# Learning Structure in Unstructured Document Bases

David Cohn
Burning Glass Technologies and
CMU Robotics Institute

www.cs.cmu.edu/~cohn

*Joint work with: Adam Berger, Rich Caruana, Huan Chang, Dayne Frietag, Thomas Hofmann, Andrew McCallum, Vibhu Mittal and Greg Schohn*

# Documents, documents everywhere!

Revelation #1: There are Too Many Documents

- email archives
- research paper collections
- the w... w... Web

*Response #1: Get over it – they're not going away*

Revelation #2: Existing Tools for Managing Document Collections are Woefully Inadequate

*Response #2: So what are you going to do about it?*

# The goal of this research

- Building tools for learning, manipulating and navigating the structure of document collections

- Some preliminaries:
  - What’s a document collection?
    - an arbitrary collection of documents
  - Okay, what’s a document?
    - text documents
    - less obvious: audio, video records
    - even less obvious: financial transaction records, sensor streams, clickstreams
  - What’s the point of a document collection?
    - they make it easy to find information (in principle...)

# Finding information in document collections

- Search engines – Google
  - studied by Information Retrieval community
  - canonical question - *"can you find me more like this one?"*
- Hierarchies – Yahoo
  - canonical question: *"where does this fit in the big picture?"*
- Hypertext – the rest of us
  - canonical question - *"what is this related to?"*

# What's wrong with hierarchies/hyperlinks?

- Lots of things!
  - manually created – time consuming
  - limited scope – author's access/awareness
  - static – become obsolete as corpus changes
  - subjective – but for wrong subject!
- What would we like? Navigable structure in a dynamic document base that is
  - automatic - generated with minimal human intervention
  - global - operates on all documents *we* have available
  - dynamic - accommodates new and stale documents as they arrive and disappear
  - personalized - incorporates our preferences and priors

# What are we going to do about it?

- Learn the structure of a document collection using
  - **unsupervised learning**
    - factor analysis/latent variable modeling to identify and map out latent structure in document base
  - **semi-supervised learning**
    - to adapt structure to match user's perception of world
- Caveats:
  - Very Big Problem
  - Warning: work in progress!
  - No idea what user interface should be

- *A few pieces of the large jigsaw puzzle...*

# Outline

- Text analysis background – structure from document contents
  - vector space models, LSA, PLSA
  - factoring vs. clustering
- Bibliometrics – structure from document connections
  - everything old is new again: ACA, HITS
  - probabilistic bibliometrics
- Putting it all together
  - a joint probabilistic model for document content and connections
  - what we can do with it

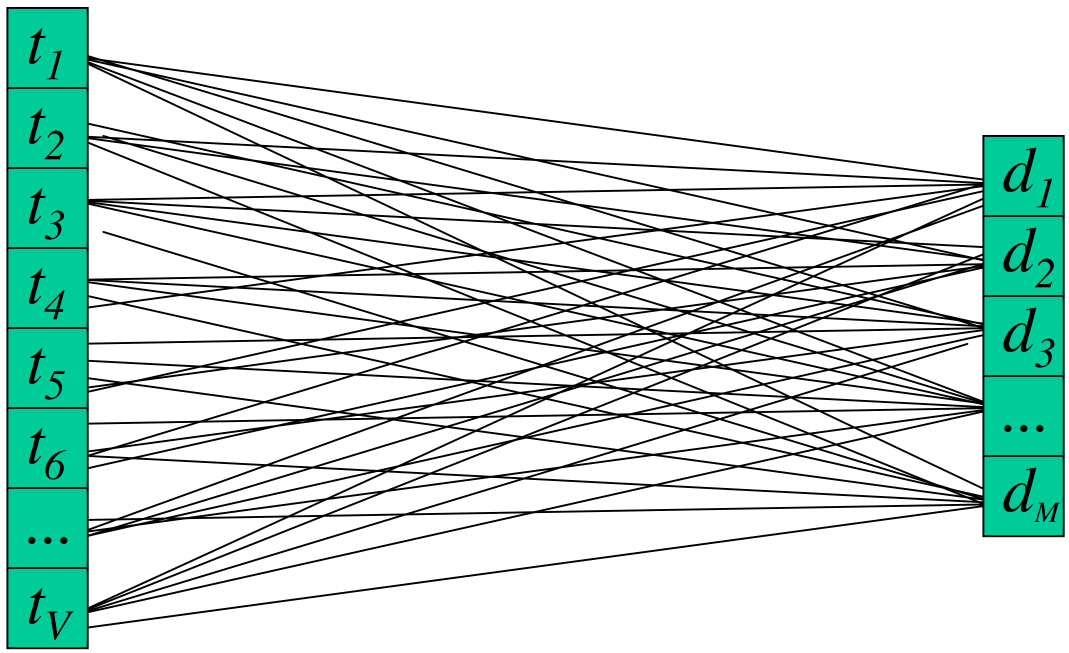
# Quick introduction to text modeling

- Begin with vector space representation of documents:
- Each word/phrase in vocabulary $V$ assigned term id $t_1, t_2, \ldots t_{|V|}$
- Each document $d_j$ represented as vector of (weighted) counts of terms
- Corpus represented as term-by-document matrix $N$

| $N$ | $d_1$ | $d_2$ | … | $d_m$ |
|---|---|---|---|---|
| $t_1$ | *0* | *1* | *4* | *2* |
| $t_2$ | *1* | *0* | *1* | *0* |
| $t_3$ | *1* | *0* | *0* | *1* |
| $t_4$ | *3* | *1* | *0* | *0* |
| $t_5$ | *1* | *0* | *1* | *4* |
| $t_6$ | *0* | *2* | *0* | *1* |
| … | *1* | *0* | *0* | *0* |
| | *0* | *1* | *0* | *1* |
| $t_v$ | *5* | *0* | *0* | *0* |

# Statistical text modeling

- Can compute raw statistical properties of corpus
  - use for retrieval, clustering, classification



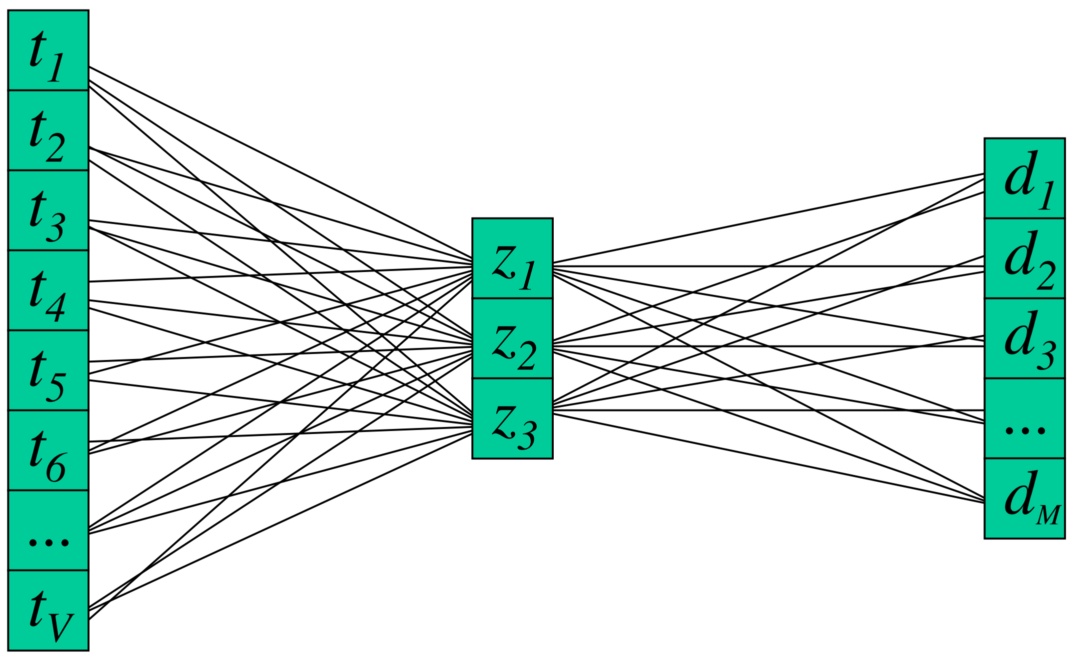

$p(t_i|d_j)$ $p(d_j|t_i)$

# Limitations of the VSM

- Word frequencies aren’t the whole story
  - Polysemy
    - *“a sharp increase in rates on bank notes”*
    - *“the pilot notes a sharp increase in bank”*
  - Synonymy
    - *“Bob/Robert/Bobby spilled pop/soda/Coke/Pepsi on the couch/sofa/loveseat”*
  - Conceptual linkage
    - *“Alan Greenspan”* ⇒ *“Federal Reserve”*, *“interest rates”*
- *Something else is going on...*

