Visual Analytics for Investigative Analysis and Exploration of Documents and Data

John Stasko

Information Interfaces Research Group School of Interactive Computing Georgia Institute of Technology

Seoul National Univ.



Text is Everywhere

- We use documents as primary information artifact in our lives
- Our access to documents has grown tremendously in recent years due to networking infrastructure
 - WWW
 - Digital libraries
 - **–** ...





Who Cares?

- Analysts and investigators from a variety of domains work with documents
 - Intelligence analysis & law enforcement
 - Academia
 - Consumers
 - Fraud
 - Investigative reporters
 - Business analysts





Example Tasks & Goals

- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that are similar to this one (so they are worthwhile)?
- How are different words used in a document or a document collection?
- What are the main themes and ideas in a document or a collection?
- Which documents have an angry tone?
- How are certain words or themes distributed through a document?
- Identify "hidden" messages or stories in this document collection.
- Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
- Find connections between documents.





Question

 Can information visualization and visual analytics help with such tasks?





Challenge

- Text is nominal data
 - Does not seem to map to geometric/graphical presentation as easily as ordinal and quantitative data
 - Bar charts, line charts, scatterplots, etc.





A Little Tour

What has been done



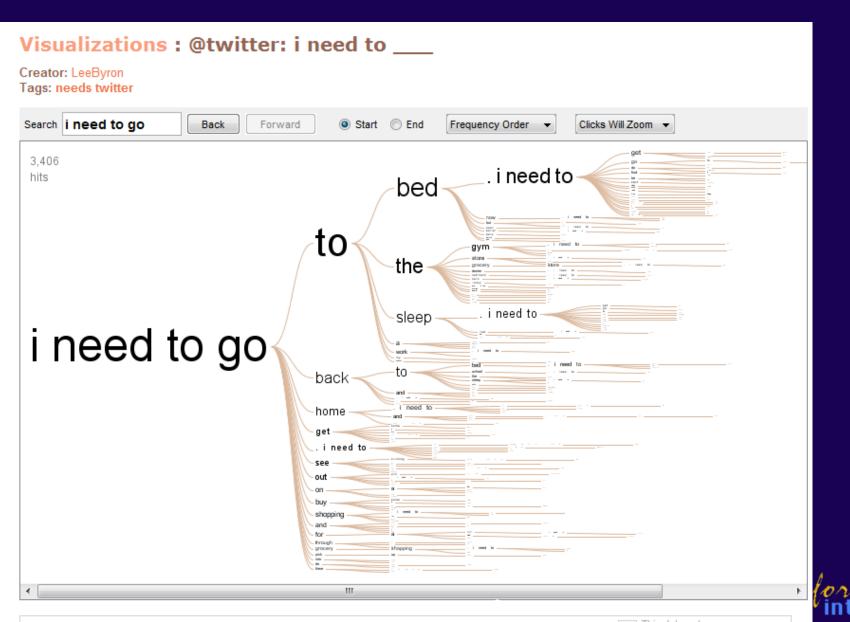


Wordle

Women's Rights Women have rights too! —macdoodle11 11 minutes ago awful 2 old

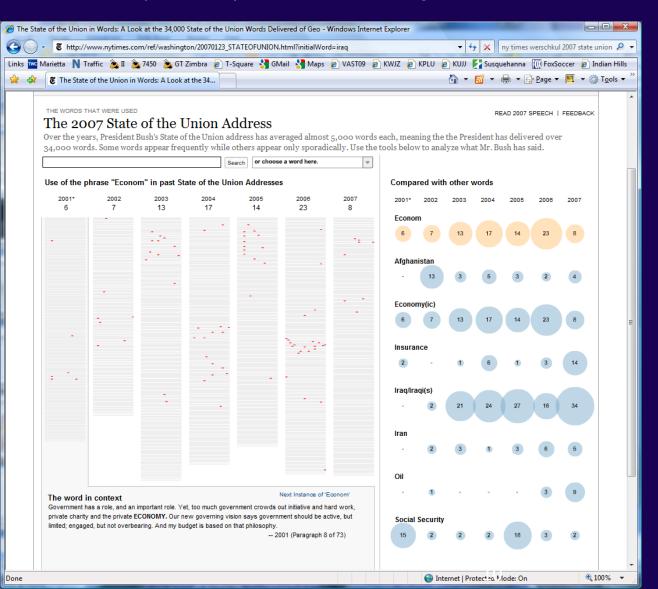


WordTree



State of the Union Addresses

http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?initialWord=iraq





Phrase Nets

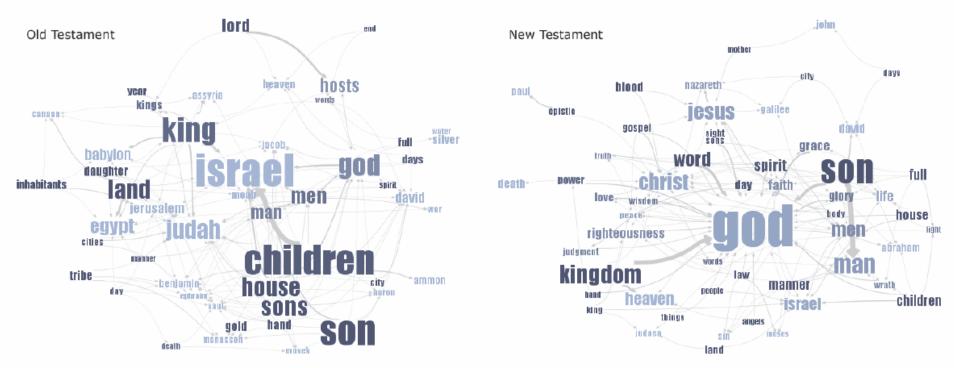


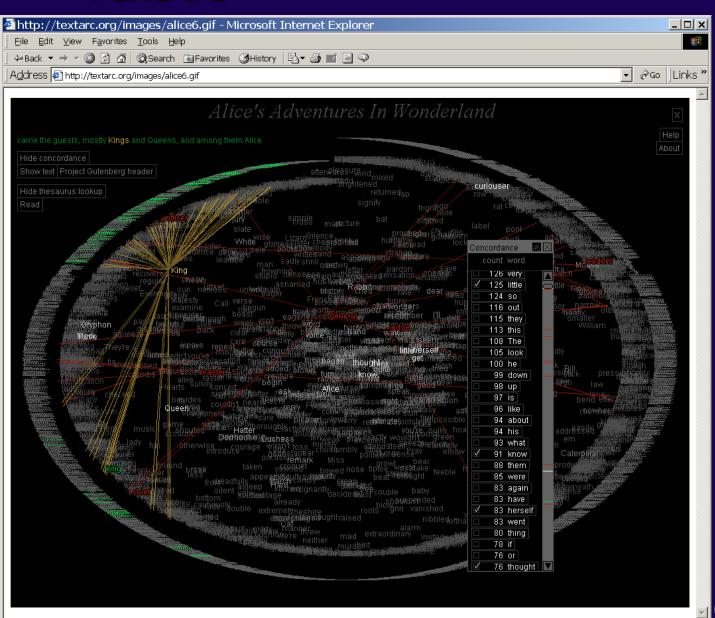
Fig 4. Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

