First publ. in: Proceedings of the International Conference on Information and Knowledge Engineering. IKE'04, June 21-24, 2004, Las Vegas, Nevada, USA. Las Vegas: CSREA Pr., 2004, pp. 275-281

Analyzing Large Collections of Email

Daniel A. Keim, Christian Panse, Jörn Schneidewind and Mike Sips
{keim,panse,schneide,sips}@inf.uni-konstanz.de
Computer Science Institute Box D78
Universität Konstanz
D-78457 Konstanz, Germany

***Abstract*— One of the first applications of the Internet was the electronic mailing (e-mail). Along with the evolution of the Internet, e-mail has evolved into a powerful and popular technology. Messages, electronically documents, pictures and even movies can be send between users of computer systems at different places all over the world within seconds. Electronic mail is a fast, a cheap and a comfortable communication method. But with the exponential increase of the Internet users and the corresponding increasing e-mail traffic, new problems arise. Examples are Spam mails or computer viruses hidden in email attachments. Therefore the analysis of email and email traffic becomes more and more important. In this paper we discuss visualization methods for analyzing large amounts of e-mails to detect interesting patterns or to identify Spam e-mails.**

Keywords: Pattern Visualization, Visual Data Mining, Data analysis

I. Introduction

Nowadays, we have to deal with heavily growing networks since network technology is used in almost every personal computer, cellular phone etc. Data communication networks such as the Internet connect millions of computers, cellular phones are used in almost every household, and personal communication networks are in commonplace. With the rapid growth of email traffic, the interest has grown in the possibility of analyzing these data, of extracting pattern and insight that might be interesting to the email receiver. Researcher at the University of Berkeley estimated, that worldwide about 31 billion emails are send daily and the annual flow of email worldwide in 2002 was about 667,585 terabytes. They expected that this amount of data will double by 2006 [1]. Finding interesting patterns in large collections of email at full scale is an important task for data analysts and business managers to recognize and respond to changing conditions quickly; within minutes when possible.

The primary data source in large collections of emails is the email header and we can extract high dimensional feature vectors from them. The attributes of an feature vector can be Source-IP, Destination-IP, Time zone of the sender, Send time or content summaries. There are many ways to analyze these data, including statistical models, pattern recognition, or machine learning techniques. Unfortunately, these approaches have not kept step with the data volumes and often fall short of providing completely satisfactory results. This situation creates new challenges in coping with scale.

Using Anti-SPAM software [2] on specialized servers can discern Spam from legitimate emails. The software can also upload potentially new forms of Spam for analysis, and develop recognition algorithms to identify and filter new types of Spam emails. The goal of our paper is to share ideas and connecting the visual exploration and data mining disciplines to allow a better, faster and more intuitive exploration of very large email collections, to give the user insight into the filter step mechanism and the public a powerful tool to optimize the used Anti-SPAM software. In this paper we present visual exploration ideas to analyze real world email data sets.

II. Visual Data Exploration

Large quantities of numerical feature vectors with thousands of records are virtually impossible to understand quickly and require the use of a interactive visual representation, since that is the fastest way for a human to finding interesting pattern communication pattern and to build deeper insight [3].

Visual data exploration often follows a three-step process: *Overview first, zoom and filter, and then details-on-demand* which has been called the Information Seeking Mantra [4]. In other words, in the exploratory data analysis (EDA) of a data set, an analyst first obtains an overview. This may reveal potentially interesting patterns or certain subsets of the data that deserve further investigation. The analyst then focuses on one or more of these, inspecting the details of the data. Visualization technology is essential for presenting overviews and selecting interesting subsets. For example, it is often helpful to simultaneously show the overview while focusing on subsets in a different type of visualization.