# Statistical text modeling

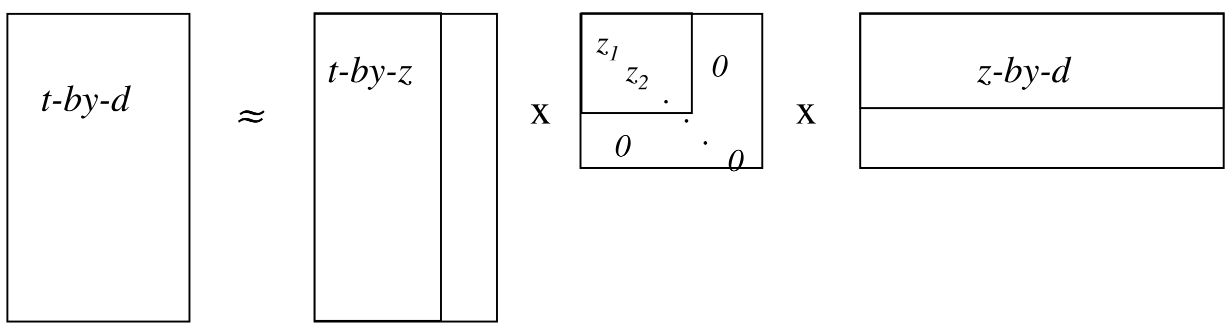
- Hypothesis: ***There's structure out there***
  - all documents can be "explained" in terms of a (relatively) small number of underlying "concepts"




$\boldsymbol{p(t_i|z_k)}$ $\boldsymbol{p(z_k)}$ $\boldsymbol{p(d_j|z_k)}$

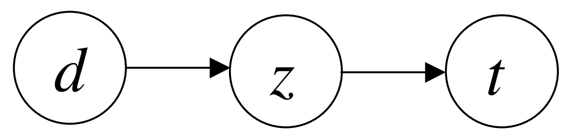
# Latent semantic analysis

- Perform singular value decomposition on term-by-document matrix *[Deerwester et al., 1990]*
  - truncated eigenvalue matrix gives reduced subspace representation
    - minimum distortion reconstruction of *t-by-d* matrix
    - minimizes distortion by exploiting term co-occurences




*Empirically, produces big improvement in retrieval, clustering*

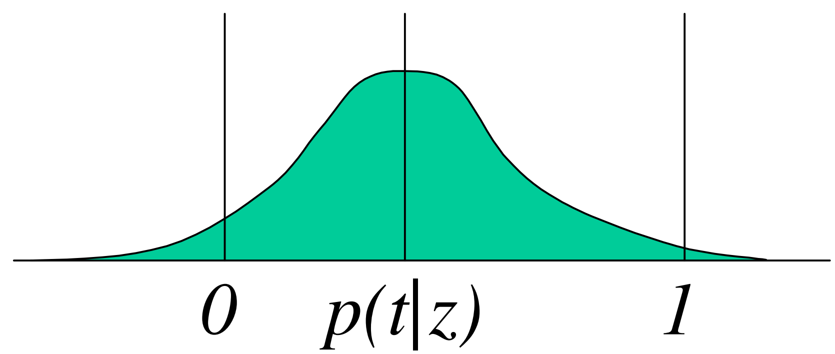
# Statistical interpretation of LSA

- LSA is performing linear factor analysis
  - each term and document maps to a point in z-space (via t-by-z' and z'-by-d matrices)
- Modeled as a Bayes net: 


  - select document $d_i$ to be created according to $p(d_i)$
  - pick mixture of factors $z_1 \ldots z_k$ according to $p(z_1 \ldots z_k | d_i)$
  - pick terms for $d_i$ according to $p(t_j | z_1 \ldots z_k)$
  - Singular value decomposition finds factors $z_1 \ldots z_k$ that "best explain" observed term-document matrix

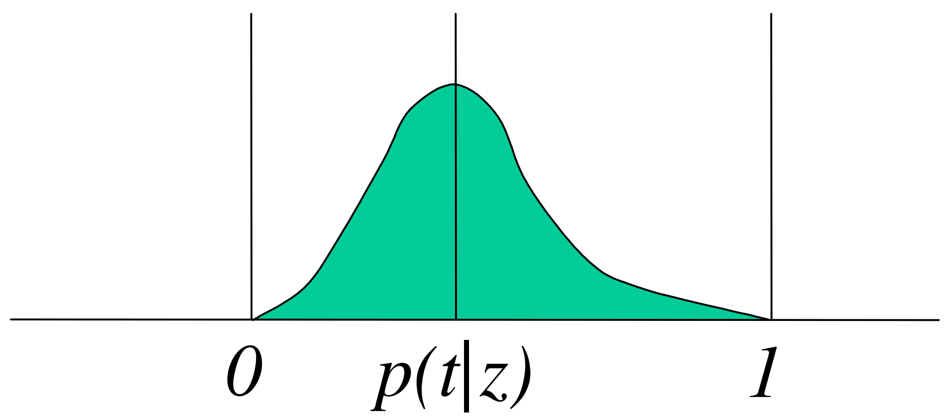
# LSA - what's wrong?

- LSA minimizes "distortion" of t-by-d matrix
  - corresponds to maximizing data likelihood assuming *Gaussian* variation in term frequencies
  - modeled term frequencies may be less than zero or greater than 1!



# Factoring methods - PLSA

- Probabilistic Latent Semantic Analysis (Hofmann, '99)
  - uses multinomial to model observed variations in term frequency
  - corresponds to generating documents by sampling from a "bag of words"



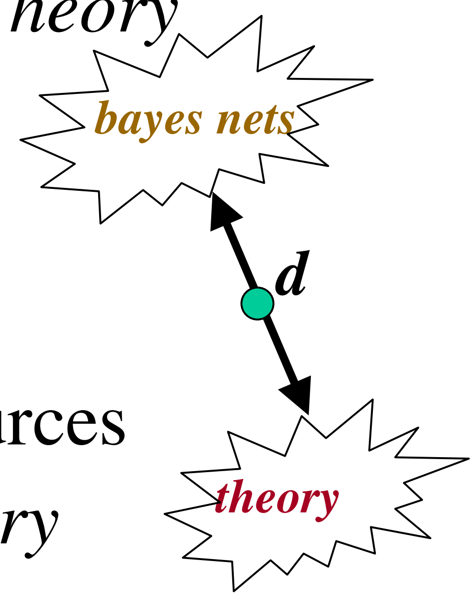
# Factoring methods - PLSA

- Perform explicit factor analysis using EM
  - estimate factors: $$p(z_k \mid t_i, d_j) = \frac{p(t_i \mid z_k)\, p(z_k \mid d_j)}{p(t_i \mid d_j)}$$
  - maximize likelihood: $$p(t_i \mid z_k) = \sum_j \frac{N_{ij}}{\sum_{i'} N_{i'j}} p(z_k \mid t_i, d_j)$$
  $$p(z_k \mid d_j) \propto \sum_i \frac{N_{ij}}{\sum_{i'} N_{i'j}} p(z_k \mid t_i, d_j)$$

- Advantages
  - solid probabilistic foundation for reasoning about document contents
  - seems to outperform LSA in many domains

# Digression: Clusters vs. Factors

- Clustered model
  - each document comes from *one* of the underlying sources
  - ***d*** is either a *Bayes net* paper or a *Theory* paper
- Factored model
  - each document comes from linear combination of the underlying sources
  - ***d*** is 50% *Bayes net* and 50% *Theory*



# Using latent variable models

- Empirically, factors correspond well to categories that can be verbalized by users
  - can use dominant factors as clusters (spectral clustering)
  - can use factoring as front end to clustering algorithm
    - cluster using document distance in *z* space
    - factors tell how they differ
    - clusters tell how they clump
  - or use multidimensional scaling to visualize relationship in factor space

| | |
|---|---|
| [**0.642** 0.100 0.066 0.079 0.114] | business-commodities |
| [**0.625** 0.068 0.055 0.126 0.125] | business-dollar |
| [**0.619** 0.059 0.098 0.122 0.102] | business-fed |
| [0.052 **0.706** 0.108 0.071 0.063] | sports-nbaperson |
| [0.093 **0.576** 0.097 0.105 0.129] | sports-ncaadavenport |
| [0.075 **0.677** 0.053 0.100 0.095] | sports-nflkennedy |
| [0.065 0.084 **0.660** 0.099 0.093] | health-aha |
| [0.059 0.124 **0.648** 0.088 0.081] | health-benefits |
| [0.052 0.073 **0.700** 0.081 0.094] | health-clues |
| [0.056 0.064 0.045 **0.741** 0.094] | politics-hillary |
| [0.047 0.068 0.062 **0.741** 0.082] | politics-jones |
| [0.116 0.159 0.125 **0.463** 0.136] | politics-miami |
| [0.078 0.062 0.045 0.170 **0.645**] | politics-iraq |
| [0.107 0.079 0.068 0.099 **0.646**] | politics-pentagon |
| [0.058 0.090 0.055 0.139 **0.659**] | politics-trade |

