

Link Analysis in Information Retrieval

Lise Getoor

<http://www.cs.umd.edu/~qinglu/summer02-reading.htm>

Link Analysis

- Finding patterns in graphs
 - Bibliometrics – finding patterns in citation graphs
 - Sociometry – finding patterns in social networks
 - Collaborative Filtering – finding patterns in rank(person, item) graph
 - Webometrics – finding patterns in web page links

Web Link Analysis

- Used for
 - ordering documents matching a user query: ranking
 - deciding what pages to add to a collection: crawling
 - page categorization
 - finding related pages
 - finding duplicated web sites

Web as Graph

- Link graph:
 - node for each page
 - directed edge (u,v) if page u contains a hyperlink to page v
- Co-citation graph
 - node for each page
 - *undirected* edge (u,v) iff exists a third page w linking to both u and v
- assumption:
 - link from page A to page B is a recommendation of page B by A
 - If A and B are connected by a link, there is a higher probability that they are on the same topic

Connectivity-Based Ranking

- Query-independent: gives an intrinsic quality score to a page
- Approach #1: larger number of hyperlinks pointing to a page, the better the page
 - drawback?
 - each link is equally important
- Approach #2: weight each hyperlink proportionally to the quality of the page containing the hyperlink

Page Rank

- PageRank $R(p)$ of page p :

$$R(p) = \varepsilon / n + (1 - \varepsilon) \cdot \sum_{(q,p) \in G} \frac{R(q)}{\text{out degree}(q)}$$

- where
 - ε is a dampening factor usually set between 0.1 and 0.2
 - n is the number of nodes in G
 - $\text{outdegree}(q)$ is the number of edges leaving page p

Alternate Formulation: Random Surfer

- The random surfer can follow any outlink from a page with equal probability. Periodically the random surfer gets bored and jumps to a random page on the Web.
- Page rank is the stationary distribution of infinite walk p_1, p_2, p_3, \dots
 - Each node is equally likely to be the starting node
 - At node p_i
 - with probability ϵ , node p_{i+1} is chosen uniformly at random from all the nodes in G
 - with probability $1-\epsilon$, node p_{i+1} is chosen uniformly at random from the nodes q in G s.t. (p_{i+1}, q)

PageRank cont.

- PageRank is the dominant eigenvector of the probability transition matrix of the random walk
- When PageRank is computed iteratively using the previous equation, the computation will eventually converge (under some weak assumptions on the values in the probability matrix)
- Typically 100 iterations suffice to converge.
- Advantage: not query specific, can be done once

Query-dependent Connectivity-Based Ranking

- Carrier and Kazman
- For each query, build a subgraph of the link graph G limited to pages on query topic
- Build the neighborhood graph
 - A start set S of documents matching query given by search engine (~ 200)
 - Set augmented by its neighborhood, the set of documents that either point to or are pointed to by documents in S (limit to ~ 50)
 - Then rank based on indegree

problem?

Idea

- We desire pages that are **relevant** (in the neighborhood graph) and **authoritative**
- As in page rank, not only the in-degree of a page p , but the quality of the pages that point to p . If more important pages point to p , that means p is more authoritative
- Key idea: Good **hub** pages have links to good **authority** pages
- given user query, compute a hub score and an authority score for each document
- high authority score \rightarrow relevant content
- high hub score \rightarrow links to documents with relevant content

HITS algorithm

- Kleinberg, 1998
- Hypertext Internet Topic Search

Kleinberg's Algorithm

- Initialize: $h_p=1, a_p=1$ for all pages in neighborhood graph
- While the vectors H and A have not converged

$$3. \quad A[i] = \sum_{(j,i) \in N} H[j]$$

$$4. \quad H[i] = \sum_{(i,j) \in N} A[j]$$

5. Normalize the H and A vectors

Does the algorithm converge?

- A is adjacency matrix of neighborhood graph
- At the i th iteration:
 - $a_i = A^t h_{i-1}$
 - $h_i = A^t a_i$
 - normalize a_i and h_i
- Under mild conditions converges, convergence rate σ_2^2 / σ_1^2
- $h^{(k)}$ converges to the leading left singular vector of A
- $a^{(k)}$ converges to the leading right singular vector of A

Problems

- Considers only a small part of web graph, adding a few edges can potentially change scores considerably, thus easier to manipulate scores
- If neighborhood graph contains more pages on a topic different from the query, then the top authority and hub pages are on this different topic. Called topic drift.

