Only connect!
Explorations in using graphs for IR and NLP

Lillian Lee

Cornell University

http://www.cs.cornell.edu/home/llee

Overview

Two specific problems:

1. Can we improve information retrieval by using link analysis, as in Web search, if the documents don't contain hyperlinks?
2. Can we create a system that determines whether speeches in a Congressional floor debate support a piece of legislation?

Theme: Use inter-item relationships, representing them via graphs.

Style: We'll focus on simple descriptions of the main ideas.

Overview

Two specific problems:

1. Can we improve information retrieval by using link analysis, as in Web search, if the documents don't contain hyperlinks?
2. Can we create a system that determines whether speeches in a Congressional floor debate support a piece of legislation?

Theme: Use inter-item relationships, representing them via graphs.

Style: We'll focus on simple descriptions of the main ideas.

> Convince [people] that your solution is trivial. .... the advantage of [them] thinking your solution is trivial or obvious is that it necessarily comes along with the notion that you are correct.
>
> — Stuart Shieber

Respect my authority! HITS without hyperlinks

Oren Kurland (Ph.D. Cornell ’06) and Lillian Lee

SIGIR 2006

Only connect! That was the whole of her sermon.

— E.M. Forster, *Howards End*

Structural re-ranking

Given: a shortlist of documents $\mathcal{D}_{\mathrm{init}}(q)$ produced by an initial retrieval engine in response to a query q.

Goal: re-order the list to improve precision at the very top of the list,

using relationships between documents, i.e., the *structure* of $\mathcal{D}_{\mathrm{init}}(q)$

Examples:

- Similarity-based re-ranking, inspired by van Rijsbergen's ('79) *cluster hypothesis* (Liu & Croft '04)
- PageRank (Brin & Page '98), which uses explicit hyperlinks (not originally proposed for re-ranking)

Influence weights (PageRank)

Fix a directed graph with non-negative edge-weights, denoted $wt(n \to v)$.

Conceptually, the set of PageRank scores $\{\text{PR}(v)\}$ is a solution to the interdependent Pinski-Narin ('76) equations

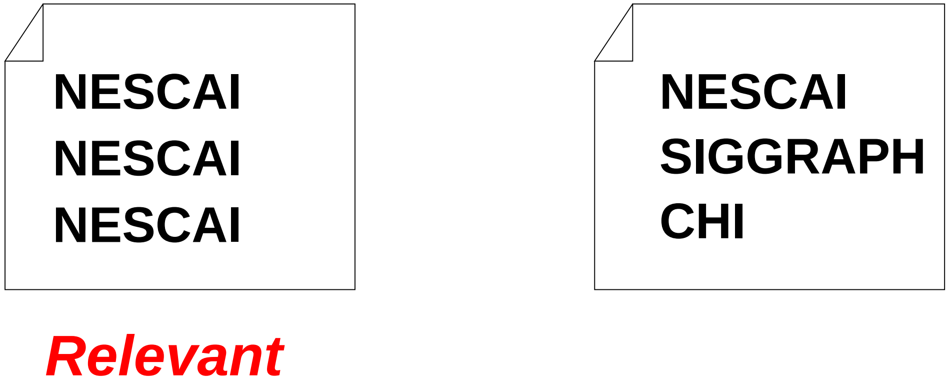
$$\text{PR}(v) = \sum_{n \in V} wt(n \to v) \cdot \text{PR}(n)$$

(the Brin-Page ('98) random jump can be folded into the edge weights).

In Web search, hyperlinks (often) “encode a considerable amount of latent human judgment” or *endorsement*. (Kleinberg '98)

⇒ **Lacking hyperlinks**, we should *infer* endorsement links — which might differ from (symmetric) similarity links.