X and Y X's Y X at Y X (is/are/was/were) Y





TextArc



Sentences laid out in order of appearance

Words near to where they appear

Significant interaction

Brad Paley

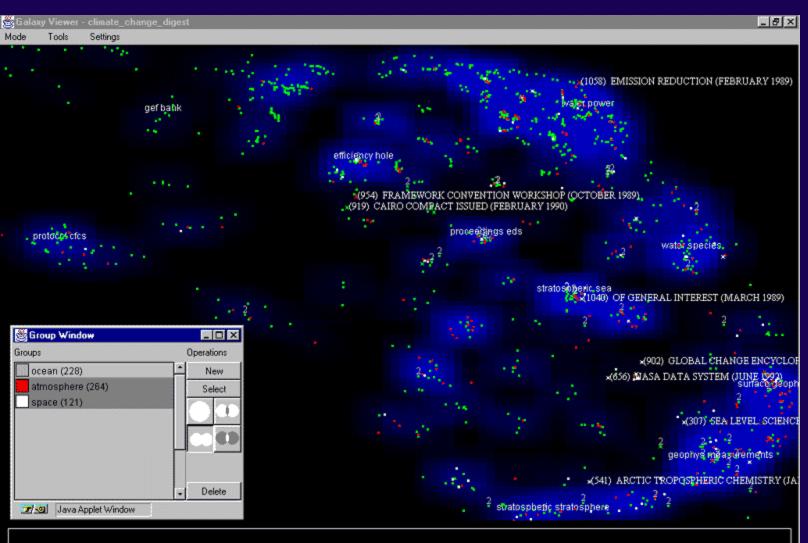


🚺 Internet



Hetzler & Turner IEEE CG&A '04

WebTheme



PNNL

Exploration Mode

Click on point to open abstract, Shift Click to view title, S to enter selection mode. Control Click to zoom in , Shift Control Click to zoom out, Click Drag to pan.

My History

What motivated the work





An Illuminating Exercise

THE INTELLIGENCE REPORTS AT HAND

Here are the folders you now have that contain the intelligence reports that will be of concern to you in your analysis. You are evaluating this intelligence information during the first two weeks of December, 2004. The near future means after 16 December, 2004 which will be the date of the final reports you have.

FBI FOLDER:

1) Report Date: 28 April, 2003 [From Canadian Security Intelligence Service]. This report concerns a bulletin issued on 2 April, 2003 by the Surgte' de l'Etat in Paris concerning a Moroccan named Abdillah Zinedine, alias: Abu Hafs [See CIA report for 17 April, 2003]. The Surgte' reported that Zinedine left Paris on a flight to Montreal, Canada on 1 April, 2003 and was traveling under a French passport in the name of Mehdi Rafiki. A Mehdi Rafiki did arrive in Montreal on 2 April, 2003 and said the purpose of his visit was to attend the funeral of an uncle. He listed his Montreal contact address as 175 Rue Durocher. This address belongs to an Irish pub operated by a man named Patrick O'Malley. O'Malley says he never heard of any person named Mehdi Rafiki. But a man using a French passport in the name Abu Hafs rented a car at the Canadauto Car Rental Agency in Montreal on 2 April, 2003. This rented car was never returned to this agency. It was discovered abandoned on Walden Ave near Schiller Park in Buffalo, NY on 6 April, 2003.

2) Report Date: 12 May, 2003 [From MI-5, UK], Riyag Yasser, a UK citizen, was arrested on 1 May, 2003 following an accident on the M4 Motorway near the Heston Service Area outside of London. Yasser has been an airtraffic controller at Heathrow Airport for the past six years. Two kilos of Semiles were found in the trunk of his car. A videocassette of a sermon given by Omar Mahmoud Othman, formerly a Salafi jihad preacher at the Baker St. Mosque in London, was found in Yasser's apartment at # 44, Northumberland Circle, East Bedfont, London. Also found in Yasser's apartment was a note containing several addresses in Canada, the USA, and in Nassau in the Bahamas. The addresses are: 721 St. Clare St., Montreal; 455 11th Street, Miami Beach, FL; 1712 Ferry Ave., Camden, NJ, and 11 Apple St. in Nassau. The Bahamas.

3) Report Date: 3 July, 2003 [Los Angeles Field office]. In response to the CIA report [29 June, 2003] regarding Abu Somad, 235 Buckthorn St, Inglewood CA, and Yazig Bafaba, 773 Flaxton St., Culver City, CA., investigators visited these two persons at their residences. Both Somad and Bafaba said they had never heard of persons named Riduan Sungkar or Omar Eyerts. Each of Somad and Bafaba was asked if they knew the other and they replied that they did not. However, it was later established that both Somad and Bafaba appeared together in a surveillance photo taken at the Callyational Bank in Culver City. Records of this bank indicated that Somad and Bafaba hold a joint account at

CIA FOLDER:

1) Report Date: 12 February, 2003 [From sources in Eqypt]. A Russian named lgor Kolokoy was arrested in Cairo on 29 January, 2003 and charged with assault on an Egyptian police officer who had attempted to arrest him for being drunk in public. Kolokoy sells medical supplies throughout the Middle East and represents a company in Moscow called Medikat. A background check on Kolokoy reveals that he was formerly an administrator at the Soviet institute Vector in Strizi, near Kiroy in Russia. When he was arrested, Kolokoy was carrying a card with a note on it reading [in Russian]: "Satrygin for H. Q., Peshawar".

2) Report Date: 15 February, 2003 [Transcript of names and information of interest taken from a captured computer hard drive in Afghanistan, just recently translated]. (a) David Loiseau, Canadian citizen from Toronto. Kanjak Training Camp, 2000-2001; wounded near Gizab, March 2002. (b) Raeed Beandali, American, Detroit MI., Mine warfare training, Al Khaldun, training camp, 2000-2001; wounded near Qalat, 26 November, 2002; sent to London 12 December, 2002. (c) Fahd Khadr, Canadian, Montreal. Light weapons training Al-Badr I training camp, 1999-2000. Sent home 23 October, 2000.

3) Report Date: 22 February, 2003, Surveillance report on Cesar Arze, whose residence is 77 Avenue Francis, Santo Domingo, Dominican Republic. Arze, who moved from Havana, Cuba to Santo Domingo in 1992, works as a medical technician in Santo Domingo. Arze is under surveillance because of information that he is associated with Cuban intelligence services. Arze was photographed in company with a man identified as Hector Lopez in Bogota, Columbia on 23 January, 2003. Lopez, a known representative of FARC, has conducted narcotics distribution activities throughout South and Central America and the Caribbean.

A) Report Date: 17 March, 2003. [Report from a source in Calamar, Columbia]. This source reports observing, near Calamar on 3 March, 2003, three men he knew were Cubans discussing the transfer of cocaine with members of the National Liberation Army [ELN] in Columbia. The source recognized one of these Cubans as Jose Escalante. This source says that he knew Escalante when they were students in 1991 at a school for medical laboratory technicians in Havana. This source said he believes that Escalante works in some capacity for the Cuban military. Our source also stated that Cuban representatives frequently visit in the tri-border area of Columbia, Ecuador and Peru to arrange for the exchange of weapons of various sorts needed by ELN for cocaine, that the ELN supplies to the Cubans who sell and distribute this narcotic throughout the Caribbean and Central American region.

