CONNECTING THE DOTS BETWEEN NEWS ARTICLES

Dafna Shahaf  Carlos Guestrin
Carnegie Mellon University  Carnegie Mellon University

Presenter:  Supervisors:
Monika Mitrevska  Maya Ramanath

Ralf Schenkel
The Idea
THE PROBLEM!!

- Information overload problem
- Easy to miss the big picture
- “Can’t Grasp Credit Crisis? Join the Club”
  - David Leonhardt for New York Times
FINANCIAL CRISIS AND ITS EFFECT ON THE HEALTH CARE REFORM

About 31,280,000 results (0.24 seconds)

Global Financial C
BBC News Global econo
Republic's financial ci
www.bbc.co.uk/news/s

Credit Crisis - New
News about the credit c
financial crisis has big
topics.nytimes.com/top

Financial Crisis N
Financial Crisis News
2319 Stories, most race
news.yahoo.com/topics

Newspaper Article
30 Sep 2008 ... Newsp:
financial crisis in THE El
www.paparticles.com

Global Financial C
The latest news, analy
... Featured articles 8 w
www.bbc.co.uk/global-f

Financial Crisis an
6 May 2010 ... Financi
Financial Crisis Org on
financialcrisis.org - Ci

How the Democrats
22 Sep 2008 ... The financi
the financial crisis of the past year has prono...
**Would be nice...**

<table>
<thead>
<tr>
<th>Date</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3.07</td>
<td>Home Prices Fall Just a Bit</td>
</tr>
<tr>
<td>3.4.07</td>
<td>Keeping Borrowers Afloat</td>
</tr>
<tr>
<td></td>
<td>(Increasing delinquent mortgages)</td>
</tr>
<tr>
<td>3.5.07</td>
<td>A Mortgage Crisis Begins to Spiral, ...</td>
</tr>
<tr>
<td>8.10.07</td>
<td>... Investors Grow Wary of Bank’s Reliance on Debt.</td>
</tr>
<tr>
<td></td>
<td>(Banks’ equity diminishes)</td>
</tr>
<tr>
<td>9.26.08</td>
<td>Markets Can’t Wait for Congress to Act</td>
</tr>
<tr>
<td>10.4.08</td>
<td>Bailout Plan Wins Approval</td>
</tr>
<tr>
<td>1.20.09</td>
<td>Obama’s Bailout Plan Moving Forward</td>
</tr>
<tr>
<td></td>
<td>(... and its effect on health benefits)</td>
</tr>
<tr>
<td>9.1.09</td>
<td>Do Bank Bailouts Hurt Obama on Health?</td>
</tr>
<tr>
<td></td>
<td>(Bailout handling can undermine health-care reform)</td>
</tr>
<tr>
<td>9.22.09</td>
<td>Yes to Health-Care Reform, but Is This the Right Plan?</td>
</tr>
</tbody>
</table>
**GOALS**

- Methods for automatically connecting the dots
  - Structured, easy way to uncover hidden connections between two pieces of information
- Given two news articles, the system automatically finds a coherent story
- Better understanding of the progression of the story
WHAT MAKES THE STORY GOOD?
But..

- Clinton’s alleged affair and the 2000 election Florida recount
- **s:** Talks Over Ex-Intern’s Testimony On Clinton Appear to Bog Down (Jan 1998)
- **t:** Contesting the Vote: The Overview; Gore asks Public For Patience (Nov 2000)
THE PROBLEM

- Locality of shortest path
- Articles related locally but no global coherent theme
WORD ACTIVATION PATTERNS
FORMALIZING STORY COHERENCE

- D – set of articles; W – set of features

\[
\text{Coherence} \ (d_1, \ldots, d_n) = \sum_{i=1}^{n-1} \sum_{w} 1(w \in d_i \cap d_{i+1})
\]

- The chain is only as strong as it’s weakest link

\[
\text{Coherence}(d_1, \ldots, d_n) = \min_{i=1 \ldots n-1} \sum_{w} 1(w \in d_i \cap d_{i+1})
\]
FORMALIZING STORY COHERENCE

- Considering only words from articles can be misleading
  - Lawyer and court => prosecution
- Some words are more important than others
COMBINING IMPORTANCE AND MISSING WORDS

\( \text{Influence}(d_i, d_j \mid w) \)

- Is high
  - If The two documents are highly connected
  - \( w \) is important for the connectivity

\[
\text{Coherence}(d_1, \ldots, d_n) = \min_{i=1\ldots n-1} \sum_w \text{Influence}(d_i, d_{i+1} \mid w)
\]
JITTERINESS

