and easy to remember since it is both spontaneous and common. The more often induced elements, **Time**, **Place**, **Co-Author**, **Other Documents**, **Exchanges**, **Tasks** and **Storage**, appear once per story. They are important but hard to remember, requiring something to jog the users' memories.

As to the nature of the elements themselves, we find a variable degree of accuracy in Time. Other Documents, in paper sometimes, are often mentioned with the help of short sub-stories. The Tasks can refer to both the real world and the computer, and Events are seldom mentioned. References to Content, rather than actual phrases or words, are often related to the overall structure or appearance of the document.

Only 36.7% of all possible transitions occurred more than once. The most common were **Time-Purpose**, **Tasks-Contents**, **Subject-Time**, **Format-Purpose**, and **Storage-Format**. Reflexive transitions are also fairly common, occurring when the user clarifies something he has just said. Different absolute

frequency values might distort the results. We confirmed that no such bias exists by computing the normalized transition frequencies. Furthermore, we calculated, for each story element, the transition probabilities to itself and the others. The most probable transitions were **Place-Place** (0.417), **Contents-Contents** (0.344), **Tasks-Contents** (0.316), and **Time-Purpose** (0.25). This is enough to build some expectations, but not for any certainties.

3.1 Archetypical Stories

From the values above, we trained a Hidden Markov Model to generate archetypical document-describing stories. Stories for the different document kinds to are fairly similar to each other, with the exception of the one for Recent Documents. However, the trends therein didn't prevail when all document kinds are considered, showing a great structure variability for those stories. An archetypical document-describing story could be as follows: **Time, Purpose, Time, Place, Storage, Co-Authors, Co-Authors, Co-Authors, Exchange, Exchange, Format**. More detailed results can be found in the technical report describing the experiment [5].

4. DISCUSSION

We found little relevance of personal factors such as gender and age to the way stories are told. In general, no large user customization will be necessary regarding what to expect from a story. An important conclusion is that older users seem able to tell stories as good as those told by younger ones. Most differences found in the stories are due to document kind (documents of the user vs. those of others). Hence, it is important to determine it early in the narrative, to correctly form expectations about what can be found ahead in the story.

A dialoguing interface is important. Some elements are almost only remembered after a question, so dialogues with users are needed to obtain all the information they can actually remember. On the other hand, the dialogues should not waste resources trying to discover information about elements that are spontaneously but rarely mentioned, showing that if they are to be remembered at all, they will be volunteered with no need for inducement.

References that are taken for granted by the storytellers are also common, relating to their particular context. It is important to take the context in which the story is told into consideration, comparing it to a model of the users' world and of the users themselves. Information such as the one in the user's agenda and address book will be important. Some ability to deal with ambiguity is also necessary and contextual information is of inestimable value in doing so.

We also found that users easily remember the overall structure of documents, suggesting that a way to identify that structure or the document's visual appearance to match it with a user's request would be useful.

Rather surprisingly, events occurring during the user's interaction with the document, are not relevant and there is no need to capture them. More important are other, related, documents. Users often make small descriptions of them. Special care should be taken to capture those recursive stories, while not confusing them with the main document.

Finally, some elements are to be expected more frequently than others, and in a given order. That information should be used to help the system know what to expect at a given point in the story, and to help direct that dialogue to try to discover information they might be prone to remember at that point.

5. CONCLUSIONS AND FUTURE WORK

Growing numbers of documents make new retrieval strategies a necessity. Our innate ability to tell stories can provide an efficient, natural way to do it. We have shown that a wealth of information can be collected from document-describing stories. We managed to discover overall trends in those stories, such as the most probable elements and narrative structures. This allowed us to infer archetypical stories, and extract several interface design guidelines. While dialogues are important in helping users build their story, users of all ages seem to be able to construct them. Context is very important, unlike events occurring during interactions with the document.

Still to be studied are the accuracy of the users' memories and stories, and the scalability of the approach. Also, low-fidelity prototype based Wizard of Oz experiments will help validate these results.

6. ACKNOWLEDGEMENTS

This work was funded in part by the Portuguese Foundation for Science and Technology, under grant POSI/34672/99.

7. REFERENCES

[1] Abowd, G. and Mynatt, E. Charting Past, Present, and Future Research in Ubiquitous Computing. *ACM Transactions on Computer-Human Interaction*, 7(1), pp 29-58, ACM Press 2000.

[2] Bälter, O., Sidner, C. Bifrost inbox organizer: giving users control over the inbox. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 111-118, ACM Press, 2002.

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

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

[5] Gonçalves, D. Telling Stories About Documents, Technical Report, Instituto Superior Técnico, 2003 (http://www.gia.ist.utl.pt/~djvg/phd/files/telling_stories.zip)

[6] Huberman, M. and Miles, M. Analyse des données qualitatives. Recuil de nouvelles méthods. Bruxelles, De Boeck, 1991.

[7] 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.