Beacon in the Dark: A System for Interactive Exploration of Large Email Corpora

Tim Repke
Ralf Krestel
Hasso Plattner Institute, University of Potsdam
Potsdam, Germany
{firstname.lastname}@hpi.uni-potsdam.de

Jakob Edding, Moritz Hartmann, Jonas Hering,
Dennis Kipping, Hendrik Schmidt,
Nico Scordialo, Alexander Zenner
Hasso Plattner Institute, University of Potsdam
{firstname.lastname}@student.hpi.uni-potsdam.de

ABSTRACT

The large amount of heterogeneous data in these email corpora renders experts' investigations by hand infeasible. Auditors or journalists, e.g., who are looking for irregular or inappropriate content or suspicious patterns, are in desperate need for computer-aided exploration tools to support their investigations.

We present our *Beacon* system for the exploration of such corpora at different levels of detail. A distributed processing pipeline combines text mining methods and social network analysis to augment the already semi-structured nature of emails. The user interface ties into the resulting cleaned and enriched dataset. For the interface design we identify three objectives expert users have: gain an initial overview of the data to identify leads to investigate, understand the context of the information at hand, and have meaningful filters to iteratively focus onto a subset of emails. To this end we make use of interactive visualisations based on rearranged and aggregated extracted information to reveal salient patterns.

CCS CONCEPTS

• **Information systems** → *Document representation*; *Retrieval tasks and goals*; • **Applied computing** → *Investigation techniques*; *Evidence collection, storage and analysis*;

ACM Reference Format:
Tim Repke, Ralf Krestel, and Jakob Edding, Moritz Hartmann, Jonas Hering, Dennis Kipping, Hendrik Schmidt, Nico Scordialo, Alexander Zenner. 2018. Beacon in the Dark: A System for Interactive Exploration of Large Email Corpora. In *Proceedings of The 27th ACM International Conference on Information and Knowledge Management (CIKM '18).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269231

1 INTRODUCTION

In today's communication within corporations, emails play an important role. This electronically stored information contains the history of relevant discussion points that drive the daily business. Often this data is analysed retrospectively by internal auditors in companies, law enforcement agencies, or journalists, getting a hold on large corpora of internal documents, e.g. released by whistleblowers. Digging through massive amounts of emails to identify fraudulent behaviour of individuals or suspicious communication patterns is a tedious task, often requiring months of labour by domain experts.

Traditional tools for this task, such as the NUIX Engine[1], usually only provide keyword or pattern search and access to a database with raw extracted data. Other interfaces to the data, such as graph databases, only support node-to-node exploration of the communication graph. Internal auditors of a large bank reported, that investigations without posterior knowledge of what to look for may result in expert users having to read lots of potentially irrelevant emails.

We propose *Beacon*, a system specialised for the exploration of large scale email corpora during an investigation. By combining communication meta-data and integrating additional information using advanced text mining methods and social network analysis, our system goes beyond traditional approaches. The objectives are to provide a data-driven *overview* of the dataset to determine initial leads without knowing anything about the data. The system also supports extensive filters and rankings of available information to *focus* on relevant aspects and finding supporting emails. At each point, the interface components are updated to provide the relevant *context* in which a certain information snippet is embedded.

To prepare a large email dataset for expert users, we use a distributed processing pipeline, which extracts structured data, such as salient phrases, from free-text email bodies. Data cleansing is crucial part in working with real world unstructured data and is partly based on a previously developed neural network based system [7]. Emails are further enriched by results from topic modelling, document clustering, and social network analysis.

We developed a web-based, interactive user interface integrates the diverse set of collected information into responsive visualisations. The structured information is aggregated at a level of granularity which fits a given context best, thus providing valuable controls to navigate the sheer amount of information.

We evaluated the performance of the system using real world datasets of different size, notably emails of the U.S. Democratic National Committee (DNC) leaked in 2016 and the well known Enron Corpus [6]. A prototype was also used in an active internal investigation of a large bank. In the following sections, we describe the applied methods and interaction concept in more detail and use Enron to demonstrate the versatility of the *Beacon* system.

