InVEST: Intelligent Visual Email Search and Triage

Jay Koven, Enrico Bertini, Luke Dubois and Nasir Memon

NYU TANDON SCHOOL Of ENGINEERING CSE Department

Presentation Outline

- Motivation
- Related Work
- New Methodology
- Initial Results
- Future Research

Motivation

Email account have grown drastically

- Free Gmail accounts are now 15GB
- Average Gmail account has >10,000 emails

Investigative datasets are even bigger

- Millions of emails

Users and Investigators are becoming overwhelmed

- Finding specific information requires persistence
- Finding a “Evidence” much worse

Motivation

Email search methods haven’t changed in 30 Years

- Grep / REGEX search still the core
- Long lists of results or worse no results

Machine learning is not the complete answer

- Emails are short and can be cryptic
- Techniques tend to work only on limited problems
- No clues to what was not found

Investigation Not Search!

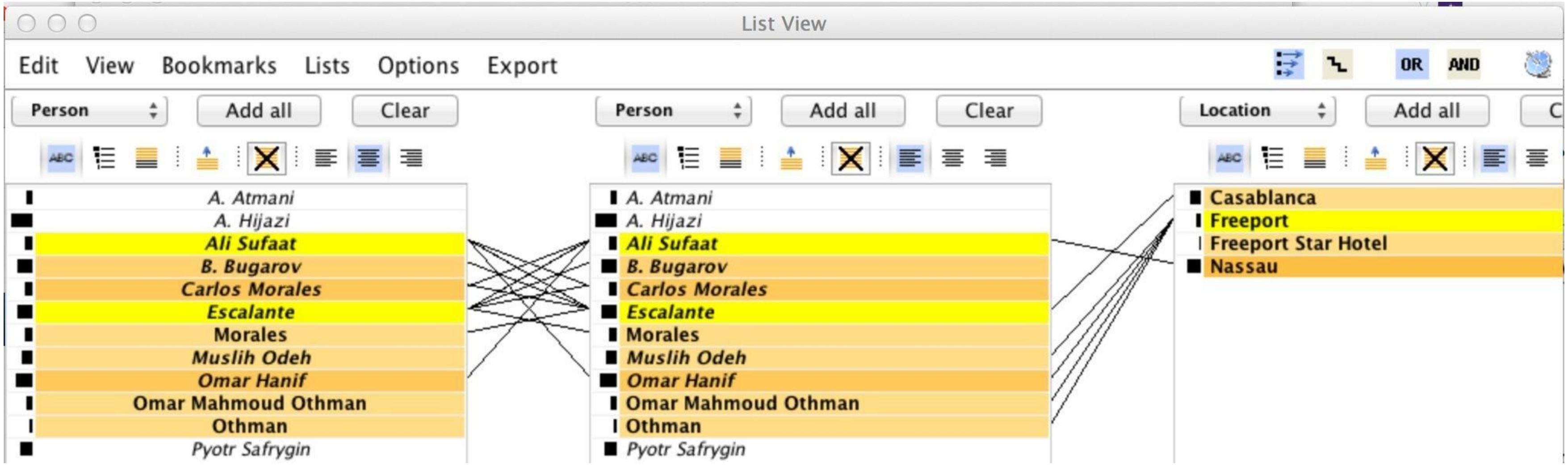
- Starts with some knowledge but incomplete
- Must find all emails related to investigation
  - Need to know more than just content
    - Fill in the “blanks”
      - Relationships between corresponders
      - Who or what were they talking about
      - When were they talking about it
      - Need to find hidden connections

Related Research

- Machine learning analysis of text and email data
  - Unsupervised (LDA)
  - Supervised (sLDA)
- Forensic Visualizations
  - Mostly work on static data (Tableau)
- Visual Analytics
  - Overview, Zoom, Filter (Wrong direction)