Examining endorsement

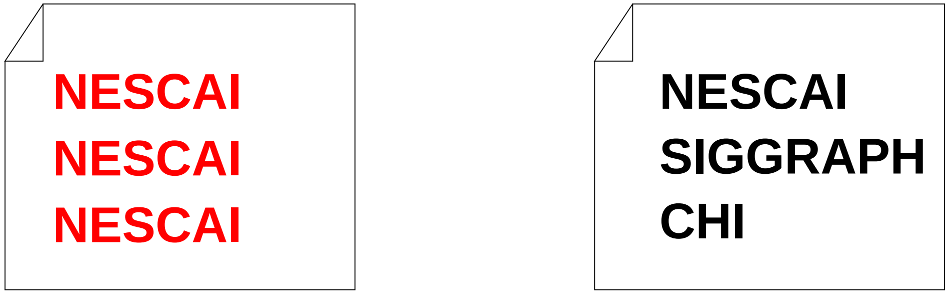


NESCAI
NESCAI
NESCAI

Relevant

NESCAI
SIGGRAPH
CHI

Examining endorsement

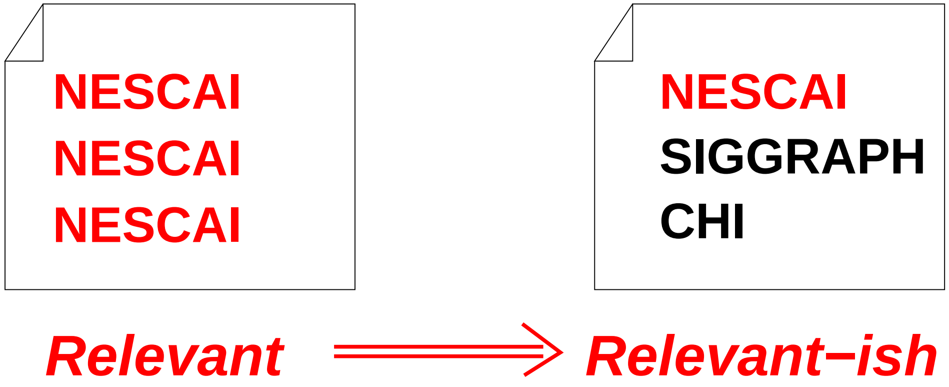


NESCAI
NESCAI
NESCAI

NESCAI
SIGGRAPH
CHI

Relevant

Examining endorsement

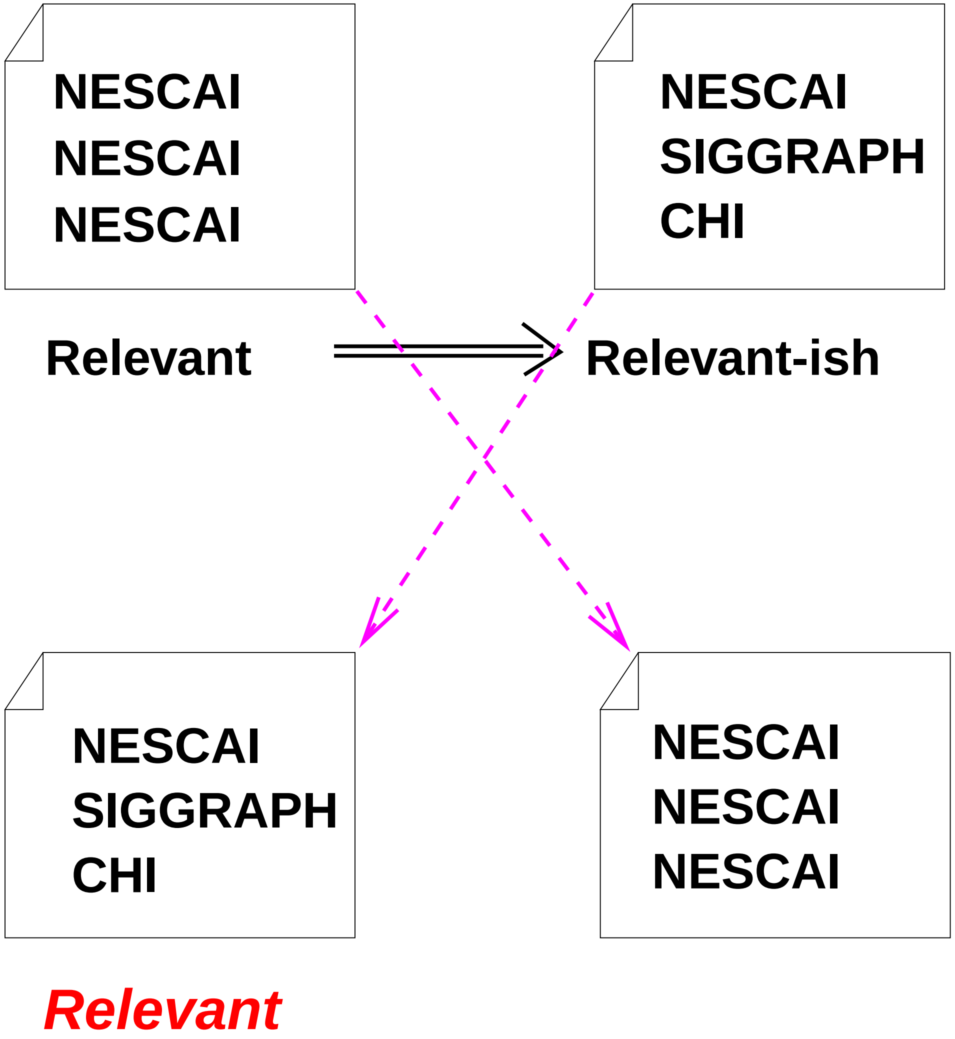


NESCAI
NESCAI
NESCAI

NESCAI
SIGGRAPH
CHI

Relevant ⟹ *Relevant−ish*

