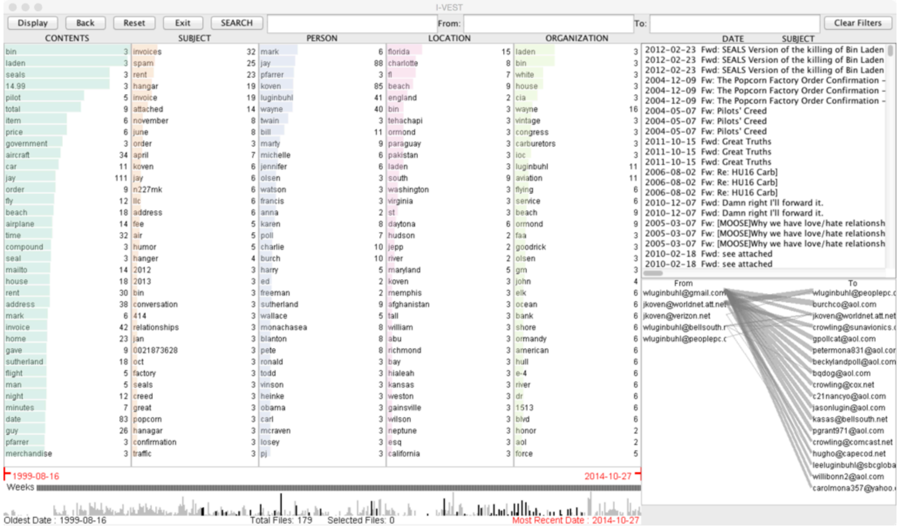
I-VEST: Intelligent Visual Email Search and Triage

Jay Koven, Enrico Bertini, Nasir Memon, R. Luke Dubois



CONTENTS		SUBJECT		PERSON		LOCATION		ORGANIZATION	
bin	3	invoices	32	mark	6	florida	15	laden	3
laden	3	spam	25	jay	88	charlotte	8	bin	3
seals	3	rent	23	pfarrer	3	fl	7	white	3
14.99	3	hangar	19	koven	85	beach	9	house	3
pilot	5	invoice	19	luginbuhl	41	england	2	cia	3
total	9	attached	14	wayne	40	bin	3	wayne	16
item	6	november	8	twain	3	tehachapi	3	vintage	3
price	6	june	8	bill	11	ormond	3	congress	3
government	3	order	3	marty	9	paraguay	3	carburetors	3
aircraft	34	april	7	michelle	6	pakistan	3	ioc	3
car	11	koven	6	jennifer	6	laden	3	luginbuhl	11
jay	111	jay	6	olsen	3	south	9	aviation	11
order	9	n227mk	6	watson	3	washington	3	flying	6
fly	12	llc	6	francis	3	virginia	3	service	6
beach	18	address	6	anna	2	st	3	beach	9
airplane	14	fee	5	karen	8	daytona	6	ormond	9
time	32	air	5	poll	7	hudson	2	faa	3
compound	3	humor	5	charlie	10	jepp	2	goodrick	3
seal	3	hanger	4	burch	10	river	2	olsen	3
mailto	14	2012	3	harry	5	maryland	5	gm	3
house	18	2013	3	ed	2	koven	3	john	4
rent	30	bin	3	freeman	2	memphis	3	elk	6
address	38	conversation	3	sutherland	9	afghanistan	3	ocean	6
mark	6	414	3	wallace	5	tall	3	bank	6
invoice	42	relationships	3	monachasea	8	william	3	shore	6
home	23	jan	3	blanton	8	abu	3	ormandy	6
gave	9	0021873628	3	pete	8	richmond	3	american	6
sutherland	18	oct	3	ronald	3	bay	3	hull	6
flight	5	factory	3	todd	3	hialeah	3	e-4	6
man	5	seals	3	vinson	3	kansas	3	river	6
night	12	creed	3	heinke	3	weston	3	dr	6
minutes	7	great	3	obama	3	gainsville	3	1513	6
date	83	popcorn	3	carl	3	wilson	3	blvd	6
guy	26	hanagar	3	mcraven	3	neptune	3	honor	2
pfarrer	3	confirmation	3	losey	3	esq	3	aol	2
merchandise	3	traffic	3	pj	3	california	3	force	5