The visualization strategy for large collection of email isn't straightforward. Almost all email data collections consist of more than three attributes and therefore do not allow a simple visualization by 2-dimensional or 3-dimensional plots to find interesting patterns. Over the last years, a large number of novel information visualization techniques have been developed, allowing visualizations of multidimensional data sets without inherent two- or three-dimensional semantics. An example for such a visualization technique is the Parallel Coordinates Technique [5] (see Figure 1). Parallel Coordinates display each multi-dimensional data item as a set of line segments that intersect each of the parallel axes at the position corresponding to the data value for the respective dimension. The parallel coordinate techniques can be used to emphasize network data in such way, that the axis represent longitude, latitude, and other attributes. In our email analysis example, the parallel coordinates show 12 attributes of Spam data (x,

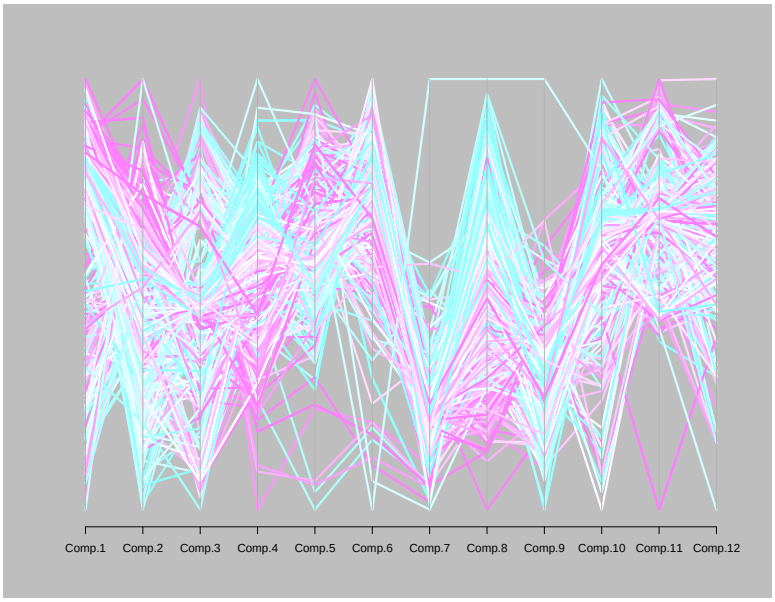

Fig. 1. Parallel Coordinates displaying a sample of 500 Spam email arrived in 2003. The bipolar colormap encodes Greenwich Mean Time(GMT) and the axis of the parallel coordinate plot show 12 attributes of Spam data (x, y, time zone, hour, attr1, attr2, ...)

y, time zone, Hour, Attr1, Attr2,...). This example shows, that there are patterns in the data, since the data values are not randomly distributed. The user can detect clusters for attributes 4-9 and correlation between attributes 6 and 7, since most high values for attribute 6 correspond to low values for attribute 7. Our goal is to detect these patterns and present them in a visual form which is useful for the user and gives him the possibility to get insight and knowledge about the data.

Overviews of other approaches for visualizing multidimensional data can be found in [6] and in a number of recent books [7] [8] [9]. A classification of these techniques can be found in [10].

III. Research Challenges

This section briefly describes the research aspects related to our work. Since our intention is not to replace Spam filters by visual data analysing techniques but bring them together, the goal is to give the user more insight into his email data. As an effect this could also lead to the improvement of existing Spam filters. Interesting points are the visual detection and the source of malicious or Spam email. Common questions are:

- When and from where do emails arrive?
- Are there patterns classifying Spam and No-Spam emails?
- Where is the sender of malicious or Spam emails located?
- Are there relationships among the different email attributes?
- Are there time patterns in the email data?

To answer all this questions there is still research necessary. We present methods which could be a first step in investigating the questions mentioned above using visualization methods.

IV. Email Data

A single email header contains different sources of interesting information. First, the email header contains information about the network. A network consists of nodes, links, and statistics about the current condition of the network, which may be the raw data or data summaries and may vary over time, are associated with the nodes and the links. An important task for finding interesting pattern in large collection of emails is to analyze the path of the emails. The path information can be derived from the email headers.

An extended header will show the path of the message from the sender to the receiver. Figure IV is an example of an extended header which uses bold printed letters for the source of attributes.

Second, the email header contains important information about the content of the email. Most of the people write some keyword in the subject field to give the receiver an idea about the content, priority, and importance of email. Extracting an individual distribution of the keyword in the subject field can give important information about the class of emails, which this email belongs to. Important classes of emails are Spam, No-Spam, private, business and many more.

