Visualising E-mail Communication to Improve E-discovery

Mithileysh Sathiyanarayanan*
giCentre, City, University of London, UK

Cagatay Turkay†
giCentre, City, University of London, UK

Jason Dykes‡
giCentre, City, University of London, UK

ABSTRACT

Electronic Discovery (E-discovery) is an investigation domain where electronic data is searched to find information and use it as an evidence in a legal case. One of the investigation areas in this domain is electronic mail (E-mail) communication. Lawyers and analysts involved in this activity are usually presented with a large E-mail dataset to manually comb through information in order to discover key information they need, expending large amounts of time, energy, effort and money in the process. We design and develop an interactive visualisation that will support our collaborators in an organisation specialising in E-discovery to unravel the multi-faceted information in the given communicated E-mails to find/discover pertinence, key information, points of interest (PoIs) and to develop evidence through which legal cases can be built.

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

1 INTRODUCTION

Currently, there are no E-discovery tools for E-mail investigation that have the ability to display and identify pertinent information in an effective way, that is to discover pertinence (relevance) between multi-facets of the data (different granularity of time, individuals, connections and content) [1]. 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 to find pertinence and key information [1]. Investigators search through E-mails, seek answers to various questions in the reports-who? what? when?-to produce it as an evidence to the judiciary. However in a data context such as E-mail, identifying "pertinence" is ambiguous and the information obtained is multi-faceted which makes investigation process tedious and complex [1]. Hence the need for visualisation empowered solutions to support the analysts with this particular task.

In this work, we aim to address the question "How can visualisation support analysts in finding pertinent and key information in a corpus of E-mail within an investigation domain?". In most of the investigation process, the exact question to investigate is not always known. In that case, it is good to consider visual analytics, as suggested by Tamara Muzner [7]. Having investigators in the analytic loop improves investigation process as the visual tool aids in identifying anomalies, changes, patterns, trends and the investigators can continuously refine the search process until the desired results are found. We design and develop an interactive visualisation that will support our collaborators in an organisation specialising in E-discovery to unravel the multi-faceted information in the given communicated E-mails to discover pertinence, key information, points of interest (PoIs) and to develop evidence through which legal cases can be built.

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

2 RELATED WORK

The Jigsaw Investigation tool developed at Georgia Tech by J. Stasko et al. [8] supports investigative process by mapping relationships (people, places, things etc.) found in datasets. Many investigative teams use this tool for identifying key information/relationships within data but sometimes the tool must be used in combination with other tools to carry out investigations. This tool does not focus or support investigating E-mail communication (archived and/or live), due to E-mail's multi-faceted nature, as they have different granularity of time, individuals, connections and content. So, finding pertinence (relevance) and key information between the multiple facets of the data is tedious and complex [1]. We aim to develop an interactive visualisation tool specifically dedicated to investigate E-mail communication to support the analysts with this particular task.

We reviewed work on E-mail visualisations and identified EmailMap, Email time, EzMail, remail, Themail, Seemail, Mailview [5]. However, we found that there is a minimal past work on visualising multi-facetedness in E-mail datasets to find pertinence. We will address this issue in our work.

3 METHOD

We used Munzner's [6] "Nested Model for Visualization Design and Validation" Methodology to frame work that was undertaken with the collaborators at Red Sift London. This model helped to structure the design process by providing the four nested levels of design: domain characterisation, data/task abstraction, visual encoding/interaction idiom and algorithm. Each levels in this nested form helps in analysing the problem and validate the solution independently. An Incremental Development combined with an Iterative Prototyping approach [4] was found to be the best solution, they work well within the nested model. Domain characterisation - we considered E-discovery [1] as an investigation domain for searching, finding relevance/information and use it as an evidence in a legal case. Our target domain users are lawyers and analysts. We specifically considered Enron [2] scandal as a case study for investigating the E-mails communicated by the employers and employees and used this publicly available dataset for the design process. Initial meetings where held with domain experts in order to understand the current workflow. Unstructured interviews were conducted to gather tasks, followed by initial requirements and they were validated using mock-ups and sketches. Task & Data Abstraction, Visual Encoding & Interaction - we used the Why?, What?, How? framework [7] to abstract the tasks, explore visualisations and develop interaction paradigms that would satisfy these tasks. After an initial evaluation of different charting libraries and Red Sift (most products are web-based), D3 is considered the most suitable technology and the current best in class platform for building interactive data visualisation which uses JavaScript. For extremely fast interaction (incremental filtering, reducing and comparing), crossfilter techniques were implemented, as they are quite helpful for exploring large multivariate and/or multi-faceted datasets in the browser. The first version of the prototypes were developed using the simulated dataset by using D3.js based on the paper sketches. We went through a several paper/prototype iterations before testing it on the real dataset, as visualisation solutions are best validated with real datasets.

