Digital Investigation 18 (2016) S138–S148

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS USA 2016 — Proceedings of the 16th Annual USA Digital Forensics Research Conference

InVEST: Intelligent visual email search and triage

Jay Koven*, Enrico Bertini, Luke Dubois, Nasir Memon

NYU Tandon School of Engineering, United States

Keywords:
Email forensics
Data forensics
Data visualization
Email search
Data analytics

A B S T R A C T

Large email data sets are often the focus of criminal and civil investigations. This has created a daunting task for investigators due to the extraordinary size of many of these collections. Our work offers an interactive visual analytic alternative to the current, manually intensive methodology used in the search for evidence in large email data sets. These sets usually contain many emails which are irrelevant to an investigation, forcing investigators to manually comb through information in order to find relevant emails, a process which is costly in terms of both time and money. To aid the investigative process we combine intelligent preprossessing, a context aware visual search, and a results display that presents an integrated view of diverse information contained within emails. This allows an investigator to reduce the number of emails that need to be viewed in detail without the current tedious manual search and comb process.

Introduction

In this paper we present InVEST, a methodology and tool which will aid in the discovery of evidence and information in a *large* email data set relevant to an investigation. Large, for the sake of this discussion, is any email data set that is so large as to prevent the examiner from conducting a manual search and review. For example, the 2002 Enron bankruptcy email data set contains over 500,000 emails. For such a large data set, the assistance of a tool is often required. We develop a visual analytic approach that is aimed at assisting the investigator in finding emails related to his case especially when the exact nature of the evidence is unclear.

In her book **Visualization Analysis and Design** (Munzner and Maguire, 2015) Muzner demonstrates the advantages to using visual analytics when the exact question is not known. This is frequently the case with forensic investigations where the term “discover” is often more appropriate than “find.” Visual analytic tools work with the human visual systems ability to identify patterns, trends, and anomalies, even in situations in which machine learning algorithms fail. This is particularly true when the content of emails may change with each investigation, making it difficult to create a training data set that would be effective across investigations. The ability to put the investigator in the middle of the analytic loop rather than just evaluating the results creates a more efficient investigative process because the investigator continuously refines the search process until the desired results are found. The InVEST pipeline was developed to support this iterative investigative approach.

In general, when conducting an investigative search through emails it is the answer to the questions *who?, what? and when?* that an investigator seeks. Currently much of the investigation of emails is done with keyword searches of both the contents and headers of each email using methods that have changed little in 30 years. Often the best keywords for these searches are unclear or ambiguous, leading to excessive numbers of emails in the results. Intelligent Visual Email Search and Triage (InVEST) offers a methodology for reducing the number of emails which need to be viewed in a

* Corresponding author.
E-mail addresses: jkoven@nyu.edu (J. Koven), enrico.bertini@nyu.edu (E. Bertini), dubois@nyu.edu (L. Dubois), memon@nyu.edu (N. Memon).
URL: http://engineering.nyu.edu/people/jay-koven

http://dx.doi.org/10.1016/j.diin.2016.04.008

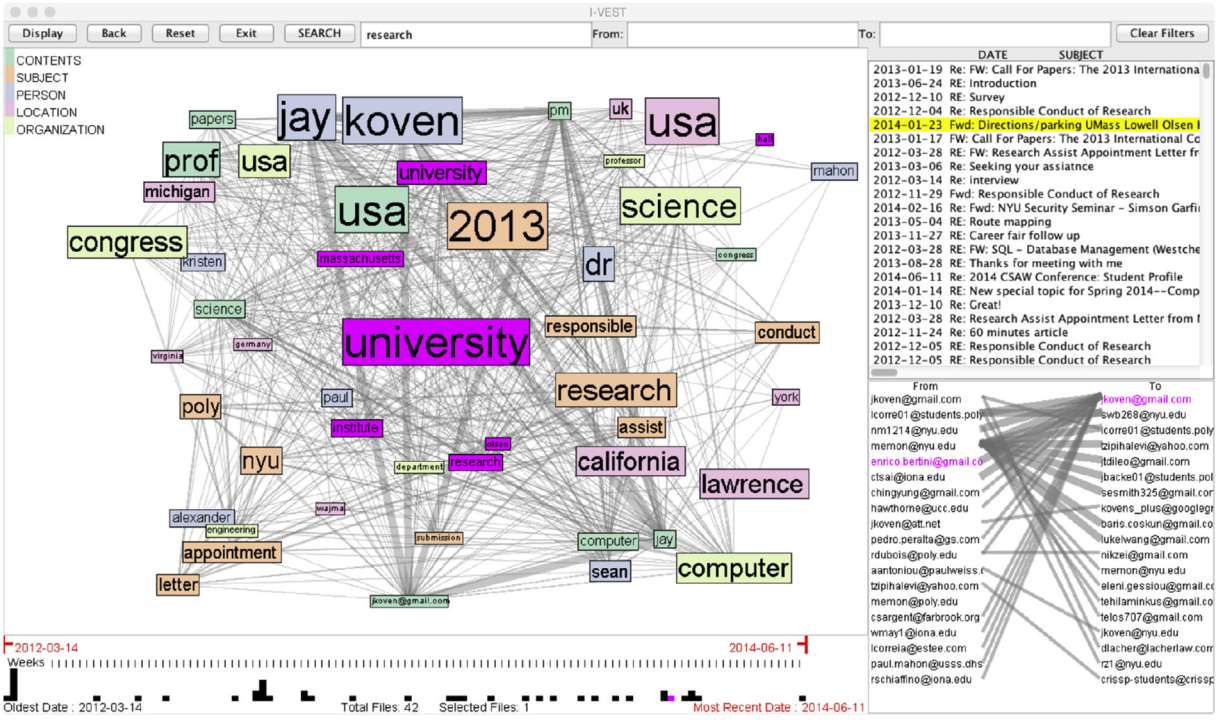

Fig. 1. InVEST display showing relationships to a selected file.

large data set while also giving the user a quick overview of the keywords, entities, correspondents and their relationships within a set of results. Data visualization uniquely allows the investigator to evaluate these relationships in the context of their surrounding data, making the detection of anomalies, trends and patterns easier.