Finally, the email header contains some geo-spatial information about the sender host. For example, it is very easy to extract the time zones from the header (see IV).

All these facts result in large collections of email data. Exploring such large collections of email data is an important but difficult task. In the next section, we will present some first steps in analyzing large collections of email.

```
Message-ID: <Pine.GSO.4.33.0109101126310.26314-100000@ucsub.colorado.edu>
Date: Tue, 11 Sep 2001 10:21:52 -0600 (MDT)
From: jim mascarelli <James.Mascarelli@Colorado.EDU>
To: Christian Panse <cpanse@dbvis.uni-konstanz.de>
Subject: Re: your mail
```

Fig. 2. An email header – The bold printed letters represent the spatial information of the sender (time and longitude)

V. Heterogeneous Attributes

Like mentioned in the last sections, the primary source of interesting data about the email communication is the email header. It is very easy to extract some basic information from the header including geo-spatial, spatio-temporal and non-spatial components.

Definition 1 (Large Collection of emails)
Formally, a large collection of emails can be seen as a massive database DB with spatio-temporal, non-spatial and possibly geo-spatial information: $DB = \{\{\lambda_1, \varphi_1, t_1, a_{1,1}, ..., a_{1,m}\}, ..., \{\lambda_n, \varphi_n, t_n, a_{n,1}, ..., a_{n,m}\}\}$ *where* $\lambda_i \in [-\pi; \pi]$ *represents longitude,* $\varphi_i \in [-\pi/2; \pi/2]$ *latitude,* t_i *the time and* $a_i \in \mathbf{R}^m, n, m \in \mathbf{N}$ *the attributes of the i-th data item.*

All these data classes have specific characteristics and it is very important to take these characteristics into account while analyzing the email data. In this section, we briefly describe the three different data classes and the problems extracting interesting pattern using these data classes.

A. Geo-spatial attributes

In a large number of application domains data is collected and referenced by its geo-spatial location, typically a pair of x/y coordinates describing a geographical region. For example, credit card purchase transactions include both the address of the place of purchase and of the purchaser, telephone records include addresses and sometimes coordinates or at least cell phone zones, and census data and other government statistics also contain addresses and/or indexes for places. In analyzing email data, geo-spatial information could be useful to locate the geographic region of the sender (e.g. trace the way of Spam emails) or to detect geo-related patterns in the email data.

Without loss of generality, geo-spatial data can be formally defined as follows:

Definition 2 (2-dimensional spatial data)
An 2-dimensional spatial data item $x_i \in DB$ *is an ordered* $d-$*tuple of the form* $x_i = x^1{}_i, \cdots, x^d{}_i$ *with*

- $x_i \in D_1 \times \cdots \times D_d$
- *Without loss of Generality:* $(x^1{}_i, x^2{}_i)$ *is a vector*
- $x^1{}_i = \lambda v_1$, $x^2{}_i = \beta v_2$, $\lambda, \beta \in R$, $(v_1, v_2) \in R^2$: $\exists \gamma : \gamma \neq 0 : \gamma v_1 + \gamma v_2 = 0$

Large spatial data sets can be seen as a result of accumulating samples or readings of phenomena in the real world while moving along two dimensions in space. In general, spatial data sets are discrete samples of a continuous phenomenon. So, how do we get spatial information from an email header? A straightforward approach is to employ a geo-locator database in order to map the sender email IP/domain to a geographic location. The basic idea of our *GeoLocator* is to scan a log-file and convert the monitored ISP network information to geo-spatial coordinates (longitude and latitude). The current main component of our *GeoLocator* is a collection of public available IP databases. These databases have a collection of IP address with their associated geo-spatial location. Examples for other Geo-Locator databases are NIC and DNIC.
A method to get information about the network without GeoLocator databases is presented in [11]. Another way to get an idea about the geo-spatial location of an email is to use the time zone attribute contained in the email header, which can be used the determine the longitude of the email sender. In our approaches we used a geo-locator as well as the time attribute to get geo-spatial email information. We present the extracted information in the next section.

