Connecting the Dots Between News Articles

Dafna Shahaf
Carnegie Mellon University

Carlos Guestrin
Carnegie Mellon University

Presenter:
Monika Mitrevska

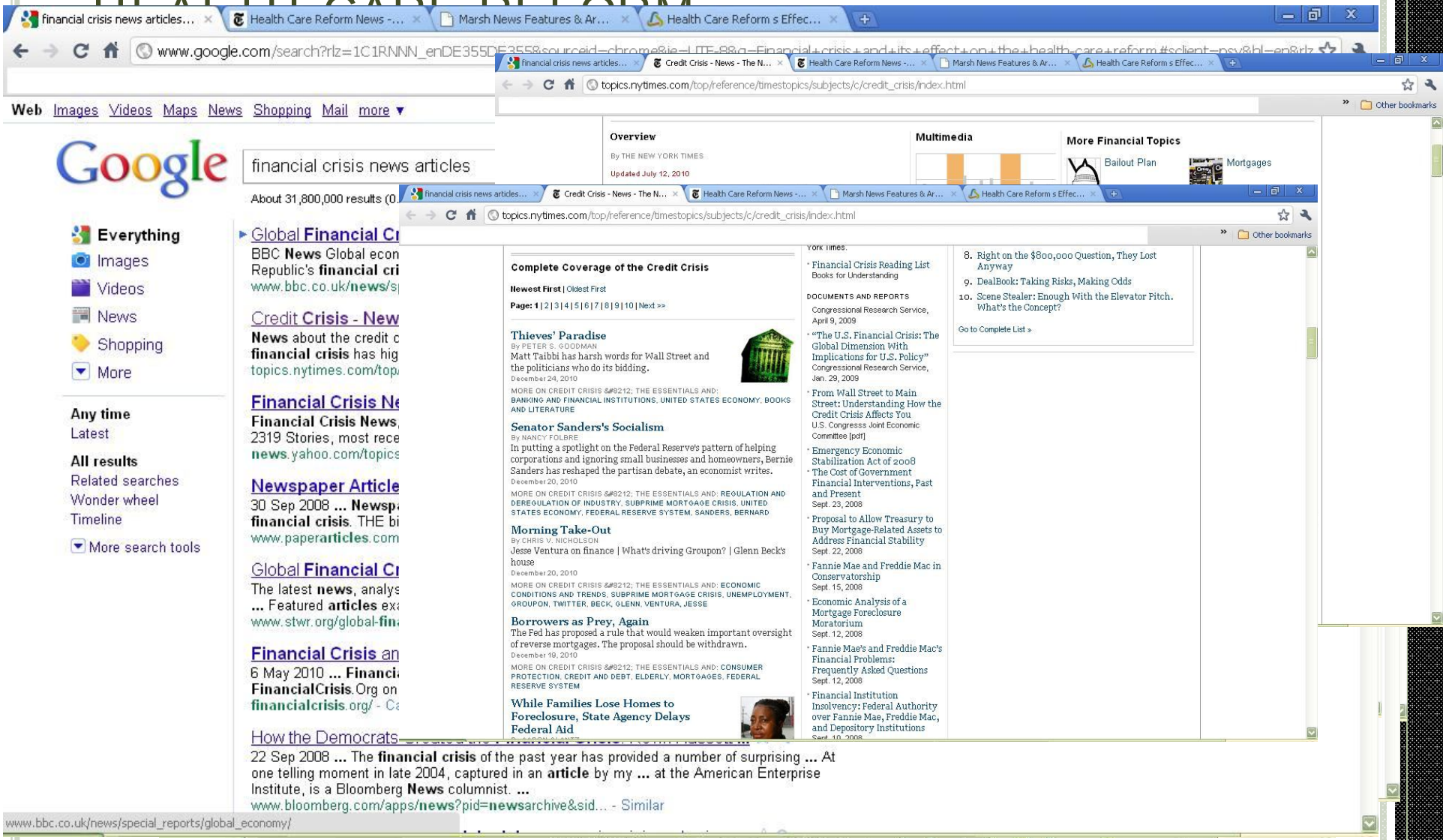
Supervisors:
Maya Ramanath
Ralf Schenkel

