Life-browsing with a Lifetime of Email

Sudheendra Hangal
Computer Science Department
Stanford University.
Stanford, CA, USA
hangal@cs.stanford.edu

Monica S. Lam
Computer Science Department
Stanford University.
Stanford, CA, USA
lam@cs.stanford.edu

CHI 2010, April 10–15, 2010. Atlanta, Georia, USA
ACM 978-1-60558-930-5/10/04.

Abstract

Many mainstream users have accumulated large email repositories, egged on by free and ubiquitous service providers exhorting them to "Never delete anything!" Over a lifetime, many of these users can expect to accumulate 50 years or more worth of email archives.

Unlike blogging or keeping a diary or journal, email silently captures our experiences and thoughts, virtually every day, week, month, or year, in situ, as they come up in our communication. Unlike most other life-logging techniques, email already has a large user base with a large volume of data, that can be used for experiments. We have implemented an exploratory email mining tool called Dunbar to understand challenges in life-browsing with a large corpus of email. Based on our experience so far, we outline several research questions in email mining that need to be addressed.

Keywords

Life-logging, Email, Data mining

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User interfaces Evaluation/ methodology

Introduction

Email as Life-Logging

The reach of email transcends geographical, cultural and linguistic boundaries; it has been estimated that there are over 1.3 billion email users worldwide and that this number will grow to 1.8 billion by 2012 [5]. Since the first email message was sent in 1971, the usage of email has evolved significantly. Today, email is used not just for person-to-person communication rich in sentiment and emotion, but also in myriad other settings: for example, to plan events and trips, maintain records, make online purchases, transfer files, track to-do items, process business workflow in corporations, and even to indulge in email wars.

Over the last decade or so, a large number of mainstream users have amassed large amounts of email, often running into several gigabytes, thanks to the availability of cheap storage, and the ubiquity of service providers offering free email storage. Today, it is not uncommon to find people, especially in universities, who have access to email archives going back 20 or more years. It is quite likely that the current generation of Internet-savvy adults will amass 50 years or more worth of email over their lifetimes, and moreover, these archives will remain reasonably accessible to them.

We believe email is a prime target for study in the area of life-logging. In the wired world, email is interwoven with many daily activities and thus contains nuggets about most of a user's thoughts and actions. Indeed, email has become a de-facto tool of record; many people consciously deposit important information they are afraid of losing into their email, knowing that it will be archived and they can look it up later. In corporations, people routinely write up emails to their colleagues containing information that has already been communicated, as a way to maintain a trail.

Ironically, while the pulse of the entire Internet can be easily captured with sophisticated tools like Google Zeitgeist or Twitter trending topics, there are relatively few tools to help individuals understand their own large-scale personal data. Further, email service providers are already mining email and identifying topics relevant to the user for the purposes of advertising, but end users do not have easy ways to mine their own email!

Email archives are useful not only for the owner of the email, but perhaps for their friends and family many years later, or for digital archaeologists or other researchers. In future it should be possible for a 30-year old to recall a field trip he made in kindergarten, along with all the details associated with planning the event that are present in his parents' email archives. Similarly, we expect that people nearing end of life will be able to summarize their life for their loved ones, including all the friends they had, the places they visited, etc. If we are successful building tools to tap the life-logging potential of email, ordinary people will be able to easily write a memoir for their friends and family. We have begun addressing some of the challenges in building these systems with a program called Dunbar. It is publicly available at the URL: http://prpl.stanford.edu/dunbar. In the rest of this paper, we describe what we have learnt so far, and the research issues that need to be tackled.

Use Models

To understand user requirements for life-browsing, we interviewed 10 students at our university and asked them when they felt the need to reflect on their lives, with or without the use of digital artifacts. Most users said they felt the need to reminisce at major life events, e.g. "I am getting married and want to create a recap of my life from childhood till now", or "when my baby grows up to be a certain age", or "when I am about to die." One user said, "If I become famous and I want to write a book." Others said they reminisce at regular intervals, typically on birthdays or at the end of the year. One person mentioned that she has to remind herself what happened in the past year when writing to her friends to whom she sends cards just once a year. Two users specifically mentioned the need to revive memories for tracking work progress: when filling in performance reviews for a manager or a progress report to an advisor, or when "getting ready to update my resume and look for a new job." One user mentioned the need to do so in preparation for a speech. The variety of answers indicates there are many settings in which life-browsing is useful.