Our interviews with both private and government lawyers and investigators revealed several consistent observations related to investigations involving email data sets. First, the data sets are very large and are growing rapidly which provides a challenge to finding relevant information. Second, our interviews revealed a lack of a good set of investigative tools to deal with many of the issues created by large email data sets. These issues include:

- Reducing the size of keyword search results.
- Removing duplicate, unimportant or unrelated emails from the data sets.
- Finding anomalies in the data.
- Inability to summarise search results or other subsets of emails.
- Finding indirect relationships between email accounts.

Finally, since these data sets contain emails from multiple related correspondents, there is a large amount of duplication of emails due to forwarding and replies that add to the volume of the data but do not provide any information gain to the investigation. The immense volume of emails leads investigators to use alternate approaches such as focusing on a single individual or greatly limiting the time range. However, this is not possible in many investigations, therefore better methods of dealing with these data sets are necessary.

Most of the subjects that we interviewed investigated the emails by importing them into a common email program such as Outlook or Thunderbird they then use the REGEX search ability supplied by these programs to manually search for relevant information using predetermined keywords or addresses. Since, they often do not know the exact nature of emails they are looking for they must make these searches as general as possible so they do not miss relevant data. They then comb through the search results to find the emails they are looking for. These methods require large amounts of manpower, time and money, all of which are usually in short supply. A well designed visual analytic tool allows the data to essentially draw itself, which allows anomalies to be quickly identified among the surrounding homogeneous data points.

In this paper we present three contributions: First, we introduce a context based visual search with extensive preprocessing of the data set that greatly improves search efficiency by removing duplicate information and junk from search results. Second, we define a visual analytic pipeline that supports user interaction with email search results and is guided by the contextual information found within these results. This information includes ranking of important keywords and entities, relationships of senders and receivers and temporal flow of search results as well as relationships between them. Third, we present a filter and expand interaction with the search results that allows for an efficient triage of data in order to produce a manageably sized set of results to be examined in detail. By bringing these contributions to the digital forensic community, we offer a way for the digital forensic examiner to be more efficient, achieve better results, and make large email data sets more manageable.

Related work

There are several areas of research related to our project. The first area is the work related to investigative analysis of email and textual based documents. Second there has been research in the visualization of emails and other large collections of text documents. Finally, there is related work in social network analysis of email collections.

Investigative techniques

The most comprehensive work on text based data for investigative analysis has been done by the Jigsaw Project at Georgia Tech by Stasko et al. (Yi et al., 2007; Kang et al., 2011; Liu et al., 2013). Their work focuses on supporting the investigative process by creating tools that help the analyst find and map relationships found in data sets. These relationships can be between people, places and things in any combination. These tools help the analyst piece together a coherent story from information contained in a document set which is limited to several thousand documents. For emails this size limit is not adequate, InVEST adopts visualization techniques that find relationships in data sets numbering in the hundreds of thousands.

Haggerty et al. (2011) propose a framework for the forensic investigation of unstructured email data that justifies the need to develop methods and tools to explore email data sets. However, their proposed framework uses visualization in only the final presentation stage which limits the advantages of visual investigation. The followup paper (Haggerty et al., 2014) shows some of the potential of visualizing the relationships of the social network combined with the email content using tag clouds for emails at the folder level. Our work takes this approach to the next level. It fully integrates visual analytics that allow the investigator to examine relationships between content, social network and time starting with the results of keyword search. We then allow the investigator to explore down to the individual email level.

Martin et al. (2005) explored large collections of emails in order to discover spam. Their methods focused on analyzing features of emails such as number of attachments, embedded images and attachment types. While they were not analyzing the content of messages their work shows that other email features such as choice of punctuation, or number and type of attachments can yield important information about the documents, such as whether or not an email is spam. InVEST expands the information available to the investigator by adding extracted entities to the visual presentation of the data thus giving him a broader overview of the contents.

Keim and Oelke (2007) demonstrated that large collections of literary text can be analyzed by looking at word usage and sentence structure to bring out hidden features of the collection thereby eliminating the need to read actual content. Their visual presentation of the books of the Bible shows striking changes in structure throughout the series and demonstrates how visualizing textual data can lead to information gain that would be difficult, if not impossible, to discover using other techniques.

Li et al. (2006) explored automated clustering of emails by feeding information derived from semantic analysis of email subject lines into the SVM classifier that was used for topic analysis. They determined the success of their analysis by how closely it clustered the emails by the folders within which they were originally filed. This method of measurement is only valid under controlled conditions since it is unclear how different filing strategies effected results. For example users may file emails by noting either whom they came from (senders) or what their subject matter is (topic). The results would differ greatly for the same data set. However, it did show that adding semantic information does make a significant contribution to understanding content of the clusters. Kulkarni and Pedersen (2005) similarly explore the content of email clusters in order to assign relevant labels to groups of emails. Thematically clustering as a method to improve forensic search results was demonstrated by Beebe and Liu (2014) and Beebe et al. (2011) in a series of work. By using machine learning tools to extract entities and important keywords and using a visualization to display the relationships between them InVEST gives the investigator an expanded view of the semantic content of the search results which improves his ability to find relevant content.

In EmailTime: visual analytics and statistics for temporal email Joorabchi et al. (2010) explore techniques for the visualization of temporal relationships of emails. Again these techniques show interesting characteristics of a data set. Temporal relationships are an important aspect of data, although, they need to be combined with more features to be a useful part of the investigative process. We integrate the temporal relationships in our work with the social network and content to form a more complete picture of the data set.