B. Spatio-temporal attributes

A next step in analyzing email data is the investigation of spatial attributes over time. This enables the user to get insight into the development of the email traffic and the distribution of the emails over a certain time period. This is especially useful to answer questions like *Is there an increase in email traffic?, Does the email geo-distribution change over time?* or *Are there spatial-temporal patterns in the email data?*. The spatio-temporal information can also be used to monitor the number of received Spam emails over time, to detect if there are special geo-regions from where most of the Spams arrive or to detect peaks of received Spams. This information can be used to improve or reconfigure the Spam filter.

C. Non-spatio-temporal attributes

The third group of attributes are the non-spatio-temporal attributes. This contains the rest of the information contained in an email like sender name, subject, email contents, email attachments or additional network data contained in the header. Spam filters use this information to detect Spam emails and malicious email attachments. There exist a lot of effective Spam filters on the market and the improvement of Spam filters is also a important research area. An overview of Spam filter techniques can be found in [12] [2]. In our approach we use the output of the Spam filter as well as the non-spatial-temporal email information for categorization of the emails. Since corporate and university networks have become increasingly clogged by email pitches for pornography, money-making schemes, and products, we are interested in

the patterns of Spam emails In our department, about one fourth of our email traffic are Spam's. In 2002, we had one Spam for every 20 legitimate email messages; today the ratio is closer to one in four. Another point of interest are the clusters contained in the non-spatio-temporal emails, i.e. is there a change in the email behaviour of the receiver. The next section will investigate all this questions and show in detail how we overcome them.

VI. Exploring Large Collections of Email

In the last section we described, which kind of attributes we can extract from each email and now we show how we used these information to analyze large amounts of email data. As example real world data set we used the email data from our department email server, collected over 7 years, which leads to an total amount of email data of approximately 10GB. A Spam filter has classified these emails in SPAM and NO SPAM. Some of the emails where also classified manually (see Figure 3). The following sections show how we used the email database and the corresponding knowledge for visual data analysing.

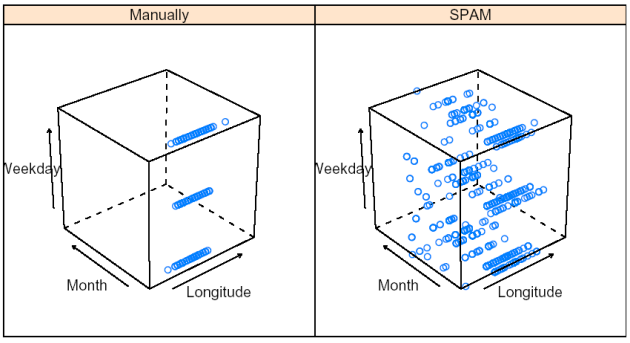

Fig. 3. Email classification in Spam and No-Spam: manually (left) and manually+Spam filter (right)

A. Exploring Email Paths

An interesting approach is to explore the path of Spam emails, classified by a Spam filter, to see interesting patterns and behaviour. The path information can be derived from the email headers. We used a Geo-Locator database to map these information to geographic locations. The path problem can be formally defined as follows.

Definition 3 (Email Path)
Let $DB = \{\{\lambda_1, \varphi_1, t_1, a_{1,1}, ..., a_{1,m}\}, ..., \{\lambda_n, \varphi_n, t_n, a_{n,1}, ..., a_{n,m}\}\}$ *a large collection of email. The goal is to explore the set of all paths in the real world* $P = \{x_0, x_1, \cdots, x_n\}$ *with* $\forall i : x_i, x_{i+1} \in GeoDB \subset DB$ *and* $(x_i, x_{i+1}) \in DB$.

Figure 4 shows the regular and Spam emails path of one of the authors. The email paths displayed in the plot have been stored since 2000. Each spatial location corresponds to a computer system from which the emails were sent. Each line segment represents the path of an email message between two computer systems. The picture on the right displays only Spam emails.