DATE	SUBJECT
2012-02-23	Fwd: SEALS Version of the killing of Bin Laden
2012-02-23	Fwd: SEALS Version of the killing of Bin Laden
2012-02-23	Fwd: SEALS Version of the killing of Bin Laden
2004-12-09	Fw: The Popcorn Factory Order Confirmation -
2004-12-09	Fw: The Popcorn Factory Order Confirmation -
2004-12-09	Fw: The Popcorn Factory Order Confirmation -
2004-05-07	Fw: Pilots' Creed
2004-05-07	Fw: Pilots' Creed
2004-05-07	Fw: Pilots' Creed
2011-10-15	Fwd: Great Truths
2011-10-15	Fwd: Great Truths
2011-10-15	Fwd: Great Truths
2006-08-02	Fw: Re: HU16 Carb]
2006-08-02	Fw: Re: HU16 Carb]
2006-08-02	Fw: Re: HU16 Carb]
2010-12-07	Fwd: Damn right I'll forward it.
2010-12-07	Fwd: Damn right I'll forward it.
2005-03-07	Fw: [MOOSE]Why we have love/hate relationsh
2005-03-07	Fw: [MOOSE]Why we have love/hate relationsh
2005-03-07	Fw: [MOOSE]Why we have love/hate relationsh
2010-02-18	Fwd: see attached
2010-02-18	Fwd: see attached

From	To
wluginbuhl@gmail.com	wluginbuhl@peoplepc.c
jkoven@worldnet.att.net	burchco@aol.com
jkoven@verizon.net	jkoven@worldnet.att.net
wluginbuhl@bellsouth.r	crowling@sunavionics.c
wluginbuhl@peoplepc.c	gpollcat@aol.com
	petermona831@aol.cor
	beckylandpoll@aol.com
	bqdog@aol.com
	crowling@cox.net
	c21nancyo@aol.com
	jasonlugin@aol.com
	kasas@bellsouth.net
	pgrant971@aol.com
	crowling@comcast.net
	hugho@capecod.net
	leeluginbuhl@sbcgloba
	willibonn2@aol.com
	carolmona357@yahoo.c

Oldest Date : 1999-08-16 Total Files: 179 Selected Files: 0 Most Recent Date : 2014-10-27

Figure 1: I-VESTdisplay with selected node and reflected highlighted emails in list view after expansion on ”From” field

Abstract

Lawyers and investigators are often presented with large email datasets that contain emails that are not all relevant to any given investigative search. They must often manually comb through information contained within these large datasets in order to find the information they need, expending large amounts of time and money in the process. Our work offers an interactive visual analytic alternative to current methodology. We introduce a method for reducing the number of emails that need to be viewed in a large dataset while also giving the user a quick overview of possible contents and relationships in a set of results.

Keywords: Visual Analytics, Data Analytics, Investigative Analysis, Unstructured Data, Email Search.

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

1 Introduction

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 using keyword searches of both the contents and headers of each email. The investigators sort through results from these searches manually and read huge numbers of documents to find pertinent information. This process is frequently accomplished by hiring external companies who specialize in looking for the proverbial needle in this haystack.

I-VEST offers a method for reducing the number of emails that need to be viewed in a large dataset while also giving the user a quick overview of possible contents and relationships in a set of results. It does this by giving effective visual feedback. Our method is based on the visual presentation and analysis of important entities, keywords, people in an email and the possible relationships between these three. The level of importance in any given set of emails can be defined in different ways depending on the field. In some cases what is important can mean the most frequent. In other cases it could be determined using a TF-IDF type of algorithm. Our visual presentation is based on a graph where nodes represent collections of emails (answering the who, what or when questions mentioned above) and the edges of the graph represent relationships between the elements in the nodes. Nodes can represent senders, receivers, entities or keywords. The edges show all relationships between the various nodes. Nodes can quickly be moved, combined, deleted or selected to create subsets or expansions.

2 Related Work

In addition to displaying a graph of terms and relationships our system also displays lists of terms in each of the fields so that a user can easily identify the most common or important senders, receivers, entities and keywords within the results. The user can chose to expand the graph based on these terms and selected nodes or edges. This allows for the discovery of hidden relationships within data that might otherwise remain undiscovered. This iterative process of selecting and expanding is key to quickly finding the emails that are relevant to an investigation. This work was highly influenced by the Jigsaw Project at Georgia Tech by J. Stasko et al [4] [5]. Jigsaw focuses on supporting the investigative process by creating tools that help an analyst find and map relationships found in datasets. 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 the document set which,

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

according to its authors, is limited in size to several thousand documents. Since jigsaw does not focus on emails we observed that differentiating between information that is contained within the email header as opposed to information contained with in the body of the letter as well as finding relationships between the two is no easy task. We also reviewed work on email visualizations [2] [3] and email analysis [1]. However, we found that there is little previous work on using visualization for analysis of both the content and relationships in email datasets.