[1] NUIX Analytics extracts and indexes knowledge from unstructured data (nuix.com)

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269231

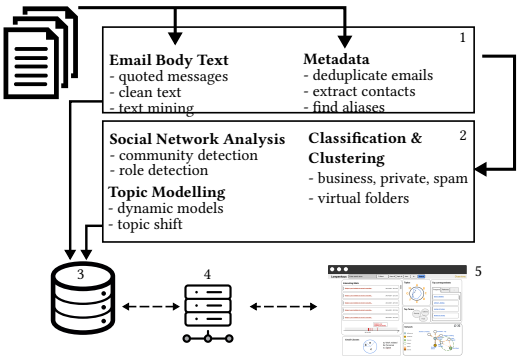

Figure 1: Overview of system architecture

2 BEACON SYSTEM OVERVIEW

An overview of the main components of the *Beacon* system can be seen in Figure 1. An Apache Spark[2] powered data ingestion pipeline (1, 2) prepares, cleans, and enriches raw email data. Results feed into Solr[2] databases (3) for efficient full-text indexing and storage of additionally extracted information about emails, as well as a Neo4j[2] database (3) containing correspondents (senders and recipients of emails), their relationships, and further information extracted from email bodies. The Flask[2] backend (4) exposes a REST API to the web-based user frontend (5) written in React[2].

The architecture of the pipeline is structured into modular tasks, which each takes an email, processes it, and returns the result. These tasks are organised into two phases as indicated in Figure 1. In the first phase, raw emails are processed, so that tasks in the second phase can rely on high quality text and an accurate representation of the communication graph. Quoted messages from email bodies are extracted and automatic in-line headers parsed. The communication graph is constructed by collapsing different name aliases of senders and recipients into individual persons, which we refer to as correspondents. Boilerplate texts, such as signatures, are removed from email bodies after extracting additional facts about correspondents (e.g. affiliation or position).

Based on the duplicate-free and cleaned dataset, tasks in the second phase of the pipeline extract salient patterns. Social network analysis is applied to the communication graph of the corporate email traffic to reveal valuable information about the organisational structure. Furthermore, we use topic models to uncover major themes. Clustering and classifying emails into work-related, private, or unsolicited mail (e.g. spam or newsletters) helps in focusing on particular types of emails.

The *Beacon* system's user interface provides powerful search capabilities and visualisations for all the structured (and unstructured) information extracted by the processing pipeline. It is organised into two views, the search view and correspondent view. Components of these views handle specific aspects, e.g. the network graph, and are tightly integrated with one another. The interface concept is tailored to the needs of expert users during an investigation. It is designed to approach a corpus with a text-driven or network-driven strategy, while keeping the option to switch context at each point.

[2]Spark: spark.apache.org, React: reactjs.org, Solr: lucene.apache.org/solr, Neo4j: neo4j.com, Flask: flask.pocoo.org

3 INGESTION PIPELINE

Before the *Beacon* user interface can be used, the raw email data is prepared, cleaned, and processed in various ways. The modular architecture allows simple domain-specific extensions, e.g. additional input pre-processors to accommodate different file formats, or adaptations of the existing tasks. In this section, we describe some of the key tasks in more detail.

Data Preparation. Emails are semi-structured text documents with protocol headers containing structured meta-data such as the time an email was sent, and the bodies in plain-text (or HTML). The body may not only contain the email the sender wrote, but also quoted messages, which are automatically inserted when replying or forwarding an email. To untangle these email threads we use Quagga [7] together with rules and heuristics to extract normalised data from these in-line headers. The extracted messages and their respective sender information are processed as separate emails.

Duplicate-free Correspondents and Emails. The process of extracting original messages from threads might lead to duplicate emails in the resulting dataset. Based on sender and time we de-duplicate the dataset while keeping track of provenance thread structures. To deal with different aliases and email addresses of correspondents, we merge aliases and email addresses refering to the same person based on extracted clues from headers and signatures.

Topic Modelling. After the raw dataset is fully processed in the first phase of the pipeline, tasks in the second phase can refer to cleaned original text messages and correspondents. We use Latent Dirichlet Allocation Topic Models [1] to find underlying semantic structures in an unsupervised fashion. Each email is modelled as a probability distribution over topics, which is added to the database for efficient querying and exploration. In future work, dynamic topic models could be considered to account for topic shifts and detecting emerging or disappearing topics in datasets which span over a long time.

