The Use of Provenance in Information Retrieval

Simone Stumpf, Erin Fitzhenry, Thomas G. Dietterich
Oregon State University
Corvallis, OR, USA
{stumpf, fitzheer, tgd}@eecs.oregonstate.edu

INTRODUCTION

The volume of electronic information that users accumulate is steadily rising. A recent study [2] found that there were on average 32,000 pieces of information (e-mails, web pages, documents, etc.) for each user. The problem of organizing and retrieving information has given rise to many personal information management (PIM) tools (e.g. [2, 4, 5]).

Despite great advances in search technology, information organization and retrieval challenges remain. Users often still navigate manually to retrieve files since generating appropriate search terms is difficult, especially when the time gaps between subsequent accesses of documents are large. Recent research [1] has shown that common search criteria, such as creation or modification time, are remembered inaccurately about 50% of the time. Even the title of a document, the most obvious search criteria, is remembered only partially correct 47% of the time and utterly incorrectly 20% of the time.

There may be other document attributes that are more easily remembered by users that could be helpful in re-finding documents. One such attribute is a document's relationship to other documents [1, 6, 7] but this is not supported in current tools. We are interested in exploiting document relationships for organizing and re-finding information.

PROVENANCE IN DOCUMENT RELATIONSHIPS

There have been various definitions of provenance. One common way of viewing provenance is the history or origin of data (e.g. [8]). We view provenance as the history of information between documents: *How did two or more separate documents come to have the same portions of content in common?*

For example, the following actions could generate document provenance relationships:

- User copies a range of cells from an Excel document to a Word document.
- User creates a copy of a Word document and modifies the copy.
- User receives an email with a Word document attachment and saves it to local storage.

USING PROVENANCE RELATIONSHIPS

Despite the evidence that provenance relationships could be useful, such information is usually not captured. Current operating systems only store document-specific information, such as title, location, time, etc. In order to capture *relationships* between documents, tools need to developed that can detect, store and use this type of information.

TaskTracer

TaskTracer [3] is a PIM tool that allows users to organize and retrieve information based on their activities. The software combines extensive data collection on user interactions, machine learning that leverage this data, and user interfaces to assist users in organizing and re-finding information. Currently, time-stamped user interactions (such as file new, open, print, save, text selection, copy/paste, windows focus, web navigation, email read/send, etc.) are detected in Microsoft Office (Word, Excel, PowerPoint, Internet Explorer, Outlook), the Windows operating system, text and pdf files. Each interaction generates events on *information resources* (such as files, web pages, emails), which are stored in a database.

Some interactions are special since they produce provenance relationships. In TaskTracer, we detect and store the following provenance relationships between resources: *Copy/Paste*, *File Copy*, *Save As*, *File Download (from a web page in Internet Explorer)*, *File Upload*, *Attachment Add*, *Attachment Save*, *Attachment Open*. (We are in the process of instrumenting *Email Reply To/Forward*.)

Provenance relationships in information retrieval

How can these stored provenance relationships help in information organization and retrieval? In our work, we primarily address how provenance relationships can assist users in re-finding documents through manual exploration[1].

A typical use case would be that a user is trying to locate a specific Word document but only remembers an "entry point" in the provenance relationship graph. We have developed a user interface component, TaskTrail (Figure 1), that displays documents and their provenance relationships. In order to see provenance relationships between documents, the user can open TaskTrail from any email, email attachment, or document known to TaskTracer as a resource. The user interface shows a visual representation of

[1] We also intend to investigate information organization by exploring provenance relationships as relational features in clustering for activity discovery.