Examining endorsement



NESCAI
NESCAI
NESCAI

NESCAI
SIGGRAPH
CHI

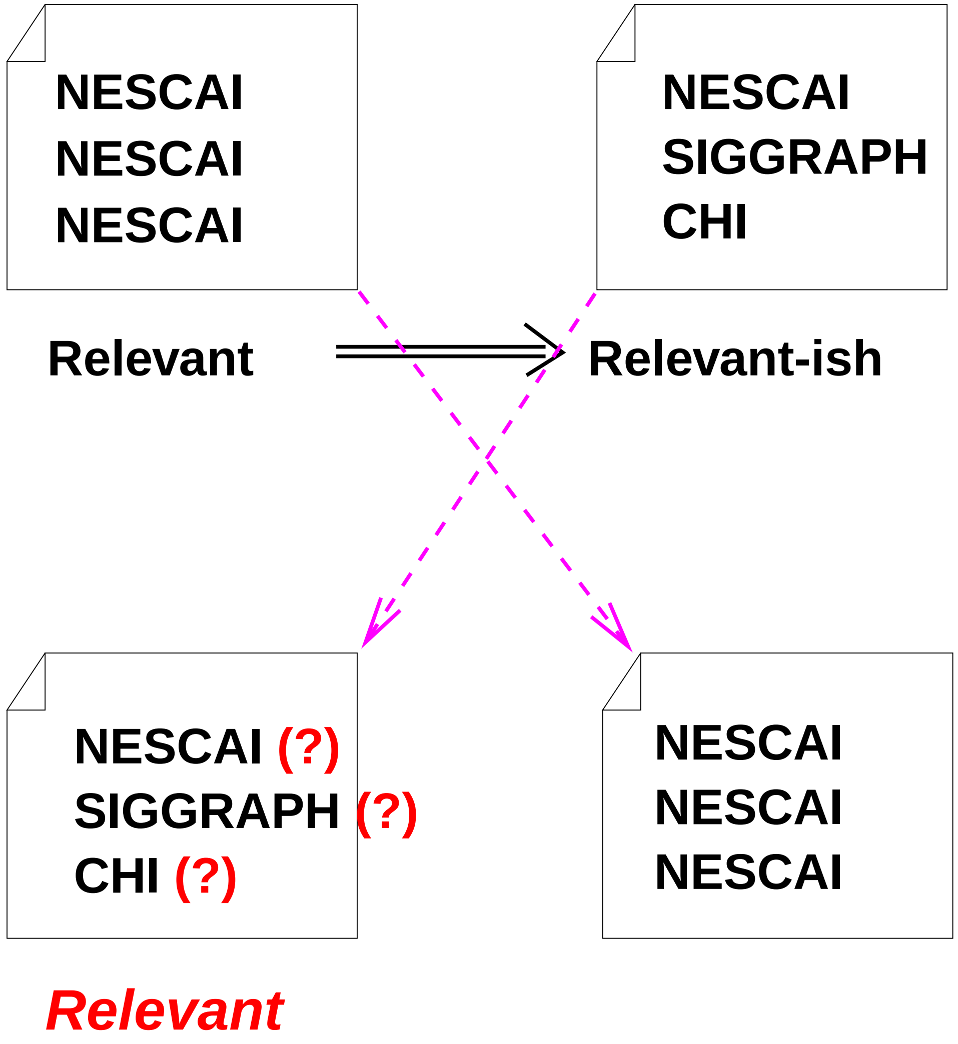
Relevant ⇒ **Relevant-ish**

NESCAI
SIGGRAPH
CHI

NESCAI
NESCAI
NESCAI

Relevant

Examining endorsement



NESCAI
NESCAI
NESCAI

Relevant

NESCAI
SIGGRAPH
CHI

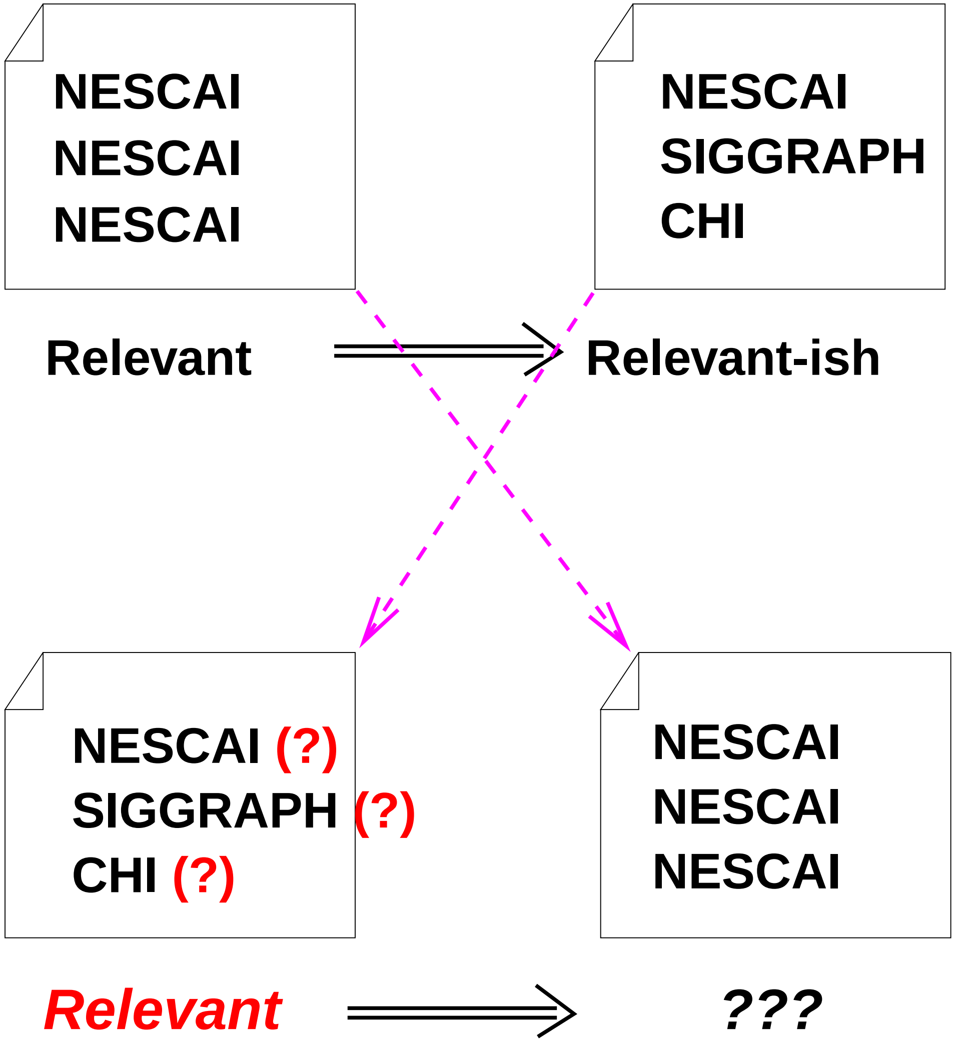
Relevant-ish

NESCAI (?)
SIGGRAPH (?)
CHI (?)

