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, as evidenced by the soaring rates rates of mailing volumes since the introduction of email as an Internet service. While considerable work has been done in improving the efficiency of email management, there is a need for improving its functionality (effectiveness). Typically, users are given few 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 have devised the *MailExplorer* 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 its users to visually analyze temporal and spatial distribution properties, as well as content-based characteristics in their email archives.

1 Introduction

In recent years, the number of email messages sent around the globe has skyrocketed. 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, marketing purposes, etc. 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 such as unsolicited email (spam) and the need for effective organization, archiving and retrieval persist. Methods from Information Visualization can help to make better use of email archives by extracting valuable information through 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. In [KMS05], we also applied Self-Organizing Maps [Koh01] for *content-based* email analysis. In this paper, we extend our *MailExplorer* System by combining and improving these approaches into an effective analysis system that unites different views on the rich email data. The system assumes that the users have organized their emails in an IMAP-based folder hierarchy. By means of linking and brushing, the users are able to interactively search for interesting temporal, geospatial, and content-based patterns in their email collections. Insight gained by this analysis may then be used to adjust personal workload patterns, to identify outlier email (e.g. spam mis-

classification), or to improve the retrieval and storage of email. The paper is organized as follows: section 2 briefly reviews related work, sections 3 through 5 detail our techniques, with the conclusions drawn in section 6.

2 Previous Work

Extensive work has been done regarding the efficiency of email management within Database and Information Retrieval research. The ways of improving the effectiveness of email usage through advanced user interfaces are not well studied, even though 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 emails 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 message contains a time stamp denoting the point in time at which it was sent. As the number of emails 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 standard time chart may fail as details could get overplotted for long sequences of observation statistics due to limited screen space. In the latest version of *MailExplorer*, we adapt the Recursive Pattern technique that was previously used to visualize the temporal distribution of email. In the new version, the rectangles for each day are grouped by weeks. Weeks are placed on top of each other, which facilitates the comparison of mail volumes on the same weekdays (Figure 1, left), similar to the calendar-based visualization in [vWvS99]. The technique is highly scalable and can be further refined (e.g. one data point per hour or minute instead of day) until each data point is represented by a single pixel. Therefore, it is well-suited for very large data sets and for a detailed analysis over long periods of time.

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 location of the respective email servers.

Originally, we employed familiar land-covering maps to display mail volumes. However, traditional maps are very limited when mapping highly non-uniformly distributed spatial locations. In email this can precisely be the case, as in our experimental email archive numerous emails originate from clusters in Europe and the US, whereas no email comes from rather exotic, geographically large locations like Antarctica for instance. 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 we simultaneously try to maintain 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 shrunk according to the fraction of weighted data points of its bin. The weights are determined by the number of emails from the corresponding location. 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 first on the country level (the weights within each country are aggregated) and then applied a second time on the data points within each country.

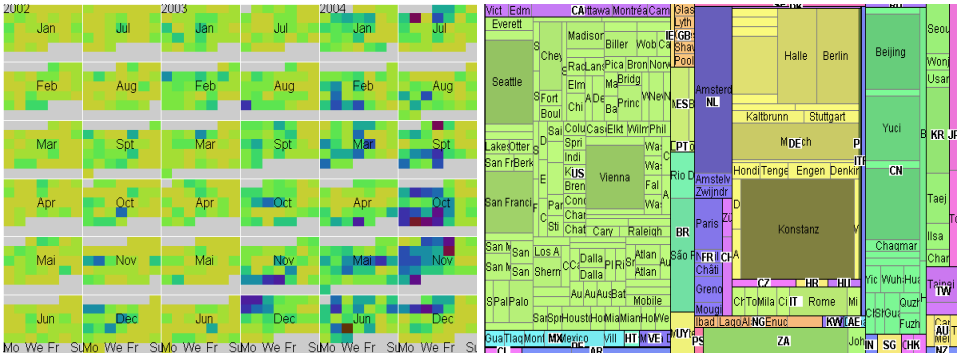

Figure 1: The left image shows a *Calendar-like Recursive Pattern* of the temporal distribution of email from selected folders. Dark colors indicate high email traffic. A clear upward trend of the mail volume is visible, especially for working days. The right figure shows the spatial distribution of 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. Each node contains a reference vector representing the input data. The projection is capable of clustering large volumes of data while approximately preserving 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 the set of terms from its respective *subject* field and mail *body*, considering the 500 most frequent subject field terms as a dictionary for the collection. This descriptor (feature vector) serves as the input for the SOM generation.

The left image in Figure 2 shows the *spam-histogram* on a SOM we generated from an archive of 9.400 emails labeled either *spam* or *non-spam*. The color scheme encodes the fraction of spam emails among all the emails 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 emails. The right image in Figure 2 shows the so-called component plane for the 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.

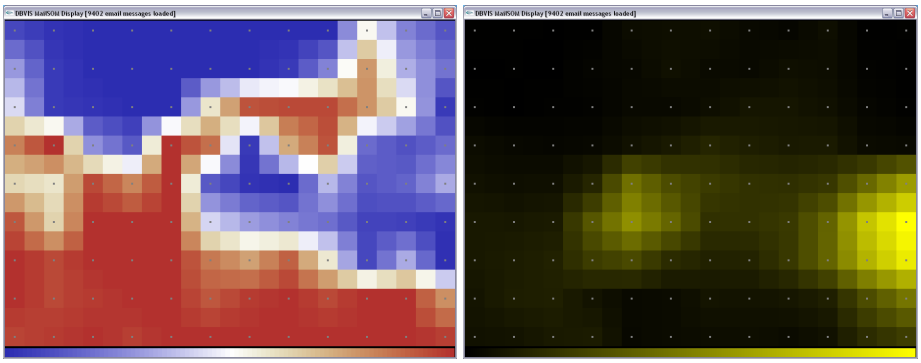

Figure 2: The left image shows a *spam-histogram* of our experimental email archive, where shades of red indicate SOM regions containing spam emails. The right image shows the component plane for term #214 ("work"), with shades of yellow indicating high term weights.

6 Results and Conclusions

In this paper, we have presented the *MailExplorer* System which aims at enabling the user to gain insight into information contained in email collections. The combination of several visual analysis techniques allows complex use cases. Using *temporal exploration*, for instance, the user could identify an interval of high email traffic, and search in the spam folder within this interval for an email that was sent from Konstanz, Germany (*spatial exploration*) and that was misclassified as being spam. Having found that email, the user might then apply the SOM (*content-based exploration*) to check whether similar misclassified messages within the spam folder exist. Future work includes the design of modules

supporting additional email attributes, and more evaluation work on several different email datasets.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG) under grant GK-1042 *Explorative Analysis and Visualization of Large Information Spaces* at the 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 the 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 Visualization 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.