Finally, Kerr (2003) explores the relationships between senders and receivers in email threads using a unique arc visualization which displays connections between senders and receivers in an email thread using a series of arcing arrows to show the connections. This work led us to thinking about the importance of tracing the sender/receiver relationships in search results in addition to threads. Using link and brush techniques InVEST connects the sender/receiver relationships with the email subjects and contents to give a more complete picture of the data.

Social network graphs

There have been projects that have used social network graphs (Diesner and Carley, 2005; Shetty and Adibi, 2005) to explore email data sets with the general goal of discovering hidden connections within the data. The Enron email Data set has been a popular choice for these analyses due to its large size and complex social interactions. A recent MIT project, Immersion[1] uses a force directed graph display to show a user the social connections within his own email accounts. Immersion is effective in showing underlying connections within a single user's email data set. InVEST takes this methodology even further by using it to identify related connections within and between multiple users' email accounts.

InVEST methodology

InVEST introduces a new approach to email investigation. The key features of our investigative methodology are:

[1] https://immersion.media.mit.edu/.

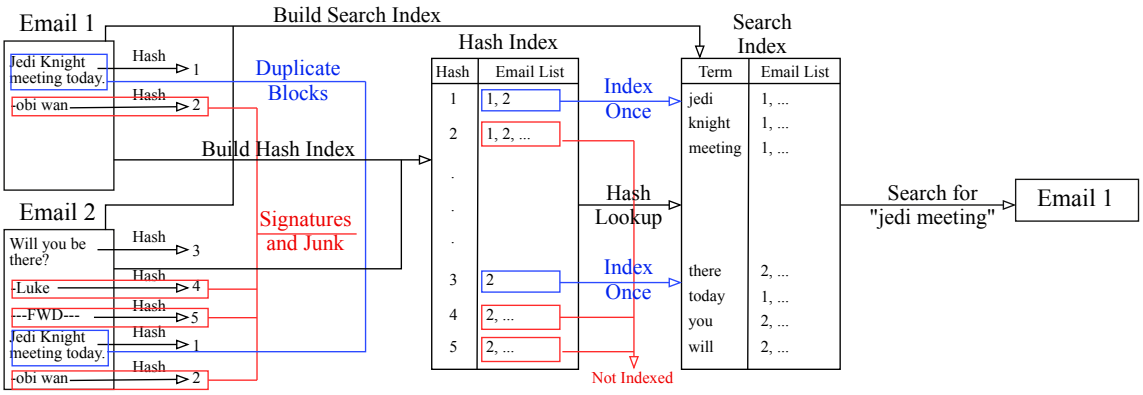

Fig. 2. Flow diagram for pre-processing and search indexing with de-duplicating, de-junking and signature removal.

1. Preprocessing of the email set to remove duplicate information and junk.
2. The extraction and identification of entities as an integrated part of the email search process.
3. Intelligent guidance based on the ranking of important terms and entities.
4. A visual analytic pipeline that allows the filtering, expansion and organization of the investigative results.

Preprocessing

The InVEST data preproscessing consists of two integrated parts. The Apache Lucene search engine library[2] is used to create the search indexes for the various email fields and extracted entities. The Enron data set used in the case studies contains approximately 517,000 emails and is preprocessed in about 33 min. The largest data set we have processed so far is just under 1,000,000 emails. The Lucene indexing is very efficient and used with other forensic tools as well as being the basis for Solr[3] and ElasticSearch[4] which are also used by some forensic tools.

Duplicates and junk

Junk, which we define as text blocks that do not add any information gain to an investigation, such as signatures, greetings and headings, are identified as chunks that appear in emails more often than a predetermined threshold number. For our testing we used a threshold of 10 which was chosen based upon trial and error indexing of five different data sets and observing the resulting indexes. These chunks are not indexed at all so they are not found in searches and do not affect the results shown to the investigators. This removes a lot of extraneous information from the result displays, making them easier to interpret. Neither duplicate blocks nor junk are removed from the actual emails so that when the investigator reads the full email all the information is still present and important relationships can be identified.

We remove duplicates and junk during the preprocessing and indexing of the emails (Fig. 2). In the first of the preprocessing passes before search indexes are created, we break up the content of each email into chunks that are separated by horizontal white space. These chunks which are usually paragraphs but can also be titles or signature blocks as well as other separators, help identify forwarded or referenced text. They are then hashed to create a unique signifier. During the indexing pass the chunks related to a specific hash are only indexed once and additional occurrences of the text are indexed via their hash. The process of creating chunks also helps identify threads to some degree. However, since many email programs add vertical bars, arrows or other characters to identify forwarded or replied to text in messages we remove these characters when they appear at the beginning of a line before the hashes are created. When searching for information only one email with an indexed chunk will be returned. However if the investigator finds a relevant email he can then find all other emails that have related chunks in them through the hash indexes.

Entities

Entities, which are defined as sequences of words in a text which are the names of things, such as persons, company names and locations, are also extracted and indexed during the preprocessing. InVEST currently uses the Stanford Named Entity Recognizer (SNER) (Finkel et al., 2005) for entity extraction. The SNER is a open source public licenced java library that uses specific language models to define and extract entities. Since, the extraction is based on loadable models, it can be trained for different languages and different entity types. The model used by InVEST extracts names, organizations, locations and dates. The extraction of entities from email contents allows for clear separation of the subjects in the emails from the correspondents. This separation combined with a display that clearly shows relationships between entities, corresponders and contents allows an investigator to quickly gain an overview of emails in the search results.

[2] https://lucene.apache.org/.
[3] https://lucene.apache.org/solr.
[4] https://www.elastic.co/.

Guidance

Intelligent guidance to the investigator through the listing of important entities, corresponders and keywords within the displayed results suggests possible avenues of search to pursue. Using the rankings for intelligent expansion of the graph by creating nodes for important terms, entities or corresponders can show meaning and relationships hidden within search results.