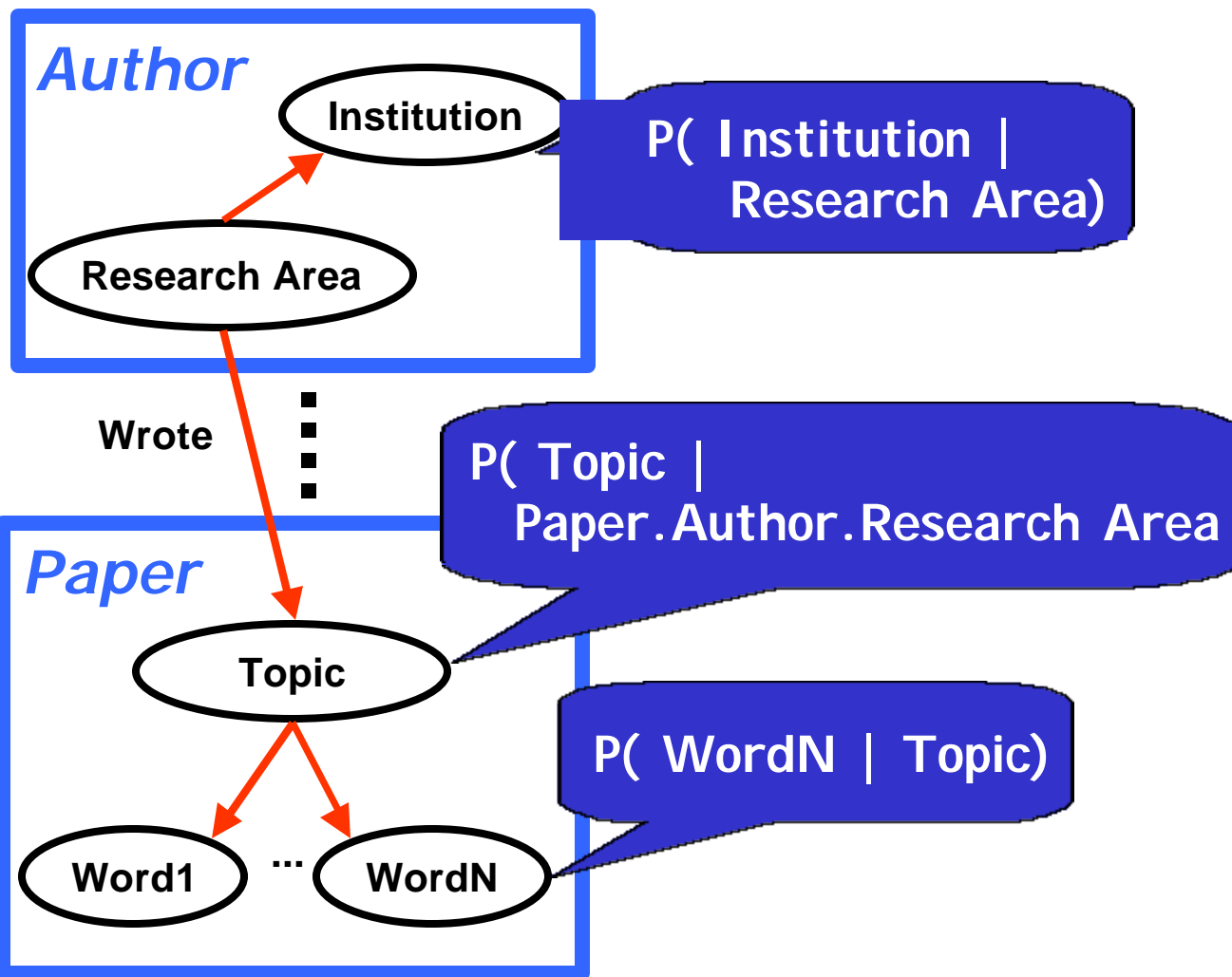
Improvements to Basic Algorithm

- Put weights on edges to reflect importance of links, e.g., put higher weight if anchor text associated with the link is relevant to query
- Normalize weights outgoing from a single source or coming into a single sink. This alleviates spamming of query results
- Eliminate edges between same domain

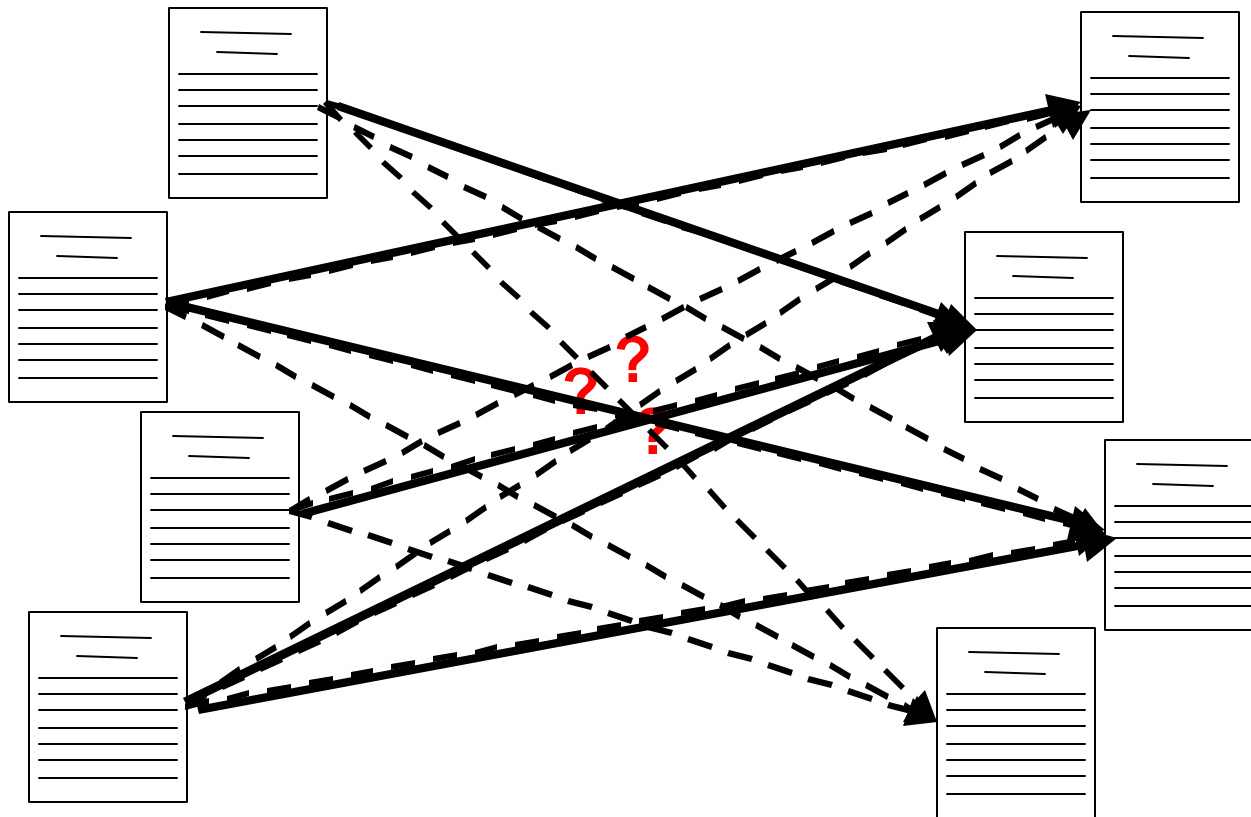
Using Link Structure for Web page classification

- Chakrabarti, et al. SIGMOD 1998.
- Mitchell and Slatterly.

