Analyzing Electronic Mail using Temporal, Spatial, and Content-based Visualization Techniques

Daniel A. Keim, Florian Mansmann, Tobias Schreck
Computer and Information Science Department, University of Konstanz, Germany
{keim, mansmann, schreck}@informatik.uni-konstanz.de

Abstract:
Email is one of the most widely-used means of communication. While mailing volumes have shown high growth rates since the introduction of email as an Internet service and considerable work has been done in improving the efficiency of email management, there is a need for improving the functionality (effectiveness) of email management. Typically, users are given little means to intelligently explore the wealth of cumulated information in their email archives. We address these shortcomings by designing Information Visualization tools for email data. We introduce the *Mail Explorer System* which aims at enabling the user to explore large quantities of email data, reflecting the rich meta data and content stored in email collections. The system allows a user to visually analyze temporal and spatial distribution properties, as well as content-based characteristics in email archives.

1 Introduction

During the last years, the amount of email messages sent around the globe has sky rocketed. Since the early days of email communication, email has evolved into an ubiquitous service used for tasks as diverse as information and file exchange, business planning, organization and scheduling, online support and marketing purposes, among others. As a result, volumes of potentially valuable information are stored in large and growing email archives. While email has reshaped personal and business communication processes in a beneficent way, certain problems like being a target for unsolicited email (spam) and the need for effective archiving and retrieval persist. Methods from Information Visualization can help making better use of email archives by extracting valuable information by means of visual analytics.

Previously, we studied *temporal characteristics* [KMP+05] of electronic mail using the Recursive Patterns technique [KAK95], and we considered *spatial characteristics* by applying geospatial map distortion techniques. Also, in [KMS05] we applied Self-Organizing Maps [Koh01] for *content-based* email analysis. In this paper, we extend our *Mail Explorer System* by combining and improving these approaches into an effective analysis system that unifies different views on the rich email data type. The system assumes the user to have organized her email in an IMAP-based folder hierarchy. By means of linking and brushing, the user is capable of interactively searching for interesting temporal,

geospatial, and content-based patterns in her email collection. Insight gained by this analysis may then be used to adjust personal workload patterns, to identify outlier email (e.g., spam misclassification), or to improve retrieval and storage of email. The paper is organized as follows. Section 2 briefly reviews related work, Sections 3 trough 5 detail our techniques, and Section 6 concludes.

2 Previous Work

Extensive work has been done regarding the efficiency of email management within Database and Information Retrieval research. How to improve the effectivity of email usage through sophisticated user interfaces is not well studied, although email archives are a rich, well-maintained and frequently used source of information. Several researchers studied social networks that can be extracted from email. Becker, Eick and Wilks [BEW95] extracted a social network graph through the analysis of the email sent within their department. Their goal was to identify key communication partners. More recently, Boykin and Roychowdhury [BR04] used graphs of co-recipients for classification of unsolicited mails. Two recent visualizations improve the email user interface by intimacy-based ratings [MK05] and by visualizing email discussions while preserving chronology within threads [Ker03].

3 Temporal Email Exploration

An email contains a time stamp denoting the point in time at which it was sent. As the amount of email being sent out or arriving at a certain time give an indication of the workload of the email account holder, analysis of the temporal attribute can reveal useful insight. The visualization as a simple time chart fails as the details will be lost in such a chart when analyzing large periods of time.

In the latest version of *Mail Explorer*, we adapt the Recursive Pattern technique that was previously used to visualize the temporal distribution of the email. In the new version, the rectangles for each day are grouped in weeks. Weeks are placed on top of each other, which makes mail volumes on same weekdays easily comparable (Figure 1, left), similar to the calendar based visualization in [vWvS99]. As we use pixel-based visualization, the technique is highly scalable and therefore well-suited for very large data sets.

4 Spatial Email Exploration

To assign a spatial location to each email, we use the domain name information and IP addresses of the traversed email servers. This information can be found in the *received* fields of an email header. A geo-IP database [Max] is then used to resolve the geographic position of the respective email servers.