Relevant

NESCAI
NESCAI
NESCAI

Examining endorsement



NESCAI
NESCAI
NESCAI

NESCAI
SIGGRAPH
CHI

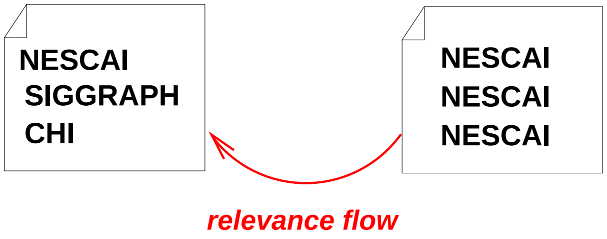
Relevant ⟹ Relevant-ish

NESCAI (?)
SIGGRAPH (?)
CHI (?)

NESCAI
NESCAI
NESCAI

Relevant ⟹ ***???***

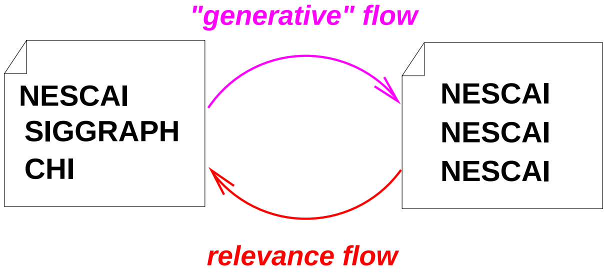
Asymmetry and language models


NESCAI
SIGGRAPH
CHI

NESCAI
NESCAI
NESCAI

relevance flow

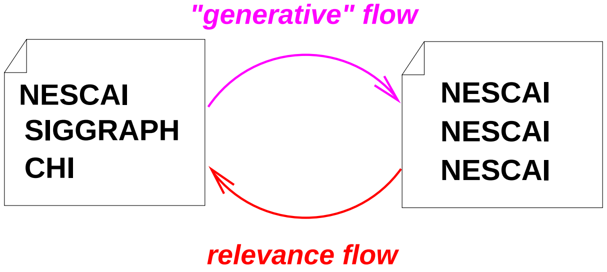
Asymmetry and language models



According to a simple language model that we can induce from the document “NESCAI SIGGRAPH CHI”, $P($“NESCAI NESCAI NESCAI”$) \propto \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$.

But, according to a simple language model induced from the document “NESCAI NESCAI NESCAI”, $P($“NESCAI SIGGRAPH CHI”$) \propto 1 \times 0 \times 0 = 0$.

Asymmetry and language models



According to a simple language model that we can induce from the document “NESCAI SIGGRAPH CHI”, $P($“NESCAI NESCAI NESCAI”$) \propto \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$.

But, according to a simple language model induced from the document “NESCAI NESCAI NESCAI”, $P($“NESCAI SIGGRAPH CHI”$) \propto 1 \times 0 \times 0 = 0$.

Idea: relevance spreads from “offspring” documents o to those documents g that “generate” them, where the “rate” of this spread is related to $P_g(o)$.

Note: Language models have been used in other ways in IR as well (Ponte & Croft ’98, Croft & Lafferty, eds, ’03)

One representative experiment

(Many details surpressed.)

Data: Three TREC corpora (170K-530K documents), 50 or 100 queries.

Preprocessing via Lemur (www.lemurproject.org).

Graph construction: Nodes = top 50 documents in the initial ranking.

Conceptually, edges connect nodes to their top k generators.

Evaluation metric: Precision of the top five documents.

We follow IR practice for parameter-setting.

One representative experiment (cont.)

Focus algorithm: $\textsf{PR}(v) = \sum_{n \in V} wt(n \rightarrow v) \cdot \textsf{PR}(n)$