Probabilistic Relational Models



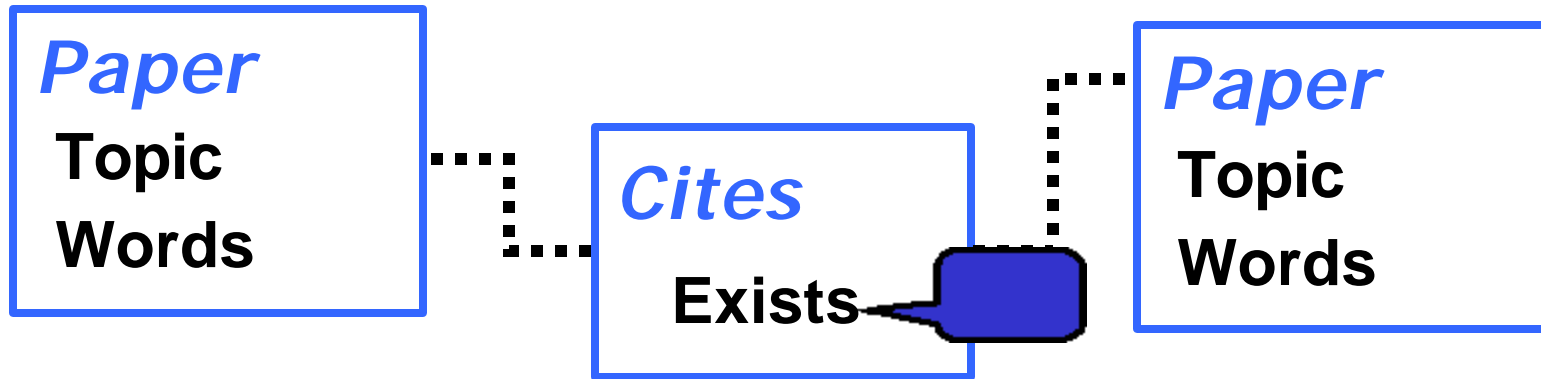
Link Uncertainty



Document Collection

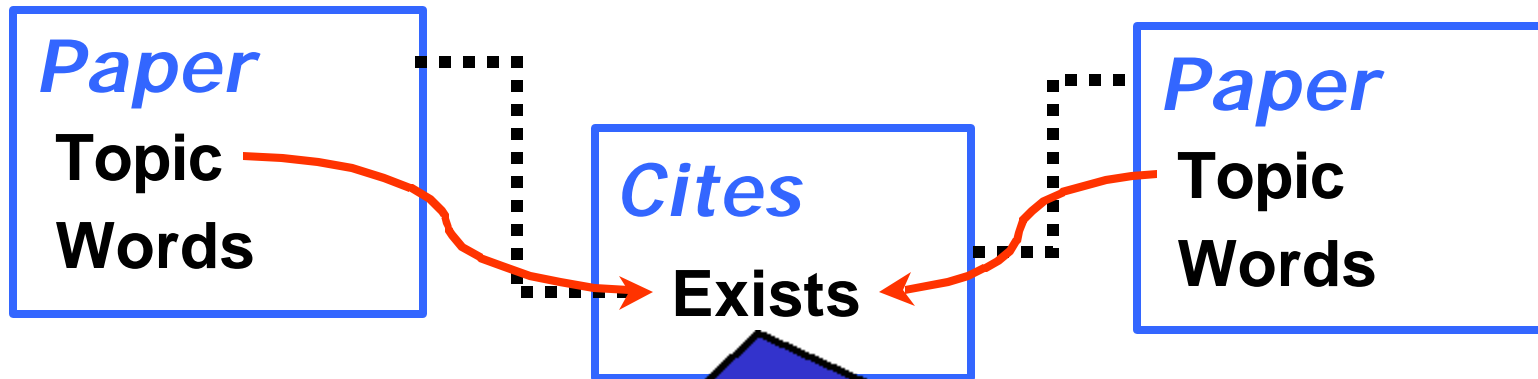
Document Collection

PRM w/ Exists Uncertainty



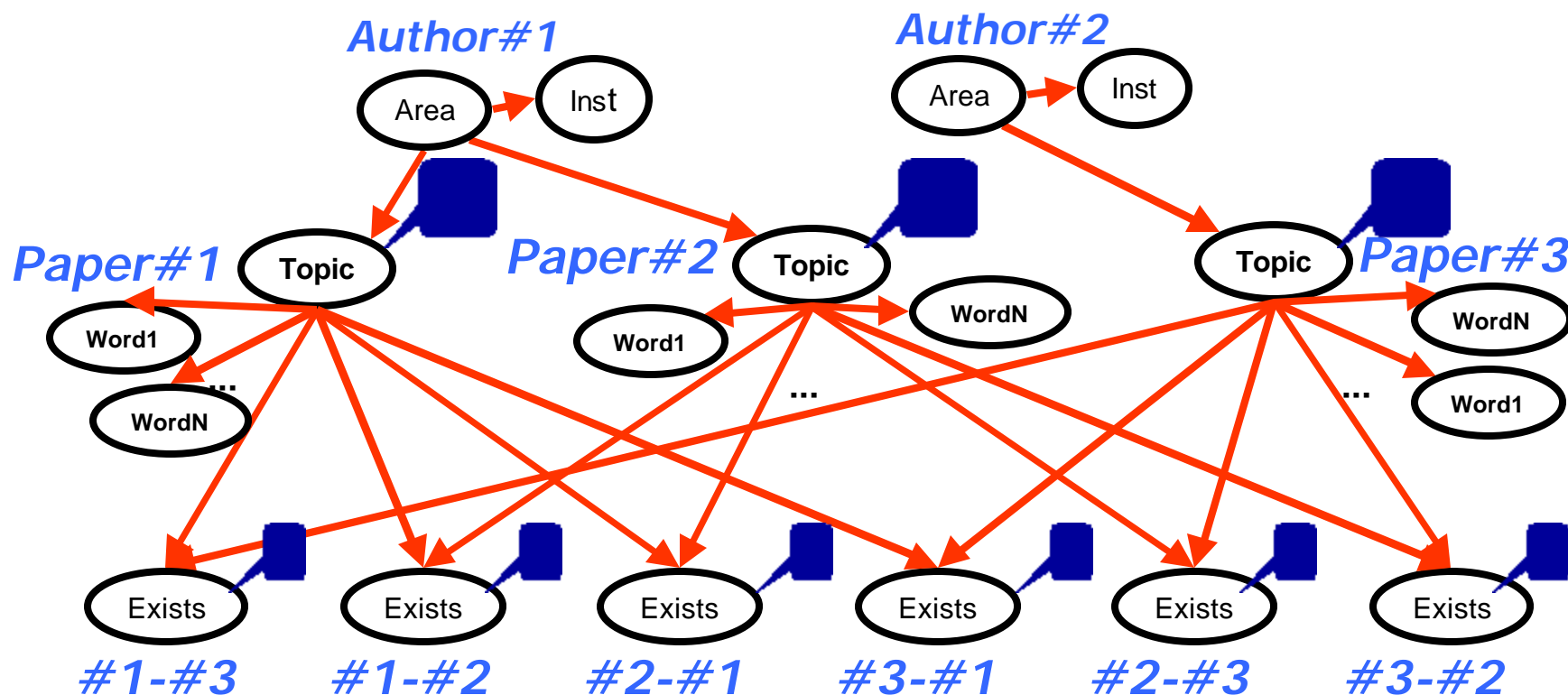