Interaction

The visual analytic pipeline integrates the entity extraction, an interactive visual environment, and intelligent guidance. By interacting with data the investigator can filter, expand and organize search results. The pipeline provides a structured environment within which emails can be searched, investigated, and eventually triaged in order to reduce the data set to a relevant subset of emails which can then be examined in detail.

Visual analytic pipeline

Visual analytics, defined by Keim et al. as "an integral approach to decision-making, combining visualization, human factors and data analysis" (Keim et al., 2008) is well suited to the process of exploring email data sets where the exact nature of a search is hard to define. This process is more closely related to discovery verses more traditional forensics methods which are aimed at finding specific pieces of data using well defined searches. A key part of the visual analytic pipeline is a feedback loop that allows the user to refine the data analysis to interactively improve the results. By including the user in the feedback loop, the InVEST tool makes full use of the human visual system, which excels at being able to identify patterns, trends, and anomalies. This is what makes the InVEST tool better than any existing text-based tool with a machine learning algorithm.

The core of our approach is to present to an analyst a graph of nodes and edges that allows exploration of the connections between people, locations and terms in emails. The graph reflects both the content of email text and the subject line. In addition, a bipartite graph shows the connections between senders and receivers. In order to explore this information our system implements an interactive visual pipeline (Fig. 3) which allows a user to explore the content and relationships between emails contained within a target data set. The information in these graphs combines to form intelligent suggestions to an investigator and guides them toward information and relationships contained within emails which may be important. The information in all of the displays are synchronized and cross linked. Highlights, selections or deletions are reflected in all of the displays to maximize relationship information in the results. This allows the user to easily identify the relationships, and it presents the information in such a way as to not overwhelm the cognitive ability of the user. Thus, the InVEST tool makes it possible for the digital forensic examiner to keep pace with the growing size of email data sets.

User search

The process of analyzing an email data set starts with a user defined keyword search that can be enhanced by filtering for specific senders or receivers. Depending on the initial results additional searches can be added to either reduce or expand the result set. The search tool, as well as all of the indexing previously mentioned, is built on top of the Apache Lucene search engine. All header fields and extracted entities are indexed as well as the email body text. This gives the investigator greater flexibility to reduce or expand the email set being considered based on the initial search results. By including the user in the feedback loop of the search algorithm, better results are achieved.

Displaying the results

Results from searches are presented in a cumulative fashion in five separate views, three of which are always visible while the remaining two displays can be viewed by the investigator individually. In addition to these views there is a color key that ties node colors to fields, such as "Subject", "Contents", "Person", "Location" or "Organization". "Subject" and "Contents" represent terms from the subject line and body of the emails while the remaining fields represent entities extracted by the SNER.

The main view is a network graph display of nodes and edges. Each node in the graph represents emails that contain a keyword represented by a node label. The field that the keyword was found in is indicated by the color of the node, while the size of the node indicates the ranking of the term or entity represented. Each edge represents emails that are in common between two nodes which are connected by the edge. The thickness of the edges represents the number of emails. The number of emails in the results as well as the number of emails in selected nodes are also displayed at the bottom of this view.

The visual display of the keywords allows the investigator to easily identify the nature of the keyword through its color. By selecting keywords of interest he can quickly

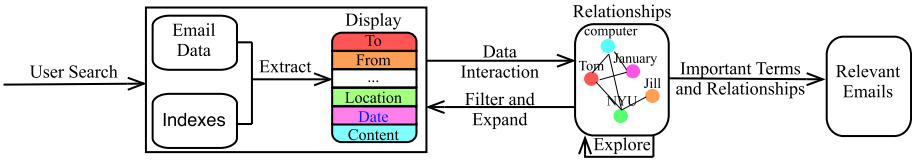

Fig. 3. InVEST analytic pipeline.

see their relationship to each other through the highlighted edges. Relationships between unconnected keywords can be identified though intermediary nodes. Email subject lines related to the selected keywords are highlighted on the list display which gives the investigator an overview of the likely content of the emails associated with the selection. The investigator can quickly try different selections to find subject lines he feels are related to his investigation.

An alternative main view, which can be toggled with the graph view displays the same information as the graph view in list form. Each field has a list of keywords associated with it that is displayed in rank order. The number of emails in the resulting set that contains each keyword is shown to the right of the word and the relative rank value for each is shown by a translucent field colored bar. When the mouse pointer is hovered over a keyword, all keywords in the display that appear in emails with that word are highlighted to show relationships.

A scrollable list of emails in the search results is displayed to the right of the graph. The list shows the date and subject line for each email and they are sorted by search rank order using a TF-IDF algorithm. When more detail is required the complete header and contents of emails can be viewed in a separate window by selecting them in the list. Emails can also be removed from the results using this view.

Below the list of emails is a bipartite graph of the most common senders and receivers in the results. The thickness of edges between senders and receivers represents the number of connections between the two in the results set. Hovering the mouse pointer over a sender or receiver will highlight the edges and connections related to that sender or receiver.

Finally, below the main display is a temporal display of the results in the form of a time line sectioned by weeks. The height of the weekly bars represent the volume of emails sent during that week. The two sliders above the time line can be used to limit results to a specific interval. By providing multiple views to the user he is better able to make use of his perceptual bandwidth which allows for the processing of more data in less time. This results in the investigator being more productive when using the InVEST tool.

Filter and expand

All views can be used to explore and refine search results by filtering and expanding. Nodes and edges can be selected and highlighted in the graph view. The results can be filtered by removing, combining and sub-setting selections.

Emails in displayed results can be filtered by deleting selected emails in the list view. When emails are deleted all views are updated to reflect the changed results. The full headers and contents of emails can be examined in a separate window with important keywords highlighted. The results can be expanded by requesting that emails related to the selection be displayed. Results can also be expanded by adding keywords or entities found in the results to the search in order to pull in additional emails that contain new terms.