It is easy to see, that most of the emails received by our department located in Konstanz(Germany) come from Europe and North America, while almost all of the emails received from South America, Africa or Asia are Spam. This information could be used to adapt the Spam filter. It is also interesting to see that a large part (25 percent) of the emails are Spam.

B. Exploring the Spread of Email Communication

Another approach is to use the time zone attribute to visualize the longitude, as geo-spatio component, from which the emails arrive. Since the email header contains the time zone, we can easily distort a familiar land-covering map in such a way that the area of the map region is proportional to the number of emails being sent. A cartogram algorithm can be run using the spatial information and the number of emails. One way of doing this has been introduced by an American geographer [13]. A more efficient algorithm, which is based on histograms, has been introduced by [14]. Figure 5 shows clearly the different office times of our partner organization around the world. A regular pattern reflects clearly the different time zones.

C. Exploring email behaviour over time

A next step in analyzing emails is the exploration of the email behaviour over time. Therefore we use spatio-temporal (time, longitude) as well as non-spatio-temporal attributes. Figure 6 shows the emails received over 12 months by one of the authors, starting from January 2001 (lower left corner) till December 2001 (upper right corner). We visualized the extracted information using Data cubes. The dimensions of the Data cube are day of the week, hour and time zone. The color represents the different clusters the emails belong to. We computed the clusters by using the sender attribute [15] . Green colors indicate that the sender is a colleague (group member) of the author, pink colors indicate that the sender is a friend of the author. Red and Blue colors indicate that this emails belong to special project groups, blue colors to Project1 and red colors to Project2.. There are some interesting information the viewer can get from Figure 6. Emails from colleagues (pink color) usually only arrive during the week and of cource from the same time zone. One can identify some regular patterns. There is one month (May 2001) where no emails from colleagues arrived. The reason is that the author has been on vacation during this month. It is also easy to identify on which months the author worked on the 2 projects, March-April Project 1 (blue color) and January, March, April, October Project 2 (red color). In July the author received only a small number of emails from friends. The reason could be that most of them have been on vacation. This example demonstrates that visualization techniques can be very useful to understand and explore the data.

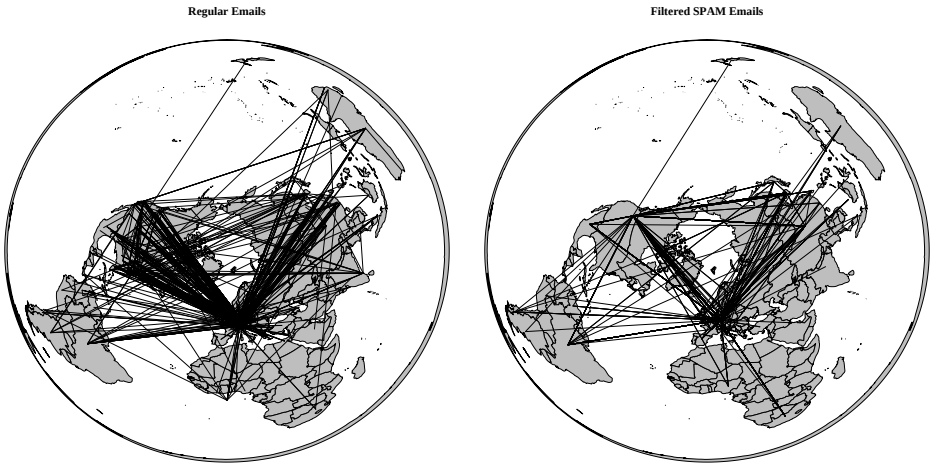

Fig. 4. The figures display the worldwide email routes of one of our *IMAP* users. The *IMAP* server is located in Konstanz, Germany (*37 41.0N / 09 08.3E*). In our department, Spam hits one fourth of our email traffic.

