1. Probabilistic Author-Topic Models for Information Discovery

M. Steyvers, P. Smyth, M. Rosen-Zvi and T. Griffiths

2. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model

C. Chemudugunta, P. Smyth and M. Steyvers

Presented by Sophia Zhao 11/1/2007

Probabilistic Author-Topic Models for Information Discovery

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths

Proceedings Knowledge Discovery and Data Mining (KDD), 2004

Outline

- Introduction
- Related Work
- The Overview of the Author-Topic Model
- Applications of the Author-Topic Model to CiteSeer
- An Online Query Interface
- Conclusions

Introduction

- Abundant text resources
  - The Web and various digital libraries
- Automatic information extraction becomes important
- Supervised learning
  - Categorizing documents into known classes or topics
- Unsupervised learning
  - No predefined topics / labeled documents available
  - Uncover hidden topic structure

Related Work

1. Dimensionality Reduction
   - Representing the high-dimensional term vectors in a lower-dimensional space associated with specific topics
   - Examples:
     - Non-linear dimension reduction via self-organizing maps – WEBSOM system
     - Linear projection – Latent semantic indexing (LSI)
2. Document Clustering
   - Each cluster is associated with a latent topic
   - Each document is associated with one cluster

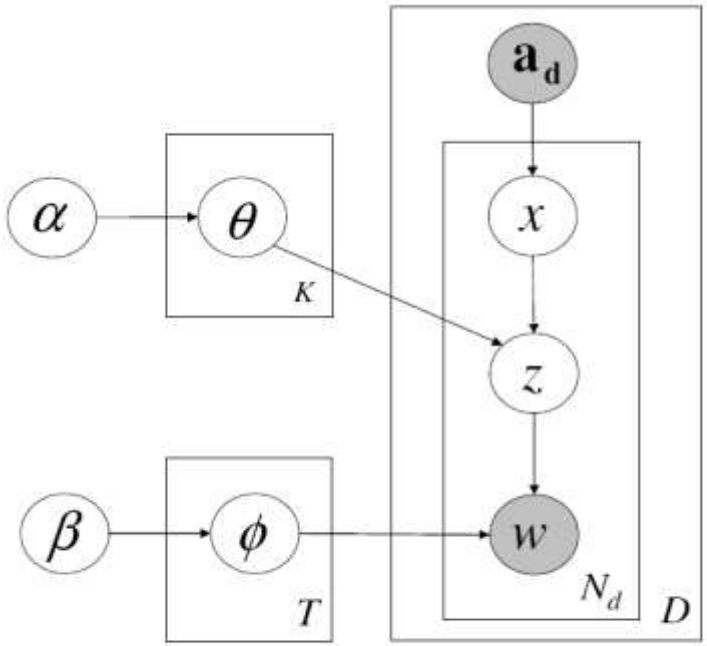
! Problem for multi-topic documents

Related Work (contd)

3. Probabilistic Topic Models
   - The aspect model (pLSI)
     - Topics ~ Multinomial probability distributions over words
     - Problem: parameter overfitting, new doc inference
   - Latent Dirichlet allocation (LDA)
     - Topic-word and document-topic distributions
   - Authorship information
     - Stylometry, authorship attribution and forensic linguistics – using stylistic features (sentence length, stop words, etc.)
     - This work – Extracting the general semantic content
       - The author-topic model and its applications to CiteSeer

The Author-Topic Model

1. The graphical model and the generative process:



Given the set of co-authors:

1. Choose an author

2. Choose a topic given the author

3. Choose a word given the topic

A document with multiple authors:

x - Author
z - Topic
w - Word
θ, ϕ - Multinomial

Sample author $x \sim A_d$
Sample topic $z \sim \theta_x$
Sample word $w \sim \phi_z$

T = 300

Figure 1: The graphical model for the author-topic model using plate notation.