While exploring a data set the user can back up (undo) through the steps taken to create each display and move forward in a different direction if desired. The iterative process of filtering and expanding is repeated until a search is successfully completed. The combination of the visualization of the relationships between the keywords and the linking of those relationships with the social network of senders and receivers is a powerful aid in the exploration and discovery of information in email data sets. This power is reenforced by the addition of keywords extracted from the email contents.

Interaction with the data

As discussed previously the analytic process begins with a familiar keyword search to establish a starting set of emails for exploration. These emails may be displayed in one or more nodes with edges showing relationships between emails represented by the nodes. At any time during the analytical process results from additional searches can be added to or removed from the displays along with any edges that may connect new results with those that already exist on display. By using a visual display to organize past and future search results, the cognitive load on the user is reduced. By reducing the cognitive load on the user, the user is able to devote more cognitive energy towards analysis rather than wasting cognitive energy keeping track of past and future search results.

In addition to supplementing the graph with additional searches an analyst can work directly with the graph. He can organize nodes on the screen in a fashion which reflects how he is thinking about the data. I.e. he can drag nodes closer together which contain related emails or move other nodes to the side that may not be relevant to a current investigation. When it is confirmed that nodes are not needed they can easily be selected and deleted. While nodes can be deleted edges cannot since they represent the intersection of the contents of two nodes. When a node is deleted, the term associated with its label is no longer considered relevant. It is no longer shown since it has no ranking and its value is not used to rank the emails that contain the term. Nodes can also be combined by dragging one node on top of another node. The combined nodes will contain the emails from both nodes and will be labeled with a new label that contains labels from both nodes separated by a vertical bar. During exploration an analyst can create new graphs to work with by selecting subsets of existing nodes and edges or by combining nodes before continuing to explore. If at some point during exploration an analyst is unhappy with his current set of results he can back up through his decision making process and proceed with a different set of choices until he is satisfied with the set of emails displayed in the graph.

Cross linking of the result displays

Since the email list and the graph display are presenting different representations of the same data they are closely linked. If a node or edge is selected and highlighted in the graph then emails that are in the selected node or edge are also highlighted in the temporal display, bipartite sender

receiver graph and email list. The email list display hides all unselected emails when there is a selection so that emails related to the selection are easier to find. The reverse is also true. When a user selects an email or group of emails in the email list, nodes that contain the selected emails are highlighted as seen in Fig. 1. Since the information given to an analyst by each of the displays is very different, using highlighting to tie these displays together allows for quicker identification of related emails. Cross linking of the different visual presentations creates an extremely flexible system that gives an investigator multiple avenues of exploration when trying to discover evidence that is in contained in the data set. Automatically changing the related information in each of the displays as the user focuses on specific relationships, terms, or time frames maximizes the information gain from each interaction with the data.

Finding important terms and entities

Defining and finding important keywords, senders, receivers and entities in email data sets is a key feature of the InVEST approach to analysis. Both the graph and keyword display lists rely on finding relevant terms in each of the email fields. In order to guide an analyst to relevant emails the system must find key terms while filtering out those which are common or irrelevant. Much of the work on this type of feature extraction has been done focusing on web search engines and on news feed types of data as well as on information retrieval systems. Because the text contained in emails is usually abbreviated and often contains incomplete sentences and thoughts it is not clear which of the current methods of feature extraction, if any, will produce the best results. InVEST currently uses a version of TF-IDF for identifying important keywords and entities where the set of emails being displayed is treated as a single document for the purpose of ranking all terms in the results. Conceptually TF-IDF determines the importance of a term by comparing its frequency in the search results with the overall frequency of the term in the data set. The more frequently the term appears in the search results the more important it is and the more frequently the term is in the overall data set the less important the term. The final ranking is a balance of these two values.

The basic equation for the TF-IDF calculations is:

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D) \quad (1)$$

Where t is the term being ranked, d is the document being ranked (in InVEST this is the set of emails in the current results) and D is the entire collection of emails. Term frequency (tf) is the number of times the term appears in the result set and inverse document frequency (idf) is the inverse of the number of emails in an entire data set that contain the term. To rank the emails the rank value of each term in the email is summed together.

Case studies and discussion

We used InVEST to explore the Enron Email data set to confirm its ability to give meaningful feedback while exploring a large and confusing data set. The Enron email data set is a set of about 500,000 emails that were collected as part of the investigation of the well-known accounting scandal that brought this energy company to bankruptcy in 2001. This publicly available data set is frequently used within the research community because it is the only publicly available data set that was used in an actual law enforcement investigation. This data set is also used because it's extreme size and complexity make it a challenge for a manual review. Thus, this data set is an ideal data set to be used to evaluate the InVEST tool. The Enron data set is also notoriously difficult among researchers when it comes to finding useful information in its email contents. Most published research based on this data set involves exploring the social network as cited in the related work section of this paper. One of the biggest impediments to finding useful information in the Enron data set are the large number of emails with subject lines like “Enron Mentions”, “Energy News”, or “Energy Issues”. of which, there are over 24,000. These emails are executive summaries that were distributed to an extended mailing list, tend to be long and contain market information or news that is related to Enron’ business. They are usually quite long and each covers an extended range of subjects, so that almost any normal email search for information related to Enron's business or top executives returns thousands of these emails, making it difficult to find real information in the data set.

Using the Enron emails for our case studies seemed appropriate since most forensic searches of data sets start with some basic assumptions about what the investigator is searching for such as an individual and/or a crime. We know that Enron committed multiple crimes related to energy trading. We therefore decided to look for information related to two different areas related to Enron's efforts to control the energy market.

The case studies were conducted by one of the authors after consultations with a professional investigator from law enforcement on possible exploration strategies. In addition, in order to understand more about the company and the original case, the author discussed the original Enron case with a lawyer who was involved in some of the original civil cases related to the Enron bankruptcy.