4 Interactive Visualisation Prototype

Our functional prototype supports visual analysis for E-mail data exploration with the combination of matrices and bar charts that provide concurrent perspectives of multiple facets of the data. Heat matrices are to visualise relationship between two different granularities, shows the number of occurrences and help identify areas for further analysis, such as peak periods of activity (patterns/trends). Bar charts are to select components of interest, find changes in the matrices, and for comparing different subsets of data within the views. These two charts uses crossfilter techniques along with D3 that help in search, navigation, drilling down and investigation, which aid in

i). identifying "pertinence" in data: to filter huge data and to fetch a relevant data using visual representations (from investigation point of view). Each bar charts will enable search towards finding relevance or subsets of data of interest based on the regular activities of individuals, for further investigation.

ii). identifying "key information" in data: the selection of components will aid in comparing two different subsets of data (for ex., email data of two different years) to find key information.

iii). identifying "points of interest (PoIs)" in data: the selections and filtering aid in discovering various PoIs in time, individuals and contents. Using the PoIs, further filtering can be done in the investigation process.

The multi-faceted data is color-coded enabling the facets to be distinguished and compared quickly to find various tasks and information. The complete visualisation provides an informative and comprehensive overview of the entire dataset and exploration opportunities using the crossfilters for interaction (incremental filtering, reducing and comparing).

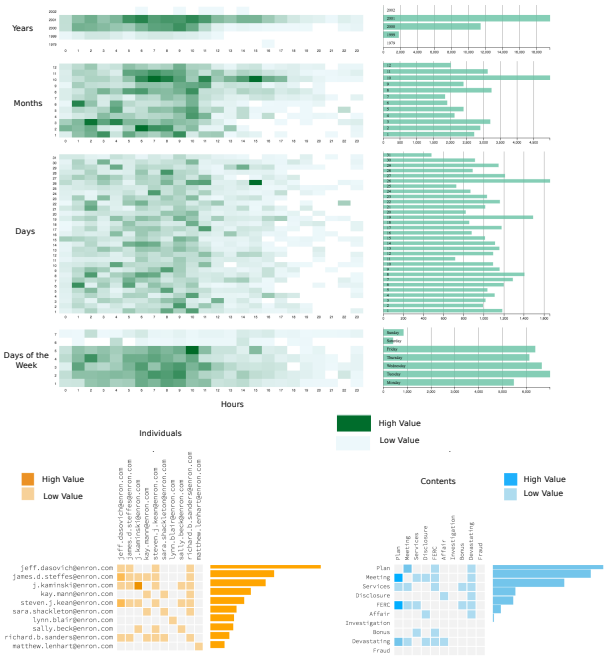

Figure 1: D3 Prototype: a combination of matrices and bar charts for investigating multi-faceted E-mail data.

Applied Context: an E-discovery Investigation - we present an example for the applicability of our model in the investigation domains that contains multi-faceted E-mail communication data in various forms: temporal, individuals/connections & context views (green, orange & blue respectively in the Figure 1). In the Enron scandal report, an US legal team had only the temporal information (October 2001) but they did not have any information about the individuals, connections and contents. The legal team had to manually comb through the E-mails to find pertinence and find individuals and keywords. Our interactive prototype helped in selecting the time-frame of interest (based on the report) using bar charts. The selected bars in the temporal view visually represent frequency of E-mails sent by all the individuals in the particular year & month of interest, which further sampled (filtered) the data to give the days, days of the week, hours, individuals' connection and contents. Our tool identified Enron employee "J.Kaminski" bearing email ID "j.kaminski@enron.com" self-emailed (cc'd) the most on October 2001 with the combination of contents "plan, meeting, investigation & ferc", which is pertinence (relationship between the facets). Our prototype not only helped in "finding" but also in "discovering" various key information, pertinence, PoIs, unexpected, unusual and interesting relationships by filtering in all the views which are consistent with the legal report of the Enron case.

5 Discussion and Future Work

Our interactive prototype can be used to explore real-archived E-mail datasets as well as our own personal E-mail accounts (live). Our work has demonstrated that searching for particular information utilising a limited search information can be done comprehensively and successfully (which would have been extremely difficult to identify using a basic email or database search). Based on the legal report or highest frequency or based on analysts' interest, the facets (time, individuals and contents) can be selected using bar charts, which further samples the data and aids in identifying various expected and unexpected information in the matrices.

We will work further to develop an effective analyst-friendly visualisation tool to explore and understand anomaly behaviours, pertinence, dynamic changes between two subsets and the underlying communication structures in E-mail communication which will help in improve E-discovery investigations. The tool will be tested by the company partners, students and legal analysts on publicly available datasets in order to observe the efficacy of finding relevance, key information and PoIs in a selected set of E-mails and also to evaluate the visualisation design choices for some of the tasks, such as aggregation, comparison, etc using Visual Data Reasoning (VDAR) [3]. The studies will help to determine the potential effectiveness of our techniques in actual ongoing email investigations.

References

[1] S. Attfield and A. Blandford. Discovery-led refinement in e-discovery investigations: sensemaking, cognitive ergonomics and system design. *Artificial Intelligence and Law*, 18(4):387–412, 2010.

[2] 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.

[3] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2012.

[4] C. Larman and V. R. Basili. Iterative and incremental developments. a brief history. *Computer*, 36(6):47–56, 2003.

[5] S. J. Luo, L. T. Huang, B. Y. Chen, and H. W. Shen. Emailmap: Visualizing event evolution and contact interaction within email archives. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific*, pages 320–324. IEEE, 2014.

[6] T. Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6), 2009.

[7] T. Munzner. *Visualization analysis and design*. CRC Press, 2014.

[8] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.