Analysing the Communication Graph. The communication graph of corporate emails contains valuable information about the organisational structure. Social network analysis can reveal these structures and can be used to calculate measures for each correspondent to provide indicators during exploration [4]. Community and role detection [2] provides a powerful tool to quickly gather insights into who is in contact with whom. Communities, for example, are more densely connected clusters of correspondents and thus suggest close professional association, for example members of departments or projects. Adjacent to that, roles are derived from these high level structures. Social hierarchies can be detected by combining several of these measures into scores [3]. This approach automatically ranks correspondents in the interface, which helps to quickly spot influencers among hundreds or sometimes thousands of people in large email corpora.

Email Classification and Clustering. Clustering and classification provides additional starting points to navigate through the sheer volume of large corpora, thus making it more manageable. Therefore, we train classifiers [9] to separate business-related emails from the occasional personal email and possibly irrelevant content such

as spam or newsletters. Domain-specific classifiers may be added as desired by the experts or use-case.

Apart form the supervised approach, we integrate previously extracted information to cluster emails. Topic models already yield means to find semantically related emails. However, adding more features than just text opens a different perspective [8]. For example, some datasets contain the information of each correspondent's folder structure. The clustering implicitly brings matching folders from different users together. Emails outside of explicit folders can be added to existing clusters.

4 USER INTERFACE

The *Beacon* user interface is structured into two main views, the search view and correspondent view (Figure 2). Components of these views are specialised widgets for different aspects and are tightly integrated with one another. When expanding a widget, more details and controls become available. Each view is tailored to a different way of approaching a dataset. The search view offers a more general approach as it provides a starting point with an overview of the entire dataset at first and later a context to the active search. To investigate one person who sent or received emails in the dataset, the correspondent view offers detailed insights. In the following paragraphs, we are highlighting the key components.

Search Panel. The search panel at the top of the interface is always available and activated filters universally apply to each component. Interactions with some of the components can also be used to set these filters, thus making their use seamless. Users can also manually set these filters and keyword searches directly from the search panel. At each point during the exploration, the context is explicitly described by the activated filters.

Communication Matrix. Visualising a complex social network graph based on the email correspondence using nodes connected by edges quickly becomes infeasible as the many overlapping connections make salient structures unclear. *Beacon* therefore features a powerful adjacency matrix, where rows and columns represent correspondents and their connections indicated by coloured dots.

Salient structures become apparent when rearranging rows and columns based on different metrics. Using the probabilistic community affiliation for example, highlights well connected correspondents. Additionally adding secondary rankings reveals structures within these communities. The colour and size of dots is used to encode primary topics of the email traffic, frequency of communication, or simply to highlight search terms. All these options help to quickly establish an understanding about who is talking to whom.

Topic Visualisation. Another way to approach a large email corpus is by focusing on the contents first and the communication network later. *Beacon* uses the concept of Topic Spaces [5] to plot emails as dots into a two-dimensional space. The position is determined by the topic distribution of a respective email which sets the forces an email is pulled towards corresponding topics positioned around the circumference. This is especially useful to determine the semantic focus of one correspondent and compare her to another. Also it can be used to see how emails containing a specific search term are placed on the topic space and therefore reflecting in which aspect these terms are discussed. Hovering the cursor over items in any other component highlights emails that information belongs to, thus allowing rapid exploration of salient dependencies or frequent co-occurrences.

Other features. Although the high-level interactive visualisations are effective tools to navigate and explore a large corpus quickly or provide context to the current selection, they only help to guide the expert user to specific emails. Therefore, *Beacon* always provides lists of emails matching the active filters to skim over or look at in more detail. For each email, there is a list of similar emails and salient phrases to quickly grasp key aspects in conjunction with a topic distribution. Additional temporal context is given by the histogram over time, which shows the general frequency of communication as well as the activity of single correspondents or keyword usage. Hierarchy scores, email classes, and associated clusters are shown as labels where applicable. We complement this by a traditional graph visualisation initialised with a selection of correspondents. Interactions link to respective correspondent views, lists of emails for an edge, or expand the graph.