4





Jigsaw

Visualization for Investigative Analysis across Document Collections

- Law enforcement & intelligence community
- Fraud (finance, accounting, banking)
- Academic research
- Journalism & reporting
- Consumer research

"Putting the pieces together"







The Jigsaw Team

Current:

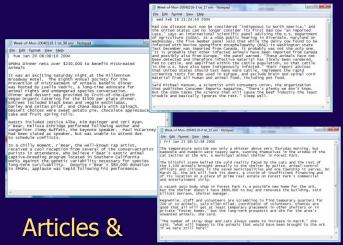
Carsten Görg Zhicheng Liu Youn-ah Kang Chad Stolper and many alumni

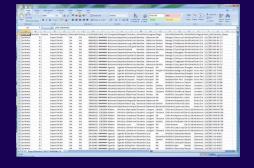




Problem Addressed

Help "investigators" explore, analyze and understand large document collections





Spreadsheets

reports





Blogs



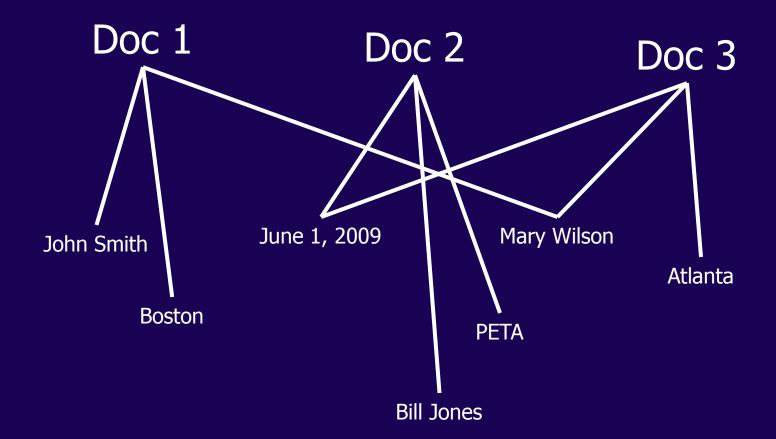


Our Focus

- Entities within the documents
 - Person, place, organization, phone number, date, license plate, etc.
- Thesis: A story/narrative/plot/threat within the documents will involve a set of entities in coordination











Entity Identification

- Must identify and extract entities from plain text documents
 - Crucial for our work
- Not our main research focus We use tools from others





Sample Document

Report: 20040510-4_16

May 14 2004

VANCOUVER, British Columbia - A Canadian immigration panel is considering whether accused environmental saboteur Tre Arrow can apply for refugee status in Canada.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in Oregon, asked the Canadian authorities to remain in Canada as a political refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether Arrow is affiliated with a terrorist group, which would immediately disqualify him from receiving refugee status in Canada, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether Arrow is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.





Entities Identified

Source:

Date: May 14, 2004

<u>VANCOUVER</u>, <u>British Columbia</u> - A Canadian immigration panel is considering whether accused environmental <u>saboteur Tre **Arrow**</u> can apply for refugee status in <u>Canada</u>.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in Oregon, asked the Canadian authorities to remain in Canada as a political refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether **Arrow** is affiliated with a terrorist group, which would immediately disqualify him from receiving refugee status in <u>Canada</u>, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether Arrow is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.





Sample Document 2

Title: Proving Columbus was Wrong

Abstract: In this work, we show the world is really flat. To

do this, we build a bunch of ships. Then we...

PI: Amerigo Vespucci

Co-PI: Vasco de Gama, Ponce de Leon

Organization: Northwest Central Univ.

Amount: 123,456

Program Mgr: Ephraim Glinert

Division: IIS

ProgramElementCode: 2860





Entities Already Identified

Text

Title: Proving Columbus was Wrong

Abstract: In this work, we show the world is really flat. To

do this, we build a bunch of ships. Then we...

PI: Amerigo Vespucci

Co-PI: Vasco de Gama, Ponce de Leon

Organization: Northwest Central Univ.

Amount: 123,456

Program Mgr: Ephraim Glinert

Division: IIS

ProgramElementCode: 2860







Connections

- Entities relate/connect to each other to make a larger "story"
- Connection definition:
 - Two entities are connected if they appear in a document together
 - The more documents they appear in together, the stronger the connection





Jigsaw

- Computational analysis of document text
 - Entity identification, document similarity, clustering, summarization, sentiment
- Multiple visualizations (views) of documents, analysis results, entities and their connections
- Views are highly interactive and coordinated

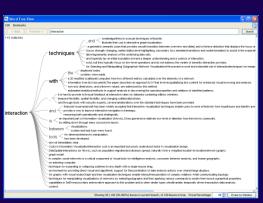


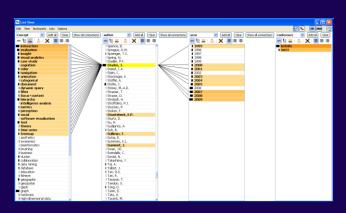


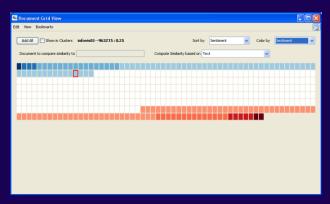
System Views



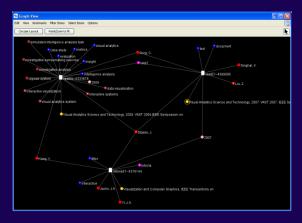


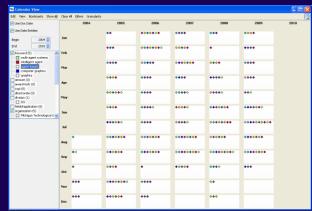


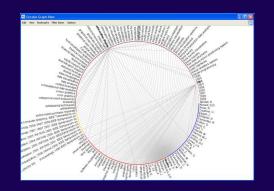




Document Cluster View					660
Edit View Bookmarks					
Makhaki Hawad Pan masks	mag, reduction, methods:	document, text, color: 26 ^f 15	etwork, used, structure: 28	3d, used, user: 30	interaction, treemacs, user: 50
Group by Filters Clear Filters	ditivariate, exploring, dimen	sicretestachors, state, using:	14design, collaboration, support sc	: 23 hemas, vanish, magnificat	abstract, viewing, uses: 25 ion: 8
Hide Unfiltered					
Clusters All Documents mag. reduction, methods: 15 document, tent, color: 26 rebrork, used, structure: 28 double, used, user: 5 methodoc, user: 50 methodoc, us	code, using, sources: 8	new, video, stories: 17	visual, tree, graphs: 35	set, trends, time	: 13 Visualang, Inshing, model 11
metaphors, state, using 14 design, conflavoration, support design, conflavoration, support design design, userial, respectation design, useria, respectation design, respectation de	analytic, dasign, system	ns: 45 quenying, exploring,	nterfaces: 36 visualizing, analytics, ins	aghtdamet, traffic, techn	daplay, graph, time: 25 lique: 14