Dependency model for existence of relationship

Exists Uncertainty Example



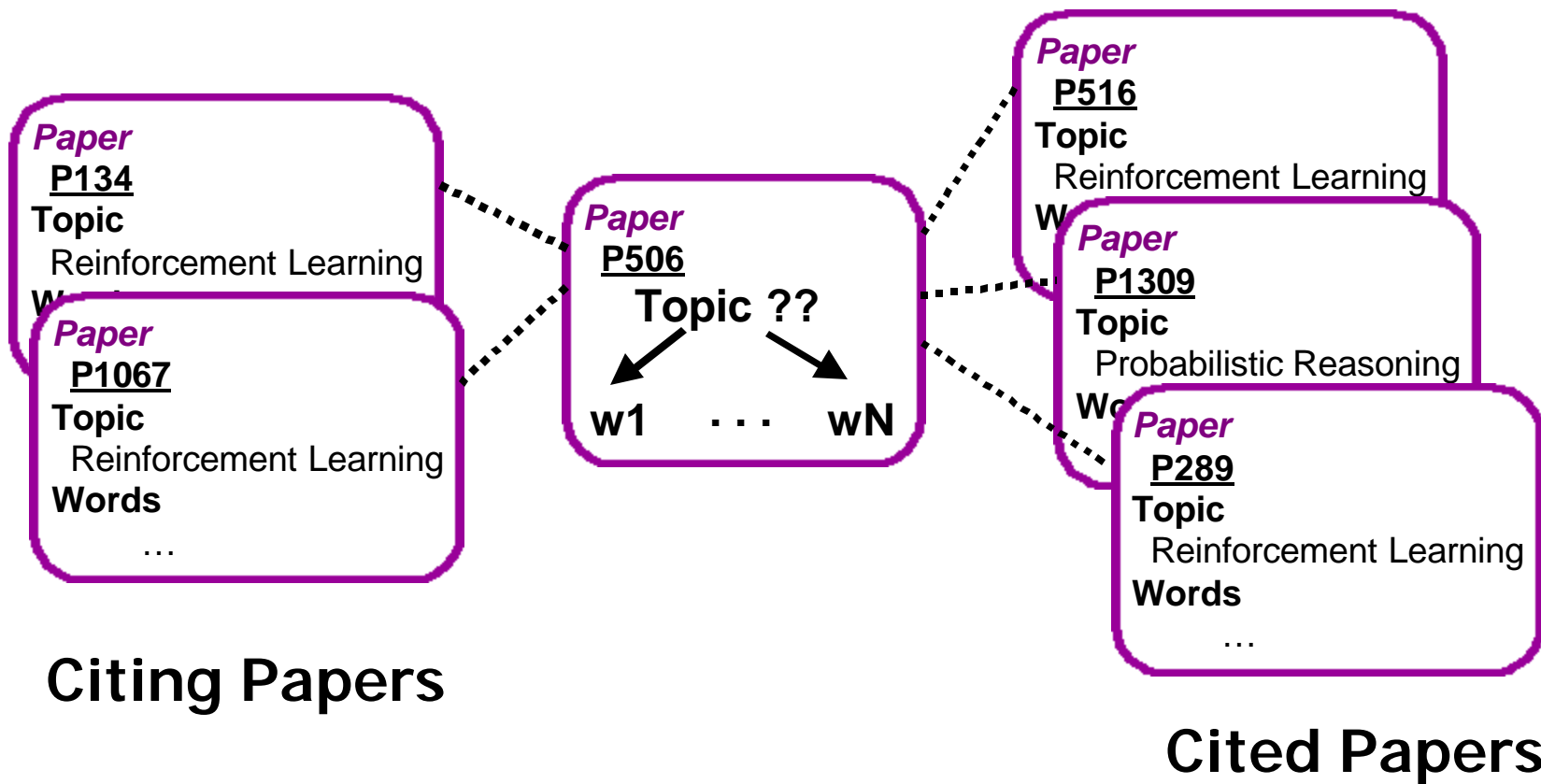
Citer.Topic	Cited.Topic	False	True
<i>Theory</i>	<i>Theory</i>	0.995	0005
<i>Theory</i>	<i>AI</i>	0.999	0001
<i>AI</i>	<i>Theory</i>	0.997	0003
<i>AI</i>	<i>AI</i>	0.993	0008