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



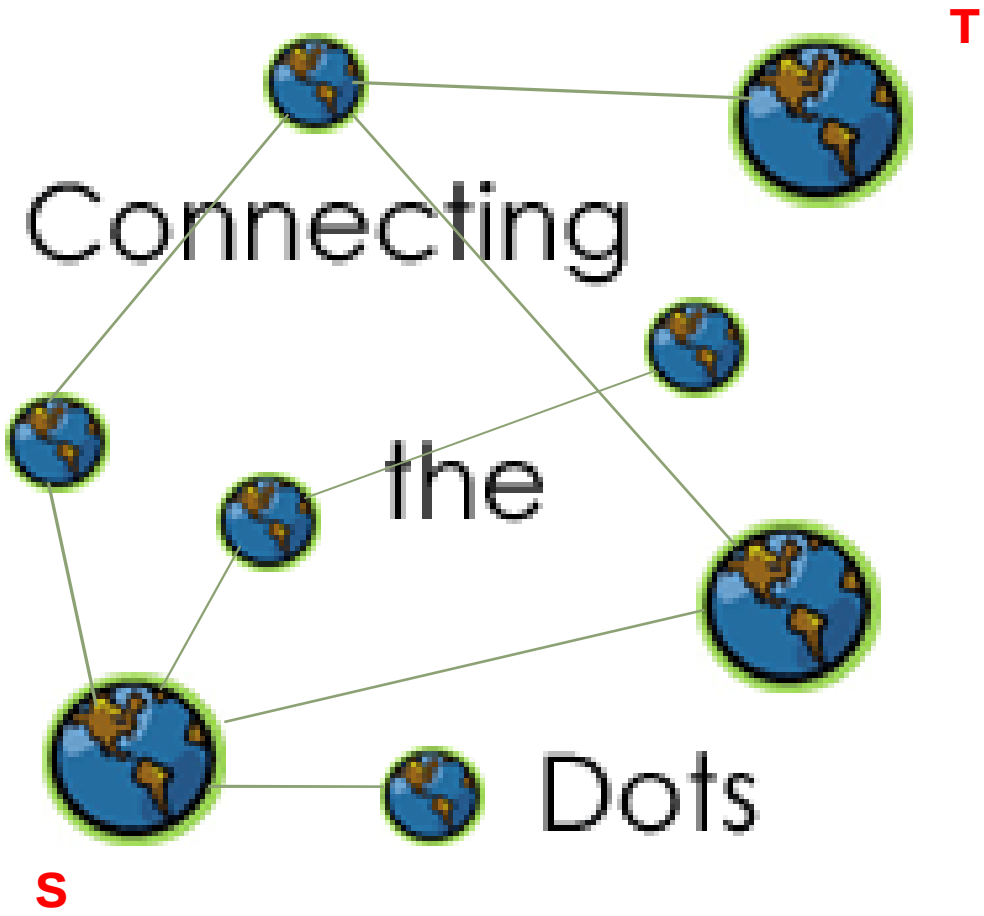
WOULD BE NICE…

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

GOALS

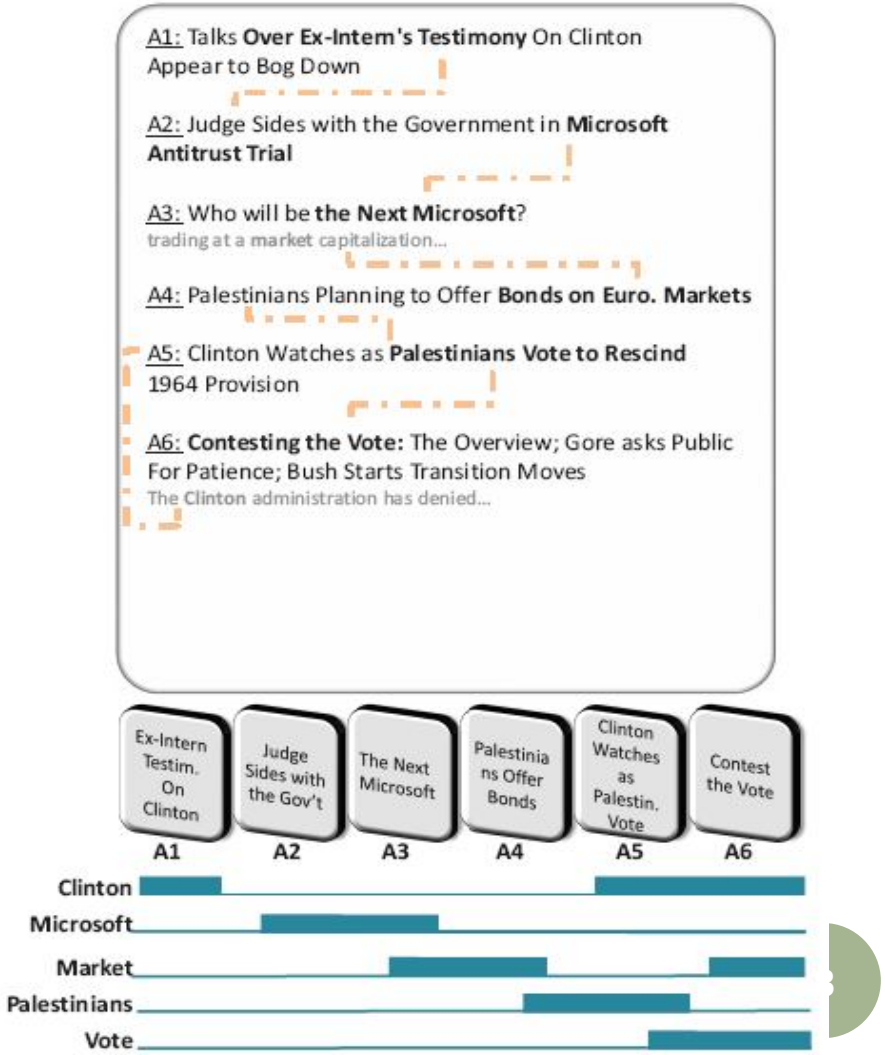
- Methods for automatically connecting the dots