3 I-VEST Pipeline

Our system creates an interactive visual pipeline which allows a user to explore the content and relationships between emails contained within a target dataset. This pipeline allows an analyst to control the exploration of emails while being guided by intelligent suggestions from our system. Visualization guides the user toward information and relationships contained within the emails which may be important.

3.1 Creating the starting point

The process of analyzing the email dataset starts with a user defined keyword search or searches of one or more email fields. Currently these fields include the usual email header fields, extracted entity fields, and the body of the email. The email header fields include To, From, Cc and Date. Extracted entities include People, Locations, Organizations and Dates from the body of the emails. By using a search to begin the investigation we give the user a familiar starting point.

3.2 Displaying the results

Results from the searches are presented in a cumulative fashion in three separate views. The first view is a list of emails that shows dates and subject lines and is sorted by date. The contents of the emails can be viewed by selecting them from this list. This display can also be used to remove emails from the results.

The second view is a network graph display of nodes and edges. Each node in the graph represents emails that contain a keyword represented by the node label. The field that the keyword was found in is indicated by the color of the node and the size of the node indicates the number of emails represented. Each edge represents files that are in common between two nodes that are connected by the edge. The thickness of the edges represents the number of documents. There is a small timeline at the bottom of the graph view that shows the density of emails sent over a timespan between the oldest and newest emails contained in the nodes. 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 emails in the list and graph views are always synchronized. The selection highlights are also mirrored to clearly show the relationships between nodes, edges and emails. The third view is a list of important keywords in each field found in the emails represented in the first two views. The importance of keywords is determined either by a TF-IDF algorithm or the frequency within which they appear depending on the field that the keyword is in. This view of the keywords acts a quick guide to the searcher about the contents of all emails in the displays.

3.3 Filter and Expand

The three views can be used to explore and refine search results by filtering and expanding. The nodes and edges can be selected and highlighted in the graph view. The nodes can be filtered by removing, combining and subsetting of the selected results.

Emails selected in the list view can be used to filter the results by deleting the emails. When emails are deleted they will be removed from all nodes that contain them. When a node becomes empty it is removed from the display. The full headers and contents of emails can be examined in a separate window with the important keywords highlighted.

The keywords in the list view can be used as a shortcut for searches. When the user clicks on one of the keywords in the list view the system executes a search for the keyword in its field which will create a new node to represent the results.

Selected nodes, edges or emails can be used for expansion. Expansion is a way to let the system intelligently show the user other emails that are related to the selection by creating a new set of nodes by examining the contents of the current selection. The expansion process can both help the user better understand the relationship between information or entities in the selected emails as well as the overall collection of emails in the dataset.

While exploring a dataset 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 the search is successfully completed.

Our work has contributed two significant inroads in email analysis. First, we facilitated viewing the relationship between the content of emails and the social network of senders and receivers as a component of the search results. Second, we made it possible to separate out extracted entity fields and show the relationships between them and the other search results.

4 Discussion and Future Work

Using I-VEST to explore the Enron Trial dataset as well as our own personal email accounts has shown that searching for specific information using a limited starting search can be done quickly and successfully. In addition when using I-VEST we found unexpected, unusual and interesting relationships within the emails we looked at. These would have been extremely difficult to find using a normal email search. The model of displaying relationships between extracted entities, keywords and metadata provides a powerful base for searching and investigating large email datasets.

We have made I-VEST available to some academic users to allow them to explore their own Gmail accounts. We did so to incorporate the feedback we receive from them to improve the interface and make it easier to navigate through email search results. In addition we will use the feedback we obtain to improve the algorithms used to determine interesting or important keywords, entities and people in search results.

Our next step is to begin testing our tool with student investigators on publicly available datasets in order to observe the efficacy of finding information in an unfamiliar set of emails. In addition we would like to expose our tool to professional investigators and attorneys in order to determine the potential effectiveness of our techniques in actual ongoing email investigations. This work was supported in part by the NSF (under grant 0966187)

References

[1] R. Bekkerman. Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. 2004.

[2] M. E. Joorabchi, J.-D. . D. Yim, and C. D. Shaw. Emailtime: visual analytics and statistics for temporal email. In *IS&T/SPIE Electronic Imaging*.

[3] H. Kang, C. Plaisant, T. Elsayed, and D. W. Oard. Making sense of archived e-mail: Exploring the enron collection with netlens. *Journal of the American Society for Information Science and Technology*.

[4] Y.-a. . A. Kang, C. Gorg, and J. Stasko. How can visual analytics assist investigative analysis? design implications from an evaluation. *Visualization and Computer Graphics, IEEE Transactions on*, 17(5):570–583, 2011.

[5] Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, X(Y), 2013.