Related Work

Jigsaw - Stasko et al



What is Intelligent Visual Email Search and Triage

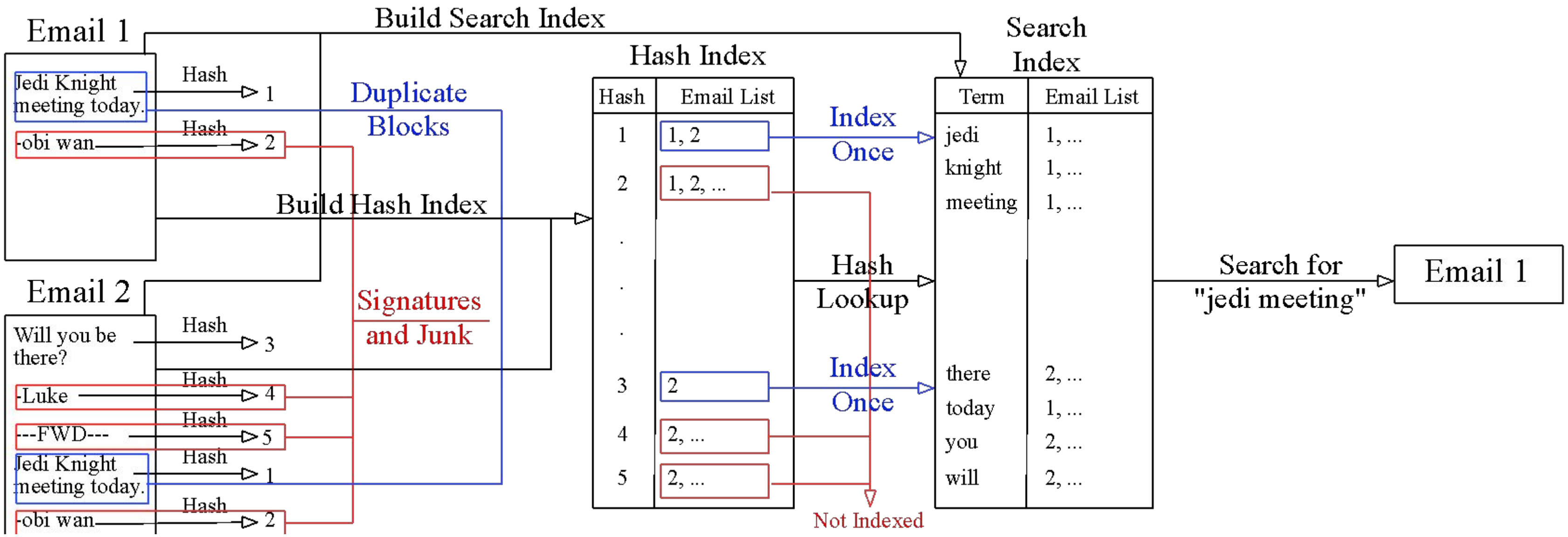
A Methodology to investigate email data sets

- Combine different types of information into clear results
  - Content (Subjects, Body Content and Entities)
  - Social Network (Corresponders)
  - Relationships between all of the above
- Find important entities and keywords
  - Ranking
  - Guidance

How?

- Create and effective visual analytic pipeline
  - Allow the analyst to explore the data interactively
  - Immediate feedback
  - Separate corresponders from Entities
  - Show relationships between keywords, entities and corresponders
  - Give intelligent guidance through ranking