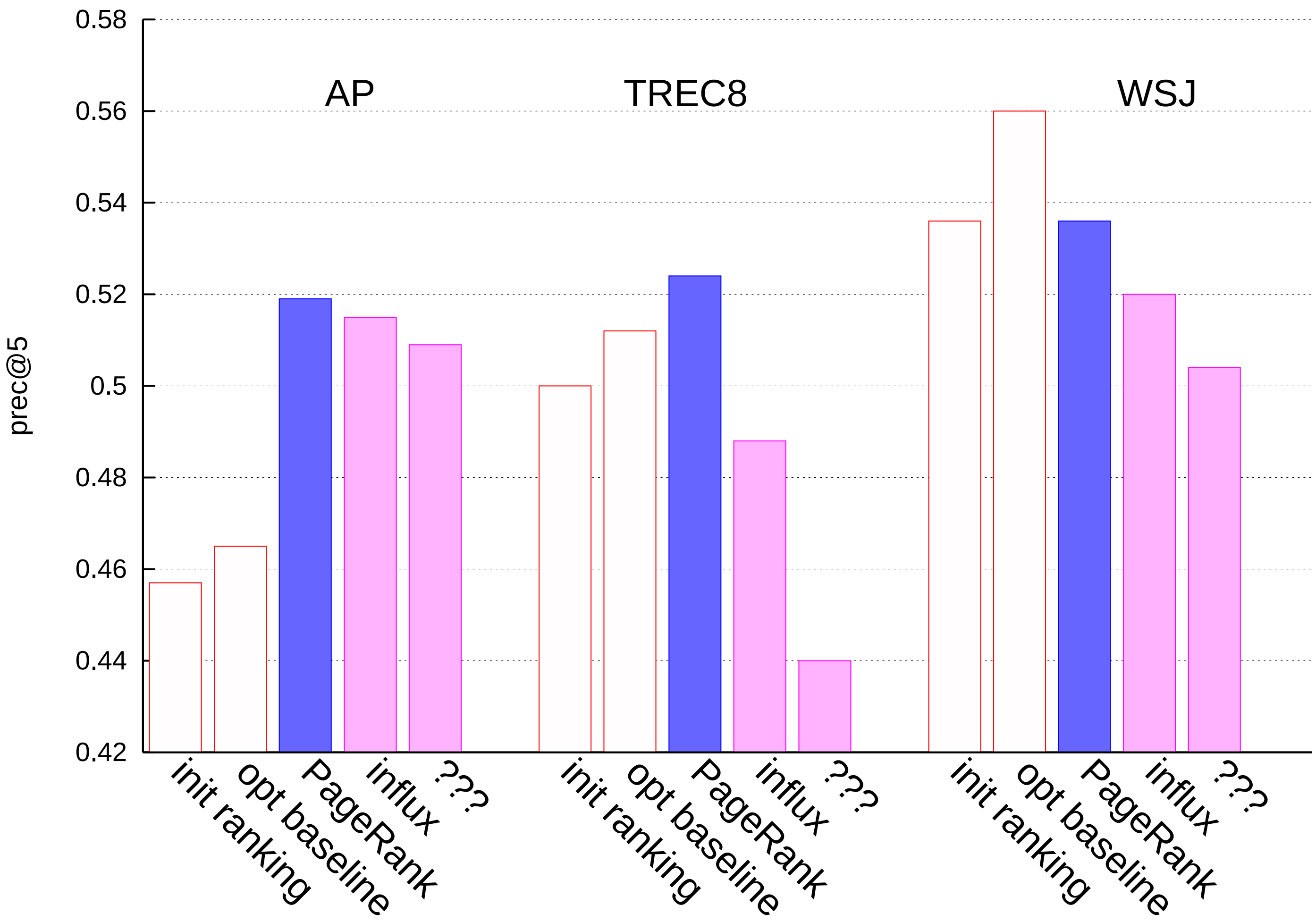
Isolated-document baselines:

- Language-model-based approach, optimized for the evaluation metric
- Initial ranking: (slightly) sub-optimal language-model-based approach

Graph-based reference comparisons

- *Influx* (weighted in-degree): $In(v) = \sum_{n \in V} wt(n \rightarrow v) \cancel{\cdot In(n)}$
- Mystery algorithm ...

PageRank comparison

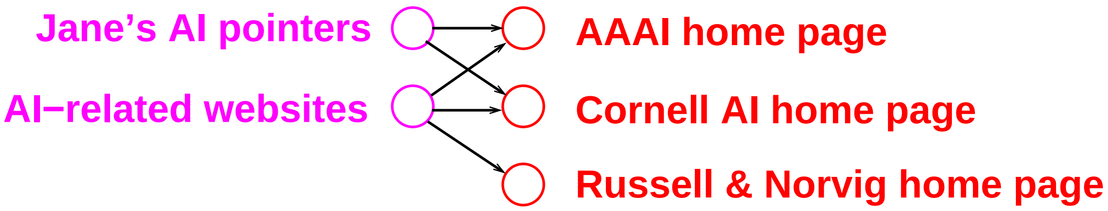


(Modulo term-weighting issues, cosine results are qualitatively similar, but lower.)

Hubs and authorities (HITS)

On the Web, there are **two mutually reinforcing kinds** of informative documents: hubs and authorities (Kleinberg '98)

"Iconic" case: a "one-way" bipartite graph:



We can "split" the equation $\text{PR}(v) = \sum_{n \in V} wt(n \rightarrow v) \cdot \text{PR}(n)$:

$$\begin{aligned} \text{auth}(v) &= \sum_{n \in V} wt(n \rightarrow v) \cdot \text{hub}(n) \\ \text{hub}(n) &= \sum_{v \in V} wt(n \rightarrow v) \cdot \text{auth}(v) \end{aligned}$$

Need for adaptation

Perhaps the problem is that in the non-Web setting, there may not be a natural hub/authority distinction between documents.

What alternative entities could help indicate document authoritativeness (and hence relevance)?