  - Structured, easy way to uncover hidden connections between two pieces of information
- Given two news articles, the system automatically finds a coherent story
- Better understanding of the progression of the story

WHAT MAKES THE STORY GOOD?



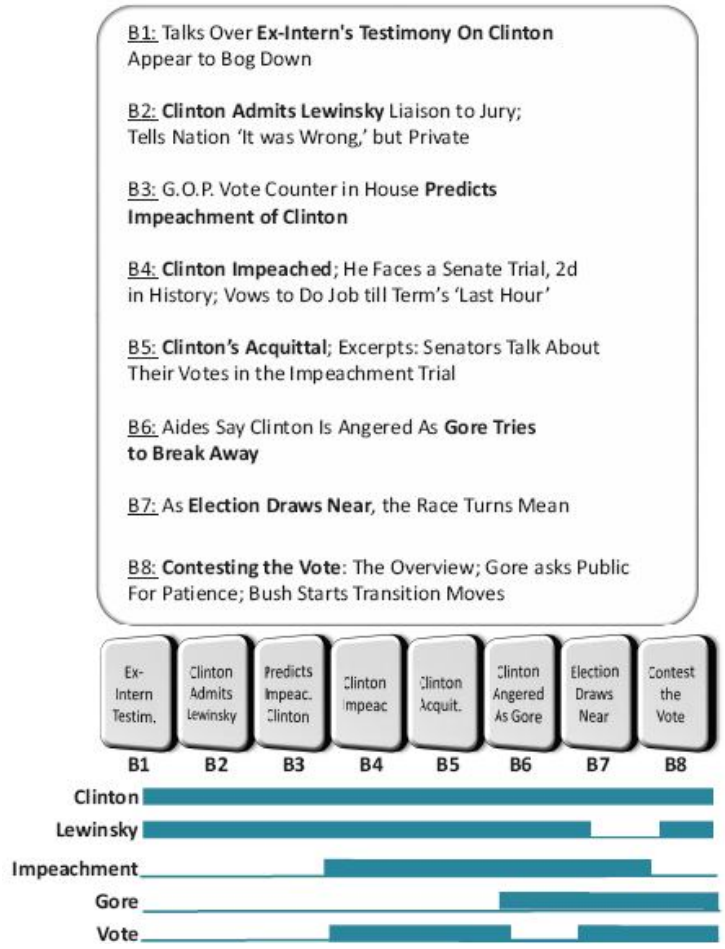
BUT..