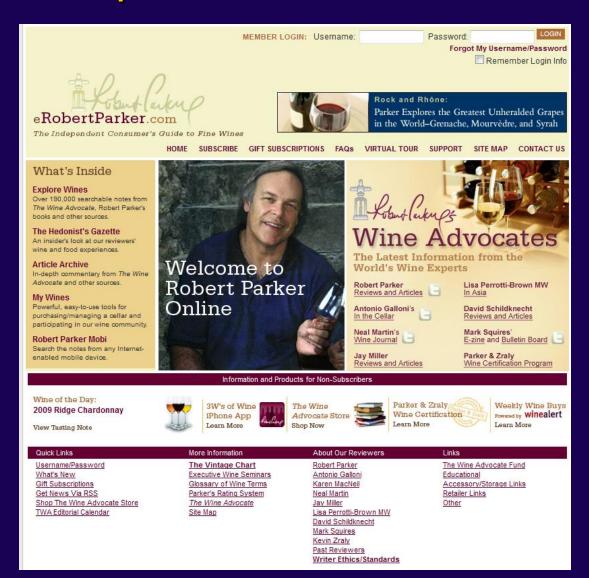


Pixels Help





Example Document Collection







One Review

```
<document>
  <docText>
 The 2007 Perlato del Bosco offers up mineral-infused blueberries,
 violets, tar and licorice. At first rather tense, the wine opens up
 nicely, with a gorgeous, muscular expression of fruit that is
 married with superb clarity and freshness throughout. Impeccably
 polished tannins inform the beautiful, long finish. Sweet scents of
 pipe tobacco add a sense of lift on the close. This is a marvelous
 effort from Tua Rita and a hugely over-achieving wine relative to
 the estate's top bottlings. Needless to say, the 2007 Perlato del
 Bosco is highly recommended. Anticipated maturity: 2012-2027.
 </docText>
 <Variety>Sangiovese</Variety>
 <ColorClass>Red</ColorClass>
 <Vintage>2007</Vintage>
 <Dryness>Dry</Dryness>
 <WineType>Table</WineType>
 <Country>Italy</Country>
 <Region>Tuscany</Region>
 <IsBarrelTasting>0</IsBarrelTasting>
 <Rating>94</Rating>
 <Source>Antonio Galloni</Source>
 <maturityShow>Young</maturityShow>
 <Producer>Tua Rita</Producer>
 <ProducerShow>Tua Rita
 <LabelName>Perlato del Bosco Vino da Tavola</LabelName>
</document>
```





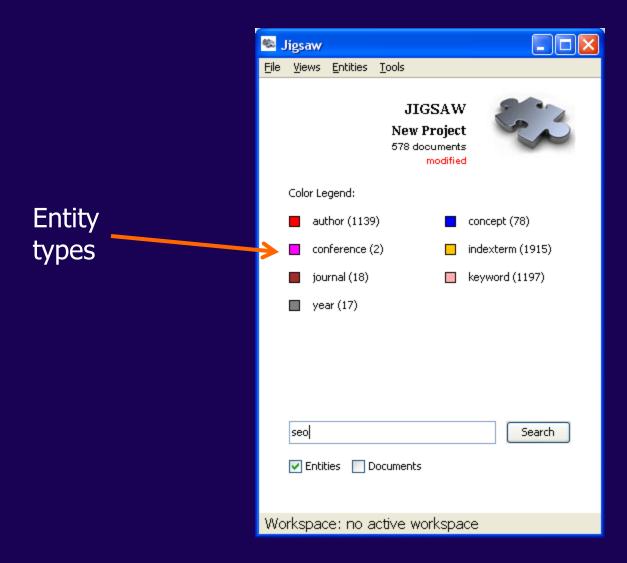
Demo

- Reviews of wines from Tuscany, '07-on
 - Text: review narrative
 - Entities: variety, producer, rating, vintage, color, location, producer, "descriptor", ...
- Descriptor (~ 9000)
 - eg: abrasive, oaky, cherry, mocha, textured
- 1132 reviews
 - From database of 150,000 reviews