# Structure within the factored model

- Can measure similarity, but there's more to structure than similarity
  - *Given a cluster of 23,015 documents on learning theory, which one should we look at?*
- Other relationships
  - authority on topic
  - representative of topic
  - connection to other members of topic

# Quick introduction to bibliometrics

- Bibliometrics: a set of mathematical techniques for identifying citation patterns in a collection of documents
- Author co-citation analysis (ACA) - 1963
  - identifies principal topics of collection
  - identifies authoritative authors/documents in each topic
- Resurgence of interest with application to web
  - Hypertext-Induced Topic Selection (HITS) - 1997
  - useful for sorting through deluge of pages from search engines

# ACA/HITS – how it works

- Authority as a function of citation statistics
  - the more documents cite document *d*, the more authoritative *d* is.
  - the more authoritative *d* is, the more authority its citations convey to other documents

- Formally
  - matrix *A* summarizes citation statistics
  - element $a_i$ of vector $a$ indicates authority of $d_i$
  - authority is linear function of citation count and authority of citer: $a = A'Aa$
  - solutions are eigenvectors of $A'A$

| *A* | $d_1$ | $d_2$ | … | $d_m$ |
|---|---|---|---|---|
| $c_1$ | *0* | *1* | *1* | *1* |
| $c_2$ | *1* | *0* | *1* | *0* |
| $c_3$ | *1* | *0* | *0* | *1* |
| $c_4$ | *1* | *1* | *0* | *0* |
| $c_5$ | *1* | *0* | *1* | *1* |
| $c_6$ | *0* | *1* | *0* | *1* |
| … | *1* | *0* | *0* | *0* |
| | *0* | *1* | *0* | *1* |
| $c_m$ | *1* | *0* | *0* | *0* |

# Let's try it out on something we know...

- Cora's *Machine Learning* subtree
  - 2093 categorized into machine learning theory, neural networks, rule learning, probabilistic models, genetic algorithms, reinforcement learning, case-based learning


- Question #1: can we reconstruct ML topics from citation structure?
  - citation structure independent of text used for initial classification
- Question #2: Can we identify authoritative papers in each topic?

# ACA authority - *Cora* citations

**eigenvector 1 (Genetic Algorithms)**
0.0492 How genetic algorithms work: A critical look at implicit parallelism. Grefenstette
0.0490 A theory and methodology of inductive learning. Michalski
0.0473 Co-evolving parasites improve simulated evolution as an optimization procedure. Hills
**eigenvector 2 (Genetic Algorithms)**
0.00295 Induction of finite automata by genetic algorithms. Zhou et al
0.00295 Implementation of massively parallel genetic algorithm on the MasPar MP-1. Logar et al
0.00294 Genetic programming: A new paradigm for control and analysis. Hampo
**eigenvector 3 (Reinforcement Learning/Genetic Algorithms)**
0.256 Learning to predict by the methods of temporal differences. Sutton
0.238 Genetic Algorithms in Search, Optimization, and Machine Learning. Angeline et al
0.178 Adaptation in Natural and Artificial Systems. Holland
**eigenvector 4 (Neural Networks)**
0.162 Learning internal representations by error propagation. Rumelhart et al
0.129 Pattern Recognition and Neural Networks. Lawrence et al
0.127 Self-Organization and Associative Memory. Hasselmo et al
**eigenvector 5 (Rule Learning)**
0.0828 Irrelevant features and the subset selection problem, Cohen et al
0.0721 Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Holte
0.0680 Classification and Regression Trees. Breiman et al
**eigenvector 6 (Rule Learning)**
0.130 Classification and Regression Trees. Breiman et al
0.0879 The CN2 induction algorithm. Clark et al
0.0751 Boolean Feature Discovery in Empirical Learning. Pagallo
**eigenvector 7 ([Classical Statistics?])**
1.5-132 Method of Least Squares. Gauss
1.5-132 The historical development of the Gauss linear model. Seal
1.5-132 A Treatise on the Adjustment of Observations. Wright

# ACA/HITS – why it (sort of) works

- Author Co-citation Analysis (ACA)
  - identify principal eigenvectors of co-citation matrix *A'A*, label as primary topics of corpus
- Hypertext Induced Topic Selection (HITS) – 1998
  - use eigenvalue iteration to identify principal "hubs" and "authorities" of a linked corpus

- *Both just doing factor analysis on link statistics*
  - *same as is done for text analysis*
- *Both are using Gaussian (wrong!) statistical model for variation in citation rates*

# Probabilistic bibliometrics (Cohn '00)

- Perform explicit factor analysis using EM
  - estimate factors: $$p(z_k \mid c_l, d_j) = \frac{p(c_l \mid z_k)\, p(z_k \mid d_j)}{p(c_l \mid d_j)}$$
  - maximize likelihood: $$p(c_l \mid z_k) = \sum_j \frac{A_{lj}}{\sum_{l'} A_{l'j}} p(z_k \mid c_l, d_j)$$
    $$p(z_k \mid d_j) \propto \sum_l \frac{A_{lj}}{\sum_{l'} A_{l'j}} p(z_k \mid c_l, d_j)$$

- Advantages
  - solid probabilistic foundation for reasoning about document connections
  - seems to frequently outperform HITS/ACA

# Probabilistic bibliometrics – *Cora* citations

**factor 1 (Reinforcement Learning)**
0.0108 Learning to predict by the methods of temporal differences. Sutton
0.0066 Neuronlike adaptive elements that can solve difficult learning control problems. Barto et al
0.0065 Practical Issues in Temporal Difference Learning. Tesauro.
**factor 2 (Rule Learning)**
0.0038 Explanation-based generalization: a unifying view. Mitchell et al
0.0037 Learning internal representations by error propagation. Rumelhart et al
0.0036 Explanation-Based Learning: An Alternative View. DeJong et al
**factor 3 (Neural Networks)**
0.0120 Learning internal representations by error propagation. Rumelhart et al
0.0061 Neural networks and the bias-variance dilemma. Geman et al
0.0049 The Cascade-Correlation learning architecture. Fahlman et al
**factor 4 (Theory)**
0.0093 Classification and Regression Trees. Breiman et al
0.0066 Learnability and the Vapnik-Chervonenkis dimension, Blumer et al
0.0055 Learning Quickly when Irrelevant Attributes Abound. Littlestone
**factor 5 (Probabilistic Reasoning)**
0.0118 Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Pearl.
0.0094 Maximum likelihood from incomplete data via the em algorithm. Dempster et al
0.0056 Local computations with probabilities on graphical structures... Lauritzen et al
**factor 6 (Genetic Algorithms)**
0.0157 Genetic Algorithms in Search, Optimization, and Machine Learning. Goldberg
0.0132 Adaptation in Natural and Artificial Systems. Holland
0.0096 Genetic Programming: On the Programming of Computers by Means of Natural Selection. Koza
**factor 7 (Logic)**
0.0063 Efficient induction of logic programs. Muggleton et al
0.0054 Learning logical definitions from relations. Quinlan.
0.0033 Inductive Logic Programming Techniques and Applications. Lavrac et al

more...

# Tools for understanding a collection

Text analysis

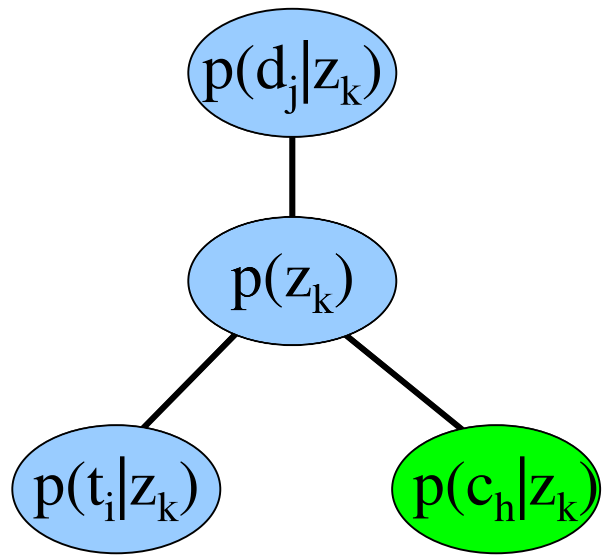
- what is the topic of this document?
- what other documents are there on this topic?
- what *are* the topics in this collection?
- how are they related?
- are there better documents on this topic?