5 DEMONSTRATION

In this section, we demonstrate how an expert can use the novel *Beacon* system. Therefore, we consider the Enron corpus [6] which consists of over 500,000 emails from the mailboxes of more than 150 Enron employees.

As described previous sections, the cleaning phase of the processing pipeline extracts quoted messages and removes duplicate emails and reduces username aliases to correspondents. During initial data preparation phase of the ingestion pipeline, the system extracts around 1.5 million quoted messages, which are reduced to about 376,000 unique emails in the cleaned dataset. The final communication graph contains around 1,500 correspondents from over 7,000 aliases found in the meta-data. The system was able to enrich 23% of those with information extracted from email signatures.

The community detection algorithm found twelve communities, which aligns with most frequent correspondent associations derived from the signatures and email address domains. A correspondent with close contact to two or more clusters is likely to be in a managing position, which requires her to coordinate with other parts of the organisation or external partners. Roles can also be defined as people who often initiate conversations, simply forward information, or mostly don't respond.

We train the classification model using publicly available annotated emails [9] and achieve F-scores over 93% for the three classes business, private, and "spam". In order to validate the quality of the clustering approach, we calculate how many emails from each folder were spread over multiple clusters. Ideally all emails from one folder would belong to the same cluster. Across the entire dataset with 40 clusters, the model achieves a score of 58%.

In order to demonstrate how to use the *Beacon* user interface, we consider an investigative journalist, who wants to write a story about the Enron scandal after the emails were released during the public trial. The main view initially provides an overview of the entire dataset. Some of the top ranked phrases refer to "California", which the journalist heard before during the trial. Clicking on the phrase "California Energy" updates the information shown in the overview. To change perspectives and see how some of the top

