eArchivarius: Accessing Collections of Electronic Mail

Anton Leuski, Douglas W. Oard, Rahul Bhagat
University of Southern California, Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina Del Rey, CA 90292
{leuski,oard,rahul}@isi.edu

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval; H.4.3 [**Information Systems Applications**]: Communications Applications—*Electronic mail*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

1. DESCRIPTION

Electronic mail is a ubiquitous communication medium that carries an enormous amount of information [3]. The number of transmitted messages and the importance of the information they carry results in a lot of email being archived. This may happen intentionally – the U.S. National Archives preserves emails as public records of government activity – or unintentionally – emails do tend to "simply accumulate" in our mailboxes. Access to historically significant email collections can provide unique insights into the actions of people who sent and received the messages. For example, imagine a social scientist looking at National Security Council emails for background information on how a policy decision was made; imagine a biographer accessing the email archive of a prominent scientist to find her role in a seminal discovery; or imagine an individual perusing his own personal email collection to remember how papers were selected for some workshop.

Providing effective access to email collections poses challenges for traditional search-based information retrieval because individual messages might not be understandable without insight into the context in which they were generated – context that might not be readily apparent in a list of emails ranked by topical relevance to a query. What is needed is a way to "retrieve" and make sense of that context. Some context is encoded in the email structure (e.g., subject-based threading), other useful clues might be provided by the identity of the sender and the recipients (if their roles are understood), and the pattern of email exchange over time might also offer insight. Thus the problem of access to email collections inherently depends on object-object and object-time dependencies, relationships that are already widely used for hyperlink-based access and for information filtering.

We present *eArchivarius* (`http://www.isi.edu/~leuski/earchivarius/`) – a system for accessing email archives that combines ranked retrieval with cluster-based and time-based navigation. We illustrate the system on a collection of emails from the National Security Council (1985-1987) [1]. The system represents two classes of objects directly: people and messages. Both are modeled as semi-structured data: a set of fields with free text content. eArchivarius automatically extracts the information about the people from the messages during the indexing stage.

A user performing a traditional topic-oriented search in a email collection will end up with a list of ranked messages that resemble a disjoint set of fragments from conversations. Suppose the user finds an interesting message and wishes to see more context. A well-known strategy is to follow up and down a thread constructed using some combination of reply-to links and subject field analysis. Threads discovered in this way can be useful, but they are often unreliable. For example, when composing a new message, replying to an old message is known as an easy way to capture the address.

In addition to threading, eArchivarius uses a cluster-based visualization that depicts messages (or people) as spheres floating in space and positioned in proportion to inter-object similarity [2]. Similar objects (according to some measure) are depicted close together, unrelated objects appear far apart. eArchivarius allows the user to choose a similarity function depending on the type of context that they wish to explore. For example, messages are clustered based on their content similarity or based on the similarity of their intended audience. Similarly, people are visualized based on content similarity in messages that they wrote, content similarity in messages that they received, or the number of messages they exchanged with each other user. The latter approach results in an activity chart that can support social network analysis to gain insight into the roles people play in an organization. The other visualization tool in eArchivarius is a timeline. Depicting messages on a timeline may, for example, help identify patterns of activity that visualization that is aggregated over an extended period would obscure.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

2. REFERENCES

[1] T. Blanton, editor. *White House E-Mail: the top secret computer messages the Reagan-Bush White House tried to destroy*. New Press, New York, 1995.

[2] A. Leuski and J. Allan. Interactive information retrieval using clustering and spatial proximity. *User Modeling and User Adapted Interaction (UMUAI)*, 2003. In Press.

[3] P. Lyman and H. R. Varian. How much information, 2000. Retrieved on 11/10/02 from `http://www.sims.berkeley.edu/research/projects/how-much-info/internet/emaildetails.html`.