Sentiment Analysis on Personal Email Archives

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 2011, May 7-12. Vancouver, Canada
ACM 978-1-4503-0268-5/11/05.

Abstract

A significant portion of a user's digital past is recorded in textual form, for example, in email messages, SMS texts, tweets, status updates and blog posts. We view this text archive as a personal informatics system that captures deep and meaningful information for the user. However, it is a challenge to efficiently browse and extract useful information from an unstructured text corpus spanning thousands of entries accumulated over many years.

We propose the use of sentiment analysis techniques on users' personal text archives to aid in the task of personal reflection and analysis . We have built and publicly released a system called Muse that processes an email archive, and slices it across different sentiment facets, such as those expressing various emotions, congratulatory messages, and messages related to family matters, religion, and health. These slices are used for visualizing the archive and as an entry point into browsing the actual messages. We describe some early experiences with this system.

Keywords

Life-logging, Personal Informatics, Email, Sentiment analysis

ACM Classification Keywords

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

Introduction

In the digital age, ordinary consumers can easily capture and store much of the information that they type. As an outstanding example, archiving of sent email messages is automatic in most email systems, which means that billions of email users have relatively easy access to their own writings in their email histories. For many users, email is a form of passive life-logging, one that requires no setup or effort, but reflects many of their daily activities. In a survey conducted by Li et al, participants reported email history as one of the top sources of information that they collected and reflected upon [3]. Users also create other textual content in the form of blog posts, SMS messages, tweets, Facebook updates, product reviews, etc. Of these, email repositories are the most convenient for the purposes of long-term archiving, since they are under the user's control, and remain accessible over relatively long periods of time. We are therefore seeing the emergence of services that use email to back up other kinds of textual data (such as Android SMS[1]), and even other semi-structured information such as comments and "likes" via a distributed social network[2].

Our overall goal is to enable the task of personal reflection and analysis and to help ordinary people identify interesting events in their pasts, write memoirs and pass on family histories to their children. In terms of the phases of personal informatics, our work falls primarily into the reflection phase [3]. Since an email archive accumulated over even a decade can easily run into several tens of thousands of messages, it is nearly impossible to review all these messages manually. We propose sentiment analysis as a way to identify messages that may have deep meaning for the user. This is particularly useful for the reflection task, since events evoking strong sentiments are particularly memorable and interesting (compared to discussion about, say, setting up a routine work meeting).

Sentiment Analysis

Sentiment analysis is a branch of natural language processing that attempts to mine the zeitgeist of sentiment and opinion around a topic. Pang and Li's survey provides a comprehensive overview of the current state of the art in sentiment analysis [4]. Sentiment analysis techniques often involve the generation and use of word lexicons such as the General Inquirer[3], Linguistic Inquiry and Word Count (LIWC)[4], and Senti-Wordnet[5]. These lexicons are matched with content in a text corpus. A common application for sentiment analysis is to gauge public opinion about a product, as expressed in news articles, blog posts, tweets, etc.

In contrast to such analysis on public data, we note that real sentiment is often expressed in personal communications. Emails are frequently used to send emotional messages of love, joy, and condolence, and reflect deeply meaningful events in people's lives like when they were in love, had a child born, completed a significant career acccomplishment, went through an illness, acquired a new hobby and so on. Therefore sentiment analysis on personal email archives seems a promising direction to pursue.

[1] http://code.google.com/p/android-sms

[2] http://mobisocial.stanford.edu/index.php#mrprivacy

[3] http://www.webuse.umd.edu:9090/tags/

[4] http://www.liwc.net/

[5] http://sentiwordnet.isti.cnr.it/

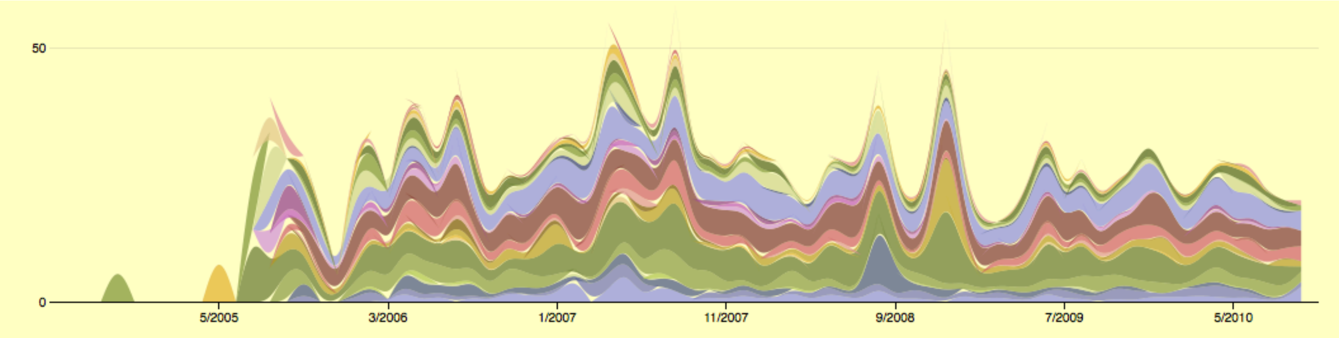

Figure 1: Sentiment analysis in Muse for one user with about 16,000 sent messages accumulated over 6 years. Each layer in the stacked graph visualization represents a different sentiment category. Clicking on the graph launches into a message browser for that category of messages. The Y-axis represents percentage of total message volume.