Link analysis

# But can they play together?

- Now have two independent, probabilistic document models with parallel formulation

**PLSA**




**PHITS**

*What happens if we put them in a room together and turn out the lights?*

# Joint Probabilistic Document Models

- Mathematically trivial to combine
  - one twist: model *inlinks c’* instead of *outlinks c*
  - perform explicit factor analysis using EM
  - estimate factors:

$$p(z_k \mid t_i, d_j) = \frac{p(t_i \mid z_k)\, p(z_k \mid d_j)}{p(t_i \mid d_j)}, \quad p(z_k \mid c'_l, d_j) = \frac{p(c'_l \mid z_k)\, p(z_k \mid d_j)}{p(c'_l \mid d_j)}$$

  - maximize likelihood:

$$p(t_i \mid z_k) = \sum_j \frac{N_{ij}}{\sum_{i'} N_{i'j}} p(z_k \mid t_i, d_j), \quad p(c'_l \mid z_k) = \sum_j \frac{A_{lj}}{\sum_{l'} A_{l'j}} p(z_k \mid c'_l, d_j)$$

  - combine with mixing parameter α

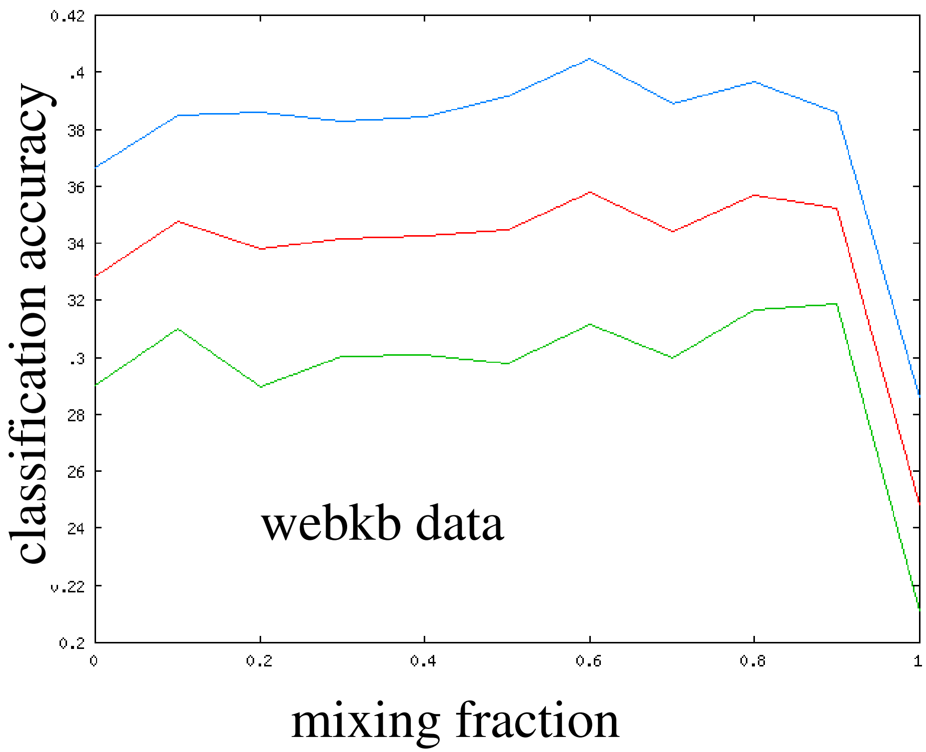
$$p(z_k \mid d_j) \propto \alpha \sum_i \frac{N_{ij}}{\sum_{i'} N_{i'j}} p(z_k \mid t_i, d_j) + (1-\alpha) \sum_l \frac{A_{lj}}{\sum_{l'} A_{l'j}} p(z_k \mid c'_l, d_j)$$

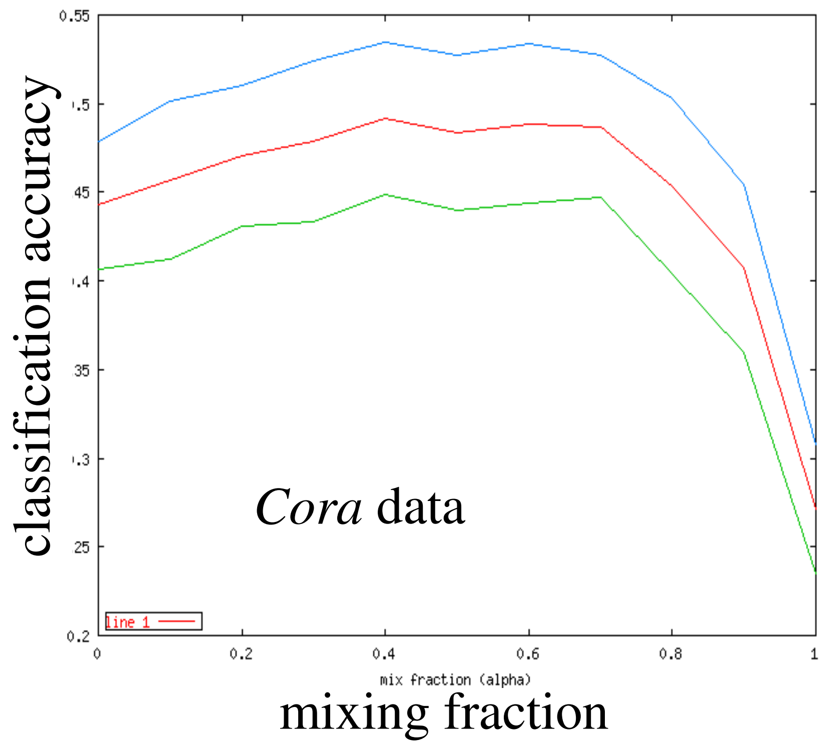
# Two domains

- WebKB data set from CMU
  - 8266 pages from Computer Science departments at US universities (6099 have both text and hyperlinks)
  - categorized by
    - source of page (cornell, washington, texas, wisconsin, other)
    - type of page (course, department, project, faculty, student, staff)
- Cora research paper archive
  - 34745 research papers and extracted references
  - 2093 categorized into machine learning hierarchy
    - theory, neural networks, rule learning, probabilistic models, genetic algorithms, reinforcement learning, case-based learning

# Classification accuracy

- Joint model improves classification accuracy
  - project into factor space, label according to nearest labeled example



# Qualitative document analysis

- What is factor $z$ "about"?

  $p(t|z)$ *[actually,* $p(t|z)^2/p(t)$*]*

•factor 1: class, homework, lecture, hours (**courses**)
•factor 2: systems, professor, university, computer (**faculty**)
•factor 3: system, data, project, group (**projects)**
•factor 4: page, home, computer, austin (**students/department**)
...

•factor 1: learning, reinforcement, neural
•factor 2: learning, networks, Bayesian
•factor 3: learning, programming, genetic
...

# Qualitative document analysis

- What is document *d* "about"? $\Sigma_k\, p(t|z_k)p(z_k|d)$

> •**Salton home page**: text, document, retrieval
> •**Robotics and Vision Lab page**: robotics, learning, robots, donald
> •**Advanced Database Systems course**: database, project, systems

- What *topics* is a document about?

**Factors for "TD Learning of Game Evaluation Functions with Hierarchical Neural Architectures," by M.A. Wiering:**

| | | |
|---|---|---|
| **0.566 Reinforcement Learning** | **0.239 Neural Networks** | **0.044 Logic** |
| **0.027 Rule Learning** | **0.026 Theory** | |
| **0.026 Probabilistic Reasoning** | **0.072 Genetic Algorithms** | |

# Qualitative document analysis

- How authoritative is a document in its field? $p(c_i|z_k)$
  (how likely is it to be cited from its principal topic?)