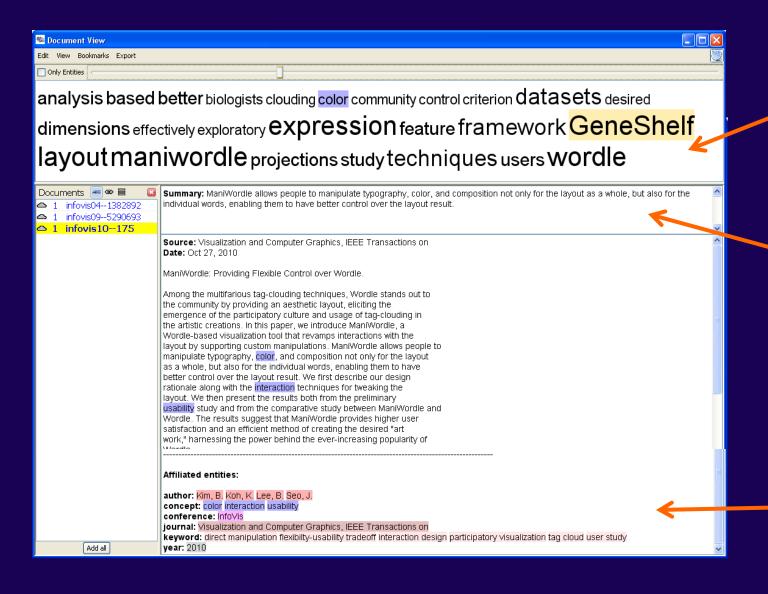
Console







Document View



Important words in loaded docs

Automatic summary

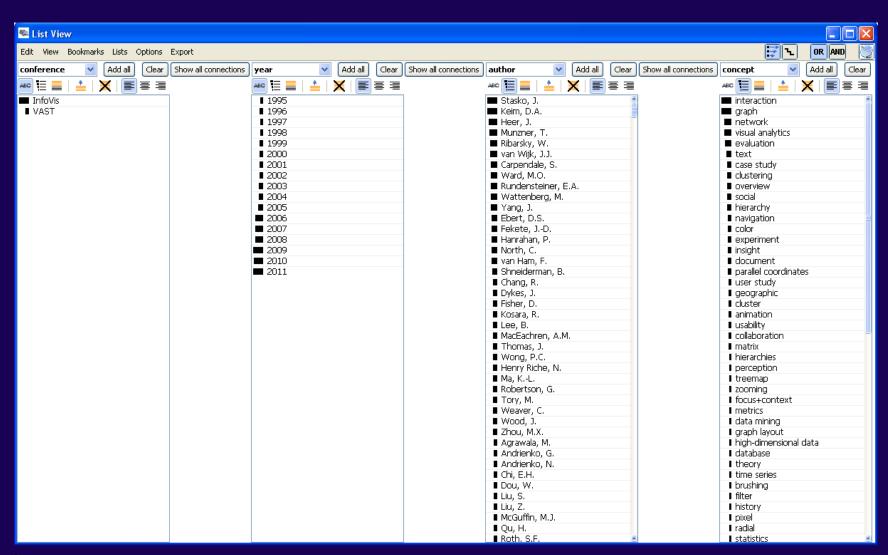
Entities identified





List View

Lists of entities by type Connections highlighted

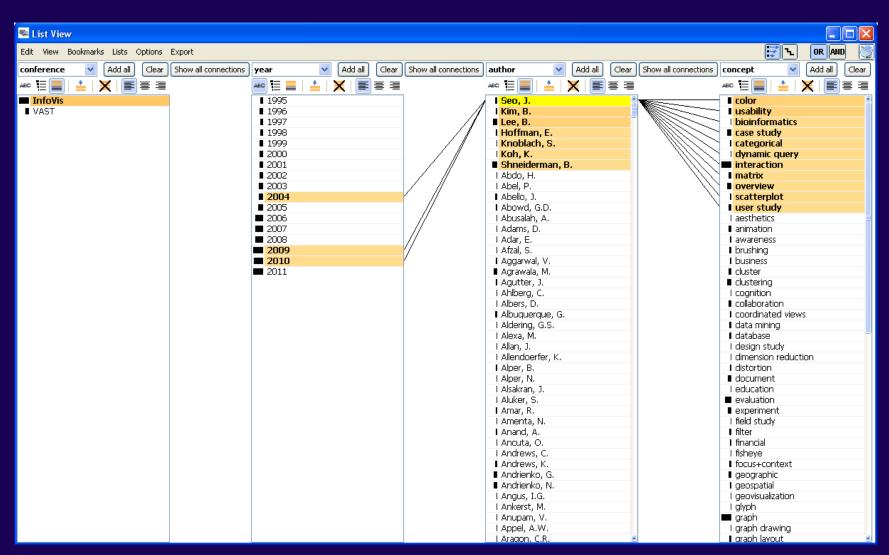






List View

Lists of entities by type Connections highlighted

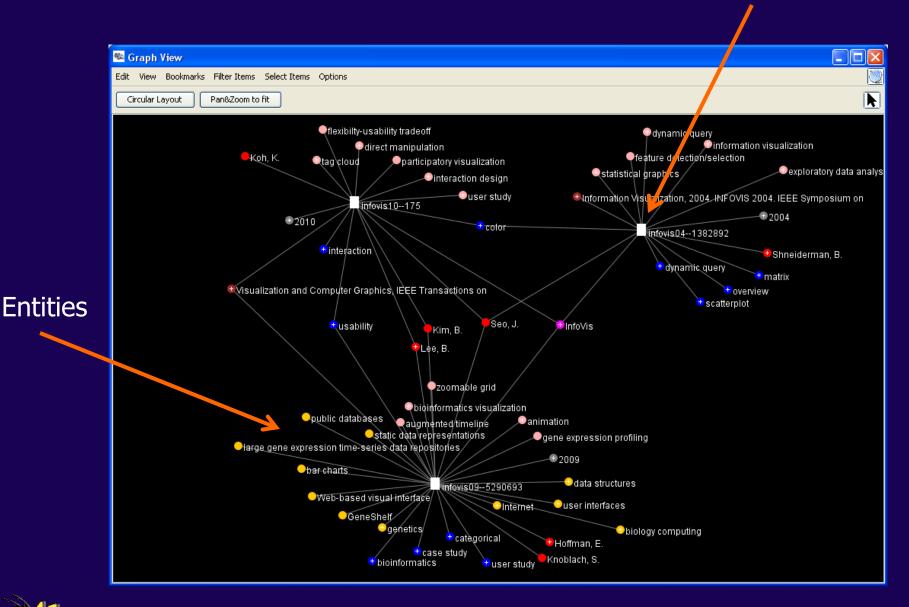






Graph View

Document



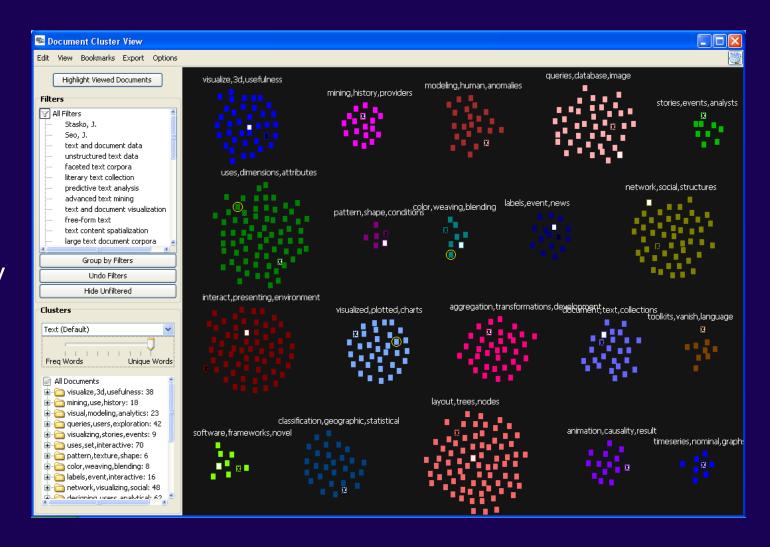




Document Cluster View

Clustered by document text or by entities

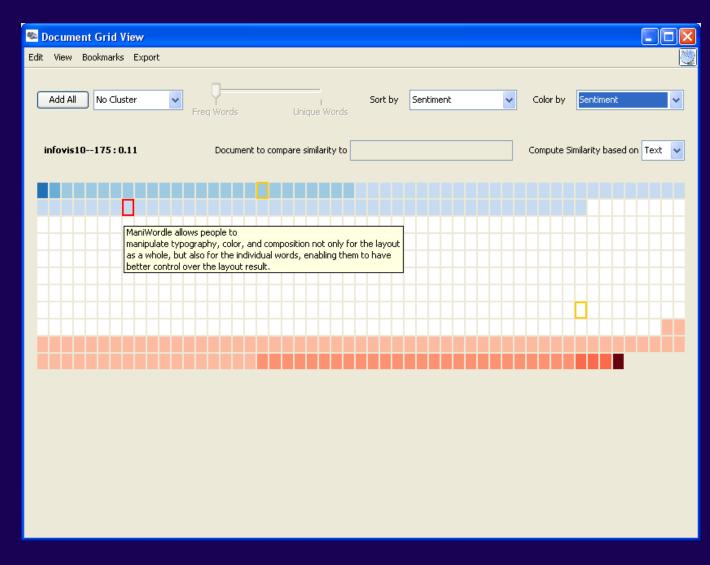
Summarized by three words







Document Grid View



