Determining and Visualising E-mail Subsets to Support E-discovery

Mithileysh Sathiyanarayanan*
giCentre, City University London, UK

Cagatay Turkay†
giCentre, City University London, UK

ABSTRACT

Electronic discovery (E-discovery) is a legal process for investigating various events in the corporate world, for the purpose of producing/obtaining evidence, one such example is an email communication (eg. Enron case). Investigating emails collected over a period of time, manually, is a strenuous process and the tools currently available on the market are based on simple keyword search and legal firms charge companies based on the volume of information produced by the search, which is then manually reviewed intensely. This results in significant costs for the company or in a number of cases settlement because they can't afford the costs of E-discovery. So, there is a great need to determine, visualise and understand whether email subsets are normal or abnormal, pertinent or privileged, relevant (interesting) or immaterial in a quick time. In order to determine relevant subsets for a legal case and to gain invaluable insight in a quick time from the email communications, we propose a multi-modal and multi-level approach which will generate automated visual representations using a manual keyword search facility that will extract the most relevant information from the email data and aids in comparing two subsets of information. In this paper, we discuss the literature review carried out, initial design process, prototypes developed and the workshops conducted. As a future work, we aim to develop a full-fledged E-discovery tool that could be implemented by the organisations to investigate email communications.

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle;

1 INTRODUCTION

Problem Statement: The avalanche of email data is expanding exponentially with different degrees of variety and complexity. With email traffic continuing to grow at 5% [1] a year in the business context more companies are now requiring time-saving and cost-effective solutions for E-discovery. The tools currently available on the market are based on simple keyword search and legal firms charge companies based on the volume of information produced by the search, which is then manually reviewed intensely [2]. The recent review report by the UK Home Office [2] states, there are no E-discovery tools that have the ability to display temporal or spatial information in an innovative way. For E-discovery compliance, experts regularly need to investigate "samples of emails" and it is important for them to select a representative (or interesting) sample. However in a data context such as email (one that is multi-modal and dynamic), the definition of "interesting" is vague and the information obtained is multi-faceted. Hence the need for visualisation empowered solutions to support the analysts with this particular task.

Research Question: How to determine a subset of a huge email data to be investigated in a quick time? How to determine and visualise whether the subset email data is normal or abnormal, pertinent (relevant) or privileged, interesting or immaterial in a quick time? Is there a tangible way to compare two subsets of email data?

Approach: To achieve the aforementioned research question and also to gather overall feedback to define to role of visualisation in E-discovery processes, we conducted workshops with a legal team and an intelligence analyst. Considering their inputs and the current E-discovery challenges, we propose a multi-modal and multi-level approach which will generate automated visual representations using a manual keyword search facility that will extract the most relevant information from the email data and aids in comparing two subsets of information. We also address the basic questions by demonstrating D3 prototypes using aggregated simulated data: 1). How a multi-modal and multi-level email communication data will be represented? 2). How are associations between the multi-modal and single-level established? 3). How can the combined approach inform about the normality, relevance and interestingness of the email communication? 4). How do our approach and designs lead us to a good subset selection?

2 RELATED WORK

Enron Case Study: Enron [3] scam is a well-known case in the Information Science and visualisation field. Enron produced fake profit reports and company's accounts which led to bankruptcy. Most of the top executives were involved in the scam, as they sold their company stock prior to the company's downfall. The Enron email is available for the public to access. In our work, we will be using the Enron data as a test case during the main design process and user-testing.

Email visualisations: There are many email visualisations developed by various researchers, but each has its own drawbacks, such as inconsistency, complexity and not very powerful tool for E-discovery investigations. Some of the well-known ones are ContactMap, ConversationMap, EmailMap, Email time, EzMail, remail, Themail, Seemail, Mailview[5]. The current E-discovery tool Jigsaw is been most preferred by the analysts though it has its own drawbacks, where the tool must be used in combination with other tools to carry out investigations. Also, two or more subsets of data cannot be compared both textually and visually. We will address this issue in our work.

E-discovery Challenges: As the email data keeps increasing exponentially, understanding the meaning contained in the data is complex, tedious, time-consuming and expensive [2]. To reduce the aforementioned and maintain high quality in the E-discovery process, an advanced and powerful visual analytic tool is in need that could visually compare two or more subsets of email data to understand what constitutes "normal", "pertinent" and "interestingness". A deep focus on analytics will help legal teams develop more insightful strategies, that is to combine keyword and context searches with visual representation used by key players and their relationships, which will elevate the analyst experience and make best decisions in the case [2].

3 CURRENT WORK

Paper Sketches: After understanding the E-discovery challenges, we came up with various paper sketches which will help analysts in investigating email communication. We went through a several iterations and had workshops with a legal team and an intelligence analyst to inform our design process.

*e-mail: Mithileysh.Sathiyanarayanan@city.ac.uk
†e-mail: Cagatay.Turkay.1@city.ac.uk

Workshops: The first workshop was with a legal team of six solicitors in Bangalore, India. They used Excel as a tool for their investigations. The key questions that the experts need addressing with our research are
Q1: How to identify "normal" dataset in a given entire dataset?
Q2: On what basis to drill-down the entire dataset?
The second workshop was with an intelligence analyst who works at the cyber investigation department, Bangalore, India. He and his team use E-discovery tools such as Jigsaw, Concordance by LexisNexis and/or IN-SPIRE to analyse unstructured data. Similarly, the key questions that need addressing with our research are
Q1: How to identify "relevant", "interesting" and "privileged" subsets in a given entire dataset?
Q2: How to generate visualisations using keyword search and time-slicer while drilling-down the entire dataset?