- Jittery activation patterns
  - Topics that appear and disappear throughout the chain
- Consider only the longest continuous stretch of each word
- Stretch – activation, not appearance

\[ \text{Coherence} \ (d_1, \ldots, d_n) = \max_{activation} \min_{s \ i=1 \ldots n-1} \sum_w \text{Influence} \ (d_i, d_{i+1} \mid w) l(w, \text{active in} \ ..d_i, d_{i+1}) \]
Scoring a chain: Linear Program Formulation

- **Smoothness**
  \[ \forall w \sum_i \text{word - init}_{w,i} \leq 1 \]
  \[ \forall w,i \text{ word - active}_{w,i} \leq \text{word - active}_{w,i-1} + \text{word - init}_{w,i} \]
  \[ \forall w \text{ word - active}_{w,0} = 0 \]

- **Activation restrictions**
  \[ \sum_{w,i} \text{word - init}_{w,i} \leq k_{Total} \]
  \[ \forall i \sum_w \text{word - active}_{w,i} \leq k_{Trans} \]

- **Objective**
  \[ \forall i \text{ minedge} \leq \sum_w \text{word - active}_{w,i} \ast \text{influence} (d_i, d_{i+1} | w) \]
  \[ \forall i, w \text{ word - active}_{w,i}, \text{word - init}_{w,i} \in [0,1] \]
Measuring influence without links

- Directed weighted graphs
  - Influence propagate through the edges
- Adding artificial edges
- No edges solution
  - Bipartite directed graph (word – document)
  - Edge weights - correlation
MEASURING INFLUENCE WITHOUT LINKS

- Intuitively: s and t are connected => short random walk starting from s reaches t frequently
- Stationary distribution for random walks

\[ \Pi_i(v) = \varepsilon \cdot 1_v(v = d_i) + (1 - \varepsilon) \sum_{(u,v) \in E} \Pi_i(u) P(v | u) \]

- w: sink node
- Stationary distribution with the new graph

\[ \Pi_i^w(d_j) \]

- The influence of d_j with respect of w:

\[ \Pi_i(d_j) - \Pi_i^w(d_j) \]
Figure 2: Word influence from an article about the OJ Simpson trial to two other documents – one about football and another about DNA evidence.
Finding a good chain

- Local search
  - Local optimum
- Optimize over words and chains
- LP problem
- All articles and edges as candidates for the chain
  - No transitions and articles known in advance
Finding a good chain

- Documents
  - node-active$_i$
  - next-node$_{i,j}$

- Words
  - word-active$_{w,i}$
  - transition-active$_{w,i,j}$

- Score of active edge

$$\sum_{w}^{transition-\text{active}_{w,i,j}} \cdot \text{Influence}(d_i, d_j \mid w)$$
Finding a Good Chain: LP

- **Objective**
  \[ \forall ij \quad \text{minedge} \leq 1 - \text{next - node}_{i,j} + \sum_{w} \text{transition - active}_{w,i,j} \ast \text{influence}(d_i, d_j | w) \]

- **Chain Restrictions**
  \[ \text{node - active}_1 = 1, \text{node - active}_n = 1 \]
  \[ \sum_i \text{node - active}_i = K, \sum_i \text{next - node}_{i,j} = K - 1 \]
  \[ \sum_i \text{next - node}_{i,j} = \text{node - active}_j \quad j \neq s \]
  \[ \sum_j \text{next - node}_{i,j} = \text{node - active}_i \quad i \neq t \]
  \[ \forall i \geq j \quad \text{next - node}_{i,j} = 0 \]
  \[ \forall i < j < k \quad \text{next - node}_{i,j} \leq 1 - \text{node - active}_k \]

- **Smoothness**
- **Activation Restriction**
ROUNDING

- LP defines fractional directed flow from s to t
- Start from s and iteratively pick the next node of the chain
- Current $d_i$, next is $d_j$ with probability
  \[
  \frac{\text{next} - \text{node}^*_{i,j}}{\sum_j \text{next} - \text{node}^*_{i,j}}
  \]
- Equivalent to decomposition of the flow into a collection of s-t paths $\{P_i\}$ and picking a path proportional to its weight
GUARANTEES

- **Claim**
  - The expected length of the path is $K$

- **Theorem:**
  - $V$ optimal value of LP
  - The lower bound of the rounded solution is $(1-c)V$ with probability at least $1-\square$
  - for $c=\square \frac{2}{v*\ln(n/\square)}$
**SCALING UP**