Preprocessing: Index and De-Junk



Hash Index

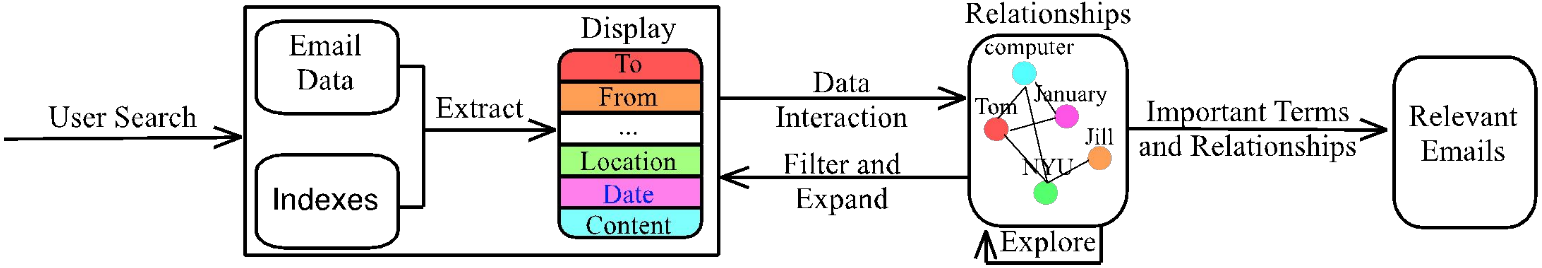
Hash	Email List
1	1, 2
2	1, 2, ...
.	
.	
.	
3	2
4	2, ...
5	2, ...

Search Index

Term	Email List
jedi	1, ...
knight	1, ...
meeting	1, ...
there	2, ...
today	1, ...
you	2, ...
will	2, ...

Visual Analytic Pipeline

- The starting point
- Filter and Expand
- Interacting with the results
- Cross Links between displays