Ground Bayes Net

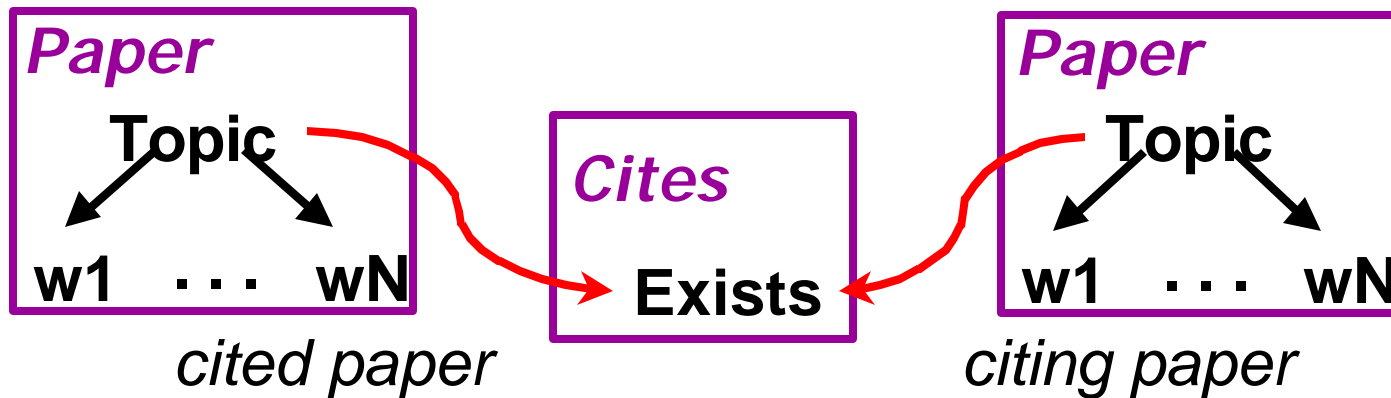


- Captures correlations between topics of related papers
- Information flows along active paths in the Bayes net