The Author-Topic Model

2. Bayesian estimation of the model parameters
   - Gibbs sampling
   - The Probability of the *i*th word *m* assigned to topic *j* and author *k*:

$$P\left(z_i = j, x_i = k | w_i = m, \mathbf{z}_{-i}, \mathbf{x}_{-i}\right) \propto \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{AT}_{kj} + \alpha}{\sum_{j'} C^{AT}_{kj'} + T\alpha}$$

   - V * T word by topic count matrix $\phi_{mj} = \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta}$
   - K * T author by topic count matrix $\theta_{kj} = \frac{C^{AT}_{kj} + \alpha}{\sum_{j'} C^{AT}_{kj'} + T\alpha}$

C_{mj}^{WT} # of times word m is assigned to topic j, excluding current instance
C_{kj}^{AT} # of times author k is assigned to topic j, excluding current instance

Application to CiteSeer

- CiteSeer Dataset Description
  - D = 162,489 abstracts, K = 85,465 authors
  - Vocabulary size V = 30,799, word tokens = 11,685,514
  - Preprocessing for punctuation and common stop words
  - T = 300, α = 0.16, β = 0.01
  - 5 independent Markov chains, 2000 iterations each
- Noise in data
  - Title fields
  - Common author names (e.g., J_Smith)

Learned from CiteSeer

Data Mining

TOPIC 52	
WORD	PROB.
DATA	0.1622
MINING	0.0657
DISCOVERY	0.0408
ATTRIBUTES	0.0343
ASSOCIATION	0.0328
LARGE	0.0279
DATABASES	0.0257
KNOWLEDGE	0.0175
PATTERNS	0.0174
ITEMS	0.0173
AUTHOR	**PROB.**
Han_J	0.0164
Zaki_M	0.0089
Liu_B	0.0071
Cheung_D	0.0066
Shim_K	0.0051
Mannila_H	0.0049
Rastogi_R	0.0049
Ganti_V	0.0048
Toivonen_H	0.0043
Liu_H	0.0043

Probabilistic Learning

TOPIC 68	
WORD	PROB.
PROBABILISTIC	0.0869
BAYESIAN	0.0791
PROBABILITY	0.0740
MODEL	0.0533
MODELS	0.0466
PROBABILITIES	0.0308
INFERENCE	0.0306
CONDITIONAL	0.0274
PRIOR	0.0273
POSTERIOR	0.0228
AUTHOR	**PROB.**
Koller_D	0.0104
Heckerman_D	0.0079
Ghahramani_Z	0.0060
Friedman_N	0.0060
Myllymaki_P	0.0057
Lukasiewicz_T	0.0054
Geiger_D	0.0045
Muller_P	0.0044
Berger_J	0.0044
Xiang_Y	0.0042

Information Retrieval

TOPIC 298	
WORD	PROB.
RETRIEVAL	0.1208
INFORMATION	0.0613
TEXT	0.0461
DOCUMENTS	0.0385
INDEXING	0.0369
DOCUMENT	0.0316
QUERY	0.0261
CONTENT	0.0256
SEARCH	0.0174
RELEVANCE	0.0171
AUTHOR	**PROB.**
Oard_D	0.0097
Hawking_D	0.0065
Croft_W	0.0057
Jones_K	0.0053
Schauble_P	0.0052
Voorhees_E	0.0050
Callan_J	0.0046
Fuhr_N	0.0042
Smeaton_A	0.0042
Sanderson_M	0.0041

