Presented at IEEE Visualization, Baltimore, MD, October 2016

Analyzing Hillary Clinton's Emails

Vasundhara Dehiya* and Klaus Mueller†
Computer Science Department, Stony Brook University, NY, USA and SUNY Korea, Songdo, Korea

Abstract

Due to the controversy regarding use of personal email on non-government servers by Hillary Clinton during her time as Secretary of State, her email data was made public. Yet, reading through them is impractical. In this poster, we provide a visual analysis of the content of these emails. Based on the 7945 emails available, we identify the relation between the textual content of the emails with world policies. We unravel how the content of these emails are reflective of US emotion and behaviour with other countries around the world along with their relative importance. We also unravel correlations in the data to predict some features of content in redacted emails based on available data.

Keywords: information visualization, visual analytics, email, text mining

1 Introduction

Hillary Clinton has been involved in controversy for using her personal email accounts on non-government servers during her time as the United States Secretary of State. It is believed that this action violated protocols and federal laws that required appropriate record keeping. Based on this claim, lawsuits were filed, and the State Department released nearly 7,000 pages of Clinton's heavily redacted emails on 31 August 2015. Kaggle [1] has cleaned and normalized the released documents and made them available for public analysis, which we used in our analysis.

2 Related Works

Email mining is similar to text mining since both of them consist of textual data. Thus, natural language processing is an integral part of analysis and will be used in this project as well. However, there as certain features with are unique to email mining. Emails includes additional information in the headers of email that can be exploited for various email mining tasks. Text in email is significantly shorter and, therefore, some Text Mining techniques might be inefficient in email data. Email is often cursorily written and may not have well-formed sentences [4]. Spelling and grammar mistakes as well as non-standard user acronyms also appear frequently. Email is personal and thus generic techniques are difficult to be effective for individuals. Email is a data stream targeted to a particular user and concepts or distributions of target classes of the messages may change over time, with respect to the messages received by that user. Email will also have noise. HTML tags and attachments must be removed in order to apply a text mining technique. In some other cases, noise is intensively inserted. In spam filtering for example, noisy words and phrases are inserted, in order to mislead machine learning algorithms [2]. All these make email mining an interesting research problem where visual analytics can be used for inferring conclusions about the data in the emails along with the latent variables (like sender information, time of sending) available.

*e-mail:vasundhara.dehiya@stonybrook.edu
†e-mail:mueller@cs.sunysb.edu

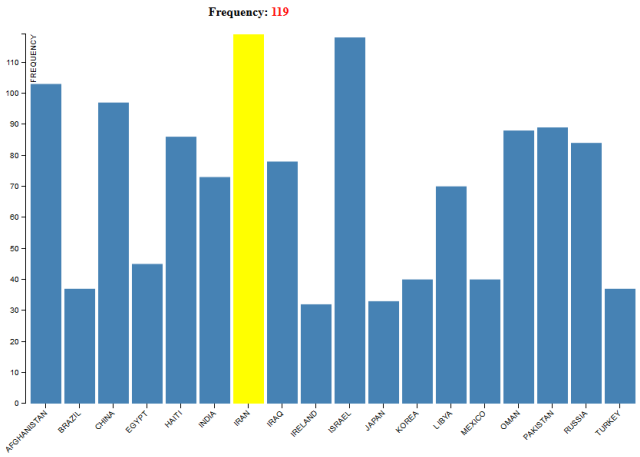


Figure 1: Most frequently discussed countries

A vastly studied email dataset is the Enron email dataset. It was used for social network and link identification, to determine hidden structure information from noisy incomplete email data. Some approaches for this include graph entropy based approach [6], graph and spectral analysis [3]. *Themail* is a visualization of relationships based on email interaction history [7], thus this is a time series approach. Email visualization was also performed by *EzMail* [5] which created a tool to be used by email client as a multi-view interface and allowed for visualizing messages, message properties and history.

3 Approach

The data available included email content, sender, receiver and time stamp, along with more metadeta. In this project, we were mainly interested in international relations and policies that can be inferred form the content. First we extract and identify emails referring to various country names. We visualize which countries are discussed most frequently. Amongst these top candidates, we visualize which countries are discussed in correlation with one another. This approach provides a single yet highly efficient visualization wherein a user can easily identify relation between countries and how policies are effected by this. Following this, we performed time series analysis for sentiment associated with each of the countries. We attempt to identify how this sentiment affected world events in that year based on the emotion expressed in the email. Lastly, even though we have a lot of redacted emails, we analyse if we can make predictions about that email content based on the length of the email.

4 Observations

Considering the text of the emails, we search the content for names of country and create a database of the names of the countries mentioned in unique email.Figure 1 depicts the 31 most frequently referred countries with Iran(119) and Iraq(118) being at top, followed by Afghanistan(103), China(97) and Pakistan(89), which are evident of importance of these countries in US policies and adminis-

tration making during that time. A correlation matrix (Figure 2) among the countries help visualize which countries are discussed together. High positive correlation is seen between Korea-Japan and Afghanistan-Iraq. Iran and Iraq have positive correlation with many other countries, denoting that these countries are talked about very often and in conjunction with multiple other countries. Negative correlation of Hondurus, Qatar, and Cuba with almost all other countries means that these countries are generally discussed in isolation with other countries and emails are exclusively centred about that country.