Originally, we employed familiar land-covering maps to display mail volumes on. However, traditional maps are very limited when mapping highly non-uniform distributed spatial locations. But in email this can be the case, as in our experimental email archive many email originate from just a few locations in Europe and the US, whereas hardly any mail comes from locations like Antarctica. Therefore, our *HistoMap* approach (Figure 1, right) is based on shrinking and enlarging geographic areas w.r.t. their importance as measured by the respective data volume, while at the same time maintaining map topology. We generate the distorted map by subsequently dividing the original world map horizontally and vertically into a fixed number of histogram bins. Then, each area is enlarged or shrieked according to the fraction of weighted data points of its bin. This partitioning is applied recursively until each rectangle represents only one weighted data point. To integrate the country hierarchy into the visualization, the algorithm is applied once on the aggregated weights of the spatial points of a country and then again on the data points within each country.

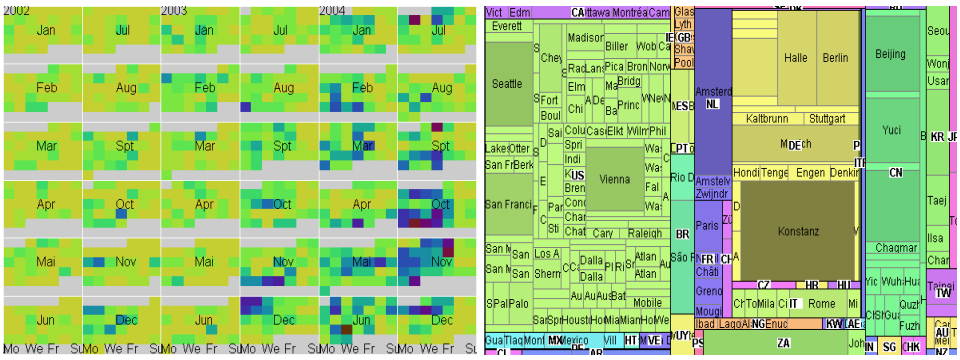

Figure 1: The left image shows a *Calendar-like Recursive Pattern* of the temporal distribution of email from selected folders. Dark colors indicate high email traffic. A clear upward trend of the mail volume is visible, especially on working days. The right figure shows the spatial distribution of spam mailers on the distorted *HistoMap*. In some cases spammers try to hide their location and we can only trace the last mail server which is our own in Konstanz, Germany.

5 Content-based Email Exploration

Besides temporal and geospatial characteristics, analyzing email *content* is an interesting task. Self-Organizing Maps are a well-known technique for projecting a distribution of high-dimensional input data onto a regular grid of map nodes in low-dimensional (e.g., 2D) output space. The nodes each contain a reference vector representing the input data. The projection is capable of clustering large volumes of data while approximately reflecting input data topology. SOMs can be visualized in various ways based on the reference vectors, or on aggregates of the input data mapped back to the grid. We have defined a simple email descriptor based on the well-known $tf \times idf$ document indexing model from Information Retrieval. We represent each email by the $tf \times idf$ weights of its *subject*

field term, considering the 500 most frequent subject field terms in the collection. This descriptor (feature vector) is then input to the SOM generation.

The left image in Figure 2 shows the *spam-histogram* on a SOM we generated from an archive of 9.400 email labeled either *spam* or *non-spam*. The color scheme encodes the fraction of spam email among all the email mapped to each SOM node. Shades of red indicate high degrees of spam, while shades of blue indicate low degrees of spam (these are the "good" email regions). Clearly, the SOM learned from our basic descriptor discriminates spam from non-spam email. The right image in Figure 2 shows the so-called component plane for term "work", where shades of yellow encode weight magnitude. Combining both images, we learn that this specific term occurs in both spam and non-spam email. SOM Email use cases include *organization* of email archives by identifying SOM cluster structure, *retrieval* by matching queries to SOM nodes followed by exploration of neighboring nodes, and *classification* by mapping incoming email to the best matching units on a pre-labeled SOM.