Database Querying

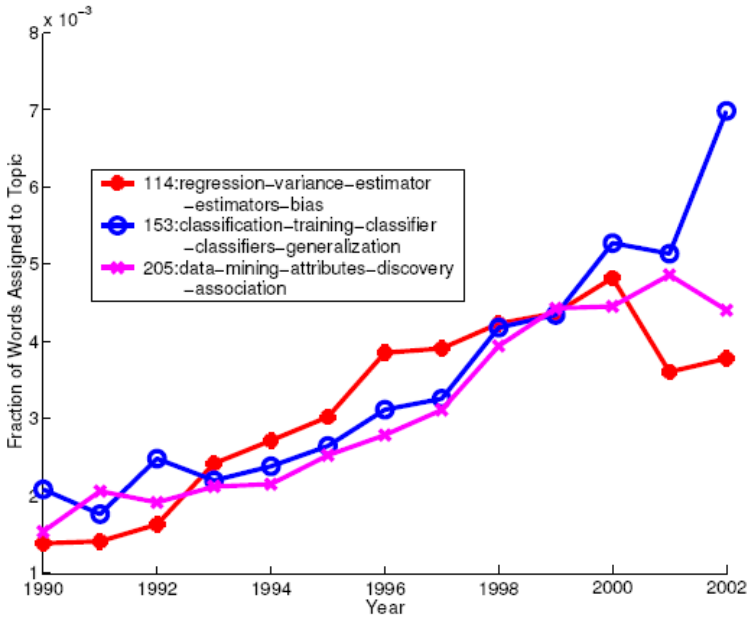
TOPIC 139	
WORD	PROB.
QUERY	0.1406
QUERIES	0.0947
DATABASE	0.0932
DATABASES	0.0468
DATA	0.0426
RELATIONAL	0.0384
JOIN	0.0188
PROCESSING	0.0165
SOURCES	0.0114
OPTIMIZATION	0.0110
AUTHOR	**PROB.**
Levy_A	0.0092
Naughton_J	0.0078
Suciu_D	0.0075
Raschid_L	0.0075
DeWitt_D	0.0062
Widom_J	0.0058
Abiteboul_S	0.0057
Chu_W	0.0055
Libkin_L	0.0054
Kriegel_H	0.0054

- 10~20% "non-research-specific" topics
- Topics are stable from different runs

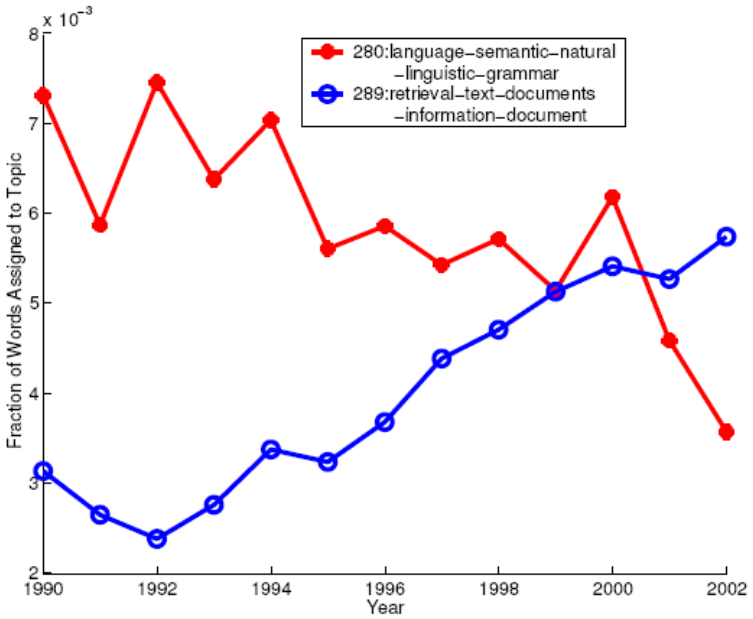
Topic Trends over Time

- The fraction of words assigned to each topic for a given year for each year 1990~2002 and each of 300 topics

Machine Learning and Data Mining



Information Retrieval and Natural Language Processing



Topics and Authors for New Documents

- Gibbs sampling only samples the words in the new documents
- Updating word assignment every 10 iterations

[AUTH1=Scholkopf_B (69%, 31%)] - Machine learning
[AUTH2=Darwiche_A (72%, 28%)] - Probabilistic reasoning

A method[1] is described which like the kernel[1] trick[1] in support[1] vector[1] machines[1] SVMs[1] lets us generalize distance[1] based[2] algorithms to operate in feature[1] spaces usually nonlinearly related to the input[1] space This is done by identifying a class of kernels[1] which can be represented as norm[1] based[2] distances[1] in Hilbert spaces It turns[1] out that common kernel[1] algorithms such as SVMs[1] and kernel[1] PCA[1] are actually really distance[1] based[2] algorithms and can be run[2] with that class of kernels[1] too As well as providing[1] a useful new insight[1] into how these algorithms work the present[2] work can form the basis[1] for conceiving new algorithms