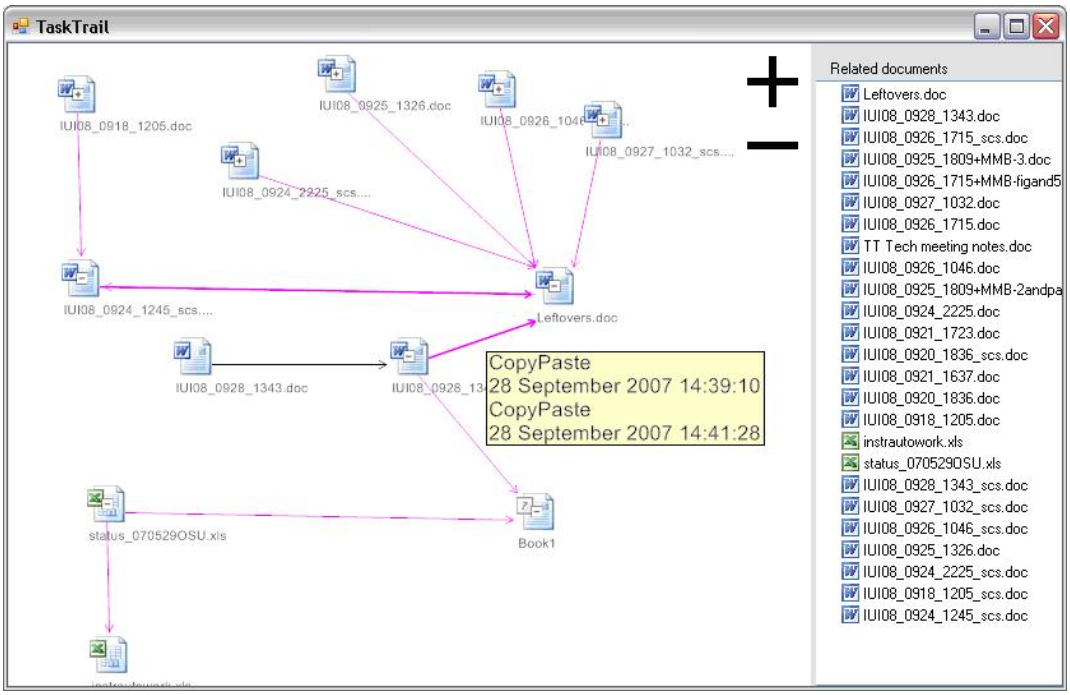

Figure 1. Screenshot of TaskTrail UI showing documents and their provenance relationships.

the entry point document and documents related to it via provenance (Figure 1, left pane), along with all documents contained in the current provenance relationship graph (Figure 1, right pane). The user can open the documents directly from TaskTrail, or explore the graphs further by expanding documents' relationships.

We have conducted a preliminary evaluation of the usefulness of this approach in terms of the amount of document relationships that can be captured. We obtained data from four faculty and student members of the TaskTracer team, covering on average 12.6 days of tracking. On average, each participant had worked with 489.5 information resources overall. On average, there were approximately six provenance graphs for each user, containing on average two resources. The majority of these provenance relationships consisted of Copy/Paste relationships, followed by SaveAs relationships. The results of our preliminary evaluation show that it is possible to capture provenance relationships but also point out the challenges of such an approach. In particular, data about provenance relationships need to be captured over a long-term period to be rich enough to be exploited.

FUTURE WORK

In the future, we would like to explore provenance relationships, their graphical representation, and their use for refinding more deeply. As a first step, we are in the process of collecting long-term data from real-world users.

From a user-perspective, it is important to understand what provenance relationships are considered important. While we have implemented some provenance relationships, it remains to be investigated whether our implementation covers the whole spectrum, and also the role that these relationships play in retrieving documents. Associated with this is how provenance relationship graphs can be visualized to encourage successful exploration. We have designed user studies to answer these questions in order to advance the role of provenance relationships in information retrieval.

REFERENCES

1. Blanc-Brude, T, Scapin, DL (2007) What do people recall about their documents?: implications for desktop search tools. *Proc. IUI.*
2. Cutrell, E, Robbins, DC, Dumais, S, Sarin, R (2006) Fast, Flexible Filtering with Phlat—Personal Search and Organization Made Easy, *Proc. CHI.*
3. Dragunov, AN, Dietterich, TG, Johnsrude, K, McLaughlin, M, Li, L, Herlocker, JL (2005) TaskTracer: a desktop environment to support multi-tasking knowledge workers. *Proc. IUI.*
4. Dumais, S, Cutrell, E, Cadiz, J, Jancke, G, Sarin, R, Robbins, DC (2003) Stuff I've seen: a system for personal information retrieval and re-use. *Proc. SIGIR.*
5. Freeman, E, Gelernter, D (1996) Lifestreams: a storage model for personal data. *SIGMOD 25*, 1.
6. Gonçalves, D, Jorge, JA (2004) Describing documents: what can users tell us?. *Proc.IUI* .
7. Rothrock, B, Myers, BA, Wang, SH (2006) Unified associative information storage and retrieval. *Ext. Abstracts CHI.*
8. Ludaescher, B (2007) A Quick Tour through the Provenance Zoo, *PrOPr*, Philadelphia, PA, 26 June 2007.