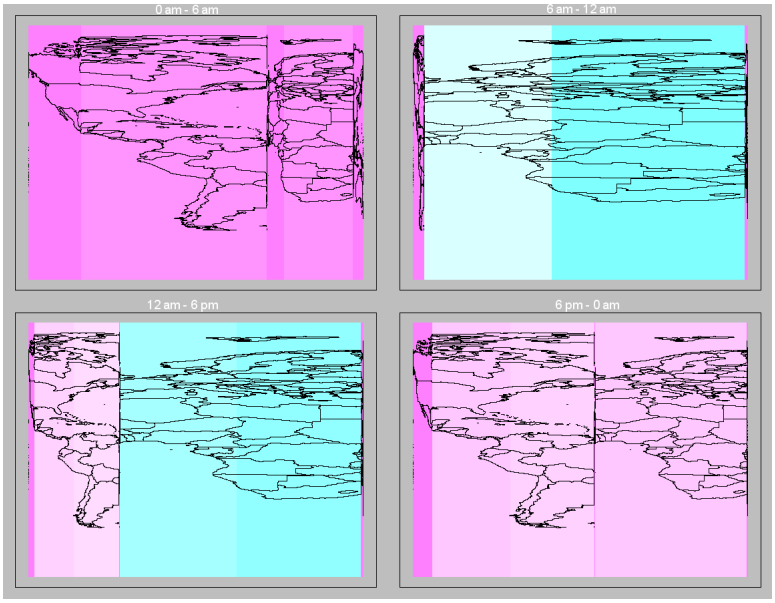

Fig. 5. HistoScale cartogram in longitude direction – The picture shows email volume patterns at four different points in time (0am–6am, 6am–12am, 12am–6pm, 6pm–0am GMT) during one day. The area of a region represents the relative number of emails and the color indicates the absolute number of emails arrived.

D. Analysis of email and Spam amount

To investigate the development of email and Spam traffic we employed pixel based level-plots. A level plot is a visualization technique for the representation of 3 dimensional tuples $t = (t_1, t_2, t_3)$ by plotting constant z slices, called contours, on a 2 dimensional format. For our email traffic analysis example, the level plots illustrate the time zone(horizontal axis), the 24 hours of a day (vertical axis) and the number of emails arrived(color) In Figure 7 the email traffic of one of the authors over 7 years (1997-2003) is visualized. The figures show that there is a strong increase in the number of received emails per year. There is also a higher spread in time zones where most of the emails came from. One can also see that most of the emails arrive between hours 10 and 14. As a phenomena the email behaviour changed in 2002, since most received emails came from time zone -5 this year. The reason for this effect is, that the author had a research year in the USA.

Figure 8 shows 2D Level Plots of email data over the years 2001-2003 classified as SPAM and NO-SPAM for one of the authors. The user can easily see that there is a strong increase in Spam emails, whereby most of them come from America. Again, in this figure becomes clear that the author had a research year in the USA since the email traffic moved to time zone -5 in 2002. In 2003 most emails arrived from Europe and America, since there are most of our research contacts located.