This paper presents[2] a comprehensive approach for model[2] based[2] diagnosis[2] which includes proposals for characterizing and computing[2] preferred[2] diagnoses[2] assuming that the system[2] description[2] is augmented with a system[2] structure[2] a directed[2] graph[2] explicating the interconnections between system[2] components[2] Specifically we first introduce the notion of a consequence[2] which is a syntactically[2] unconstrained propositional[2] sentence[2] that characterizes all consistency[2] based[2] diagnoses[2] and show[2] that standard[2] characterizations of diagnoses[2] such as minimal conflicts[1] correspond to syntactic[2] variations[1] on a consequence[2] Second we propose a new syntactic[2] variation on the consequence[2] known as negation[2] normal form NNF and discuss its merits compared to standard variations Third we introduce a basic algorithm[2] for computing consequences in NNF given a structured system[2] description We show that if the system[2] structure[2] does not contain cycles[2] then there is always a linear size[2] consequence[2] in NNF which can be computed in linear time[2] For arbitrary[1] system[2] structures[2] we show a precise connection between the complexity[2] of computing[2] consequences and the topology of the underlying system[2] structure[2] Finally we present[2] an algorithm[2] that enumerates[2] the preferred[2] diagnoses[2] characterized by a consequence[2] The algorithm[2] is shown[1] to take linear time[2] in the size[2] of the consequence[2] if the preference criterion[1] satisfies some general conditions

Figure 5: Automated labeling of a pseudo-abstract from two authors by the model.

Detecting Surprising Papers

$$\text{Perplexity}(\mathcal{W}_d|a) = \exp\left(-\frac{\log p(\mathcal{W}_d|a)}{|\mathcal{W}_d|}\right)$$

$p(W_d|a)$ - probability assigned to the word W_d given author a;

$|W_d|$ - # of words

Table 2: Papers ranked by perplexity for M. Jordan, from 33 documents.

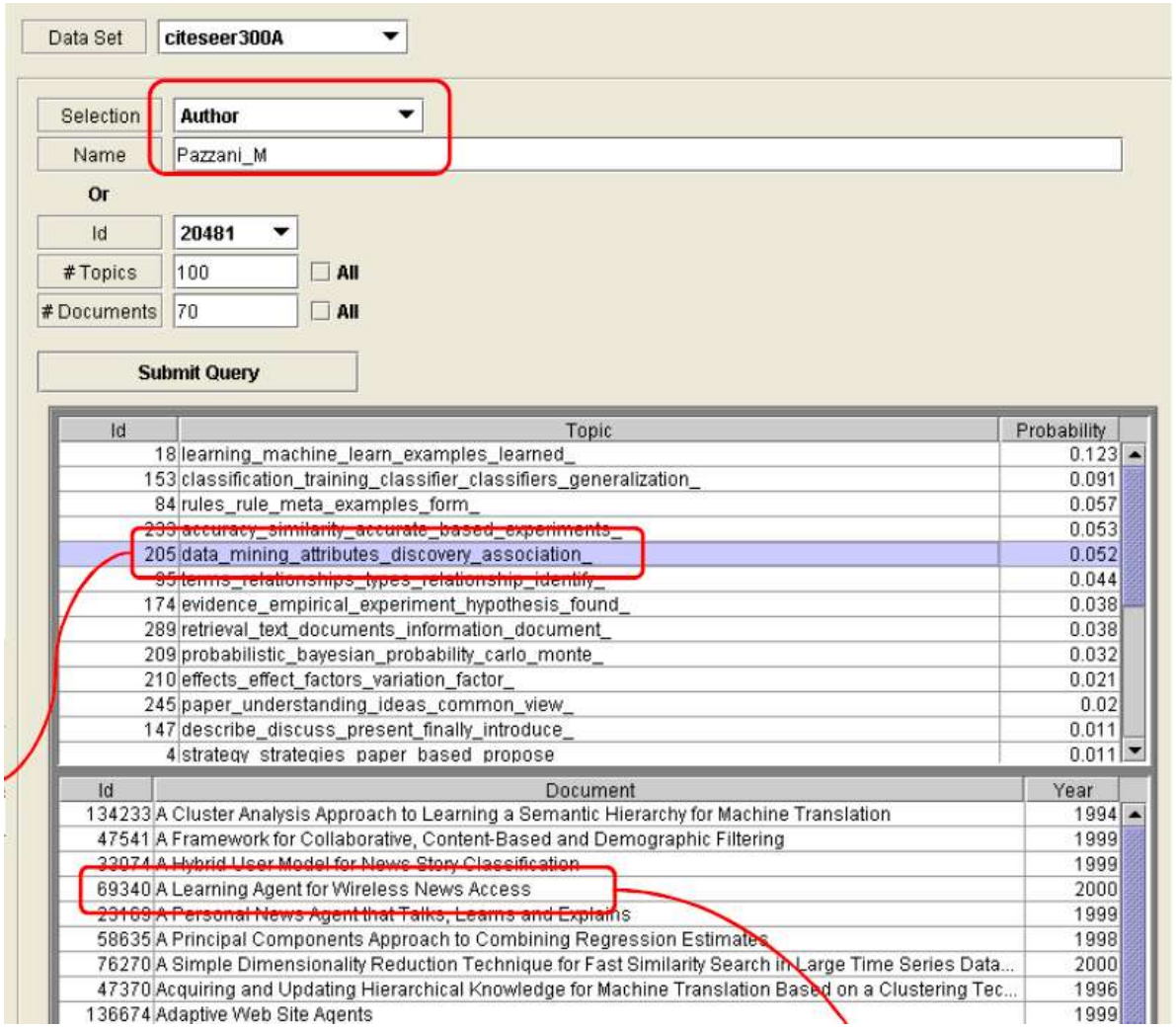
Paper Title	Perplexity Score
Software configuration management in an object oriented database	1386.0
Are arm trajectories planned in kinematic or dynamic coordinates? An adaptation study	1319.2
MEDIAN SCORE	372.4
On convergence properties of the EM algorithm for Gaussian mixtures	180.0
Supervised learning from incomplete data via an EM approach	179.0