User controls order and color

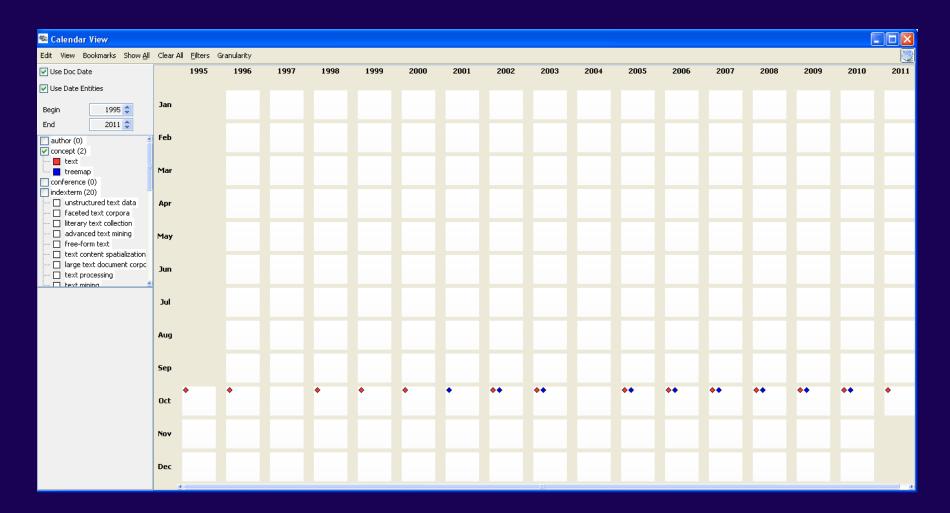
Sentiment analysis shown here





Calendar View

Showing connections between entities and dates

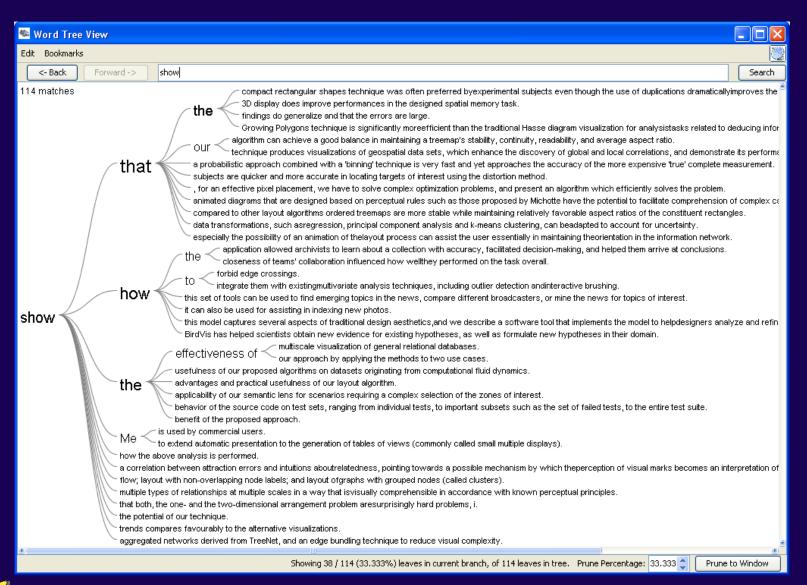






WordTree View

Context of a word in the collection

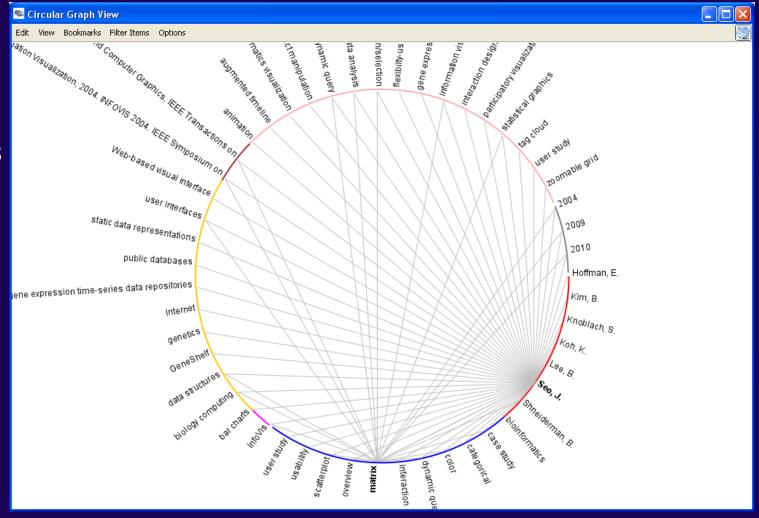






Circular Graph View

Connections between entities

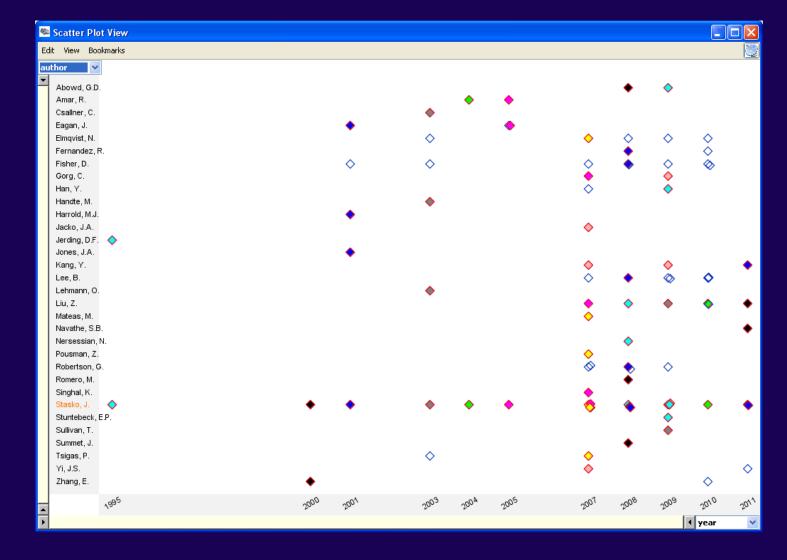






Scatterplot View

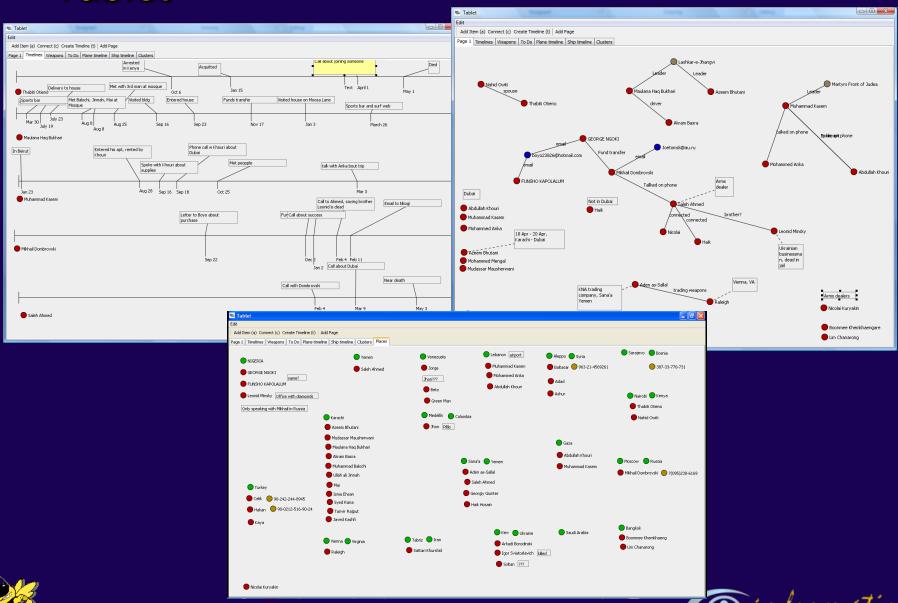
Documents containing pairs of entities







Tablet



- Preprocessing
 - Bag of words
 - Stop words, stemming
 - Disregard words/entities occurring less than three times
 - TFIDF