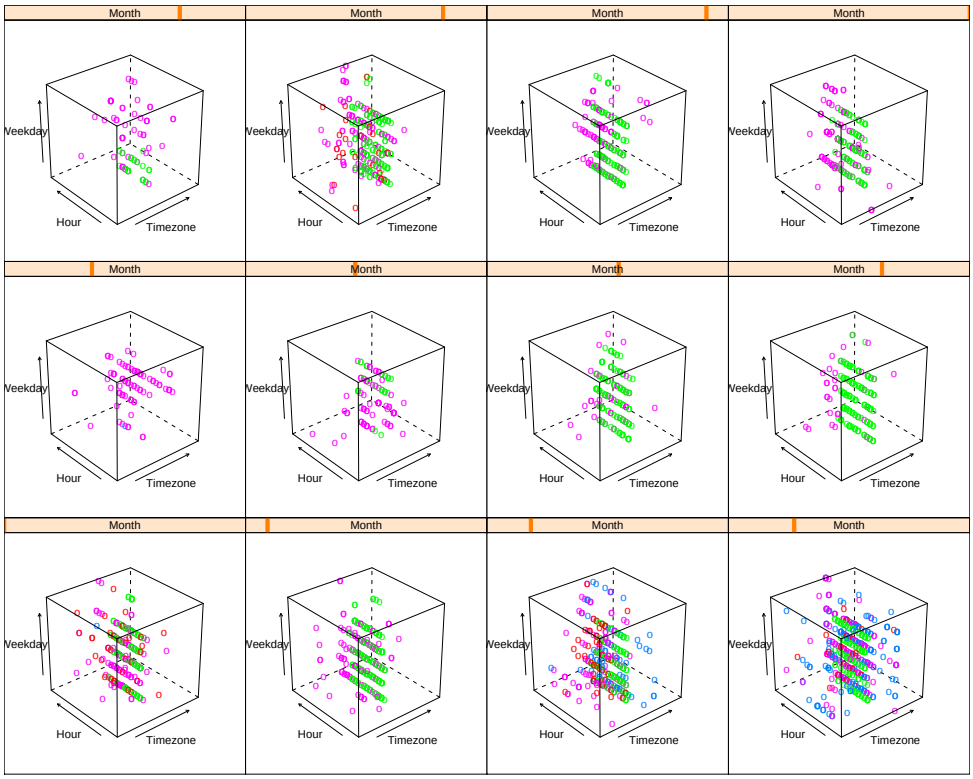

Fig. 6. Data cubes representing email traffic of one of the authors over one year from 1/2001 (lower left corner) till 12/2001 (upper right corner). Colors represent the different email categories (colleagues-green, friends-pink, project1-blue, project2-red). The data cube dimensions are day of the week, time zone and hour.

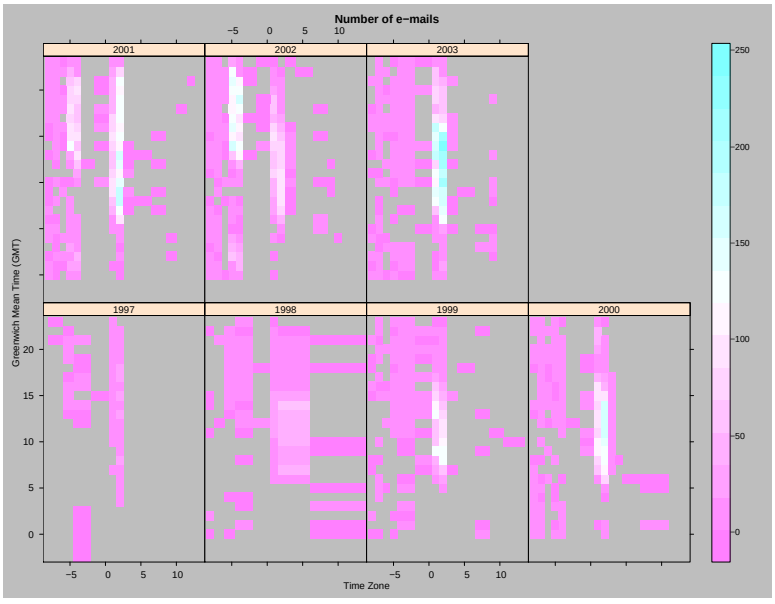

Fig. 7. 2D Level Plot of the email over the years 1997-2003 .

