News and Blog Analysis with Lydia

Steven Skiena
Dept. of Computer Science
SUNY Stony Brook
http://www.cs.sunysb.edu/~skiena
Large-Scale News Analysis

- Our Lydia news analysis system does a daily analysis of over 1000+ online English and foreign-language newspapers, plus blogs, RSS feeds, and other news sources.
- We currently track over 1,000,000 news entities, providing spatial, temporal, relational and sentiment analysis.
- We believe our data and analysis should be of great interest in political science and related fields.
www.textmap.com
George W. Bush

PERSON
22898 references in 8326 articles [Show Articles]
[Web Query] [Popularity Time Series] [Textmap]

Complete
Name
Iraq
White House
Republican
Harriet Miers
Democrats
Bush
Cindy Sheehan
Karl Rove
Al Gore
Washington, DC
Social Security
Americans
John Roberts
Bill Clinton

Coref / Ref.

Also Known As
George W. Bush, George Bush

30 Days
Name
Bush
Democrats
Iraq
Republican
Congress
Washington, DC
Samuel Alito
Americans
Cindy Sheehan
U.S
United States
Social Security
Supreme Court
Clinton
Al Gore

Coref / Ref.

7 Days
Name
Bush
Democrats
Iraq
Congress
Republican
Washington, DC
Social Security
Americans
Samuel Alito
Jesse Samuel
United States
U.S
Bill Clinton
America
Clinton

Coref / Ref.

Popularity Time Series (30 Days) (What does popularity time series mean?)

[Graph showing popularity time series over 30 days]
Jack Abramoff

PERSON
17446 references in 3564 articles [Show Articles]
[Web Query] [Popularity Time Series]

Complete
Name
Tom DeLay
Bob Ney
Congress
Delay
Adam Kidan
Republican
Nev
Indian
Michael Scanlon
House
SunCruz
Washington, DC
John Dodds
House Administration Committee

Also Known As
Jack Abramoff, Abramoff, Lobbyist Jack Abramoff

Popularity Time Series (30 Days) (What does popularity time series mean?)
President Bush

PERSON
2439 references in 1341 articles [Show Articles]
[Web Query] [Popularity Time Series]

Complete
Name
Bush
Congress
Iraq
Americans
Vice President
United States
Jack Abramoff
State
Republican
Confederate States
Constitution
America
Washington, DC
U.S

Also Known As
President Bush, President

30 Days
Name
Bush
Congress
State
Jack Abramoff
Vice President
Iraq
Americans
Republican
Confederate States
United States
Union
America
Constitution
Washington, DC

7 Days
Name
Bush
Union
State
America
Iraq
Cindy Sheehan
United States
Congress
Republican
Social Security
Americans
Washington, DC
Ayman Al-Zawahri
Constitution

Popularity Time Series (30 Days) (What does popularity time series mean?)

Graph showing references per million over time.
Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions
System Architecture

**Spidering** – text is retrieved from a given site on a daily basis using semi-custom spidering agents.

**Normalization** – clean text is extracted with semi-custom parsers and formatted for our pipeline.

**Text Markup** – annotates parts of the source text for storage and analysis.

**Back Office Operations** – we aggregate entity frequency and relational data for a variety of statistical analyses.

Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. Lydia: A System for Large-Scale News Analysis. In *String Processing and Information Retrieval: 12th International Conference (SPIRE 2005).*
Text Markup

- We apply natural language processing (NLP) techniques to annotate interesting features of the document.
- Full parsing techniques are too slow to keep up with our volume of text, so we employ shallow parsing instead.
- We can currently markup approximately 2000 newspapers per day per CPU.
- Analysis phases include…
Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York.

She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president. Mr. Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.
Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York.

She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president. Mr. Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.
Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York.

She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president.

Mr. Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.
Dr./NNP Judith/NNP Rodin/NNP, the/DT former/JJ president/NN of/IN the/DT University/NNP of/IN Pennsylvania/NNP, will/MD become/VB president/NN of/IN the/DT Rockefeller/NNP Foundation/NN next/JJ year/NN, the/DT foundation/NN announced/VBD yesterday/RB in/IN New/NNP York/NNP. She/PRP will/MD take/VB over/in IN March/NNP 2005/CD, succeeding/VBG Gordon/NNP Conway/NNP, the/DT foundation/NN's/POS first/JJ non-American/JJ president/NN. Mr./NNP Conway/NNP announced/VBD last/JJ year/NN that/IN he/PRP would/MD retire/VB at/IN 66/CD in/IN December/NNP and/CC return/NN to/TO Britain/NNP, where/WRB his/PRP$ children/NNS and/CC grandchildren/NNS live/VBP.
Actor Classification

<p>Dr./NNP Judith/NNP Rodin/NNP, the/DT former/JJ president/NN of/IN the/DT University/NNP, will/MD become/VB president/NN of/IN the/DT Rockefeller/NNP Foundation/NN next/JJ year/NN, the/DT foundation/NN announced/VBD yesterday/RB in/IN New/NNP York/NNP. She/PRP will/MD take/VB over/IN in/IN March/NNP 2005/CD, succeeding/VBG Gordon/NNP Conway/NNP, the/DT foundation/NN's/POS first/JJ non-American/JJ president/NN. Mr./NNP Conway/NNP announced/VBD last/JJ year/NN that/IN he/PRP would/MD retire/VB at/IN December/NNP and/CC return/NN to/TO Britain/NNP, where/WRB his/PRP$ children/NNS and/CC grandchildren/NNS live/VBP.</p>
Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York. She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president. Mr. Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.
Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York.