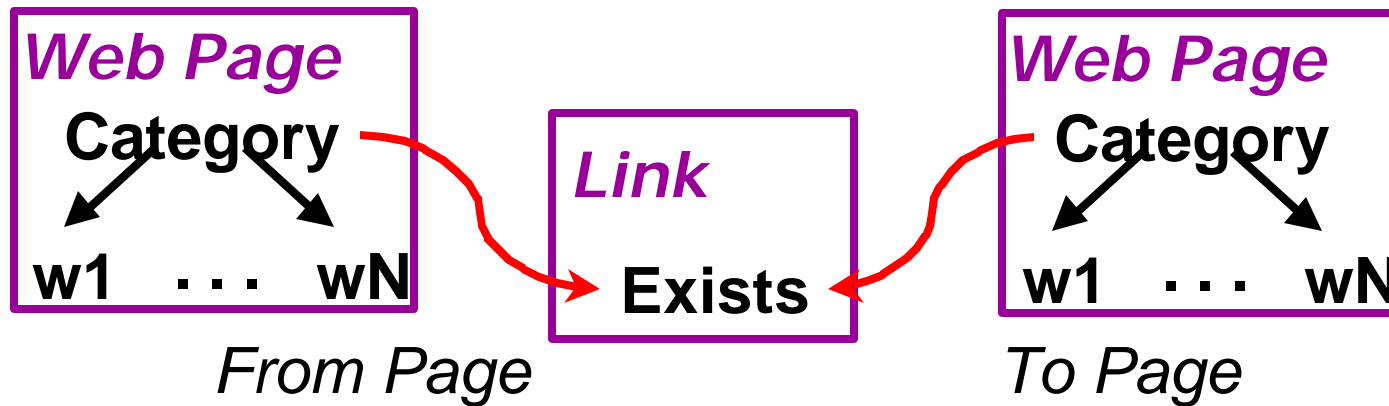
Task: Predict Topic/Category



Domains

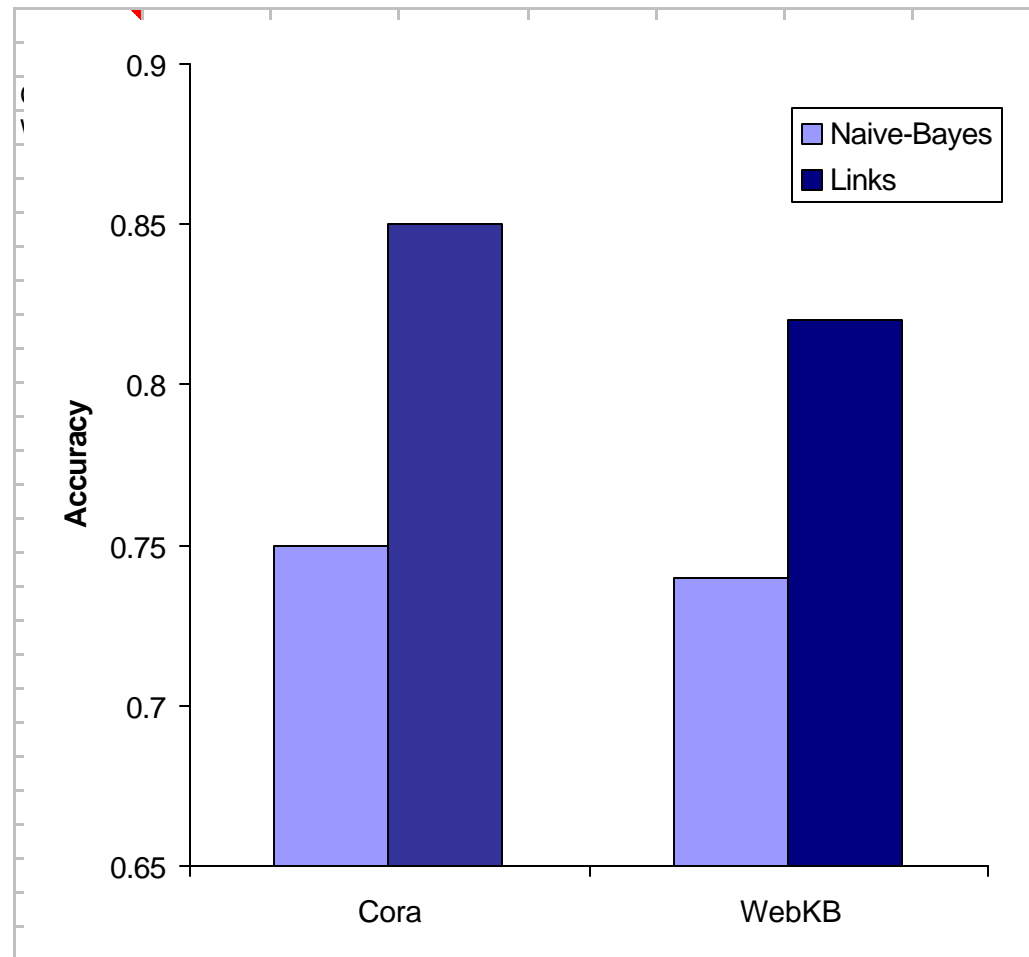


Cora Dataset, McCallum, et. al



WebKB, Craven, et. al

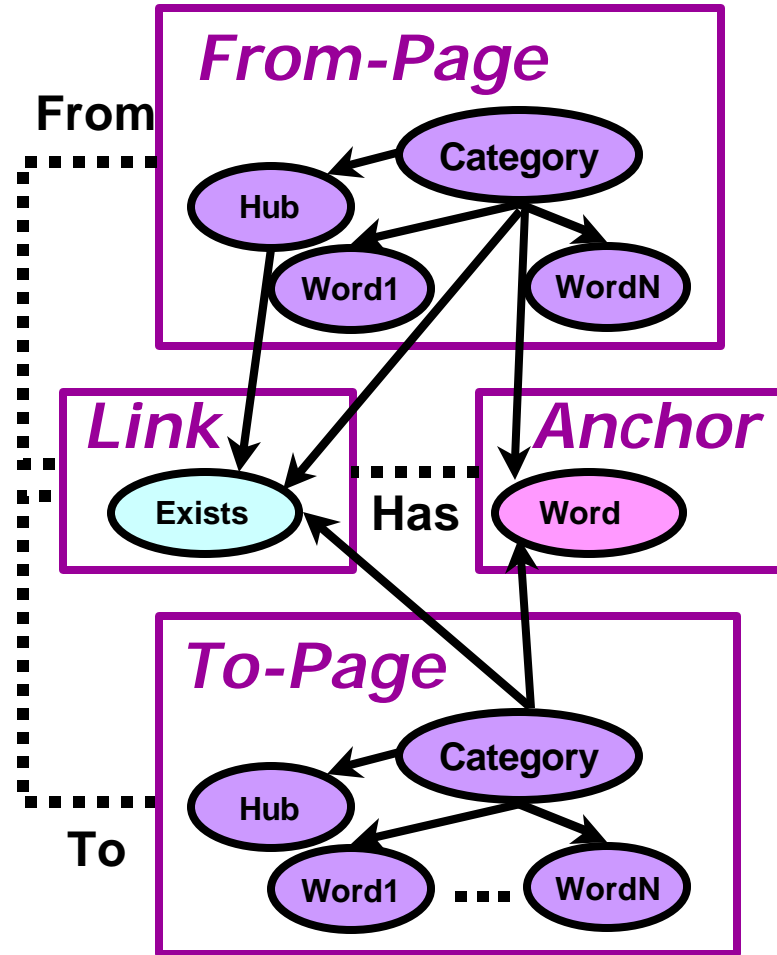