Low perplexing papers are written by Michael Jordan at Berkeley

High perplexing papers are written by Mick Jordan of Sun Microsystems

An Author-Topic Browser

- Authors
- Topics
- Documents
- Words



Conclusions

- Introduced the probabilistic author-topic model
- Demonstrated that Bayesian estimation can be used to learn such models from very large text corpora
- The application to CiteSeer was shown to extract substantial novel hidden information
  - Topic time-trends
  - Author-topic relations
  - Unusual papers for specific authors
- Other application for future study
  - Recommending potential reviewer for a paper

Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model

Chaitanty Chemudugunta, Padhraic Smyth and Mark Steyvers

Proceedings Neural Information Processing Systems (NIPS), 2007

Outline

- Introduction and Motivation
- Contribution
- The Topic Model for Special Words
- Illustrative Example
- Experimental Results
- Conclusions

Introduction and Motivation

- A query to a historical archive of news articles:
  - **election + campaign + Camejo***
- The problem of LSI and topic models
  - Highly rank articles are related to presidential elections and do not necessarily include the name Camejo
- Opposite problem with word-based retrieval (TF-IDF)
- The ad hoc LDA approach
  - Linearly combined with doc-spec word distribution
- To Trade-off generality and specificity in a principled way

* Peter Camejo – who ran as vice-presidential candidate along side independent Ralph Nader in 2004

Contribution

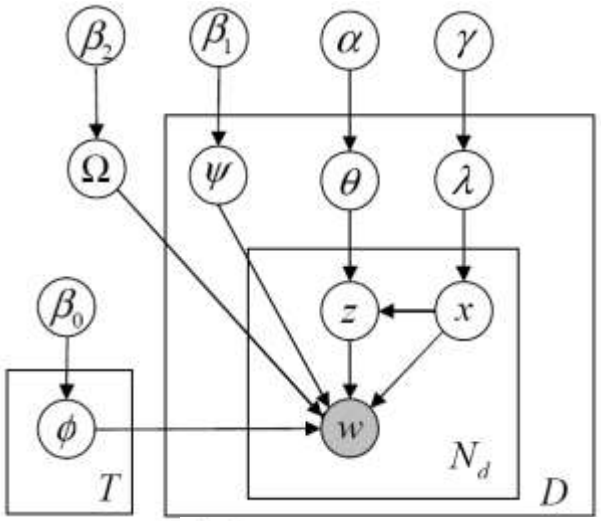
- The Special Words with Background (SWB) Model
  - Generating words
    - From general topics
    - From document-specific word distributions
    - From a corpus-wide background distribution
- Applications in information retrieval
- Applications to any large sparse matrix of count data
  - Transaction data sets
  - Web sites visited