Sentiment analysis is performed to identify US sentiment associated with each country and how it changes with time. The sentiment score is calculated as the difference of positive and negative terms in the email content referring to that country. We plot the sentiment year-wise and overall. Figure 3a shows the cumulative sentiment score for each country with different bar colors representing the year in which the email was sent. Interestingly, for year 2012, we have the few emails publically available, yet all of them show positive sentiment for all top countries. Our goal was to identify how these sentiments effected world events. In 2009, high positive sentiment is associated with Afghanistan and India. This may be represented in signing of a Defence Pact among India and US in 2009. In 2010, an overall negative sentiment was associated, especially with Germany, Israel and Oman. Negative sentiment with Israel correlates with Israeli Policy changes which was strongly condemned by the US government. For years 2011 and 2012, a very few number of emails are available to us. This leads us to our next goal of predicting the content of the email. Based on emails with known content, we perform correlation analysis among the length of an email with its sentiment. Figure 3b shows that there is a strong positive correlation for 2011 and 2012 and thus, given a retracted email, we can predict its emotion based on the length of the document. Also, Figure 2 can be used to identify which countries are being discussed in case we have partial content or names of one or more country in the text available. This information is useful in order to predict the type of content in a redacted document based on only the length of document and the names of countries present in the text.

5 Conclusion

We successfully applied visual analytics concepts for information analysis to this set of email data of Hillary Clinton during her term as Secretary of State. We identified which countries were discussed in maximum number of emails and identified correlations between countries mentioned together. We also identified the sentiment associated with the emails and obtained correlation among sentiments

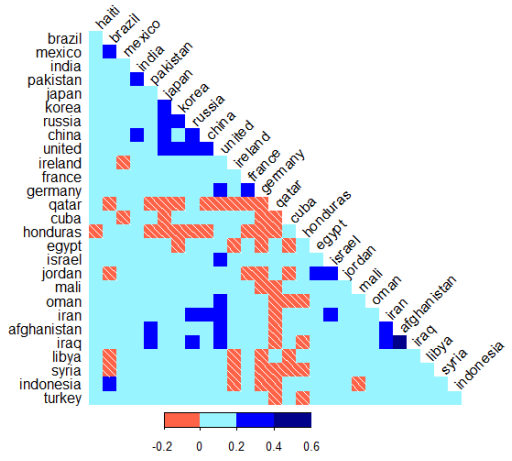
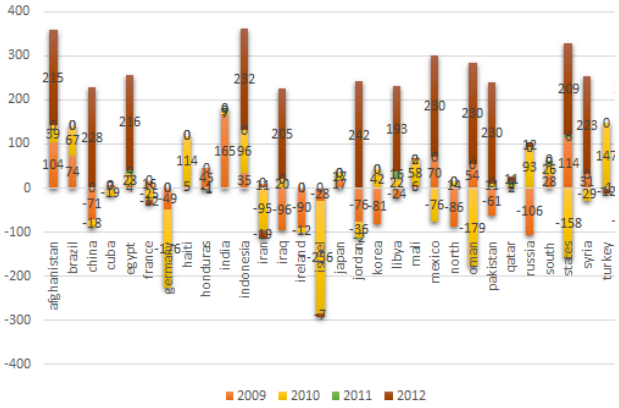
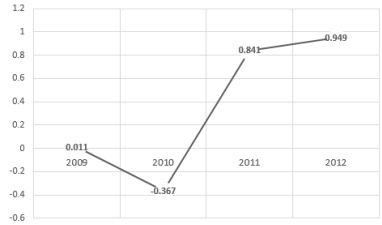


Figure 2: Countries discussed together



(a) Net Sentiment for each country



(b) Correlation of sentiment with length of email

Figure 3: Sentiment Analysis

in different years. We answered the question: Is the length of email predictive of its sentiment? We were able to identify sentiment variation associated with different countries in different years and world events associated with email sentiment and identified key factors that may be useful to predict context of partially available emails based on correlation with length, sentiment and the countries that are discussed together.

Acknowledgements

This research was partially supported by NSF grant IIS 1527112 and the Korean Ministry of Science, ICT and Future Planning, under the IT Consilience Creative Program (ITCCP) supervised by NIPA.

References

[1] Hillary clinton's emails. https://www.kaggle.com/kaggle/hillary-clinton-emails. Accessed: 2015-09-30.

[2] P. S. Bogawar and K. K. Bhoyar. Email mining: a review. *IJCSI International Journal of Computer Science Issues*, 9(1), 2012.

[3] A. Chapanond, M. S. Krishnamoorthy, and B. Yener. Graph theoretic and spectral analysis of enron email data. *Computational & Mathematical Organization Theory*, 11(3):265–281, 2005.

[4] I. Katakis, G. Tsoumakas, and I. Vlahavas. E-mail mining: Emerging techniques for e-mail management. *Web Data Management Practices: Emerging Techniques and Technologies: Emerging Techniques and Technologies*, page 219, 2006.

[5] M. Samiei, J. Dill, and A. Kirkpatrick. Ezmail: using information visualization techniques to help manage email. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, pages 477–482. IEEE, 2004.

[6] J. Shetty and J. Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, pages 74–81. ACM, 2005.

[7] F. B. Viégas, S. Golder, and J. Donath. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988. ACM, 2006.