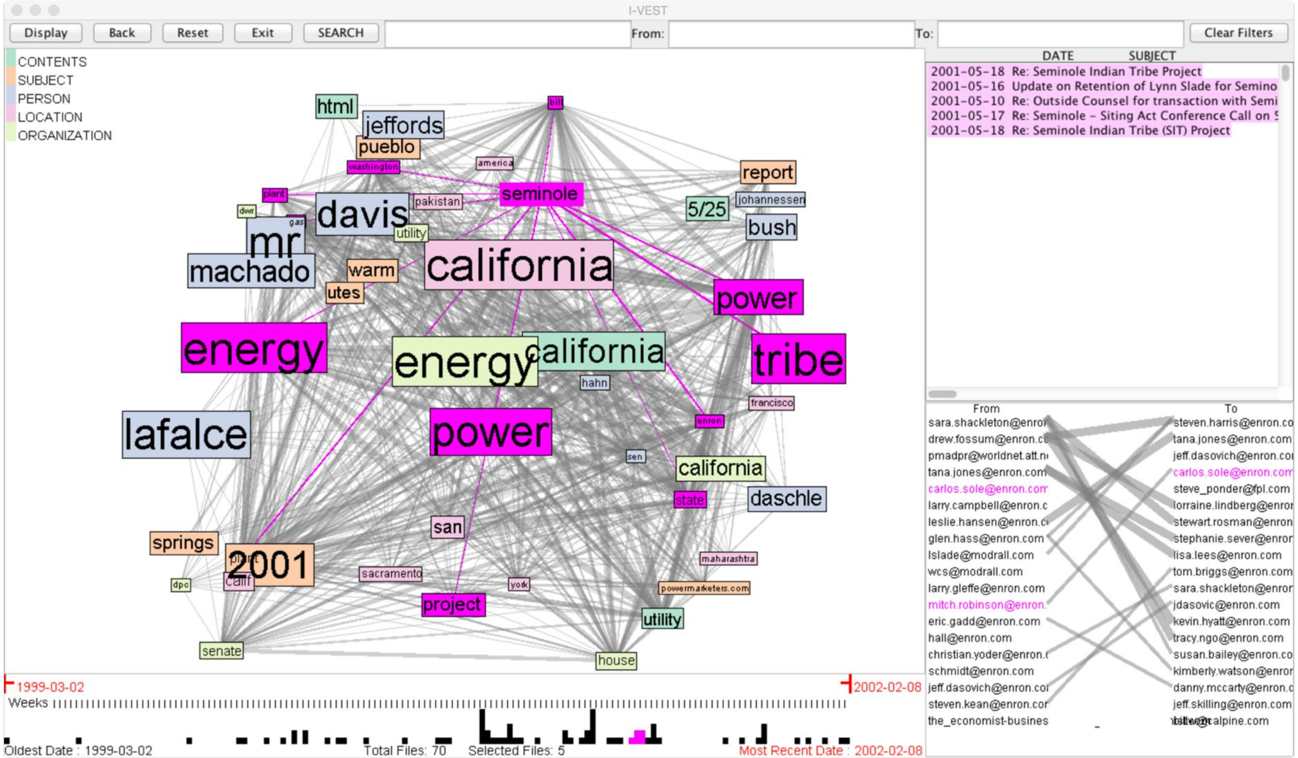
Entity Extraction and Separation

- Entity Integration
    - Why separate entities from corresponders?
    - Entity Extraction Methodology (Stanford NER)
    - Entity Identification (When is Tom also Thomas?)

Intelligent Guidance

- Ranking of terms (Finding “Important”)
  - For guiding analysts with list
  - Graph expansion
  - TF-IDF
    - Works (Sort of)
    - Problems
      - Should term frequency be measured by email or by returned set?
      - What makes a term important in Investigation?

InVEST Views



DATE	SUBJECT
2001-05-18	Re: Seminole Indian Tribe Project
2001-05-16	Update on Retention of Lynn Slade for Semino
2001-05-10	Re: Outside Counsel for transaction with Semi
2001-05-17	Re: Seminole – Siting Act Conference Call on 5
2001-05-18	Re: Seminole Indian Tribe (SIT) Project

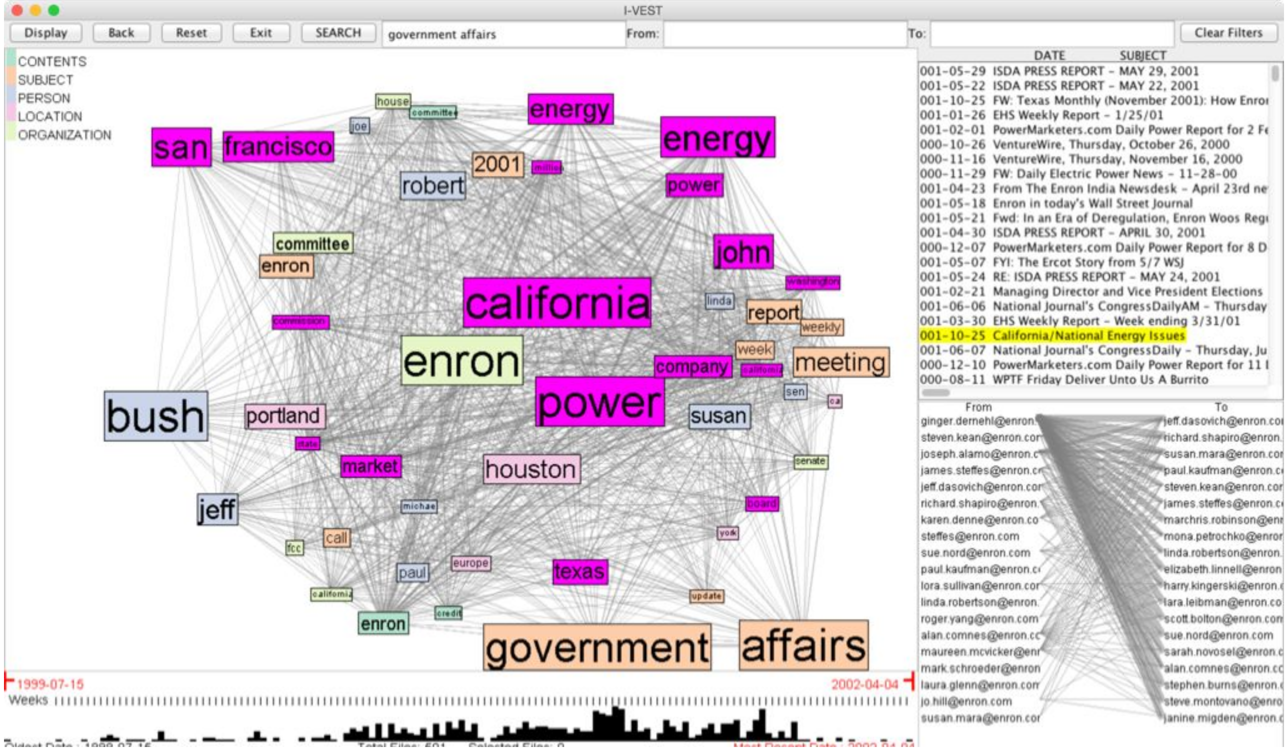
From
sara.shackleton@enron
drew.fossum@enron.cc
pmadpr@worldnet.att.ne
tana.jones@enron.com
carlos.sole@enron.com
larry.campbell@enron.c
leslie.hansen@enron.c
glen.hass@enron.com
lslade@modrall.com
wcs@modrall.com
larry.gleffe@enron.com
mitch.robinson@enron.
eric.gadd@enron.com
hall@enron.com
christian.yoder@enron.c
schmidt@enron.com
jeff.dasovich@enron.cor
steven.kean@enron.cor
the_economist-busines