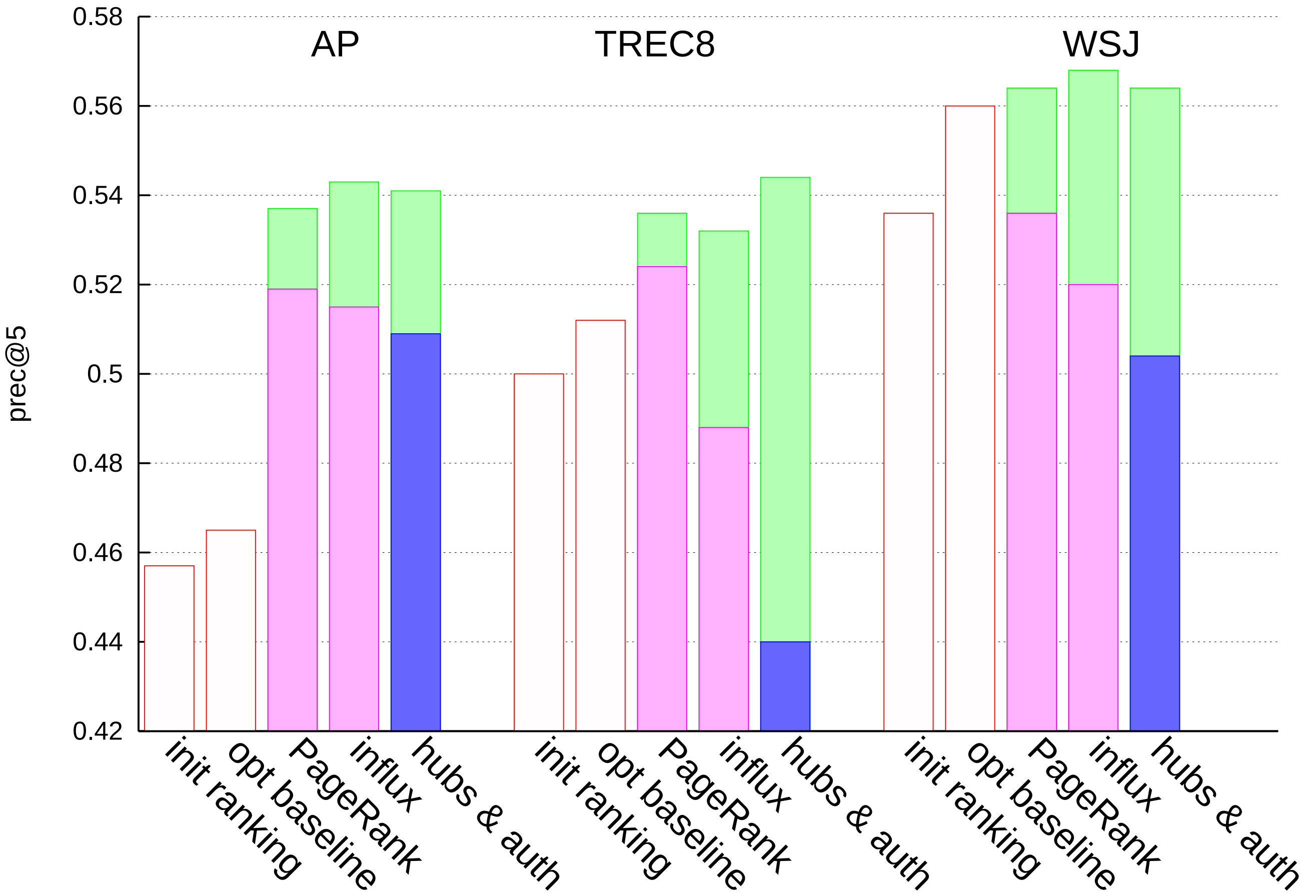
Document clusters as hubs

Recall the cluster hypothesis (van Rijsbergen ’79): closely associated documents tend to be relevant to the same requests.

Idea: “good” *clusters* of documents are those that contain relevant documents; and relevant documents are those contained in “good” clusters.

Construction: one-way bipartite graphs with *overlapping* clusters on the left, documents on the right, and links based on language models

Hubs-and-authorities comparison using clusters



There are stronger results for hubs and authorities, but the set-up is hard to explain.

Related work

Everyone already knew that graphs are useful ...

Most specifically related to what we have just discussed:

- Applications of PageRank and hubs-and-authorities to non-hyperlinked entities of the same type (documents, sentences) (Erkan & Radev ’04, Mihalcea & Tarau ’04, Mihalcea & Tarau ’05, Otterbacher, Erkan & Radev ’05, Zhang et al. ’05, Feng, Shaw, Kim & Hovy ’06, ...)
- Explicit use of mutually-reinforcing entities: terms and documents, words and context, queries and URLs, etc. (Dhillon ’01, Karov & Edelman ’98, Beeferman & Berger ’00, ...)
- Query-dependent clustering to directly improve ad hoc retrieval (Preece ’73, Leuski ’01, Tombros, Villa & van Rijsbergen ’02, Liu & Croft ’04, Kurland ’06, ...)

Get out the vote:
Pro-vs.-con classification of Congressional speeches

Matt Thomas (B.S. Cornell ’06),

Bo Pang (Ph.D. Cornell ’06),

Lillian Lee (B.A. Cornell ’93, and/but still here)

EMNLP 2006

Only connect the prose and the passion and both will be exalted.

— E.M. Forster, *Howards End*

Hype(r) links

The on-line availability of politically-oriented documents, both official (e.g., full text of laws) and non-official (e.g., blogs), means ...

The “[alteration of] the citizen-government relationship” (Shulman & Schlosberg 2002)

“The transformation of American politics” (*The New York Times*, 2006)

“The End of News?” (*The New York Review of Books*, 2005)

An opportunity for NLP?

Hype(r) links

The on-line availability of politically-oriented documents, both official (e.g., full text of laws) and non-official (e.g., blogs), means ...

The “[alteration of] the citizen-government relationship” (Shulman & Schlosberg 2002)

“The transformation of American politics” (*The New York Times*, 2006)

“The End of News?” (*The New York Review of Books*, 2005)

An opportunity for NLP?

> One ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end.

Hype(r) links

The on-line availability of politically-oriented documents, both official (e.g., full text of laws) and non-official (e.g., blogs), means ...

The “[alteration of] the citizen-government relationship” (Shulman & Schlosberg 2002)

“The transformation of American politics” (*The New York Times*, 2006)

“The End of News?” (*The New York Review of Books*, 2005)

An opportunity for NLP?

> One ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end.
>
> — George Orwell, “Politics and the English language”, 1946

NLP for opinionated politically-oriented language

Sentiment analysis — a hot NLP research area focused on subjective or opinion-oriented language (Pang & Lee, book-length survey, summer '08) — applied to this domain can enable:

- *eRulemaking*: the "electronic collection, distribution, synthesis, and analysis of public commentary in the regulatory rulemaking process" (Shulman & Schlosberg '02)
- the summarization of un-solicited commentary and evaluative statements, such as editorials, speeches, and blog posts

  (these may contain complex language, but not as complex as in the legislative proposals themselves ...)