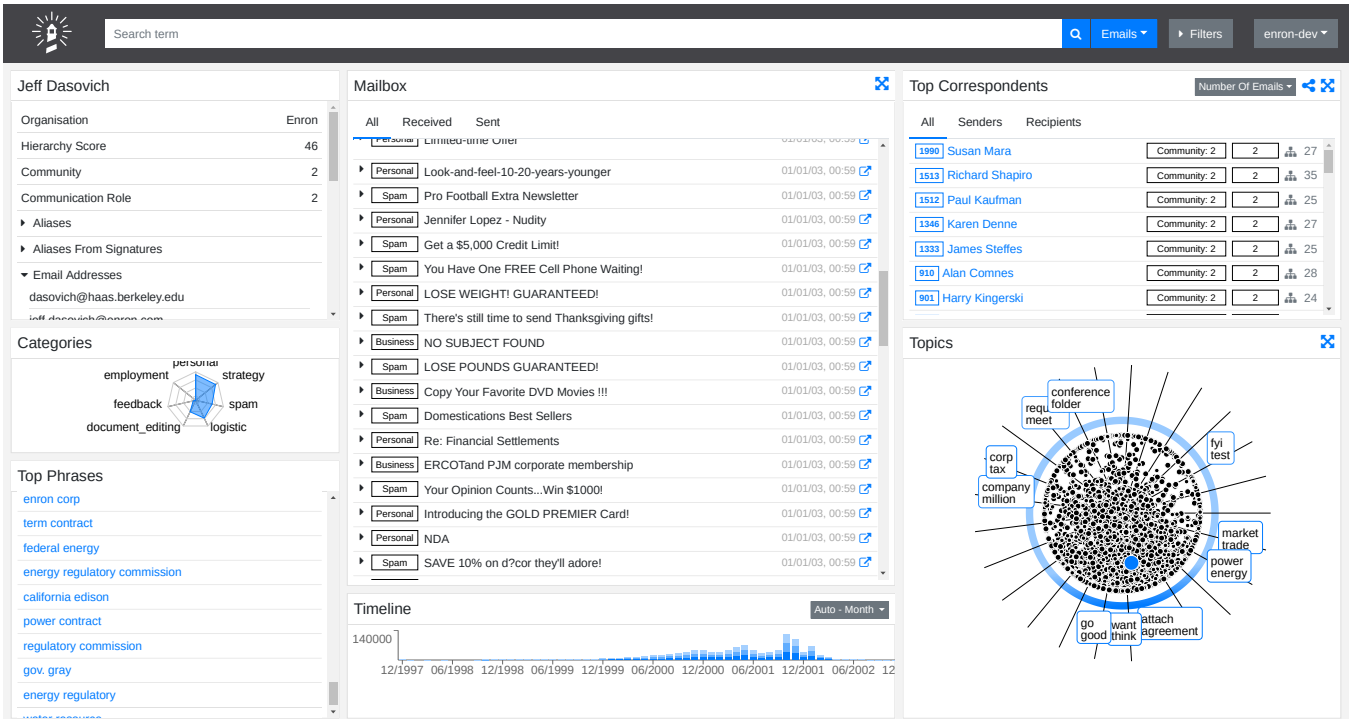

Figure 2: Correspondent View showing one mailbox, top phrases, extracted details, correspondents, topic distribution, etc.

correspondents using these keywords in emails relate, she opens the correspondent view for two of them. Topic spaces, email categories, and top phrases in this view make it easy to quickly compare them. For example, one often deals with meetings, appointments, and projects, thus suggesting a managing position, the other often uses topics for transactions, stock markets, and investments, which could indicate a position as a financial expert. Search filters can be applied by interacting with the visualisations, leaving only a few emails to read. The journalist found, that these two correspondents discussed profitable stock options after manipulations of the Californian energy market.

In this example using Enron emails, we've shown that Beacon makes it very easy to gain an overview of thousands of emails, identify key aspects, and navigate around the dataset seamlessly to find interesting emails. Our website[3] provides further information along with the source code for this demonstration.

6 CONCLUSION

In this paper we presented our *Beacon* system for exploration of large email corpora. The distributed processing pipeline cleans the raw dataset and combines text mining methods with social network analysis. It extracts structured information about topics, salient phrases, classes and clusters of emails, communities, scores, and roles of correspondents. The user interface provides interactive visualisations in which the extracted information is rearranged and aggregated to reveal salient patterns. With the example of the Enron corpus we demonstrated how expert users such as journalists or auditors can interact with such large corpora. *Beacon* was also evaluated with leaked emails from the U.S. Democratic National Committee (DNC) and a prototype is currently used by a large bank for an internal legal investigation. We have shown, that *Beacon* is a powerful tool to discover meaningful leads to investigate or focus on specific aspects and quickly find relevant emails. In future work we hope to develop large interactive visualisations that put social interactions into their semantic context and explore how email attachments can provide additional cues to investigators.

[3] https://hpi.de/naumann/projects/web-science.html

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *JMLR* 3, Jan (2003), 993–1022.

[2] Ting Chen, Lu-An Tang, Yizhou Sun, Zhengzhang Chen, Haifeng Chen, and Guofei Jiang. 2016. Integrating community and role detection in information networks. In *SDM*. SIAM, 72–80.

[3] Germán Creamer, Ryan Rowe, Shlomo Hershkop, and Salvatore J Stolfo. 2009. Segmentation and Automated Social Hierarchy Detection Through Email Network Analysis. In *WebKDD*. Springer, 40–58.

[4] Jana Diesner and Kathleen M Carley. 2005. Exploration of communication networks from the Enron email corpus. In *SDM*. SIAM.

[5] Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. 2016. ConToVi: Multi-Party Conversation Exploration using Topic-Space Views. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 431–440.

[6] Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *ECML*. Springer, 217–226.

[7] Tim Repke and Ralf Krestel. 2018. Bringing Back Structure to Free Text Email Conversations with Recurrent Neural Networks. In *ECIR'18*. Springer, 114–126.

[8] Ehsan Sherkat, Seyednaser Nourashrafeddin, Evangelos E Milios, and Rosane Minghim. 2018. Interactive Document Clustering Revisited: A Visual Analytics Approach. In *IUI*. ACM, 281–292.

[9] Aston Zhang, Lluis Garcia-Pueyo, James B Wendt, Marc Najork, and Andrei Broder. 2017. Email Category Prediction. In *WWW*. 495–503.