Design Approach: Our approach is based on the E-discovery challenges and the basic questions we came up with, and the workshop observations. In simple words, to be able to evaluate the "relevance" and "interestingness" of a sample, we decided to provide: 1). a multi-perspective (refers to the modalities in the data (multi-modality), i.e., individuals in the form of a network, temporal changes, and the text for the emails), and 2). a multi-level (characterised by several levels) investigation system. Each level will enable an expert in his/her search towards a representative sample.
a). Identifying "normal" data: the *first level* visually represent an *automated overview* of the entire data which helps in identifying which part of data (abnormal) to sample, based on the regular activities of individuals, for further investigation.
b). Identifying "relevance" in data: the same level also include a *manual key search* and *time-slicer* to filter the huge data and to fetch a relevant data using visual representations (from investigation point of view).
c). Identifying "interestingness" in data: the *second level* visually represent frequency of keywords used by *individuals*. Based on the highest frequency, key individuals can be selected, which further samples the data. The *third level* visually represent *contact relationships*. This will help in identifying interesting key individuals if they have exchanged those keywords with internal and/or external contacts. This level samples the data further (we call it as "complete sampling") for manual reading based on the investigation report.
d). Identifying "privilegedness" in data: The *fourth level* will aid in comparing two different subsets of data (for ex., email data of two different years), which are the most privileged information that must be protected. This level also aids in creating a list of emails to read (after complete sampling process), exports the visuals of all the levels and generates visual summaries.

D3 Prototypes: Based on the workshop insights, we have developed D3 prototypes in isolation. All the figures are screenshot. LEVEL1: Figure 1(a) represents an overview of an aggregated simulated data showing a sudden burst of emails during 1999 and 2002 (time-area graph). This visualisation has focus, panning and zooming which will help in identifying which particular year and month had high volume of messages sent/received. Figure 1(b) represents top 10 keywords used in the emails exchanged (bar graph). Using a manual key-search and time-slicer, more visuals can be generated. Figure 1(c) represents a matrix-bar chart which shows a combination of keywords (dots) used the most (bars). This is based on the UpSet [4]. Figure 1(d) represents a time-line graph showing which word was used the most during which part of the month. LEVEL 2: Figure 1(e) represents a stacked-bar graph showing individuals that used the combination of keywords in that particular time-frame. Figure 1(f) represents a matrix-bar chart which shows a combination of keywords used by individuals the most. LEVEL 3: Figure 1(g) represents again a matrix-bar chart to show which individual had sent a particular keyword to their internal contacts. Similarly Figure 1(h) represents a matrix-bar chart to show which internal contact(s) had sent a particular keyword to the externals contacts. LEVEL 4: Again all the three levels can be iterated but with a different sample to compare and investigate (not shown in the figure).

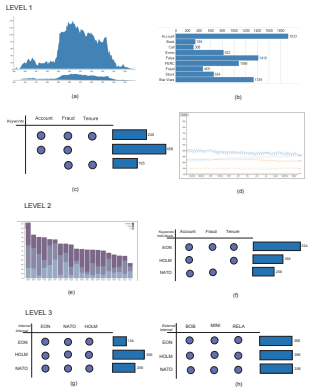

Figure 1: D3 Prototypes: multi-perspective and multi-level design

4 Future Work

We aim to develop a simple, powerful, effective, efficient and analyst-friendly visualisation tool which will be tangible and feasible to understand anomaly behaviours and what constitutes "normal" and "pertinent", then understand the dynamic changes between two subsets and the underlying communication structures in email communication which will help in E-discovery investigations. We aim to deliver innovation on several fronts: Developing novel combinations of visual and algorithmic analysis: The complexity of the data requires us to not only utilise and improve state-of-the-art intelligent algorithms in data analysis but also calls for novel techniques where humans' cognitive capabilities are fostered. The potential of such novel combinations in information discovery and decision making within E-discovery domain is not yet investigated thus an innovation we want to exploit in this project. Text analytics such as automated Named Entity Recognition or Classification of Email categories will aid in providing valuable data preprocessing/analysis. Also, we will consider text visualization in order to provide effective views for the processed data. The complete version of the tool will have user testing using AmazonTurk to evaluate the visualisation design choices for some of the tasks, such as aggregation, comparison, etc. So, our proposed methodologies will help analysts in their E-discovery tasks through interactive and visual analytics and lead to faster and effective processes.

References

[1] http://www.radicati.com/wp/wp-content/uploads/2013/04/Email-Statistics-Report-2013-2017- Executive-Summary.pdf.
[2] D. Lawton and R. Stacey and G. Dodd. Uk home office. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/3947 digital-forensic-investigations-3214.pdf, 2014.
[3] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer, 2004.
[4] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 20(12):1983–1992, 2014.
[5] S. Whittaker, Q. Jones, B. Nardi, M. Creech, L. Terveen, E. Isaacs, and J. Hainsworth. Contactmap: Organizing communication in a social desktop. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(4):445–471, 2004.