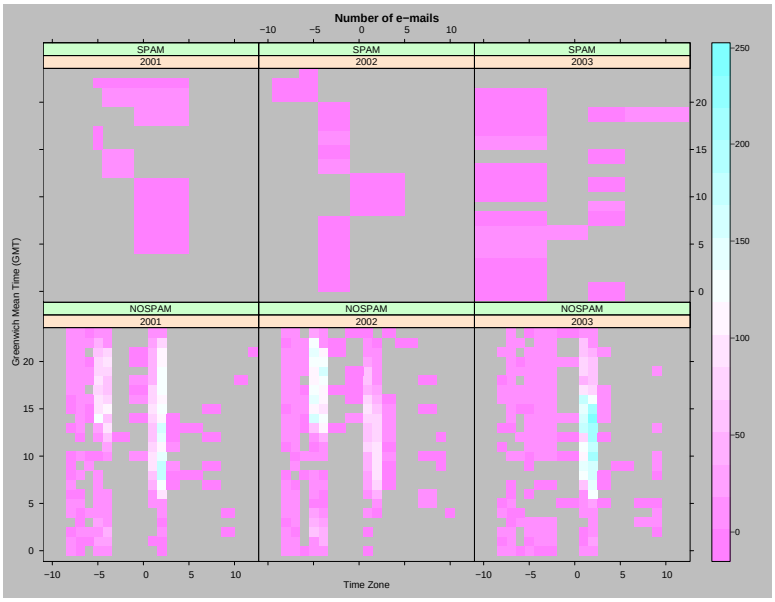

Fig. 8. 2D Level Plot of the email data over the years 2001-2003 classified as SPAM (upper row) and NO-SPAM (lower row).

VII. Conclusion and Future Work

Exploring and analyzing large collections of email data is an important research area. One of the challenges today is to find out how to deploy efficient analysis strategies for large collections of emails. In this article, we described some first steps in analyzing massive email data to find interesting patterns. We achieved this goal by employing Visual data exploration techniques like Data cubes, 2D plots and Histograms. Using these visualization techniques, we were able to detect interesting patterns and anomalies in the data. We showed also how the email traffic changed over time. The results that we presented show, that this techniques have a high potential in analyzing large collections of email. In future work, we expect to investigate:

- related approaches for visualizing large geographical data sets
- user studies to identify strengths and weaknesses of these approaches
- developing new mapping methods in our *GeoLocator* to improve the quality of the spatial location mapping
- visualization interface for Spam-filter tuning

This work will include the improvement of our approaches and the development of a framework to visually analyse and explore large collections of emails.

VIII. Acknowledgement

This work was partially funded by the Information Society Technologies programme of the European Commission, Future and Emerging Technologies under the IST-2001-33058 PANDA project (2001-2004). The authors also thank Roland Heilmann for his great support.

References

[1] P. Lyman and H. Varian, "How much information?" *Retrieved from: http://www.sims.berkeley.edu/how-much-info-2003 on 5/2004*, 2003.

[2] A. Schwartz and S. Garfinkel, *Stopping Spam, stamping out unwanted email news postings*. O'Reilly, 1998.

[3] D. A. Keim, "Information visualization and visual data mining," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 8, no. 1, pp. 1–8, January–March 2002.

[4] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proc. IEEE Visual Languages*, 1996, pp. 336–343.

[5] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in *Proc. Visualization 90, San Francisco, CA*, 1990, pp. 361–370.

[6] D. A. Keim, C. Panse, J. Schneidewind, M. Sips, M. C. Hao, and U. Dayal, "Pushing the limit in visual data exploration: Techniques and applications," in *Advances in Artificial Intelligence, 26th Annual German Conference on AI, KI 2003, Hamburg, Germany, September 15-18, Lecture Notes in Artificial Intelligence, Vol. 2821*, September 2003.

[7] S. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization*. Morgan Kaufmann, 1999.

[8] B. Spence, *Information Visualization*. Pearson Education Higher Education publishers, UK, 2000.

[9] C. Ware, *Information Visualization: Perception for Design*. Morgen Kaufman, 2000.

[10] D. A. Keim and M. Ward, *Intelligent Data Analysis, an Introduction by D. Hand and M. Berthold*, 2nd ed. Springer Verlag, 2002, ch. Visual Data Mining Techniques.

[11] N. Spring, R.Mahajan, and D.Wetherall, "Measuring isp topologies with rocketfuel," in *Proc. SIGCOMM*, 2002.

[12] L. Cranor and B. LaMaccia, "Spam!" in *Communications of the ACM, 41(8)*, October 1998, pp. 74–83.

[13] W. Tobler, "Cartograms and cartosplines," *Proceedings of the 1976 Workshop on Automated Cartography and Epidemiology*, pp. 53–58, 1976.

[14] D. A. Keim, C. Panse, J. Schneidewind, and M. Sips, "Geo-spatial data viewer: From familiar land-covering to arbitrary distorted geo-spatial quadtree maps," in *The 12-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG 2004, Plzen, Czech Republic*, February 2004.

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.