MailSOM - Visual Exploration of Electronic Mail Archives Using Self-Organizing Maps

Daniel A. Keim
University of Konstanz
Computer Science Department
78457 Konstanz, Germany
keim@inf.uni-konstanz.de

Florian Mansmann
University of Konstanz
Computer Science Department
78457 Konstanz, Germany
mansmann@inf.uni-konstanz.de

Tobias Schreck
University of Konstanz
Computer Science Department
78457 Konstanz, Germany
schreck@inf.uni-konstanz.de

Abstract

Systems for handling large electronic mail archives can leverage Information Visualization techniques to facilitate explorative data analysis. In this paper, we propose to use Self-Organizing Maps as an appropriate tool to manage large volumes of email in personal email archives.

1 Introduction

Electronic mail has become one of the most important means of communication, and emailing volumes are rising steadily. Much work has been done improving the *efficiency* of email management, while the *effectiveness* of email management from a user perspective has not received a comparable amount of research attention. Information Visualization techniques can be adapted in order to devise effective email handling systems. To facilitate the interactive management of large volumes of email, we previously investigated techniques for visualizing temporal and geo-related attributes of emal archives [KMP*05]. In this paper, we extend our work with a visualization module based on Self-Organizing Maps (SOMs) [Koh01] generated from a term occurrence email descriptor. We apply this module on an email archive and find it suitable to enhance the functionality of a email management system by offering powerful visual analysis facilities.

2 Feature Extraction

To obtain feature vectors from email data, we employ a well-known scheme from Information Retrieval. First, we determine a number n of most frequent terms from the subject fields of all emails in the archive, after having filtered the subject fields using a stop-word list to avoid the inclusion of non-discriminating terms in the description. Then, we apply the $tf \times idf$ document indexing model [BYRN99], considering each email as a document represented by its subject field. The model assigns to each document and each of the n terms a weight indicating the relevance of the given term in the given document with respect to the document collection. By concatenating the term weights for a given document we obtain a feature vector (descriptor) for that document. The set of all feature vectors of the collection is then input to the SOM generation. We note that more sophisticated email descriptors can be thought of. Specifically, in addition to body text, email data usually contains a wealth of meta data and attributes which are candidates for inclusion in the description. In this paper we chose to start with a basic feature extractor, leaving the design of more complex descriptors for future work.

3 Self-Organizing Maps

The Self-Organizing Map [Koh01] is a neural network algorithm that is capable of projecting a distribution of high-dimensional input data onto a regular grid of map nodes in low-dimensional (usually, 2-dimensional) output space. This projection is capable (a) to cluster the data, and (b) to approximately preserve input data topology. The algorithm is therefore especially useful for data visualization and exploration purposes. Attached to each node on the output SOM grid is a reference (codebook) vector. The SOM algorithm learns the reference vectors by iteratively adjusting them to the input data by means of a competitive learning process. SOMs have previously been applied in various data analysis tasks. An example of the application on a large collection of text documents is the well-known *WebSom* project. Several visualization techniques supporting different SOM-based data analysis tasks exist [Ves99]. E.g., the *U-matrix* visualizes the distribution of inter-node dissimilarity, supporting cluster analysis. *Component planes* are useful for visualizing the distribution of individual compo-

nents in the reference vectors, supporting correlation analysis. If the input data points are mapped to their respectively best matching map nodes, histograms of map population, e.g., the distribution of object classes on the map, are possible.

4 Use-cases and results

Conceptually, we identify several interesting use-cases for SOM-based visualization support in an email client.

- *Classification.* Using either automatic or manual methods, the SOM can be partitioned into regions representing different types of email, e.g., spam and non-spam email, business and private mail, and so on. For incoming email, the best matching region can then be identified, and the mail be classified as belonging to the label of that region.

- *Retrieval.* The user can search for emails by mapping a query to the SOM node that best matches the query, followed by exploring the emails mapped to the neighborhood of that node, using e.g., U-matrix or histogram-based visualizations to guide the search.

- *Organization.* The user can employ the SOM generated from her email archive to learn about the overall structure of the emails contained in the archive. The user might then create a directory hierarchy for organizing emails reflecting SOM structure information.

Due to space constraints, here we only give two proof of concept results from our experiments. We generated a SOM from an archive of 9.400 emails, using the 500 most frequent subject field terms in the $tf \times idf$ descriptor. We labeled all emails as belonging to either the spam or the non-spam class, as judged by a spam filter in combination with manual classification. The left image in Figure 1 shows a *spam-histogram* on the generated SOM. For each map node, the coloring indicates the fraction of spam emails among all emails mapped to the respective node. Shades of red indicate high degrees of spam, while shades of blue indicate low degrees of spam (the latter are the "good" email regions on the SOM). Clearly, the SOM learned from our basic descriptor discriminates spam from non-spam email. The right image in Figure 1 illustrates the component plane for the $tf \times idf$ term "work", with shades of yellow indicating high weight magnitude. Combining both images, we learn that this specific term occurs in emails both of type spam and non-spam. The rightmost "work" cluster belongs to the "good" region and compounds university-related emails from one PhD student in our working group.

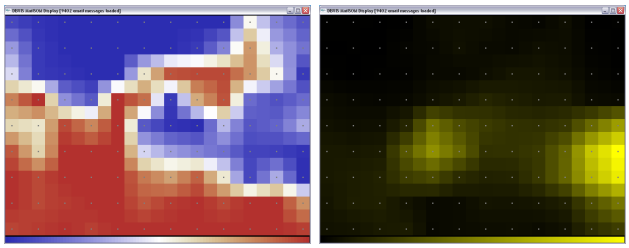

Figure 1: The left image shows a *spam-histogram* of our experimental email archive, where shades of red indicate SOM regions compounding spam emails. The right image shows the component plane for term #214 ("work"), with shades of yellow indicating high term weights.

5 Future Work

Future work involves identifying more use cases for SOM-based email data analysis. Also, more advanced descriptors should be designed. Furthermore, the tailoring of automatic labeling and text summarization algorithms for sets of emails is a promissing idea to improve the SOM display.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG) under grant GK-1042, Explorative Analysis and Visualization of Large Information Spaces, University of Konstanz. We thank Professor Kohonen and his group at the Helsinki University of Technology for providing their SOMPAK software.

References

[BYRN99] Baeza-Yates R., Ribeiro-Neto B.: *Modern Information Retrieval*. Addison-Wesley, 1999.

[KMP*05] Keim D. A., Mansmann F., Panse C., Schneidewind J., Sips M.: Mail explorer - spatial and temporal exploration of electronic mail. In *Eurographics/IEEE VGTC Symposium on Visualization* (2005), pp. 247–254.

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

[Ves99] Vesanto J.: SOM-based data visualization methods. *Intelligent Data Analysis 3*, 2 (1999), 111–126.