- 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)*

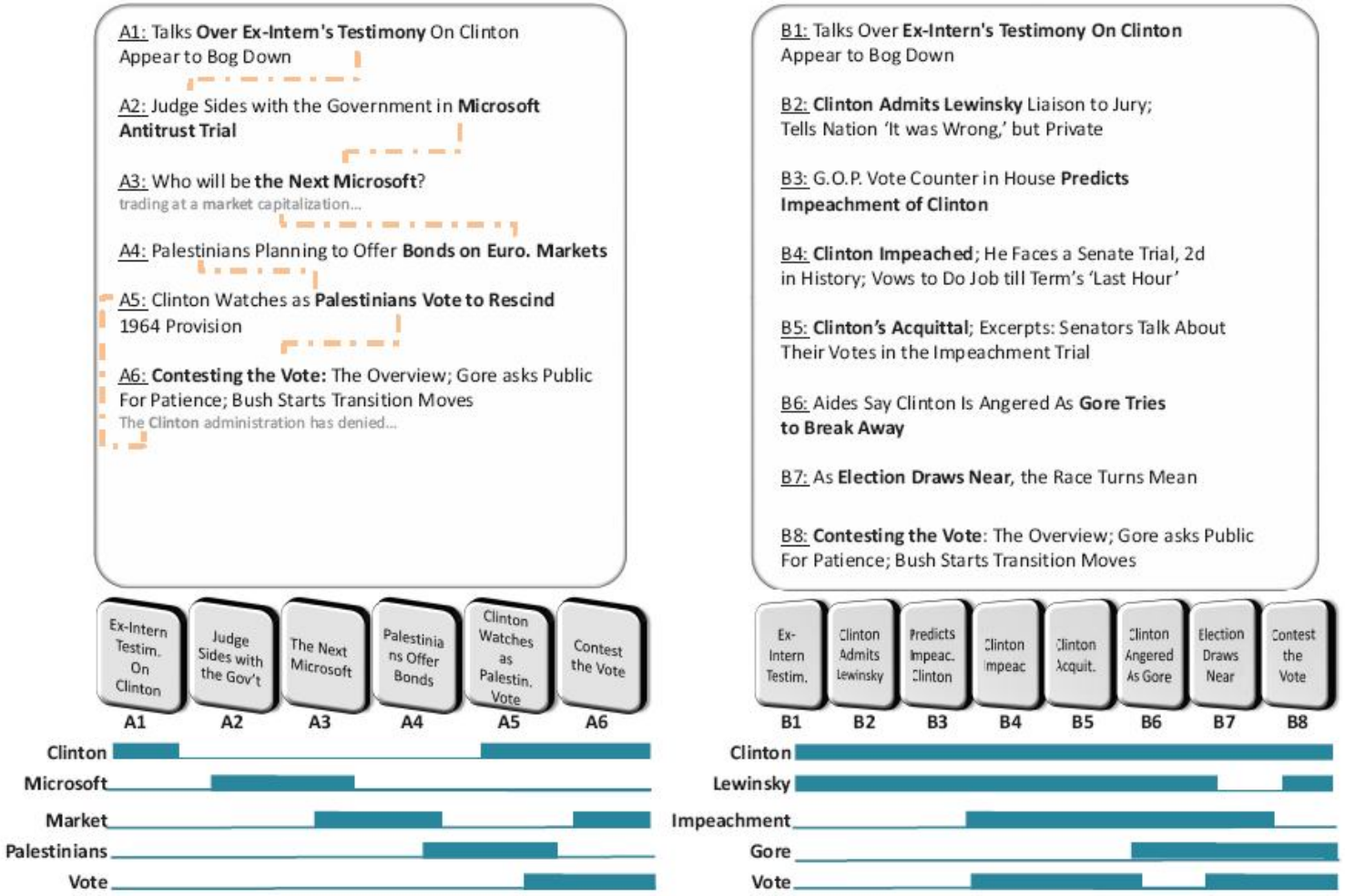


THE PROBLEM

- 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

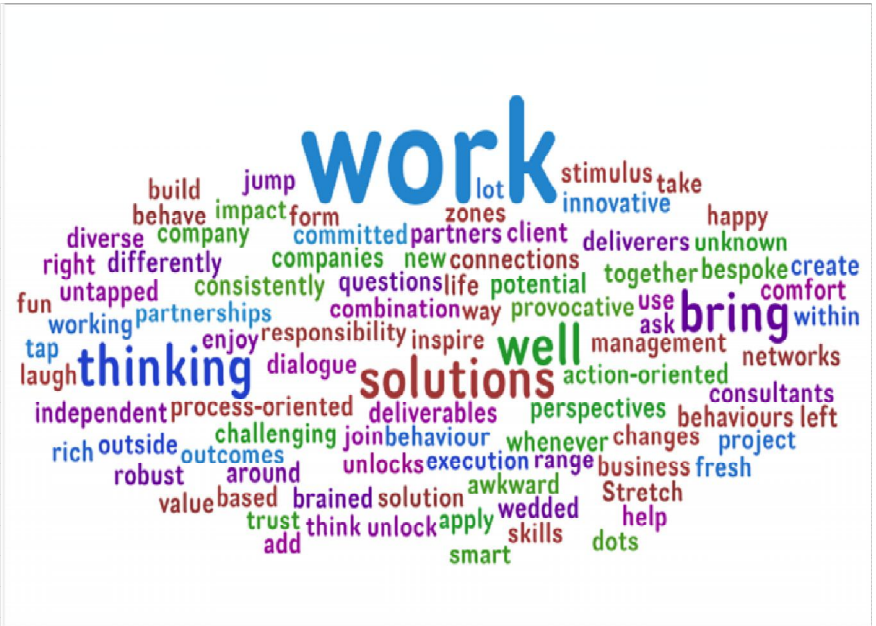
$$Coherence\ (d_1,..., d_n) = \sum_{i=1}^{n-1} \sum_{w} 1(w \in d_i \cap d_{i+1})$$

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

$$Coherence(d_1,..., d_n) = \min_{i=1...n-1} \sum_{w} 1(w \in d_i \cap d_{i+1})$$

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*

$$Influence(d_i, d_j \mid w)$$

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

$$Coherence(d_1,...,d_n) = \min_{i=1...n-1} \sum_{w} Influence(d_i, d_{i+1} \mid w)$$

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

$$Coherence\ (d_1,...,d_n) = \max_{activation\ s} \min_{i=1...n-1} \sum_{w} Influence\ (d_i, d_{i+1} \mid w) 1(w..activein\ ..d_i, d_{i+1}$$