Prediction Accuracy



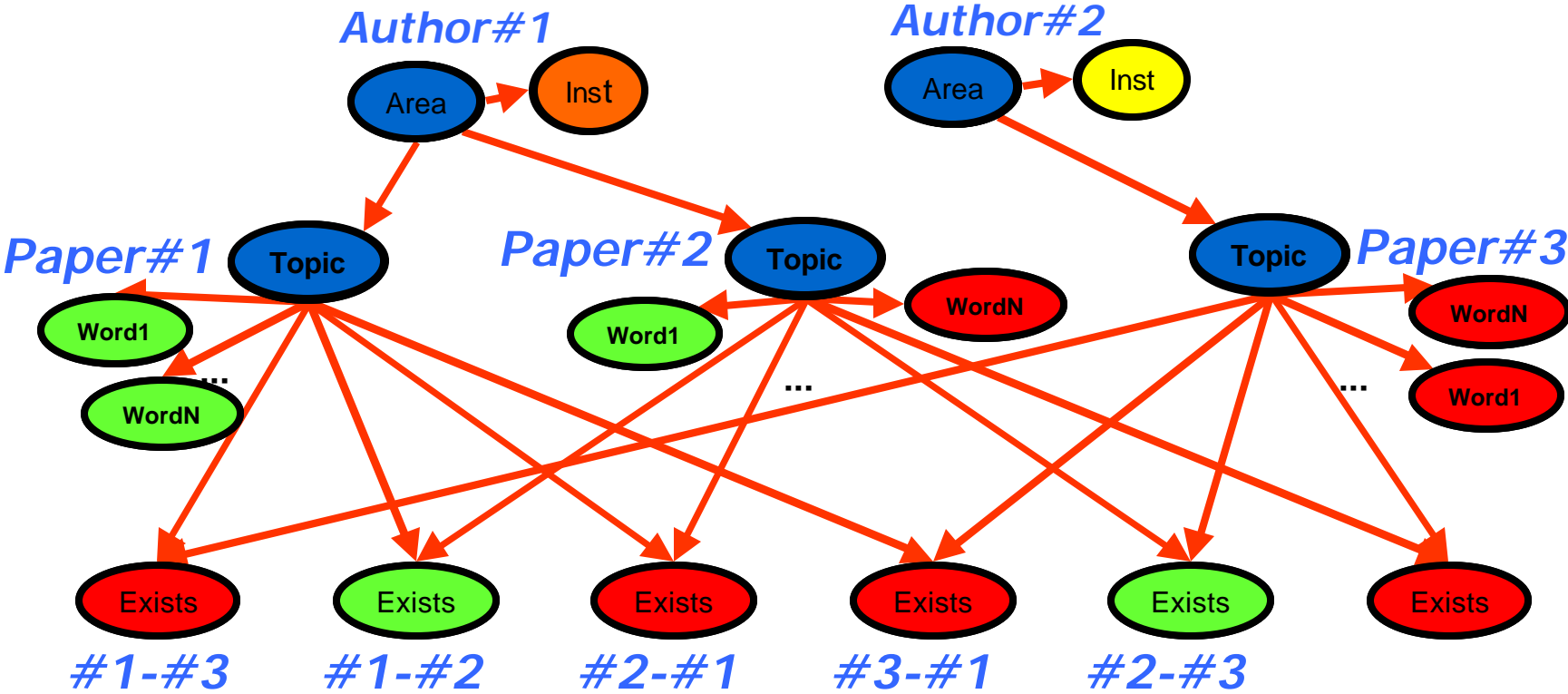
Prediction Accuracy

	Cora	WebKB
baseline	75 \pm 2.0	74 \pm 2.5
RU Citing	81 \pm 1.7	78 \pm 2.3
RU Cited	79 \pm 1.3	77 \pm 1.5
EU	85 \pm 0.09	82 \pm 1.3