- LP has $O(|D|^2 * |W|)$ variables
- Not feasible for large number of articles
- Carefully and efficiently selected subset of documents
  - Documents similar to s and t
  - Use the bipartite graph, run random walks from s and t and pick the top-ranked articles
- If the chain is not strong enough => Iteratively add articles to the set
  - Articles from time period of the weakest part of the chain
Speeding up influence calculation

- $O(|D| |W|)$ calculations of stationary distributions to calculate the influence
- One set of random walks for all $w$
- For each document simulate random walk on the original graph
- Keep track of word-nodes encountered
- When calculating the influence take the same random walk, without using $w$
**Evaluation**

- Standard methods do not apply
  - No labeled dataset suitable for the task
- Methods evaluated by running them on real data
- Preprocessed half a million articles
- OJ Simpson trial; the impeachment of Clinton; the Enron scandal; September 11th;
- 500 – 10000 doc; name entities and noun phases;
- Users to evaluate
EVALUATION

- Connecting the dots
  - K: 6 or 7; kTotal: 14; kTrans: 4
  - 10 min for chain
- Shortest-path
  - Connect each document with the nearest neighbors
  - Cosine similarity
- Google News Timeline
  - GNT – organizes news search result
  - Query string input based on s and t
  - K equally – spaced documents
- Event threading
  - Finds sub – clusters in a news event and structure them
  - Creates a graph
  - Path from cluster including s to cluster including t
  - Pick representative documents from each cluster along the path
EVALUATION

- 18 users
- Measure **familiarity** in the beginning
- The users were asked to indicate:
  - Relevance
  - Coherence
  - Redundancy
- Effectiveness – fraction of the familiarity gap closed
- Simple stories – focus around same event
- Complex stories – connected through one or more events
RESULTS

Google News Timeline:
Osama bin Laden is denounced by his family // Osama Family’s Suspicious Site (Web designer from LA buys a bizarre piece of Internet history) // Are you ready to dance on Osama’s grave? (How should one react to the death of an enemy?) // Al-Qaeda behind Karachi blast // LIVE FROM AFGHANISTAN: Deadline of Death Delayed for American Journalist // Killed on Job But Spared ‘Hero’ Label (About Daniel Pearl)

Connect the Dots: Dispatches From a Day of Terror and Shock // Two Networks Get No Reply To Questions For bin Laden (Coverage of September 11th) // Opponents of the War Are Scarce on Television (Coverage of the war in Iraq and Afghanistan) // ‘Afghan Arabs’ Said to Lead Taliban’s Fight // Pakistan Ended Aid to Taliban Only Hesitantly // Pakistan Officials Arrest a Key Suspect in Pearl Kidnapping (Pearl abducted in Paksitan while investigating links to terror) // The Tragic Story of Daniel Pearl
INTERACTION MODELS

- What if the user does not find the resulting chain satisfactory?
- Usually – Users revise their queries
- More expressive form of interaction
- Types of user feedback
  - Refinement of a chain
  - Tailoring to user interests
REFINEMENT OF A CHAIN

- Mechanism to indicate areas for refinement
  - Adding new article
  - Replacing an article
- All possible replacement/insertion action
- Pick the best one

Simpson Defense Drops DNA Challenge
Issue of Racism Erupts in Simpson Trial
Ex-Detective’s Tapes Fan Racial Tensions in Los Angeles
Many Black Officers Say Bias Is Rampant in LA Police Force
With Tale of Racism and Error, Lawyers Seek Acquittal
* In the Joy Of Victory, Defense Team Is in Discord *
* (Defense lawyers argue about playing the race card) *
The Simpson Verdict
INCORPORATE USER INTERESTS

- Mechanism to focus the chains around “important” concepts
- Add importance weight to each word

\[ \sum_{w} \pi_{w} \text{Influence} \ (d_i, d_{i+1} \mid w) 1(w..activein ..d_i, d_{i+1}) \]

- Importance increases/decreases by multiplicative factor
- Word co-occurrence information
  - With DNA, blood and evidence increase too
**Main Contributions**

- Structured, easy way to navigate between topics
- Formalizing characteristics of a good story and the notion of coherence
- Formalizing influence with no link structure
- Connecting two fixed endpoints while maximizing chain coherence
- Incorporating feedback and interaction mechanisms into the system, tailoring stories to user preferences
- Evaluating the algorithm over real news data
PLACE FOR IMPROVEMENT

- Richer forms of input and output
- More complex task
- Roadmap: Set of Chains covering different aspects
- Behavior under different query characteristics
  - News articles with less coverage
**Discussion points**

- It is not clear how they find the words that are important but do not appear in the documents.
- It is not clear how they present the results (links, article titles or important parts from the articles).
- The quality of the result depends of the users choice of articles.
THANK YOU