The Muse email mining system

We have built a system called Muse[6], short for Memories Using Email, that helps users to analyze, mine and visualize their own long-term email archives. For sentiment analysis within Muse, we have developed our own lexicon appropriate for the domain of long-term personal archives. We have used a few parts of the lexicon from the General Inquirer and the LIWC, though large parts (such as identifying pronouns) are irrelevant to us. We have derived a part of our lexicon from the psychology literature on emotions, notably by using lists of basic and complex emotions. We have also incorporated additional categories such as those expressing congratulations, life events (births, deaths, marriages, etc), health and religion. Our current lexicon spans 45 categories and about 500 terms.

Using this lexicon, Muse processes the text of email messages to identify all messages that are related to a particular sentiment category. The message frequencies for each sentiment type across time are visualized using a stacked graph visualization (see Figure 1 for the screenshot with one user's sent email data spanning 6 years and about 16,000 messages. A stacked graph visualization works well given that there are typically a few tens of sentiment categories. When the user hovers over a layer, it changes alpha value slightly to provide visual connectedness and displays a tool-tip with the name of the sentiment.

Clicking anywhere on the visualization opens up a message browsing view in another browsing tab, with all messages related to that sentiment loaded up in the tab. When viewing a long sequence of messages, an on-screen jog dial provides a way to quickly flip through successive messages, attempting to provide an experience similar to flipping through the pages of a book. Terms related to the sentiment are highlighted while displaying the mes-

[6] publicly available at http://mobisocial.stanford.edu/muse

sage contents, and hyperlinked to other messages with the same term.

From our experience, we have found that the visualization allows users to quickly trace relative volume of sentiment across a long period of time. We see users drill down from the visualization into the actual messages in two ways. One way is to browse all messages of a certain type, such as all congratulatory messages. Second, users notice spikes in some sentiment categories, and click on the spike to launch into the message browser. We therefore launch the browser into a view that has all the messages with the selected sentiment, but with the view open at the point along the time axis that the user clicked. Users find this useful to quickly explore the region of interest.

Since email is typically highly personal and sensitive, Muse typically runs on the end-user's own machine and performs all analysis locally. The program runs a local web server on the machine, and the user interacts with Muse via their regular web browser. This lets users use a familiar browser-based interface with attendant benefits that users are accustomed to, such as hyperlinks, browser tabs and plugins, while preserving privacy.

Our work is an initial attempt at bringing sentiment analysis into the domain of personal data. There are several challenges remaining to be tackled. For example, we currently measure only whether a sentiment is active or not; we do not measure its extent of activation. What alternate visualizations are possible? What sentiments occur with each other? Are specific sentiments associated with particular people or objects in the corpus? More work is needed to answer these questions.

Related work

There is some prior work on visualizing email archives by Viegas et al [5], though they focus on displaying words scoring highly on the TF-IDF metric. We Feel Fine[2] is a system for visualizing aggregated trends of feelings expressed in public blogs. Systems like Phlat [1] let users filter and search across heterogeneous text corpora, but are not targetted for browsing purposes.

Conclusion

We have outlined the utility of sentiment analysis in the reflection phase of personal informatics. Real sentiment is often expressed in personal communications. Our approach of using a custom lexicon and a stacked graph visualization as an entry point into browsing the archives appears to be promising.

Acknowledgments

We thank Dan Jurafsky and Jeff Heer for useful discussions. This work was supported in part by NSF Expedition Grant 0832820.

References

[1] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin. Fast, flexible filtering with Phlat. In Proceedings of CHI '06. ACM, 2006.

[2] S. D. Kamvar and J. Harris. We feel fine and searching the emotional web. In Proceedings of WSDM '11. ACM, 2011.

[3] I. Li, A. Dey, and J. Forlizzi. A stage-based model of personal informatics systems. Proceedings of CHI '10. ACM, 2010.

[4] B. Pang and L. Lee. Opinion mining and sentiment analysis. volume 2 of Foundations and Trends in Information Retrieval, pages 1–135, 2008.

[5] F. B. Viegas, S. Golder, and J. Donath. Visualizing email content: portraying relationships from conversational histories. In Proceedings of CHI '06. ACM, 2006.