Case study 1: finding exploitation of a market

The first question we tried to answer was “Find a previously unknown (to the user) energy market exploitation and find the emails that define its history”. Using the search terms power, energy, source, and market lead the user to identify the high value term India. An additional search for Indian power generation eventually lead to the discovery of an exploitation of power generation by American Indian tribes.

The subject line from a single email stood out as different and relevant to the question. “Seminole Indian Tribe Project”. Reading the email confirmed that power generation on Tribal Indian Lands was indeed potentially a new market exploitation previously unknown to the author. Without the visual presentation of the keywords this anomaly would have been unlikely to have been

discovered. Further searches into the data set armed with the specific target "Tribal Indian Projects" using different variations of terms displayed graphs that were rich with information. They uncovered a number of projects either in the negotiation stage or started with various American Indian Tribes including the Seminoles, Navajo, Utes and Warm Springs tribe. Relationships in the graph helped locate emails referring to negotiations with various government agencies including federal and state. The combination of this information led to a key email that explained that the basis for this exploitation is the fact that American Indian Tribes are recognized as "Independent Nations" and not subject to either state regulation or (some) federal regulation. This gave Enron and the Tribes strong leverage when negotiating contract and distribution agreements for energy generated on tribal lands. This email was found by combining related search terms shown in the graph.

After the focus on American Indian related energy was established, the author found approximately 100 emails which related to various tribal Indian projects. The subjects included contract negotiations, prospective project discussions, legislative lobbying and projected profits. As can be seen in the results graph in Fig. 4 emails related to the Seminole Tribe discussed a new power plant as well as strategies for dealing with both state and federal regulatory issues. The time to make the discovery and exploration including the skimming of at least 30 of the discovered emails was approximately 1 h. The iterative search procedure which is supported by the InVEST analytic pipeline allowed the user to quickly focus on relevant emails by emphasizing important terms and allowing the filtering out of unrelated emails that cluttered the search results. Key features used in the discovery of the exploitation of Indian Tribal lands were the linking between the graph view and the email list and the ability to quickly remove irrelevant terms from the graph which brought new terms into display.

Case study 2: efforts to influence legislation

Another question we answered was "Find and verify Enron's attempts to influence legislation which allowed them to exploit the US energy market." On the surface this is an easier question to answer than the previous one since Enron's exploitation of the California Electric Market was, if not the direct cause of their eventual downfall, certainly the focus of their vilification in the public eye. The executive summaries contained countless references to Enron and its dealings with state and federal government agencies and representatives. Therefore several initial searches were not initially useful. However, a search for "California legislative agenda" turned up 89 emails, a manageable number, and after removing the executive summaries by finding terms associated only with the summaries and removing them from the graph, there were 23 emails discussing various legislative issues. At this point the bipartite graph proved useful. There was a clear pattern of communication in the remaining emails that suggested a large amount of communication by a relatively small number of correspondents. By reading only a few of these

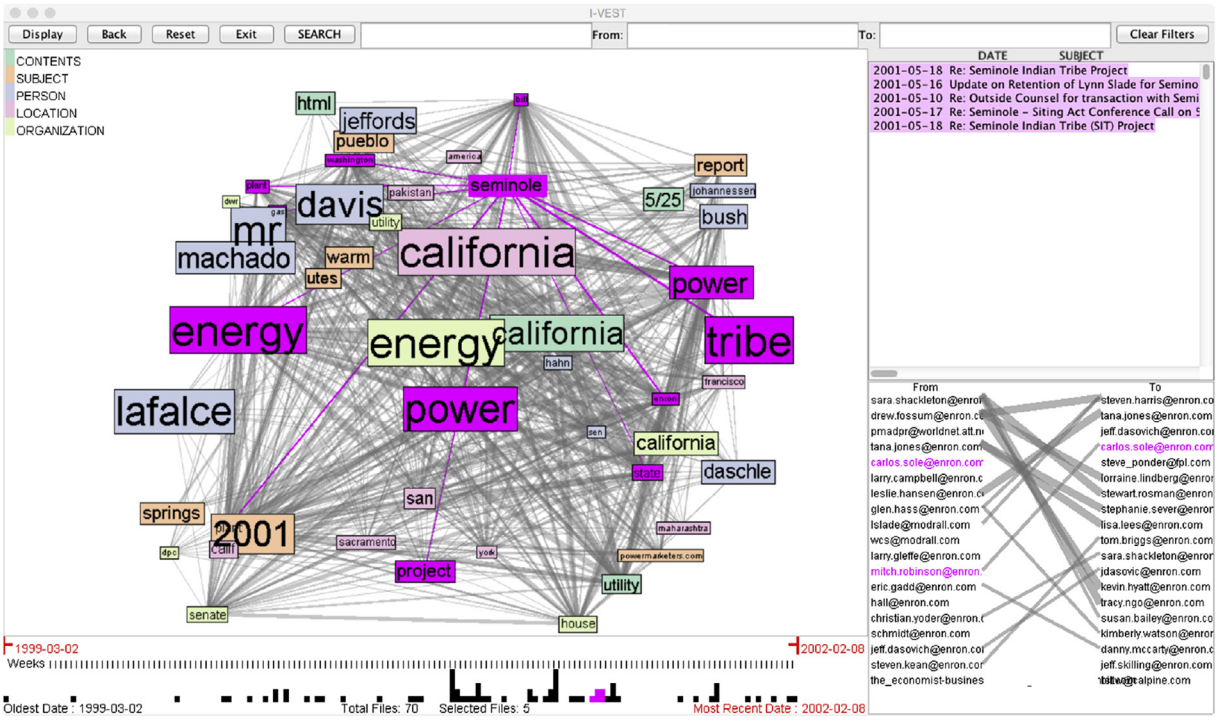

Fig. 4. Results graph for Enron Tribal Search. Highlighted are the relationships and emails related to the selected term subject: Seminole. The Cross linking of the different areas of the display give the investigator a detailed picture of the content and communications related to the selection.

emails isolated by the communication links we quickly discovered that Enron had a Government Affairs group or committee and the correspondents identified in the bipartite graphs were the group members as can be seen in Fig. 5. The cross link feature of the visualization made finding the communication pattern easy to find. By moving the mouse over various keywords the senders and receivers related to those keywords are highlighted which makes different patterns of communication stand out clearly. A followup search on "Government Affairs" turned up 501 emails after removing the summary emails. The California power and energy concerns were heavily discussed in the group's emails which was evident in the graph display. Detailed reading of some of the emails, which were chosen by subject, revealed that Enron had a comprehensive approach to lobbying government agencies on the common themes of opening markets to Enron's companies and deregulating energy markets that Enron traded in. Using the related email feature then allowed us to gather over 300 emails directly related to the Government Affairs committee and their plans on how to manage the California PUC and legislature.