Our task

Given: transcripts of Congressional floor debates

Goal: classify each *speech segment* (uninterrupted sequence of utterances by a single speaker) as supporting or opposing the proposed legislation

Important characteristics:

1. Ground-truth labels can be determined automatically (speaker votes)
2. Very wide range of topics: flag burning, the U.N., "Recognizing the 30th anniversary of the victory of United States winemakers at the 1976 Paris Wine Tasting"
3. Presentation of evidence rather than opinion

   "Our flag is sacred!": is it pro-ban or contra-ban-revocation?

4. Discussion context: some speech segments are responses to others

Using discussion structure

Two sources of information (details suppressed):

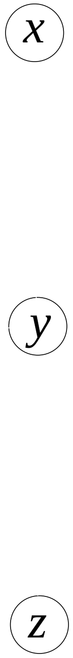
- An individual-document classifier that scores each speech segment x in isolation
- An *agreement classifier* for *pairs* of speech segments, trained to score by-name references (e.g., “I believe Mr. Smith’s argument is persuasive”) as to how much they indicate agreement

Optimization problem: find a classification c that minimizes:

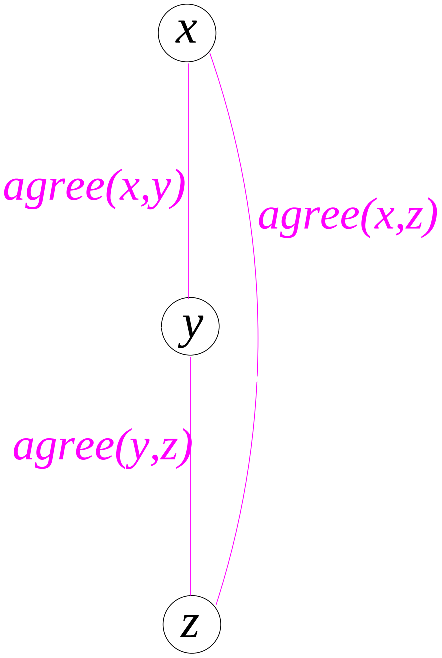
$$\sum_{x} \textit{ind}(x, \overline{c}(x)) + \sum_{x,x' :\, c(x) \neq c(x')} \textit{agree}(x, x')$$

(the items’ desire to switch classes due to individual or associational preferences)