- **factor 1** Learning to predict by the methods of temporal differences. Sutton
- **factor 2** Explanation-based generalization: a unifying view. Mitchell et al
- **factor 3** Learning internal representations by error propagation. Rumelhart et al
- **factor 4** Classification and Regression Trees. Breiman et al
- **factor 5** Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Pearl.
- **factor 6** Genetic Algorithms in Search, Optimization, and Machine Learning. Goldberg
- **factor 7** Efficient induction of logic programs. Muggleton et al

# Qualitative document analysis

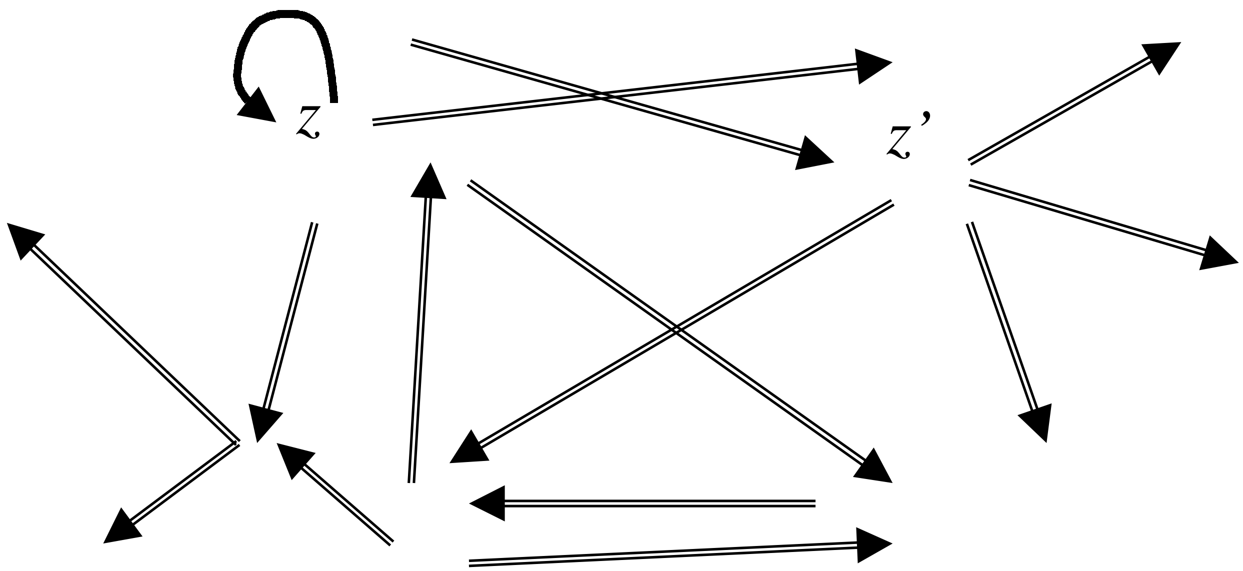
- Compute cross-factor authority
  - *"Which theory papers are most authoritative with respect to the Neural Network community?"*

$$\arg\max_c \, p(c \mid z = nn) : \frac{p(z = theory \mid c)}{\sum_z p(z \mid c)} > 0.9$$

("Decision Theoretic Generalizations of the PAC Model for Neural Net and other Learning Applications," by David Haussler)

# Analyzing document relationships

- How do these topics relate to each other?
  - words in document are signposts in factor space
  - links are a directed connection
    - between two documents
    - between two points in factor space

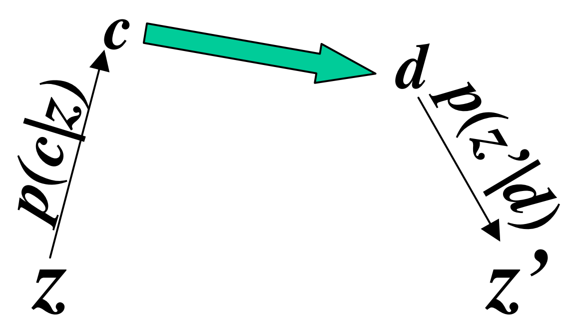


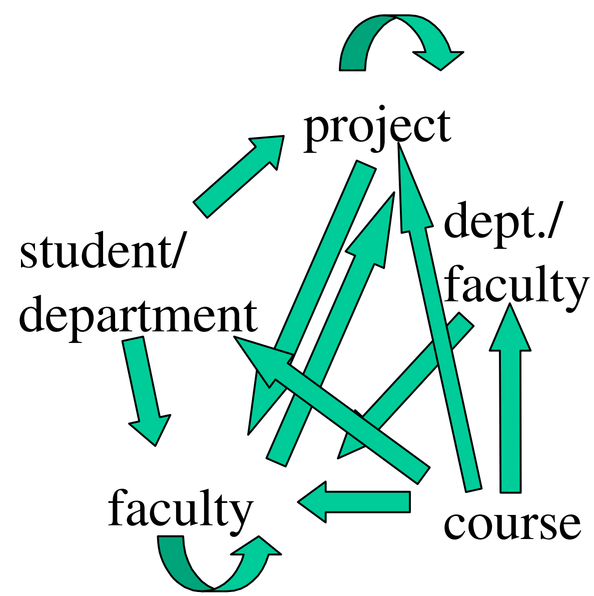
# Analyzing document relationships

- Each link can be evidence of reference between arbitrary points $z$ and $z'$ in topic space
- Integrate over all links to compute "reference flow" from $z$ to $z'$

$$f(z,z') = \sum_{links} p(c \mid z)\, p(z' \mid d)$$

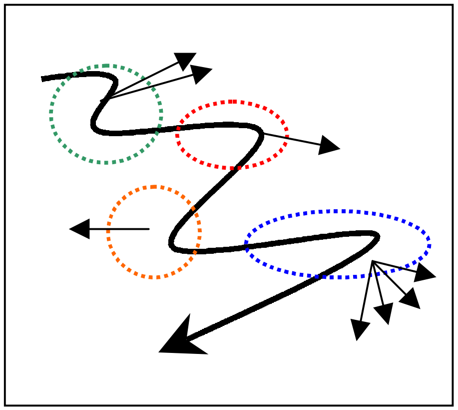
- Build a "Generalized Reference Map" over document space

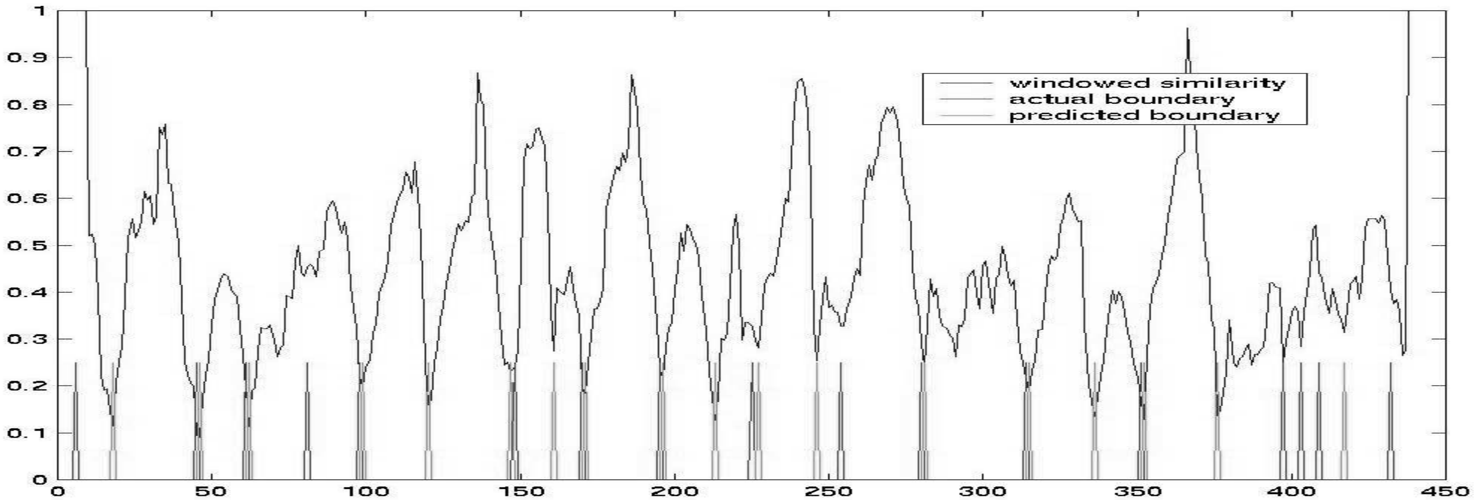


# One use: Intelligent spidering

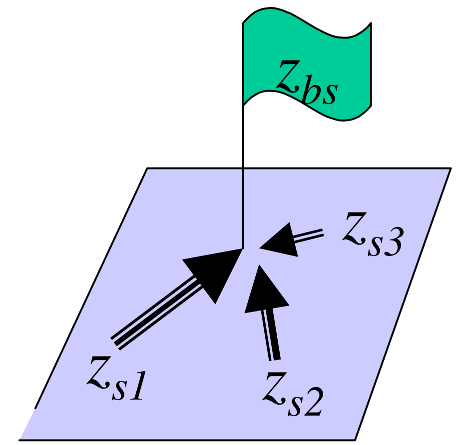
- Each document may cover many topics
  - follows trajectory through topic space
- Segment via factor projection
  - slide window over document, track trajectory of projection in factor space
  - segment at ‘jumps’ in factor space