- Document similarity
 - Latent semantic analysis (LSA)
 - Groups semantically similar terms
 - 20% of original
 - Uses Singular Value Decomposition (SVD)
 - Cosine similarity





- Document clustering by content
 - Text or entities
 - K-means
 - 20 clusters
 - Initial seed docs chosen via dissimilarity
 - Summarized by 3 terms
 - Summarization algorithm
 - User can choose number of clusters & seeds





- Document summarization
 - Most "important" sentence
 - Mutual reinforcement learning algorithm
 - Document decomposed into terms and sentences
 - Weighted bipartite graph between the two, perform power iteration



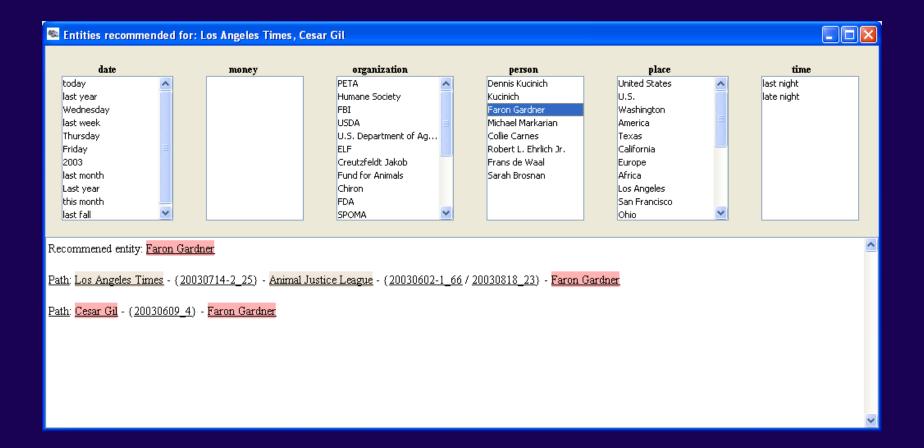


- Sentiment analysis
 - Lexicon-based approach
 - Initial set of +/- words, iterate and test
 - Allow lexicons to be augmented





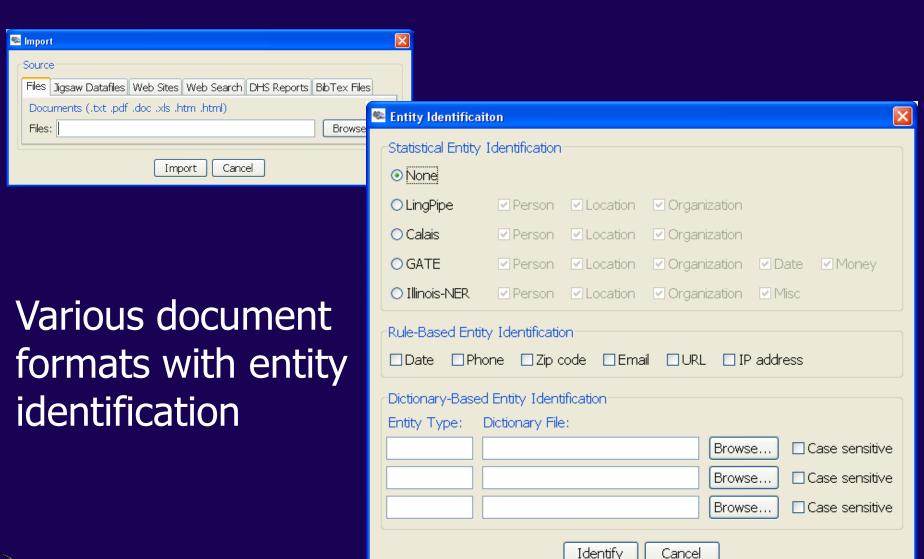
Recommend Related Entities







Document Import







Input Data Formats

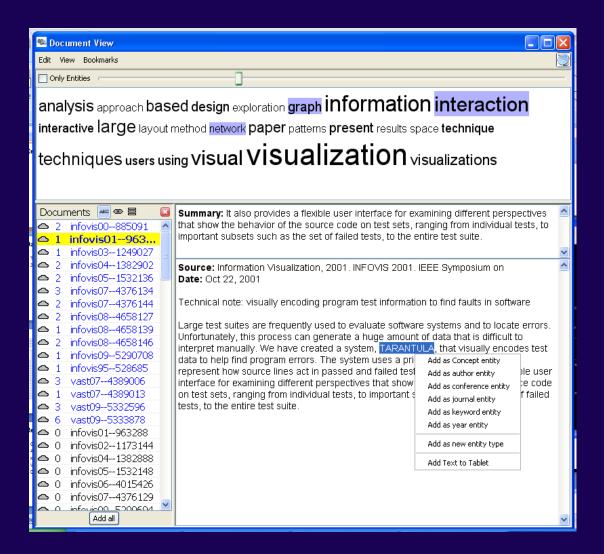
- Text, pdf, Word, html, Excel
- Jigsaw data file format
 - Our own xml