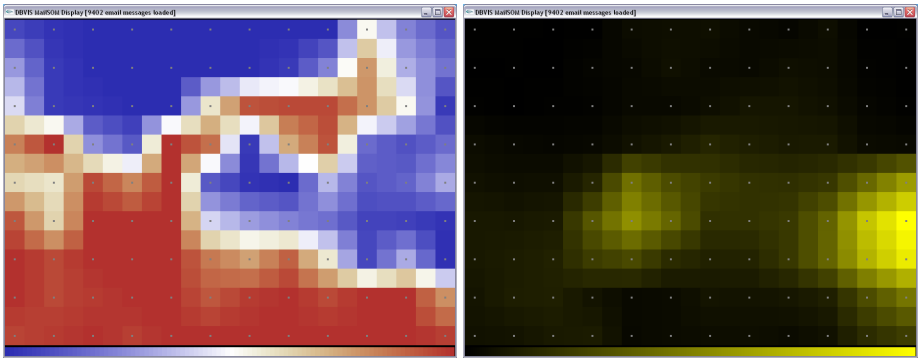

Figure 2: The left image shows a *spam-histogram* of our experimental email archive, where shades of red indicate SOM regions compounding spam email. The right image shows the component plane for term #214 ("work"), with shades of yellow indicating high term weights.

6 Results and Conclusions

In this paper, we presented the *Mail Explorer System* which aims at enabling the user to gain insight into information contained in collections of email. The combination of several visual analysis techniques allows complex use cases. E.g., using *temporal exploration* the user could identify an interval of high email traffic, and within this interval search for an email within a certain folder that was sent from Konstanz, Germany (*spatial exploration*) and that was misclassified as being Spam. The user might then apply the SOM (*content-based exploration*) to check whether there exist similar misclassified messages within the given folder. Future work includes the design of modules supporting additional email attributes, and more evaluation work on several different email datasets.

Acknowledgments

This work was partially funded by the German Research Foundation (DFG) under grant GK-1042 *Explorative Analysis and Visualization of Large Information Spaces* at University of Konstanz, and by the University of Konstanz under grant FP 06/03 *Network Profiling, Network Intrusion Visualization and Network Security*. The authors thank Christian Panse, Joern Schneidewind and Mike Sips from the Databases, Data Mining and Visualization group at University of Konstanz for their valuable comments.

References

[BEW95] Richard A. Becker, Stephen G. Eick, and Allan R. Wilks. Visualizing Network Data. *IEEE Transactions on Visulization and Computer Graphics*, 1(1):16–21, March 1995.

[BR04] P. Oscar Boykin and Vwani P. Roychowdhury. Personal Email Networks: An Effective Anti-Spam Tool. *Condensed Matter*, cond-mat/0402143, 2004.

[KAK95] Daniel A. Keim, Michael Ankerst, and Hans-Peter Kriegel. Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data. In *IEEE Visualization*, pages 279–286, October 1995.

[Ker03] Bernhard Kerr. Thread Arcs: an email thread visualization. In *IEEE Symposium on Information Visualization*, pages 211–218, October 2003. IBM Research.

[KMP+05] Daniel A. Keim, Florian Mansmann, Christian Panse, Joern Schneidewind, and Mike Sips. Mail Explorer - Spatial and Temporal Exploration of Electronic Mail. In *Eurographics/IEEE-VGTC Symposium on Visualization, Leeds, United Kingdom June 1st-3rd 2005*, pages 247–254, 2005.

[KMS05] Daniel Keim, Florian Mansmann, and Tobias Schreck. MailSOM - Exploration of Electronic Mail Archives Using Self-Organizing Maps. In *Conference on Email and Anti-Spam, July 21-22 at Stanford University*, July 2005. To appear.

[Koh01] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 3rd edition, 2001.

[Max] MaxMind. GeoIP City Database. MaxMind LLC, http://www.maxmind.com.

[MK05] Mirko Mandic and Andruid Kerne. Using intimacy, chronology and zooming to visualize rhythms in email experience. In *CHI Extended Abstracts*, pages 1617–1620, 2005.

[vWvS99] Jarke J. van Wijk and Edward R. van Selow. Cluster and Calendar Based Visualization of Time Series Data. In *INFOVIS*, pages 4–9, 1999.