# Intelligent spidering

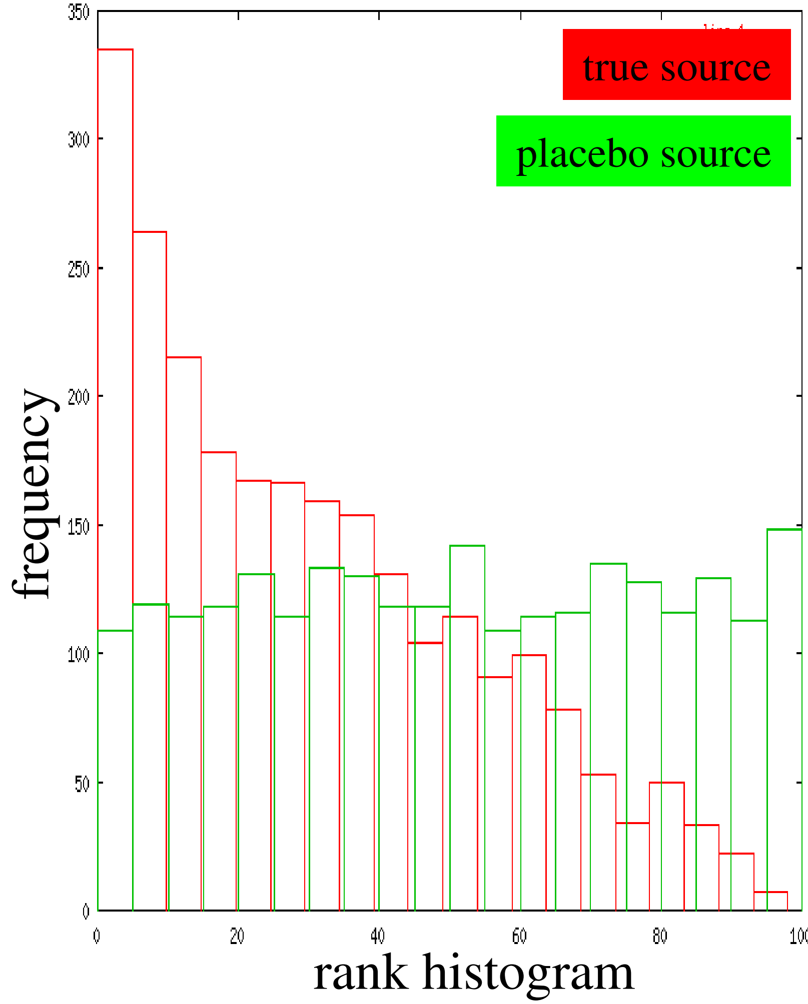
- Example: want to find documents containing phrase "Britney Spears"
  - Compute point $z_{bs}$ in factor space most likely to contain these words




  - Examine segments $s_1$, $s_2$, $s_3$... of current document, project them into factor space points: $z_{s1}$, $z_{s2}$, $z_{s3}$...
  - Compute reference flow $f(z_{si}, z_{bs})$ to determine which is most likely to contain transition to $z_{bs}$
  - Solve with greedy search, or
  - Continuous-space MDP, using normalized GRM for transition probabilities

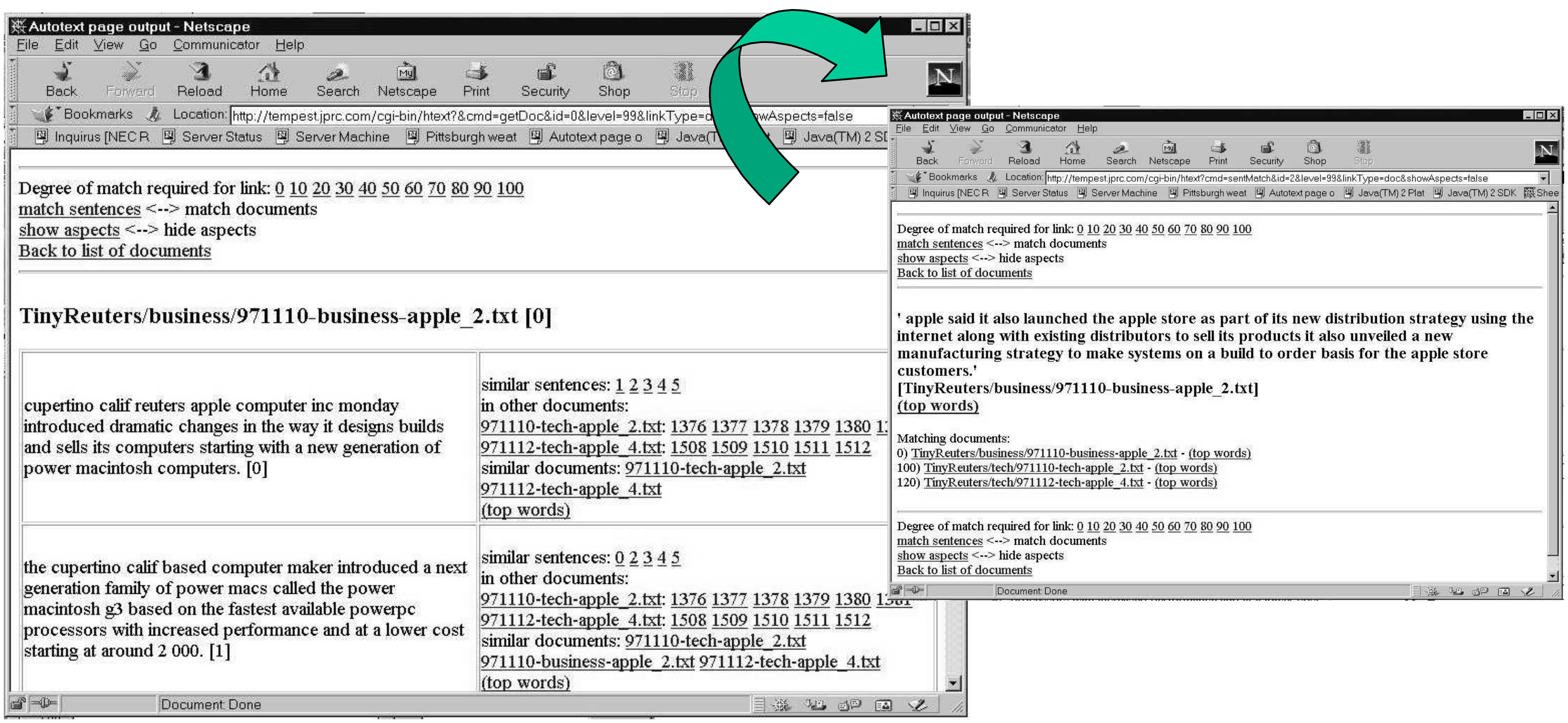
# Intelligent spidering

- WebKB experiments
  - choose target document at random
  - choose source document containing link to target
  - rank against 100 other "distractor" sources and a "placebo" source
    - median source rank: 27/100
    - median placebo rank: 50/100



# Another use: Dynamic hypertext generation

- Project and segment plaintext document
  - for each segment, identify documents in corpus most likely to be referenced



# Back to the big picture

- Recall that we wanted structure that was
  - ✓ automatic - learned with minimal human intervention
  - ✓ global - operates on all documents *we* have available
  - ✓ dynamic - accommodates new and stale documents as they arrive and disappear
  - personalized - incorporates our preferences and priors

    (subject of a different talk, on semi-supervised learning)
- What are we missing?
  - umm, any form of user interface?
  - a large-scale testbed (objective evaluation of structure and authority is notoriously tricky)

# Things I've glossed over

- Lack of factor orthogonality in probabilistic model
  - ICA-like variants?
- Sometimes you do only have one source/document
  - penalized factorizations
- Other forms of document bases
  - audio/visual streams
    - visual clustering, behavioral modeling *[Brand 98, Fisher 00]*
    - applications
      - nursebot, smart spaces
  - data streams
    - clickstreams
    - sensor logs
    - financial transaction logs