The SWB Model

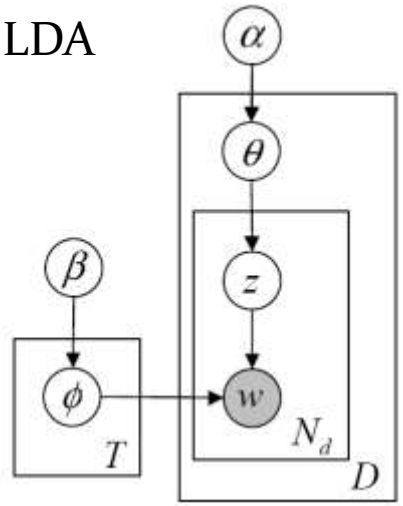
x – Latent random variable
z - Topic
w - Word
λ, θ, ϕ, ψ, Ω - Multinomial

Sample x ~ document-specific λ_d
Sample topic z ~ document-topic θ_d
Sample word w
- x= 0 ~ topic-word ϕ_z
- x= 1 ~ document-specific ψ_d
- x= 2 ~ corpus-specific Ω

$\alpha = 0.1$, $\beta_0 = \beta_2 = 0.01$, $\beta_1 = 0.0001$, $\gamma = 0.3$



LDA



The SWB Model (contd)

- The conditional probability of a word w given a doc d

$$p(w|d) = p(x=0|d)\sum_{t=1}^{T} p(w|z=t)p(z=t|d) \;+\; p(x=1|d)p'(w|d) \;+\; p(x=2|d)p''(w)$$

- Gibbs sampling equations

$$p\left(x_i = 0, z_i = t \,|\mathbf{w}, \mathbf{x}_{-i}, \mathbf{z}_{-i}, \alpha, \beta_0, \gamma\right) \;\propto\; \frac{N_{d0,-i}+\gamma}{N_{d,-i}+3\gamma} \times \frac{C^{TD}_{td,-i}+\alpha}{\sum_{t'} C^{TD}_{t'd,-i}+T\alpha} \times \frac{C^{WT}_{wt,-i}+\beta_0}{\sum_{w'} C^{WT}_{w't,-i}+W\beta_0}$$

$$p\left(x_i = 1 \,|\mathbf{w}, \mathbf{x}_{-i}, \mathbf{z}_{-i}, \beta_1, \gamma\right) \;\propto\; \frac{N_{d1,-i}+\gamma}{N_{d,-i}+3\gamma} \times \frac{C^{WD}_{wd,-i}+\beta_1}{\sum_{w'} C^{WD}_{w'd,-i}+W\beta_1}$$

$$p\left(x_i = 2 \,|\mathbf{w}, \mathbf{x}_{-i}, \mathbf{z}_{-i}, \beta_2, \gamma\right) \;\propto\; \frac{N_{d2,-i}+\gamma}{N_{d,-i}+3\gamma} \times \frac{C^{W}_{w,-i}+\beta_2}{\sum_{w'} C^{W}_{w',-i}+W\beta_2}$$

- The Special Word (SW) Model
  - θ, ϕ, ψ - Multinomial (no background distribution Ω)
  - λ is a document-specific Bernoulli, $\gamma = 0.5$