Scoring a chain: Linear Program Formulation

- Smoothness

$$\forall w \sum_i word - init_{w,i} \leq 1$$

$$\forall w,i \;\; word - active_{w,i} \leq word - active_{w,i-1} + word - init_{w,i}$$

$$\forall w \;\; word - active_{w,0} = 0$$

- Activation restrictions

$$\sum_{w,i} word \;\; - init_{w,i} \leq kTotal$$

$$\forall i \sum_w word \;\; - active_{w,i} \leq kTrans$$

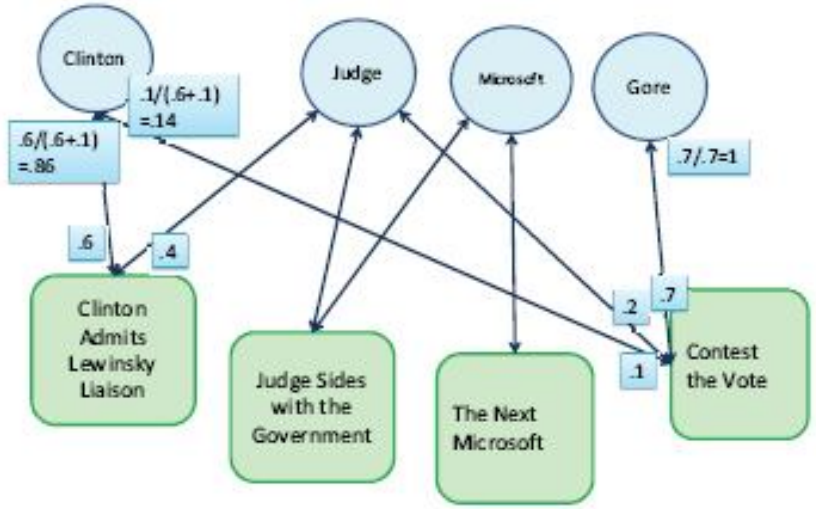
- Objective

$$\forall i \;\; \text{minedge} \;\; \leq \sum_w \text{word} \;\; - \text{active}_{w,i} * \text{influence} \;\; (d_i, d_{i+1} \mid w)$$

$$\forall i, w \;\; word \;\; - active_{w,i}, word \;\; - init_{w,i} \in [0,1]$$

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

$$\prod_i(v) = \varepsilon * 1 * (v = d_i) + (1-\varepsilon)\sum_{(u,v)\in E}\prod_i(u)P(v\,|\,u)$$