To
steven.harris@enron.co
tana.jones@enron.com
jeff.dasovich@enron.cor
carlos.sole@enron.com
steve_ponder@fpl.com
lorraine.lindberg@enron
stewart.rosman@enron
stephanie.sever@enron
lisa.lees@enron.com
tom.briggs@enron.com
sara.shackleton@enron
jdasovic@enron.com
kevin.hyatt@enron.com
tracy.ngo@enron.com
susan.bailey@enron.co
kimberly.watson@enron
danny.mccarty@enron.c
jeff.skilling@enron.com
[illegible]@calpine.com

1999-03-02
2002-02-08
Weeks

Oldest Date : 1999-03-02
Total Files: 70
Selected Files: 5
Most Recent Date : 2002-02-08

InVEST Views



DATE	SUBJECT
001-05-29	ISDA PRESS REPORT - MAY 29, 2001
001-05-22	ISDA PRESS REPORT - MAY 22, 2001
001-10-25	FW: Texas Monthly (November 2001): How Enror
001-01-26	EHS Weekly Report - 1/25/01
001-02-01	PowerMarketers.com Daily Power Report for 2 Fe
000-10-26	VentureWire, Thursday, October 26, 2000
000-11-16	VentureWire, Thursday, November 16, 2000
000-11-29	FW: Daily Electric Power News - 11-28-00
001-04-23	From The Enron India Newsdesk - April 23rd ne
001-05-18	Enron in today's Wall Street Journal
001-05-21	Fwd: In an Era of Deregulation, Enron Woos Regu
001-04-30	ISDA PRESS REPORT - APRIL 30, 2001
000-12-07	PowerMarketers.com Daily Power Report for 8 D
001-05-07	FYI: The Ercot Story from 5/7 WSJ
001-05-24	RE: ISDA PRESS REPORT - MAY 24, 2001
001-02-21	Managing Director and Vice President Elections
001-06-06	National Journal's CongressDailyAM - Thursday
001-03-30	EHS Weekly Report - Week ending 3/31/01
001-10-25	California/National Energy Issues
001-06-07	National Journal's CongressDaily - Thursday, Ju
000-12-10	PowerMarketers.com Daily Power Report for 11 I
000-08-11	WPTF Friday Deliver Unto Us A Burrito

InVEST Future Research

- User experiments
  - HCI usability testing
  - Understand the users mental models
- Improve Ranking algorithms
  - Probabilistic approach using priors?
  - Interactively Driven Clustered (Faceted) Search
- User Defined Entities

Acknowledgements

- My advisors and co-authors
- Anonymous Reviewers
- My Shepard Timothy Leschke
- NSF