Personal email explorations

In addition to the Enron case studies InVEST was used to explore personal email accounts. Exploration of personal accounts was implemented to confirm the quality of the results display. Since users were very familiar with the contents of their own accounts they were able to confirm whether or not the graph display accurately reflected information contained in resulting emails.

We made early versions of InVEST available to 5 academic researchers familiar with forensics, visualization and human computer interaction to allow them to explore their own Gmail accounts. We did so to incorporate feedback from experienced users in order to improve the interface and to make the interaction with data more intuitive. In addition we will use feedback we obtain to improve the algorithms used to determine interesting or important keywords, entities and people in search results. Initial feedback from these users has proven to be thought provoking. For example the need to remove or hide duplicate information or junk such as common signatures became clear as result of feedback from these early users. In addition some users suggested adding the ability to select a link between terms or senders and receivers in order to isolate emails relevant to only those terms or accounts to improve the filtering process. Feedback also lead to improvements in the term ranking and expansion algorithms. Other uses for the tool were suggested by these users as well. It was suggested that the tool could be made useful to help remove unwanted emails from personal accounts or to help organize the emails in a more logical fashion.

Discussion

These case studies, which are somewhat contrived due to the contents of the Enron data set, are examples that demonstrate the utility of the InVEST tool. Actual

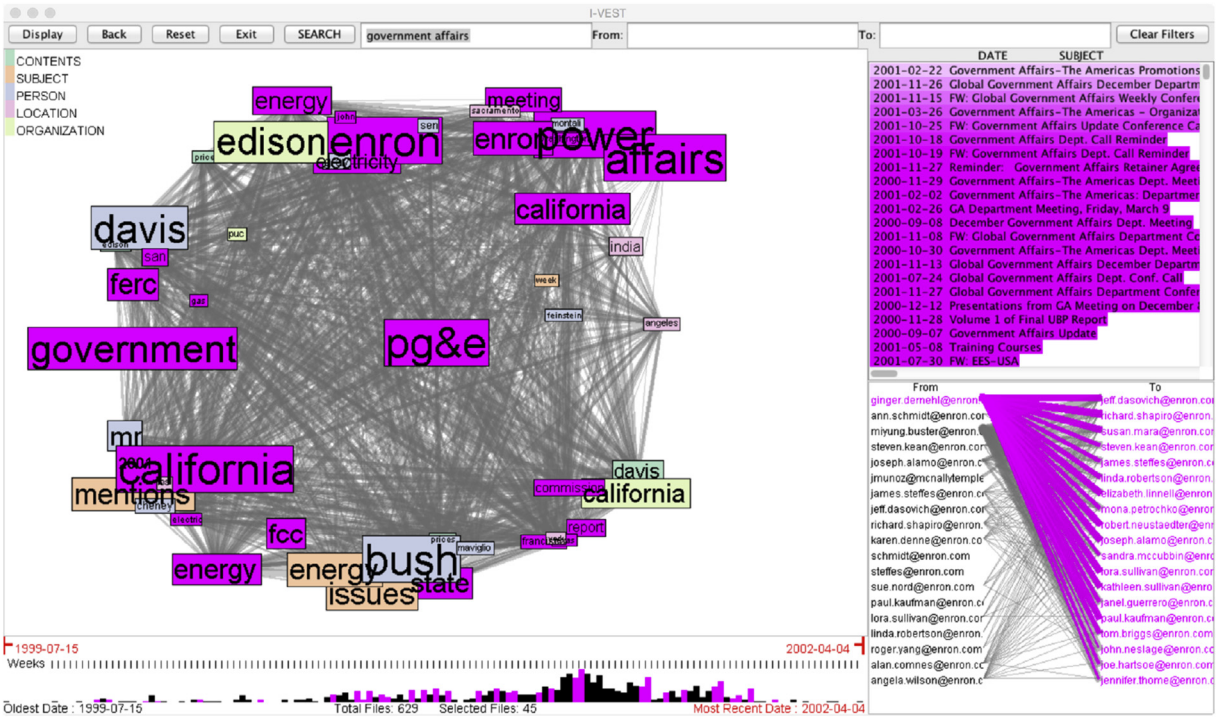

Fig. 5. Results Graph for Government Affairs. The Sender Receiver Graph Pattern on the lower right shows a strong relationship between members of "government" Affairs group. This pattern was highlighted by the selection of the term Government in the graph display.

investigative email data sets with ground truths are not currently available to researchers due to legal and privacy concerns. The case studies allow us to validate the design of the tool and also to verify that users can get meaningful results with the tool.

Additional forensics applications

The InVEST methodology and pipeline is not limited to the investigation of emails. In the current age of social networks, investigators are also faced with SMS messages, Twitter tweets and Facebook accounts that my contain evidence of a crime which may be even harder to find than in emails given the broad scope and large size of the data generated. Data from these and other social media sources is similar to emails in that it is semi structured with information about the owners, senders and even viewers (likes) as well as having text subjects and content. These attributes make InVEST highly suited for exploring these data sets with only slight modifications.