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

$$\prod_i^w(d_j)$$

- The influence of d_j with respect of w:

$$\prod_i(d_j) - \prod_i^w(d_j)$$

EXAMPLE

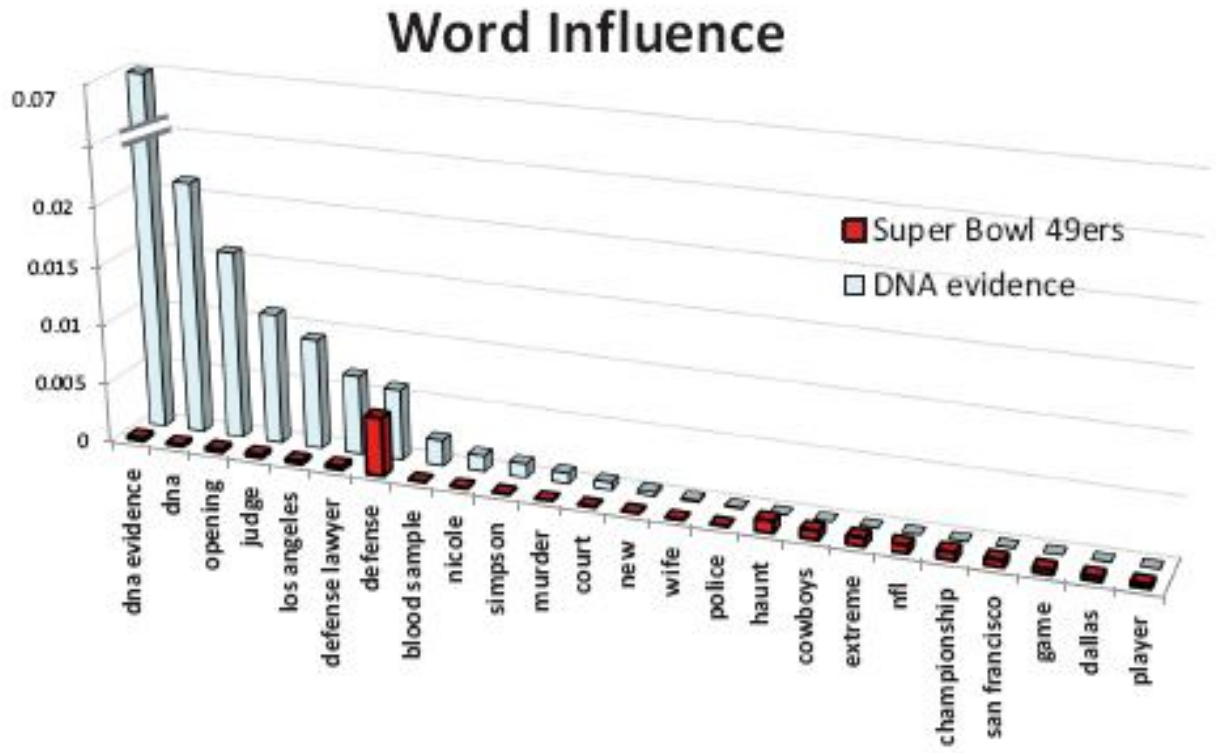
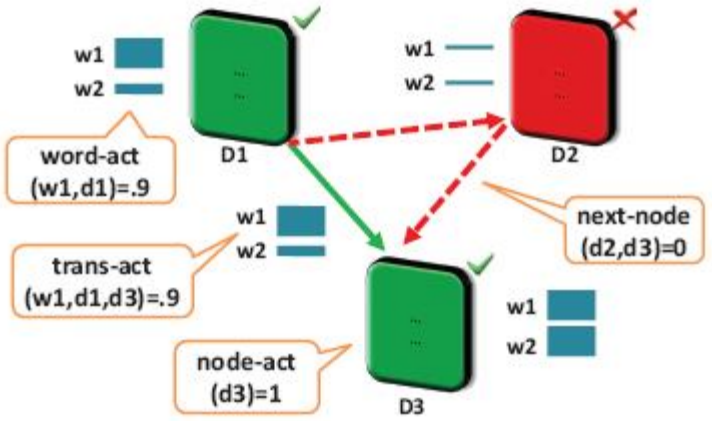


Figure 2: Word influence from an article about the OJ Simpson trial to two other documents – one about football and another about DNA evidence.