Archival Challenges

Email has several desirable properties from the point of view of long term data storage. First, email archiving is virtually automatic and requires less effort compared to other forms of digital data such as videos and pictures. Second, email formats have been relatively stable with a handful of formats that can be translated between one another. For example, the text-based mbox format has been around for a long time, and is easily inter-operable with other email storage formats. Third, email data size is relatively small for the richness of information content it carries, (on the order of a few gigabytes per decade for most users) compared to data-heavy formats like pictures and videos. Thus it is easy to backup and less prone to being lost over time as physical formats change, computers are upgraded, and so on.

We performed an informal survey of about 15 users to determine the status of their long-term email archives. A majority of them had stored a large volume of email messages, but had trouble easily locating and accessing the data. A typical comment was: "I know it is lying zipped up on a CD somewhere in the basement, and I think I could get to it if it was really important." However, we found that most of the people we interviewed did take the trouble of backing up their email when they moved jobs or changed their email accounts; they wanted the safety of knowing that they could get to it "somehow." We believe that it would be useful to develop tools that can intelligently ferret out email messages (and other types of data) from an unorganized or semi-organized digital store, organize it uniformly, remove duplicates, and then process it for presentation. There is also a need to educate users about easy and effective practices to archive their personal email in a way that is amenable for future remembrance.

Mining Techniques

We discuss below some of the key challenges that we have encountered in using email archives for the purposes of life-browsing.

Identifying people

Since our goal is to tackle email corpora built over years or decades, it is important to handle identities correctly. Email addresses as well as the way names are spelt (e.g. with or without a middle initial) are prone to change over time and thus some form of entity resolution is useful.

Text mining

Mining the text of messages and selecting appropriate summaries for presentation to the user is one of the most important features of an email mining system. Traditional text mining metrics like TF-IDF [2] are useful, but there are many ways in which they should be customized for analyzing a large corpus of email. For example, temporal information can be exploited to highlight a common term when it first emerges. Ordinary TF-IDF metrics would de-emphasize such terms due to their high frequency. We have found that single-word terms or tag-clouds are often ambiguous and inadequate when reviewing a corpus amassed over many years; therefore, longer phrases need to be presented to the user to establish context.

User Interface

What is the best means of presenting summarized email archives? An archive spanning a lifetime weaves together many different facets of a person's life, for example, work, family, travel, school, hobbies, etc. How should these threads be identified and presented? How should users be provided drill-down and interactive "zooming" capabilities?

Attachments

A lot of important information is embedded in email in the form of attachments. However, it is not easy to browse these attachments using current email clients. A system that makes the task of exploring attachments as easy as browsing files in a file system browser would be useful.

Related work

The closest prior work on processing email archives is by Viegas et al [6], though it focuses mostly on visualization aspects. TheMail focuses on the user's relationship with one person at a time, so it is probably not suitable for life-browsing which involves exchanges with hundreds of people over many years. Xobni (from xobni.com) is a plugin for Outlook that provides email frequency statistics about a contact (e.g. which hour of day most emails were sent), and lets the user see a list of exchanged attachments, and ranks contacts. Other work in this area focuses on the visualization of the patterns of email communication [1, 3, 4]. However, none of these systems apply any mining to the content of email messages.

Conclusion

We have outlined the utility of considering personal email archives as repositories for life-logging. Based on our experiments so far, life-browsing with the help of email archives appears to be a useful way for people make sense of their lives and relive their memories. There are several challenges that need to be addressed by the HCI community before life-browsing based on email can be practical and useful for the masses.

References

[1] M. Mandic and A. Kerne. Using intimacy, chronology and zooming to visualize rhythms in email experience. In CHI '05: CHI '05 extended abstracts on Human factors in computing systems, pages 1617–1620, New York, NY, USA, 2005. ACM.

[2] C. D. Manning, P. Raghavan, and H. Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[3] A. Perer, B. Shneiderman, and D. W. Oard. Using rhythms of relationships to understand e-mail archives. J. Am. Soc. Inf. Sci. Technol., 57(14):1936–1948, 2006.

[4] A. Perer and M. A. Smith. Contrasting portraits of email practices: visual approaches to reflection and analysis. In AVI '06: Proceedings of the working conference on Advanced visual interfaces, pages 389–395, New York, NY, USA, 2006. ACM.

[5] Radicati Group. http://www.radicati.com/?p=638.

[6] F. B. Viegas, S. Golder, and J. Donath. Visualizing email content: portraying relationships from conversational histories. In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 979–988, New York, NY, USA, 2006. ACM.