She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president. Mr. Gordon Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.
Geography Normalization

<p>Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York City, New York, USA. She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president. Mr. Gordon Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.</p>
The most interesting analysis occurs after markup, using our MySQL database of all occurrences of interesting entities.

Each day’s worth of analysis yields about 10 million occurrences of about 1 million different entities, so efficiency matters...

Linkage of each occurrence to source and time facilitates a variety of interesting analysis.
Duplicate Article Elimination

Supreme Court Justice David Souter suffered minor injuries when a group of young men assaulted him as he jogged on a city street, a court spokeswoman and Metropolitan Police said Saturday.

Hashing techniques can efficiently identify duplicate and near-duplicate articles appearing in different news sources.
Synonym Sets

- JFK, John Kennedy, John F. Kennedy, and John Fitzgerald Kennedy all refer to the same person.

- We need a mechanism to link multiple entities that have slightly different names but refer to the same thing.

- We say that two actors belong in the same synonym set if:
  - Their names are morphologically compatible.
  - If the sets of entities that they are related to are similar.

Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions
Juxtaposition Analysis

- We want to compute the significance of the co-occurrences between two entities.
- Similar to collaborative filtering, determining which customers are most similar in order to predict future buying preferences.
- Just counting the number of co-occurrences causes the most popular entities to be related to everyone.
Time Series Analysis

Martin Luther King

Samuel Alito
Heatmaps

- Where are people talking about particular topics?

- Newspapers have a *sphere of influence based on*:
  - *Power of the source – circulation, website popularity*
  - *Population density of surrounding cities*

- The *heat* a given entity generates in a particular location is a function of the frequency it is mentioned in local sources

Donde Esta Mexico?
Who is running for president?

New Orleans – Animation

Monthly Spatial News Analysis

New Orleans, LA

11/01/2004 - 12/01/2006

http://www.textmap.com
Comparative Entity Maps

Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions
Blog Analysis with Lydia

• Blogs represent a different view of the world than newspapers.
  □ Less objective
  □ Greater diversity of topics

• We adapted Lydia to process *Livejournal* blogs, and compared blog content to that of newspapers.

In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 
Who Gets the Scoop?
Sentiment Analysis

Sentiment analysis lets us to measure how positively/negatively an entity is regarded, not just how much it is talked about.
Most Positive Actors in News and Blogs

- **News**: Felicity Huffman, Fernando Alonso, Dan Rather, Warren Buffett, Joe Paterno, Ray Charles, Bill Frist, Ben Wallace, John Negroponte, George Clooney, Alicia Keys, Roy Moore, Jay Leno, Roger Federer

- **Blogs**: Joe Paterno, Phil Mickelson, Tom Brokow, Sasha Cohen, Ted Stevens, Rafael Nadal, Felicity Huffman, Warren Buffett, Fernando Alonso, Chauncey Billups, Maria Sharapova, Earl Woods, Kasey Kahne, Tom Brady
Most Negative Actors in News and Blogs

**News:** Slobodan Milosevic, John Ashcroft, Zacarias Moussaoui, John Allen Muhammad, Lionel Tate, Charles Taylor, George Ryan, Al Sharpton, Peter Jennings, Saddam Hussein, Jose Padilla, Abdul Rahman, Adolf Hitler, Harriet Miers, King Gyanendra

**Blogs:** John Allen Muhammad, Sammy Sosa, George Ryan, Lionel Tate, Esteban Loaiza, Slobodan Milosevic, Charles Schumer, Scott Peterson, Zacarias Moussaoui, William Jefferson, King Gyanemdra, Ricky Williams, Ernie Fletcher, Edward Kennedy, John Gotti
How Do We Do it?

- We use large-scale statistical analysis instead of careful NLP of individual reviews.
- We expand small seed lists of +/- terms into large vocabularies using Wordnet and path-counting algorithms.
- We correct for modifiers and negation.
- Statistical methods turn these counts into indices.

Good to Bad in Three Hops

Paths of WordNet synonyms can lead to contradictory results, requiring careful path selection.
What Does it Mean?

- Our scores correlate very well with financial, political, and sporting events.
Who is the American Idol?
The low point is not September 2001 but April 2004, with the Madrid bombings and war in Iraq.
Social Network Analysis
Relationship Identification

- We use verb-frames and template-based methods to try to identify the nature of statistically-significant relationships, e.g.
  - devastated <Hurricane Katrina:Louisiana>
  - killed-in <Diana:Paris, FRA>
  - became <Joseph Ratzinger:Pope Benedict XVI>
  - not-watch <Dalai Lama:`` The Simpsons ``>
Description Extraction

- We use template-based methods and WordNet sense analysis to extract meaningful descriptions, such as:
  - Warren Buffett, billionaire investor
  - Giacomo, Kentucky Derby winner
  - Kim Jong Il, North Korean leader
Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions
Future Directions

- Entity-oriented (instead of document-based) search engines
- Foreign-language news analysis
- Event-focused relation extraction
- Financial modelling and analysis
- Social network analysis

We actively seek collaboration with social scientists
The Lydia Team
... and the Lydia Cluster