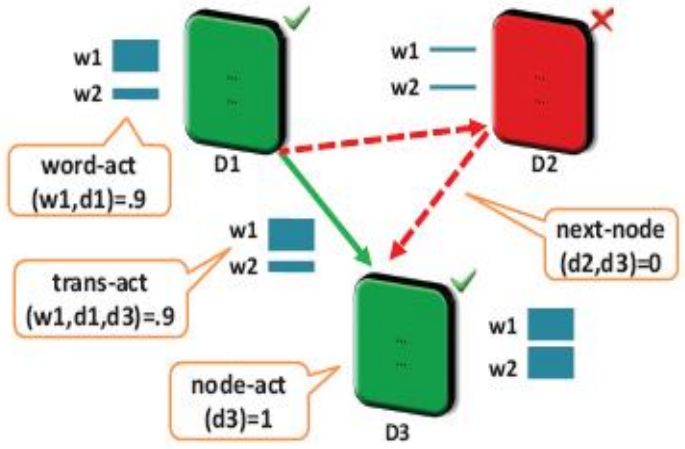
FINDING A GOOD CHAIN

- 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



$$\sum_{w} transition - active_{w,i,j} * Influence(d_i, d_j \mid w)$$

FINDING A GOOD CHAIN: LP

- Objective

$$\forall ij \quad \text{minedge} \le 1 - next - node_{i,j} + \sum_{w} \text{transition} - \text{active}_{w,i,j} * \text{influence}(d_i, d_j \mid w)$$

- Chain Restrictions

$$node - active_1 = 1, node - active_n = 1$$

$$\sum_i node - active_i = K, \sum_i next - node_{i,j} = K - 1$$

$$\sum_i next - node_{i,j} = node - active_j \quad j \neq s$$

$$\sum_j next - node_{i,j} = node - active_i \quad i \neq t$$

$$\forall_{i \geq j} \quad next - node_{i,j} = 0$$

$$\forall_{i<j<k} \quad next - node_{i,j} \le 1 - node - active_k$$

- 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

$$\frac{next-node^*_{i,j}}{\sum_j next-node^*_{i,j}}$$

- 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- □
- for c=□2/v*ln(n/□)

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

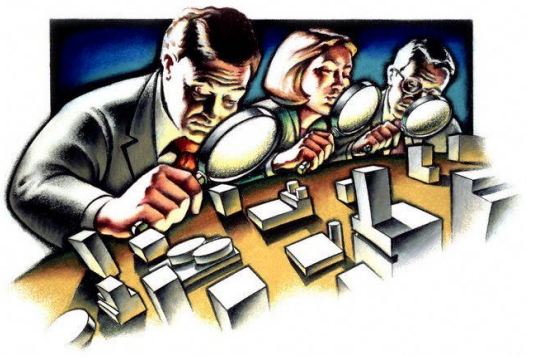
EVALUATION

- Standard methods do not apply
  - No labeled dataset suitable for the task
- Methods evaluated by running them on real data
- New York Times and Reuters dataset (1995 - 2003)
- 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