- DB?
 - Go to Excel
 - Go to text, transform to Jigsaw data file





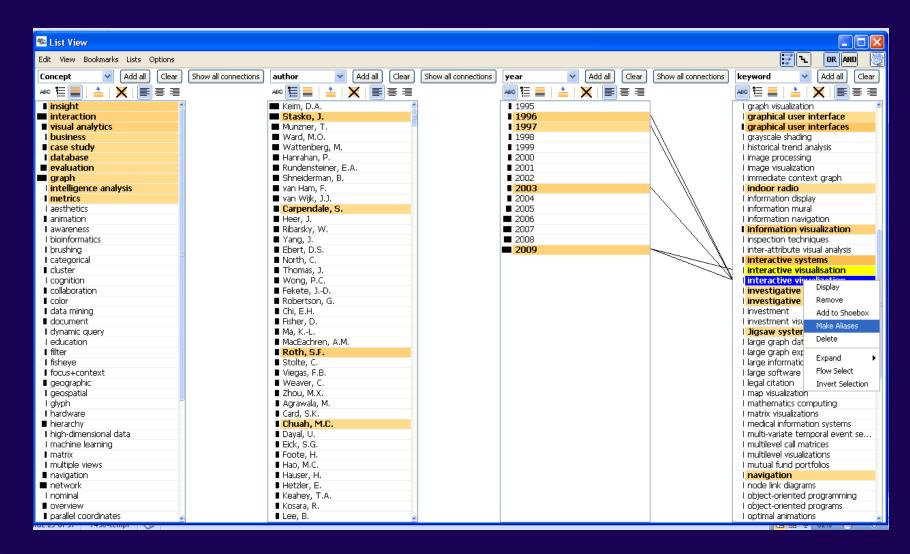
EI Correction







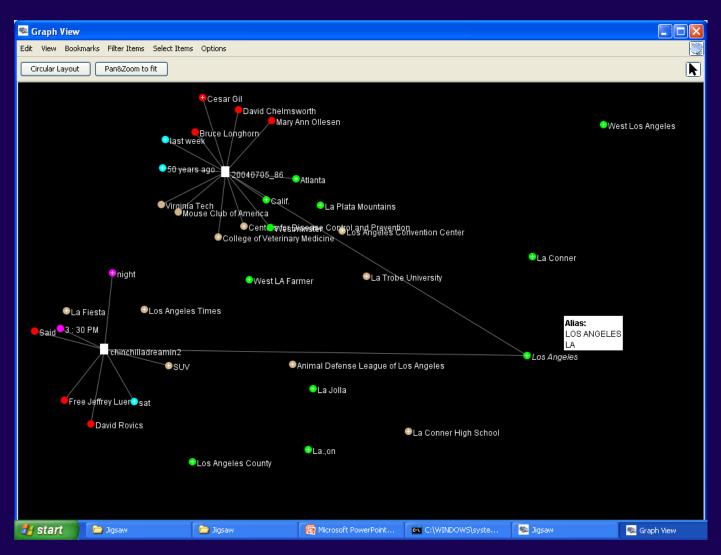
Entity Aliasing







Alias Representation







Application Domains

- Intelligence & law enforcement
 - Police cases
 - Won 2007 VAST Contest
 - Stasko et al, *Information Visualization* '08
- Academic papers, PubMed
 - All InfoVis & VAST papers
 - CHI papers
 - Görg et al, KES '10
- Investigative reporting
- Fraud
 - Finance, accounting, banking
- Grants
 - NSF CISE awards from 2000

- Topics on the web (medical condition)
 - Autism
- Consumer reviews
 - Amazon product reviews, edmunds.com, wine reviews
 - Görg et al, HCIR '10
- Business Intelligence
 - Patents, press releases, corporate agreements, ...
- Emails
 - White House logs
- Software
 - Source code repositories
 - Ruan et al, SoftVis `10





Potential Jigsaw Future Work

- Collaborative capabilities
- Improved evidence marshalling
- Present/browse investigation history
- Scalability upward
- Web document ingest
- Implement network algorithms
- DB import
- LDA

- Wikipedia & Intellipedia
- Geospatial view
- Better timeline capabilities
- Reliability/uncertainty
- Other types of data
- Active crawling/RSS ingest
- Try it on display wall
- Deployment to real clients





Room to Improve

- What Jigsaw doesn't do so well now
 - The end-part of the Pirolli-Card model
 - Helping the analyst take notes, organize evidence, generate hypotheses, etc. (The Tablet is a first step)
 - Sometimes called "evidence marshalling"

Others have focused more on that aspect...





Evaluation

 How does one evaluate the effectiveness of such a system?





Lab Study

- Objectives
 - How do people use such a system?
 - What system characteristics matter?

- Explore evaluation methods
 - Utility evaluation
 - What should we measure/observe?





Study Design

- Task and dataset
 - 50 simulated intelligence case reports
 - Each a few sentences long
 - 23 were relevant to plot
 - Identify the threat & describe it in 90 minutes

Source: doc017 **Date**: Oct 22, 2002

Abu H., who was released from custody after the September 11 incidents and whose fingerprints were found in the U-Haul truck rented by Arnold C. [see doc033] holds an Egyptian passport. He is now known to have spent six months in Afghanistan in the summer of 1999.





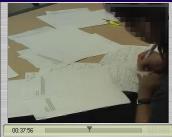
Study Design - Settings

1. Paper









2. Desktop









3. Entity









4. Jigsaw











Performance Measures

- Task sheets (like VAST Contest)
 - Three components (relevant people, events, locations)
 - +1 for correct items, -1 for a misidentified items
- Summary narrative
 - Subjective grading from 1 (low) to 7 (high)
- Two external raters
- Normalized, each part equal, mapped to 100-point scale





Results

	Paper				Desktop				Entity				Jigsaw			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Final Score	22.87	65.00	24.26	87.08	62.08	67.13	42.13	29.41	52.23	15.00	29.26	81.19	95.05	58.07	75.20	90.00
Performance	Fair	Very good	Fair	Excel- lent	Very good	Very good	Good	Fair	Good	Poor	Fair	Excel- lent	Excel- lent	Good	Very good	Excel- lent
Average Score	49.80				50.19				44.42				79.59			
Documents Viewed	50	50	50	50	50	50	50	50	49	31	45	50	31	50	46	23
# of Queries					19	18	48	8	23	61	59	91	44	4	26	8
First Query					40:49	19:55	2:47	12:41	1:31	0:29	0:59	3:12	0:18	5:35	25:37	4:18
Amount of Notes	Many	None	Many	Some	Many	Some	Few	Some	Some	None	None	Few	Some	Few	Few	Few
First Note Taking	0:07		0:05	0:16	1:53	19:57	2:47	8:20	2:37			3:14	0:48	0:32	5:15	78:45
First Task Sheet	43:20	32:53	70:13	3:25	61:35	20:26	7:33	64:11	28:09	0:52	2:55	7:20	48:26	41:48	43:00	5:33





Investigative Analysis Strategies

- 1. Overview, filter and detail (OFD)
- 2. Build from detail (BFD)
- 3. Hit the keyword (HTK)
- 4. Find a clue, follow the trail (FCFT)





Design Implications for IA Tools

- Support finding starting points/clues
- Guide the analyst to follow the right trail
- Support different strategies of SM process
- Support smooth transition between SM stages
- Provide a workspace
- Allow flexibility in organizing
- Support to find next steps when dead-end
- Facilitate further exploration





Case Studies

- Interviewed six analysts who have been using Jigsaw for 2-14 months
 - 3 intelligence analysts
 - 2 academic researchers
 - 1 business analyst
- Gain better understanding of benefits, limitations, utility
 - Inform next generation tools



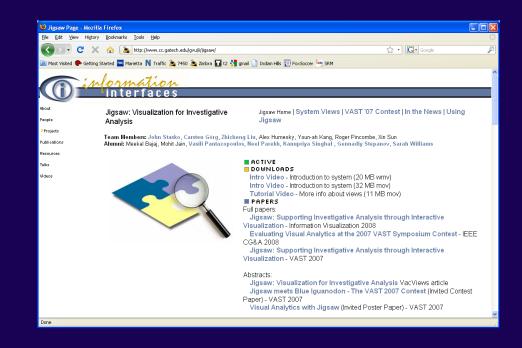


To Learn More about Jigsaw & Availability

http://www.gvu.gatech.edu/ii/jigsaw

Available for (free) trial use

Send email to: stasko@cc.gatech.edu







Acknowledgments

Work conducted as part of the Southeastern Regional
 Visualization and Analytics Center, supported by DHS and
 NVAC and the DHS Center of Excellence in Command, Control
 & Interoperability (VACCINE Center)









 Supported by NSF IIS-0414667, CCF-0808863 (FODAVA lead), NSF IIS-0915788







Thanks!

http://www.cc.gatech.edu/gvu/ii