Future research

Initial work on InVEST has identified several areas for future research. User experiments need to be conducted in two areas. The first area is improving the user interface. We will design and perform an experiment that will focus on improving the ease with which the system is used. User feedback is needed to better understand how a user expects to view and manipulate email search results so that a more intuitive and effective interface can be implemented. The first step of this process has already been started by making the tool available to a group of researchers to use with their own email accounts. Our next step will be to design a formal experiment which will have subjects find specified information from a test data set containing a ground truth. The goal will be to observe the efficacy of finding information in an unfamiliar set of emails using InVEST. We are also in the process of arranging for professional investigators to use InVEST so that we can get their feedback on real world investigative email data sets which are almost impossible to obtain for research purposes.

Research needs to be done to improve the performance of term ranking. TF-IDF and related algorithms for term ranking can be improved when working with short texts forms. Emails and other short text forms of communication such as text messages and tweets present a unique challenge due to the nature of those types of communication. Not only are they short and numerous, but also they often contain incomplete information which may rely on previous messages that may not be available. We need to explore different approaches that combine TF-IDF with more information either by combining information from threads or from users. Since most if not all investigative searches start with some prior information, one approach to solving the ranking problem might be to develop a probabilistic method that takes advantage of these priors as part of the frequency calculations.

Another area for improvement associated with term ranking is the scope of an email set being used. Should the scope include an entire email set or just the set of emails in the current search results? There are situations where each scope is appropriate. However, in order to efficiently rank terms for a subset of emails in the results an index needs to be created for the subset. We are not aware of any indexing and search systems that have implemented an efficient method for creating a new index for a subset of documents from the index created for a full document set. An elegant solution to this problem may not be possible since it has been examined in the past by the web search community, however, a memory intensive brute force solution might be possible and acceptable for the smaller size data sets created by InVEST search results.

Conclusion

In this paper we have introduced a visual analytic methodology to aid in the search and triage of large email data sets. InVEST makes it easier for investigators to identify anomalies and hidden keyword relationships within email data sets. This helps to both speed up and improve the results of the investigative process. Our interactive visual pipeline allows the investigator to find relevant emails, entities and correspondents in the email data set through an exploratory process. Once these emails are found and isolated, all related emails that were hidden from the initial search through our search result reduction techniques can be brought back into the results set for a final detailed examination. Our case studies have demonstrated that this visual interaction can be effectively used to reduce the size of a results set so that the remaining emails can then be examined in greater detail to find those pertinent to the investigation.

References

Beebe NL, Clark JG, Dietrich GB, Ko MS, Ko D. Post-retrieval search hit clustering to improve information retrieval effectiveness: two digital forensics case studies. Decis Support Syst 2011;51(4):732–44.

Beebe NL, Liu L. Clustering digital forensic string search output. Digit Investig 12, 2014;11(4):314–22.

Diesner J, Carley KM. Exploration of communication networks from the enron email corpus. In: SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA. Citeseer; 2005.

Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL; 2005. p. 363–70.

Haggerty J, Haggerty S, Taylor M. Forensic triage of email network narratives through visualisation. Info Mngmnt Comp Secur 10, 2014; 22(4):358–70.

Haggerty J, Karran AJ, Lamb DJ, Taylor M. A framework for the forensic investigation of unstructured email relationship data. Int J Digital Crime Forensics 2011;3(3):1–18.

Joorabchi ME, Yim J-DD, Shaw CD. Emailtime: visual analytics of emails. IEEE Xplore; 10, 2010. p. 233–4.

Kang Y-a, Gorg C, Stasko J. How can visual analytics assist investigative analysis? design implications from an evaluation. Vis Computer Graph IEEE Trans 2011;17(5):570–83.

Keim D, Andrienko G, Fekete J-DD, Görg C, Kohlhammer J, Melançon G. Visual analytics: definition, process, and challenges. Springer; 2008.

Keim DA, Oelke D. Literature fingerprinting: a new method for visual literary analysis. IEEE Xplore; 10, 2007. p. 115–22.

Kerr B. Thread arcs: an email thread visualization. In: Information visualization, 2003. INFOVIS 2003. IEEE symposium on. IEEE; 2003. p. 211–8.

Kulkarni A, Pedersen T. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. In: IICAI; 2005. p. 703–22.

Li H, Shen D, Zhang B, Chen Z, Yang Q. Adding semantics to email clustering. In: Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE; 2006. p. 938–42.

Liu Z, Kihm J, Choo J, Park H, Stasko J. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. In: IEEE transactions on visualization and computer graphics X (Y); 2013.

Martin S, Nelson B, Sewani A, Chen K, Joseph AD. Analyzing behavioral features for email classification. In: CEAS; 2005.

Munzner T, Maguire E. Visualization analysis and design. AK Peters visualization series. Boca Raton, FL: CRC Press; 2015. URL, https://cds.cern.ch/record/2001992.

Shetty J, Adibi J. Discovering important nodes through graph entropy the case of enron email database. In: Proceedings of the 3rd International Workshop on link discovery. LinkKDD '05. ACM, New York, NY, USA; 2005. p. 74–81. URL, http://doi.acm.org/10.1145/1134271.1134282.

Yi JS, Kang Ya, Stasko JT, Jacko JA. Toward a deeper understanding of the role of interaction in information visualization. Vis Comput Graph IEEE Trans 11, 2007;13(6):1224–31.

Further reading

Bradel L, North C, House L, Leman S. Multi-model semantic interaction for text analytics. In: Visual analytics science and technology, 2014. VAST 2014. IEEE symposium on. IEEE; 2014. p. 163–72.

Keim DA, Ankerst M, Kriegel H-PP. Recursive pattern: a technique for visualizing very large amounts of data. In: Proceedings of the 6th Conference on Visualization'95. IEEE Computer Society; 1995. p. 279.

Viégas FB, Golder S, Donath J. Visualizing email content: portraying relationships from conversational histories. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2006. p. 979–88.