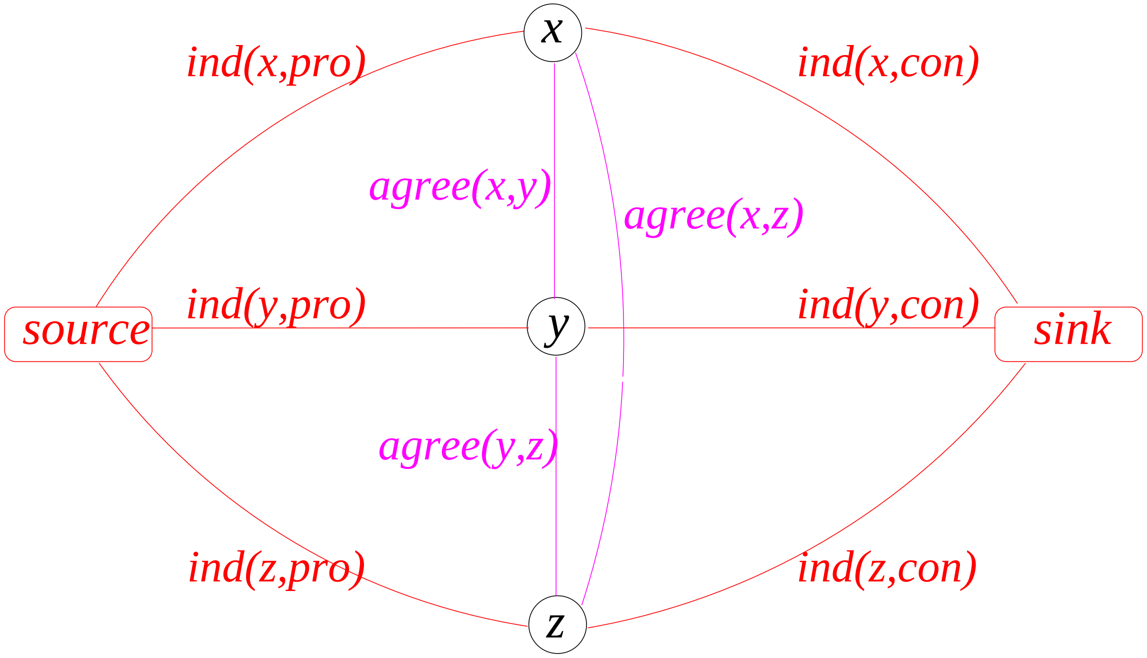
A “mitosis” encoding



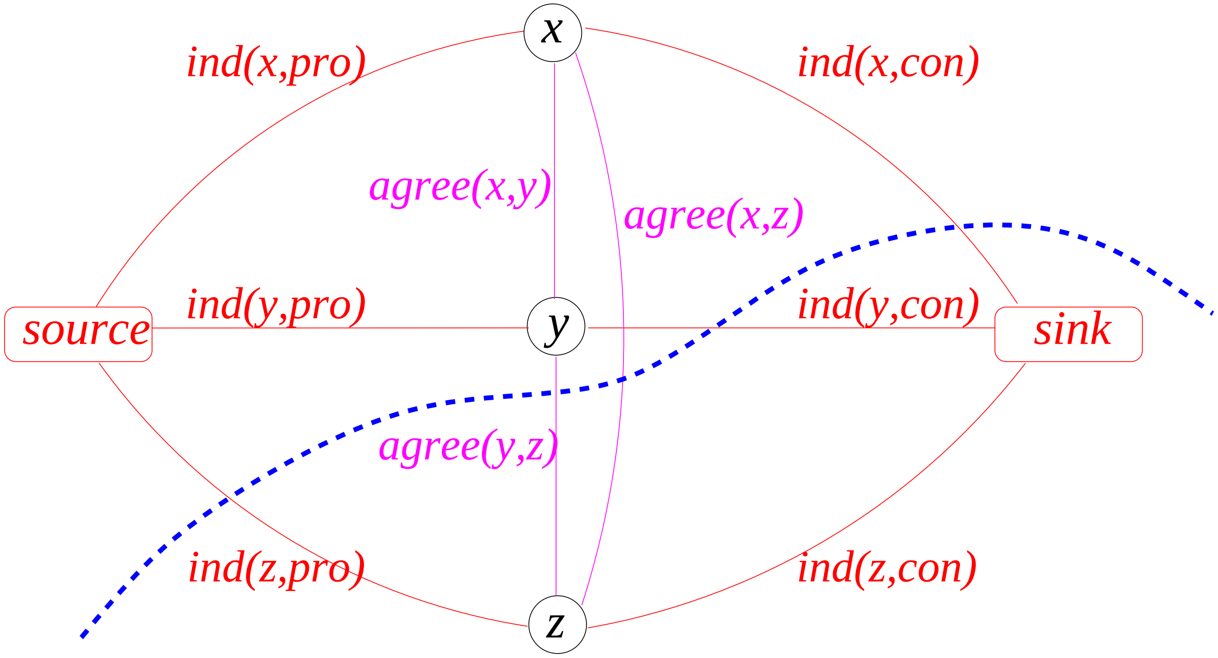
A “mitosis” encoding



A “mitosis” encoding



A “mitosis” encoding



The cost of a source/sink cut is the sum of the weights of the links that it breaks, and is thus equal to the value of our optimization function for the corresponding classification [Greig, Porteous & Seheult ’89].

Solution via max fbw

When the edge weights are non-negative, network-flow techniques find the min-cost cut, and hence solve our optimization problem ...

exactly (Greig, Porteous & Seheult ’89), and

efficiently, both in theory and practice (Ahuja, Magnanti & Orlin ’93)

Previous uses in NLP: sentiment analysis and generation (Pang & Lee ’04, Agarwal & Bhattacharrya ’05, Barzilay & Lapata ’05)

Previous uses in other areas: transductive learning, vision, computational biology, Web mining (Blum & Chawla ’01; Greig, Porteous & Seheult ’89, Boykov, Veksler, and Zabih ’99; Gupta & Tardos ’00, Kleinberg ’99, Xu, Xu, & Gabow ’00, Aspnes et al. 01; Flake, Lawrence & Giles ’00; ...)

Sketch: one set of experiments

Corpus: 53 “controversial” debates = 3857 speech segments from GovTrack, split into train, test and development sets, preserving debate integrity.

Available at www.cs.cornell.edu/home/llee/data/convote.html

Support/oppose classifier accuracy	Test
majority-class baseline	58.37
#(“support”) − #(“oppos”)	62.67
SVM [speaker]	70.00
SVM, concatenating “agreeing” speakers (no graph)	64.07
SVM with weighted agreement links	**76.16**

Using extra-textual info like political affiliation would boost performance even more, but our focus is on the NLP aspects of the problem.

Related work

Sentiment analysis on politically oriented text dealing with eRulemaking or determining political “leaning” (Laver et al.’03, Efron ’04, Grefenstette, Qu, Shanahan & Evans ’04, Shulman, Callan, Hovy, & Zavestoski ’05, Cardie et al. ’06, Lin, Wilson, Wiebe & Hauptmann ’06, Kwon, Shulman & Hovy ’06, Mullen & Malouf ’06, Hopkins and King ’08, Lerman, Gilder, Dredze & Pereira ’08...)

Using discussion structure for other types of classification (Carvalho & Cohen ’05, Feng, Shaw, Kim & Hovy ’06)

Collective classification using less tractable formulations, e.g., relational and associative Markov networks, max-margin Markov networks, conditional random fields, max cut, etc. (Neville & Jensen ’00, Lafferty, McCallum & Pereira ’01, Getoor, Friedman, Koller & Taskar ’02, Tasker et al. ’02, ’03, ’04, Agrawal et al. ’03, McCallum & Wellner ’04, ...). See chapter 6 of Zhu ’05 for a survey of graph-based semi-supervised learning.

Hour’s End

Summary:

- PageRank and hubs-and-authorities on non-hyperlinked corpora
- Classification of speeches in Congressional floor debates using techniques for finding minimum cuts in graphs

Moral: graph techniques work better on sensibly constructed graphs.

Examples in this work:

- directed endorsement links based on language models
- use of both documents and clusters, especially for hubs and authorities
- use of discussion structure such as agreement clues

Hour's End

Summary:

- PageRank and hubs-and-authorities on non-hyperlinked corpora
- Classification of speeches in Congressional floor debates using techniques for finding minimum cuts in graphs

Moral: graph techniques work better on sensibly constructed graphs.

Examples in this work:

- directed endorsement links based on language models
- use of both documents and clusters, especially for hubs and authorities
- use of discussion structure such as agreement clues

Only connect (correctly)! That is the end of my sermon.