# The take-home message

- We need tools that let us learn, manipulate and navigate the structure of our ever-growing document bases
- Documents can’t be understood by contents or connections alone

**statistical document analysis**

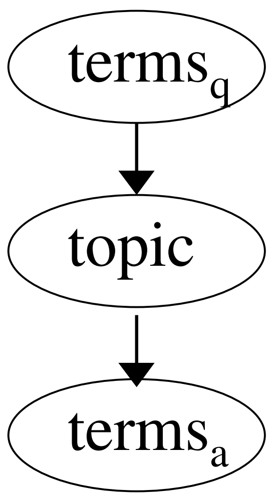
# Extra slides

# Application: What's wrong with IR?

- What we want: Ask a question, get an answer
- What we have: "Cargo cult" retrieval
  - imagine what answer would look like
  - build "cargo cult" model of answer document
    - guess words that might appear in answer
    - create pseudo-document from guessed words
  - select document that most resembles pseudo-document

# A machine learning approach to IR

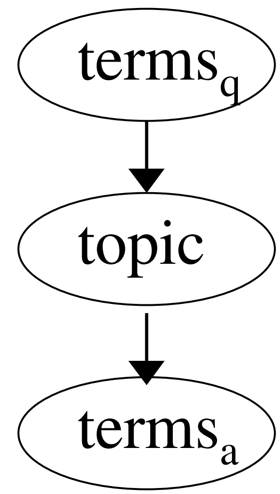
- Two distinct vocabularies: questions and answers
  - overlapping, but distinct
- Learn statistical map between them
  - question vocabulary ⇒ topic ⇒ answer vocabulary
  - build latent variable model of topic
  - learn mapping from matched Q/A pairs
    - USENET FAQ sheets
    - corporate call center document bases




- Given new question, want to find matching answer in FAQ

# A machine learning approach to IR

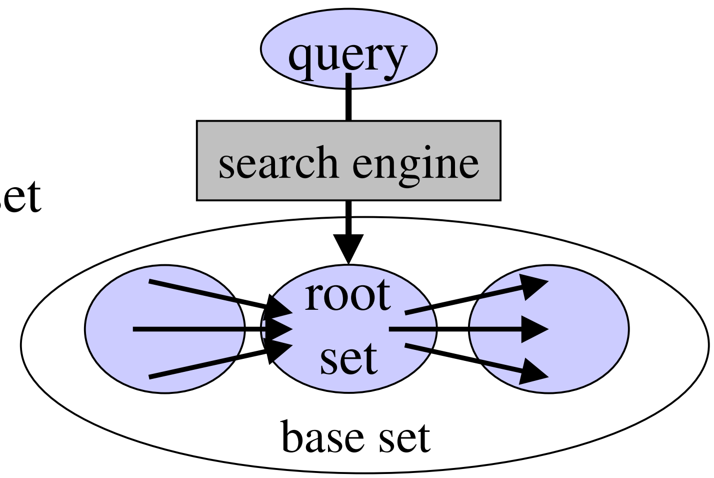
- Testing the approach:
  - take 90% of q/a pairs, build model
  - remaining 10% as test cases
    - map test question into pseudo-answer using latent variable model
    - retrieve answers closest to pseudo-answer, ranking according to tf-idf
  - score: mean and median rank of correct answer, averaged over 5 train/test splits




| db\rank | median TFIDF | median LVM | mean TFIDF | mean LVM |
|---|---|---|---|---|
| AirCanada | 10.8 (1.78) | 1.8 (0.22) | 86.4 (1.0) | 7.6 (0.68) |
| Ben&Jerry | 16 (1.46) | 5 (1.41) | 98.9 (4.1) | 25.0 (5.1) |
| USENET | 2 (0) | 2 (0) | 25.9 (2.9) | 3 (0.35) |

# PACA on web pages

- Given a query to a search engine, identify
  - principal topics matching query
  - authoritative documents in each topic
- Build co-citation matrix *M* following Kleinberg:
  - submit query to search engine
    - responses make up the "root set"
    - retrieve all pages pointed to by root set
    - retrieve all pages pointing to root set




- Example query "Jaguars"

# PACA on web pages

**ACA**
**eigenvector 1: 729.84**
0.224 www.gannett.com
0.224 homefinder.cincinnati.com
0.224 cincinnati.com/freetime/movies
0.224 autofinder.cincinnati.com

**eigenvector 2: 358.39**
0.0003 www.cmpnet.com
0.0003 www.networkcomputing.com
0.0002 www.techweb.com/news
0.0002 www.byte.com

**eigenvector 3: 294.25**
0.781 www.jaguarsnfl.com
0.381 www.nfl.com
0.343 jaguars.jacksonville.com
0.174 www.nfl.com/jaguars

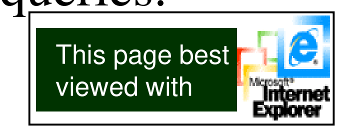
**PACA - sorted by p(c|z)**
**Factor 1**
0.0440 www.jaguarsnfl.com
0.0252 jaguars.jacksonville.com
0.0232 www.jag-lovers.org
0.0200 www.nfl.com
0.0167 www.jaguarcars.com

**Factor 2**
0.0367 www.jaguarsnfl.com
0.0233 www.jag-lovers.org
0.0210 jaguars.jacksonville.com
0.0201 www.nfl.com
0.0161 www.jaguarcars.com

# PACA on web pages

- Identifies authorities, but mixes principal topics
- What's going on?
  - web citations aren't as "intentional"
    - "most authoritative" page for many queries:

      `www.microsoft.com/ie`




  - components aren't orthogonal - data likelihood maximized by sharing some components

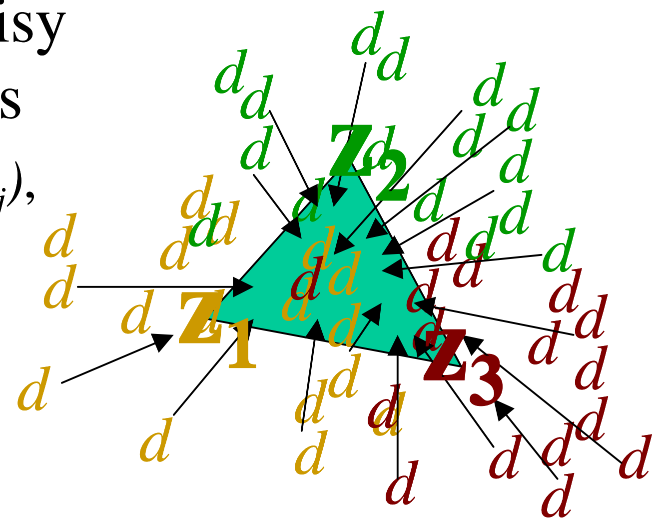
- In this case, clustered model *is* more appropriate than factored model

# Some thoughts...

- Win:
  - clear probabilistic interpretation
  - easily manipulated to estimate quantities of interest
  - authorities correspond well to human intuition
- Lose:
  - without enforced orthogonality, doesn't cleanly separate topics on web pages
  - requires specifying number of topics/factors *a priori*
    - ACA can extract successive orthogonal factors
- Draw: Computational costs approximately equivalent

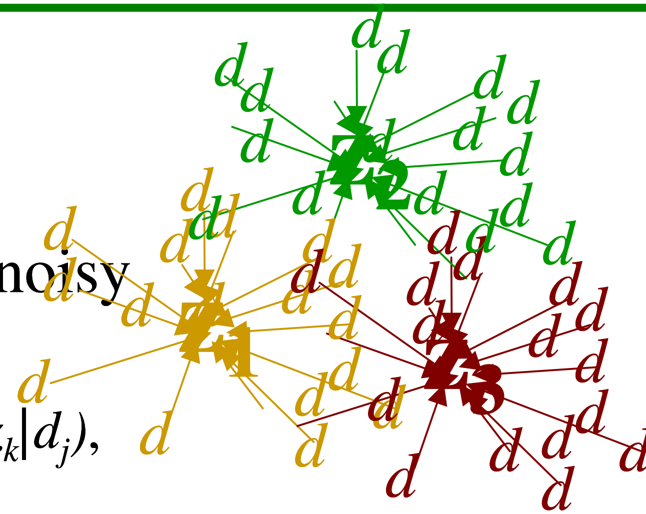
# Clustering vs. Factoring

- Factoring:
  - $z_k$ is a factor
  - each document assumed to be noisy instantiation of *mixture* of sources
    - select source with probability $p(z_k|d_j)$,
    - select one term in $d_j$ according to selected $z_k$, repeat
  - Arrange factors to minimize "distance" of data to hyperplane defined by $z_1 \ldots z_k$