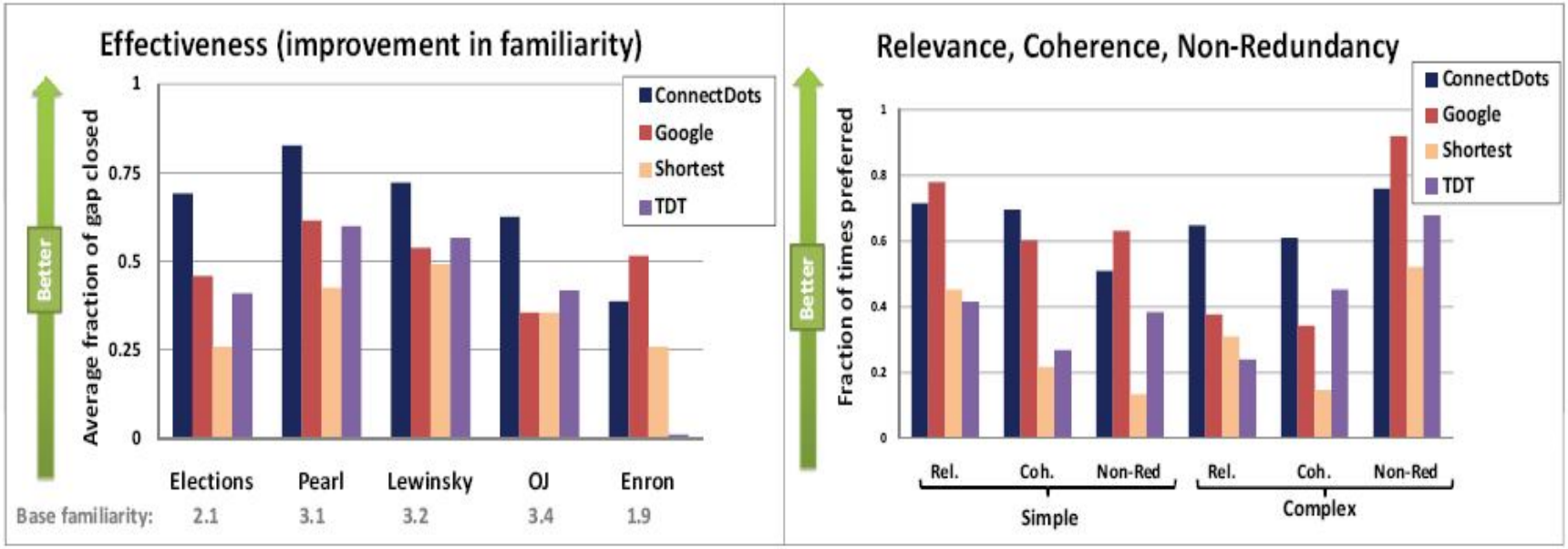


- 18 users
- Measure **familiarity** in the beginning
- The users were asked to indicate:
  - Relevance
  - Coherence
  - Redundancy
- Effectiveness – fraction of the familiarity gap closed
- Simple stories – focus around same event
- Complex stories – connected through one or more events

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

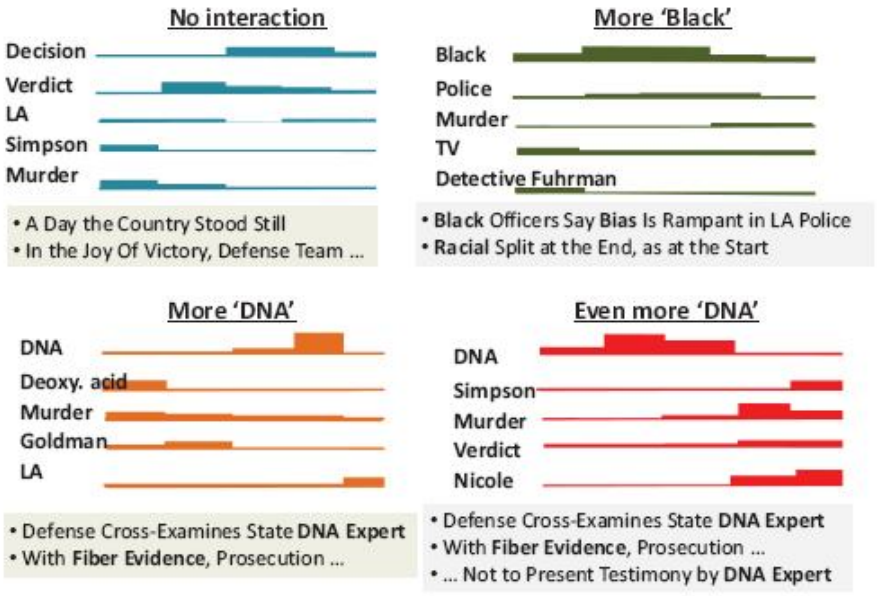
Simpson Defense Drops DNA Challenge
Issue of Racism Erupts in Simpson Trial
Ex-Detective's Tapes Fan Racial Tensions in Los Angeles
Many Black Officers Say Bias Is Rampant in LA Police Force
With Tale of Racism and Error, Lawyers Seek Acquittal
⋆ **In the Joy Of Victory, Defense Team Is in Discord** ⋆
⋆ (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

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

- 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