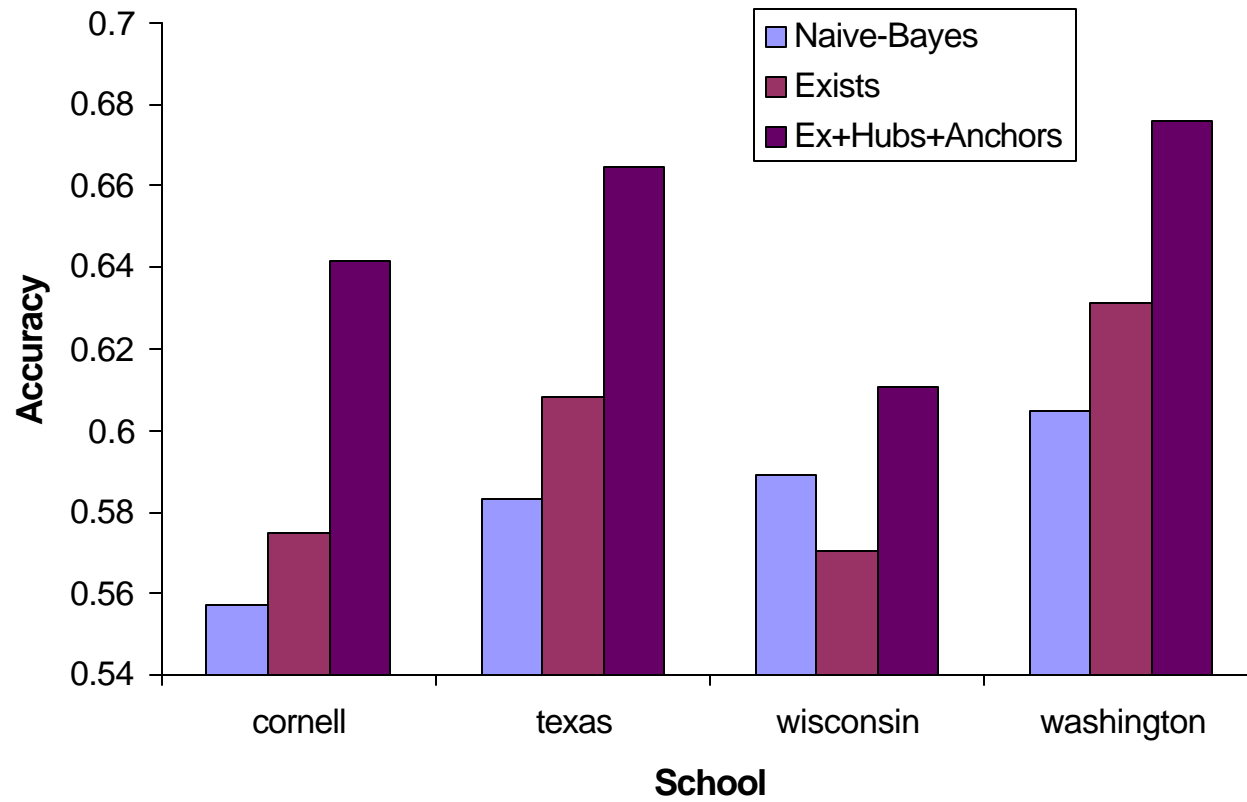
Web Domain



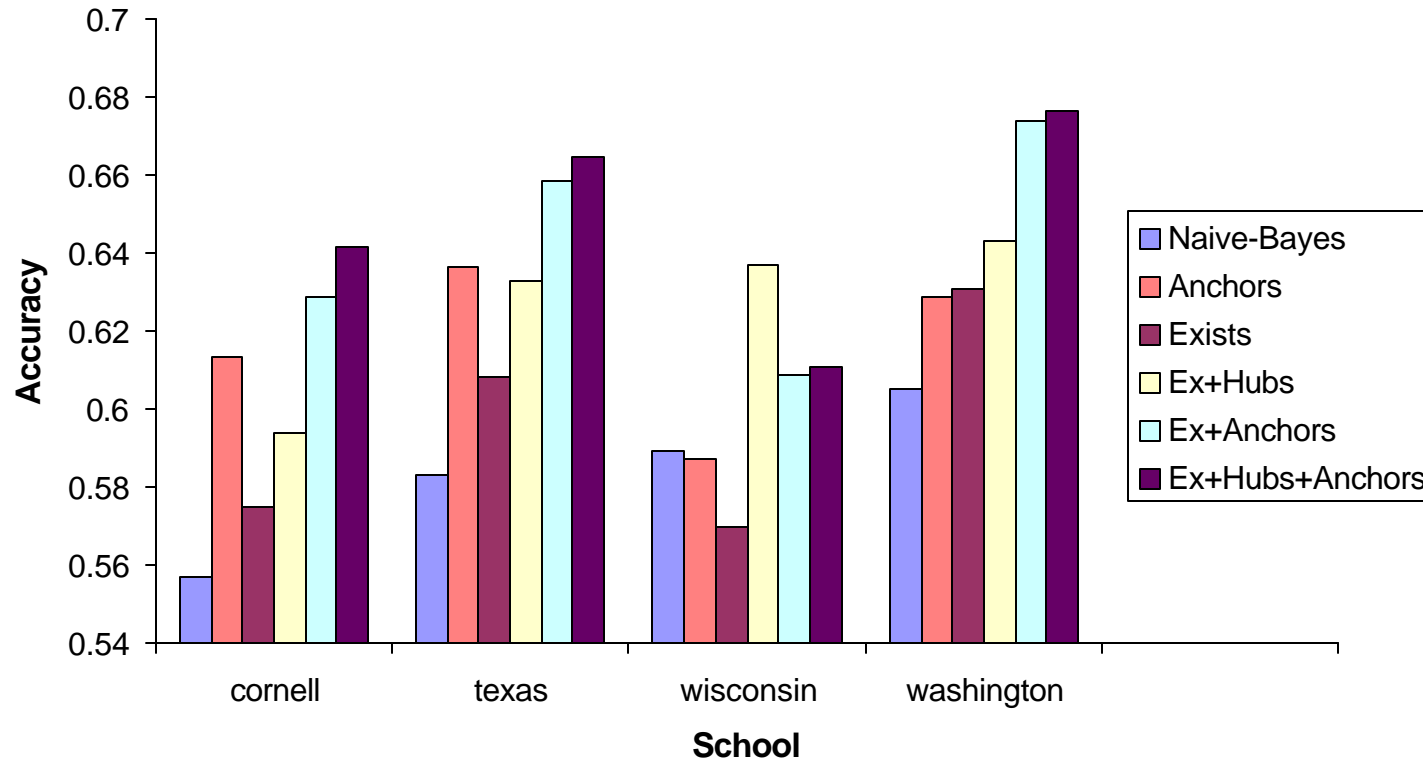
Using Unlabelled Data



WebKB Results



WebKB Results



Graph structure in the Web

- Understanding the various properties of the web graph – diameter, degree distributions, connected components – useful for:
 - designing better crawl strategies on the web
 - analyzing behavior of web algorithms that use link information
 - prediction of web structures, such as bipartite cores and better algorithms to compute them
 - predict emergence of new, yet unexploited, phenomena in the web graph

Example Web Graph Properties

- Six degrees of separation between any 2 web pages
 - almost true in the strongly connected component part of web graph if allow traversal of both out-links and in-links
 - not true in general. small-world network models for graphs
- probability that a web page has in-degree i is proportional to $1/i^x$. Latest estimate for $x=2.1$
- Average path length between two web pages is 19 (Barabasi). Disputed by (Raghavan, et. al).

References

- *Principles of Data Mining*, Hand, Mannila, Smyth. MIT Press, 2001.
- Notes from Dr. M.V. Ramakrishna
<http://goanna.cs.rmit.edu.au/~rama/cs442/info.html>
- Notes from CS 395T: Large-Scale Data Mining, Inderjit Dhillon
<http://www.cs.utexas.edu/users/inderjit/courses/dm2000.html>
- Link Analysis in Web Information Retrieval, Monika Henzinger. Bulletin of the IEEE computer Society Technical Committee on Data Engineering, 2000.
research.microsoft.com/research/db/debull/A00sept/henzinge.ps
- slides from *Data Mining: Concepts and Techniques*, Jan and Kamber, Morgan Kaufman, 2001.