# Clustering vs. Factoring

- Clustering:
  - $z_k$ is a prototype
  - each document assumed to be noisy instantiation of *one* source
    - select source with probability $p(z_k|d_j)$,
    - select all terms in $d_j$ according to selected $z_k$
  - Arrange prototypes to minimize "distance" of data to "nearest" $z_i$



# Manipulating structure

- Okay we’ve got structure - what if it doesn’t match the model inside our head?
  - clustering, bibliometric analysis are unsupervised
  - include some prior that may not match our own

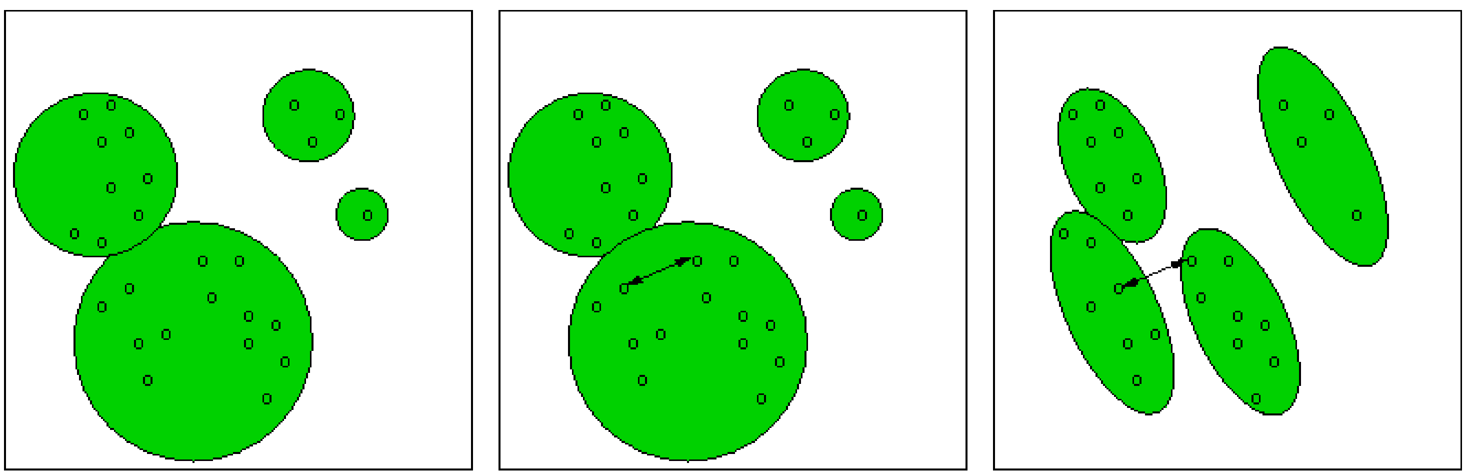
- Two approaches
  - labeled examples
    - supervised learning - absolute specification of categories
  - constraints
    - semi-supervised learning - relationships of examples

# Structure as "art"

- Labels and structure are frequently hard to generate
  - *"where should I file this email about phytoplankton?"*
- Easier to criticize than to construct
  - *"that document does **not** belong in this cluster!"*
- Forms of criticism
  - same/different clusters
  - good/bad cluster
  - more/less detail (here/everywhere)
  - many, many others...

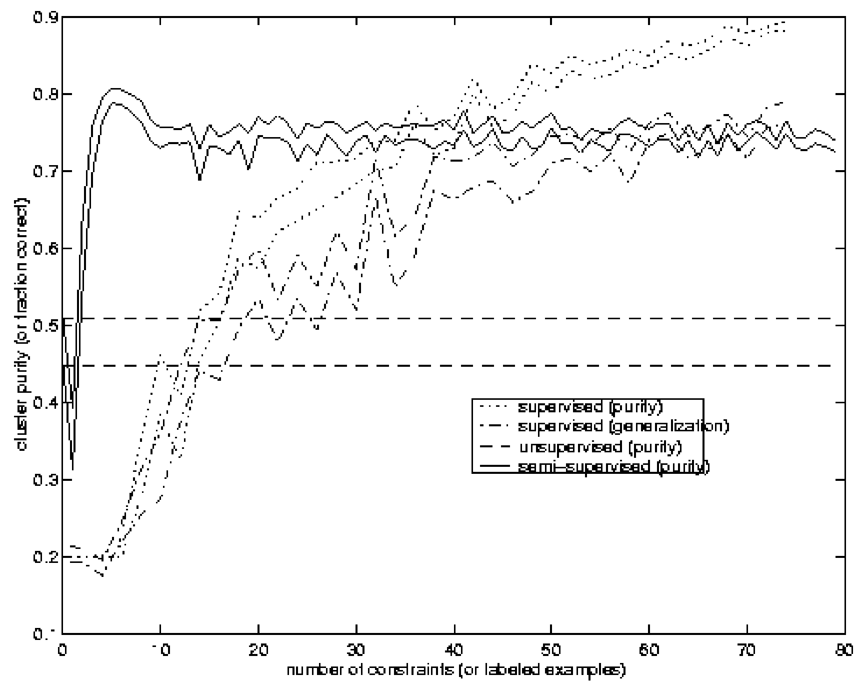
# Semi-supervised learning

- Semi-supervised learning:
  - derive structure
  - let user criticize structure
  - derive new structure that accommodates user criticism



# Semi-supervised learning - re-clustering

- Example, using mixture of multinomials
  - add separation constraints at random
  - use term reweighting, warping metric space to enforce constraints



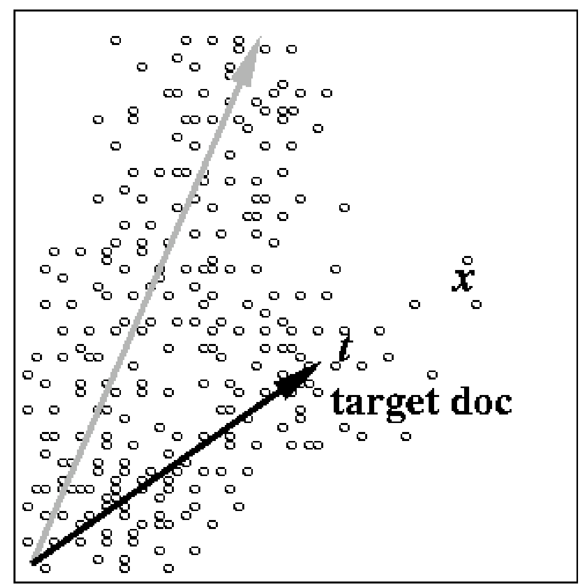

- Why is it so powerful?
  - equivalent to query by counterexample *[Angluin]*
    - user only adds constraints where something's broken

# Semi-supervised learning - realigning topics and authorities

- Given document set, may disagree with statistics on principal topics, authorities
  - want to give feedback to “correct” the statistics

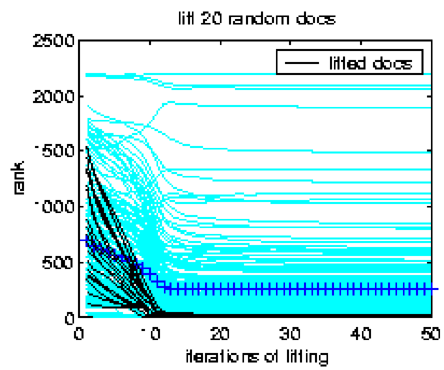
- HITS example
  - user feedback to realign principal eigenvectors
  - link matrix reweighting by gradient descent

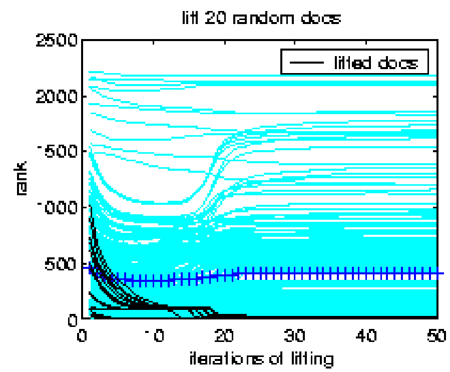
original eigenvector



# Semi-supervised learning - realigning topics and authorities

- Ex: learning “what’s *really* important in my field”
  - “lift” authoritative documents in one subfield, see how others react
  - cohesion of subfield




- Automatically creating customized authority lists
  - “lift” things you’ve cited/browsed, see what else is considered interesting