Illustrative Examples – The New York Times

3140 articles; 1,399,488 word tokens; T = 100; 10 Gibbs samples

e mail krugman nytimes com memo to critics of the media s liberal bias the pinkos you really should be going after are those business reporters even i was startled by the tone of the jan 21 issue of investment news which describes itself as the weekly newspaper for financial advisers the headline was paul o neill s sweet deal the blurb was irs backs off closing loophole averting tax liability for execs and treasury chief it s not really news that the bush administration likes tax breaks for businessmen but two weeks later i learned from the wall street journal that this loophole is more than a tax break for businessmen it s a gift to biznesmen and it may be part of a larger pattern confused in the former soviet union the term biznesmen pronounced beeznessmen refers to the class of sudden new rich who emerged after the fall of communism and who generally got rich by using their connections to strip away the assets of public enterprises what we ve learned from enron and other players to be named later is that america has its own biznesmen and that we need to watch out for policies that make it easier for them to ply their trade it turns out that the sweet deal investment news was referring to the use of split premium life insurance policies to give executives largely tax free compensation you don t want to know the details is an even sweeter deal for executives of companies that go belly up it shields their wealth from creditors and even from lawsuits sure enough reports the wall street journal former enron c e o s kenneth lay and jeffrey skilling both had large split premium policies so what other pro biznes policies have been promulgated lately last year both houses of …

Intentionally misspelled (biznesmen, beeznessmen), rare (pinkos)

john w snow was paid more than 50 million in salary bonus and stock in his nearly 12 years as chairman of the csx corporation the railroad company during that period the company s profits fell and its stock rose a bit more than half as much as that of the average big company mr snow s compensation amid csx s uneven performance has drawn criticism from union officials and some corporate governance specialists in 2000 for example after the stock had plunged csx decided to reverse a 25 million loan to him the move is likely to get more scrutiny after yesterday s announcement that mr snow has been chosen by president bush to replace paul o neill as the treasury secretary like mr o neill mr snow is an outsider on wall street but an insider in corporate america with long experience running an industrial company some wall street analysts who follow csx said yesterday that mr snow had ably led the company through a difficult period in the railroad industry and would make a good treasury secretary it s an excellent nomination said jill evans an analyst at j p morgan who has a neutral rating on csx stock i think john s a great person for the administration he as the c e o of a railroad has probably touched every sector of the economy union officials are less complimentary of mr snow s performance at csx last year the a f l c i o criticized him and csx for the company s decision to reverse the loan allowing him to return stock he had purchased with the borrowed money at a time when independent directors are in demand a corporate governance specialist said recently that mr snow had more business relationships with members of his own board than any other chief executive in addition mr snow is the third highest paid of 37 chief executives of transportation companies said ric marshall chief executive of the corporate library which provides specialized investment research into corporate boards his own compensation levels have been pretty high mr marshall said he could afford to take a public service job a csx program in 1996 allowed mr snow and other top csx executives to buy…

Document specific names: last name (snow), corporation name (CSX)

Figure 2: Examples of two news articles with special words (as inferred by the model) shaded in gray. (a) upper, email article with several colloquialisms, (b) lower, article about CSX corporation.

Experiments

Collection	# of Docs	Total # of Word Tokens	Median Doc Length	Mean Doc Length	# of Queries
NIPS	1740	2,301,375	1310	1322.6	N/A
PATENTS	6711	15,014,099	1858	2237.2	N/A
AP	10000	2,426,181	235.5	242.6	142
FR	2500	6,332,681	516	2533.1	30

Table 1: General characteristics of document data sets used in experiments.

- LDA/SW/SWB, T = 200
- Perplexity Comparisons – NIPS and PATENTS
- Information Retrieval – AP and FR

Background Component for SWB

% of words assigned to the special words dist. : % of the background distribution

NIPS	PATENTS	AP	FR
25:10	58:5	11:6	50:11

NIPS		PATENTS		AP		FR	
set	.0206	fig	.0647	tagnum	.0416	nation	.0147
number	.0167	end	.0372	itag	.0412	sai	.0129
results	.0153	extend	.0267	requir	.0381	presid	.0118
case	.0123	invent	.0246	includ	.0207	polici	.0108
problem	.0118	view	.0214	section	.0189	issu	.0096
function	.0108	shown	.0191	determin	.0134	call	.0094
values	.0102	claim	.0189	part	.0112	support	.0085
paper	.0088	side	.0177	inform	.0105	need	.0079
approach	.0080	posit	.0153	addit	.0096	govern	.0070
large	.0079	form	.0128	applic	.0086	effort	.0068

Figure 3: Examples of background distributions (10 most likely words) learned by the SWB model for 4 different document corpora.

Intuitive results: The background words are commonly used across a broad range of documents within each corpus

Experimental Results

- Perplexity Comparisons

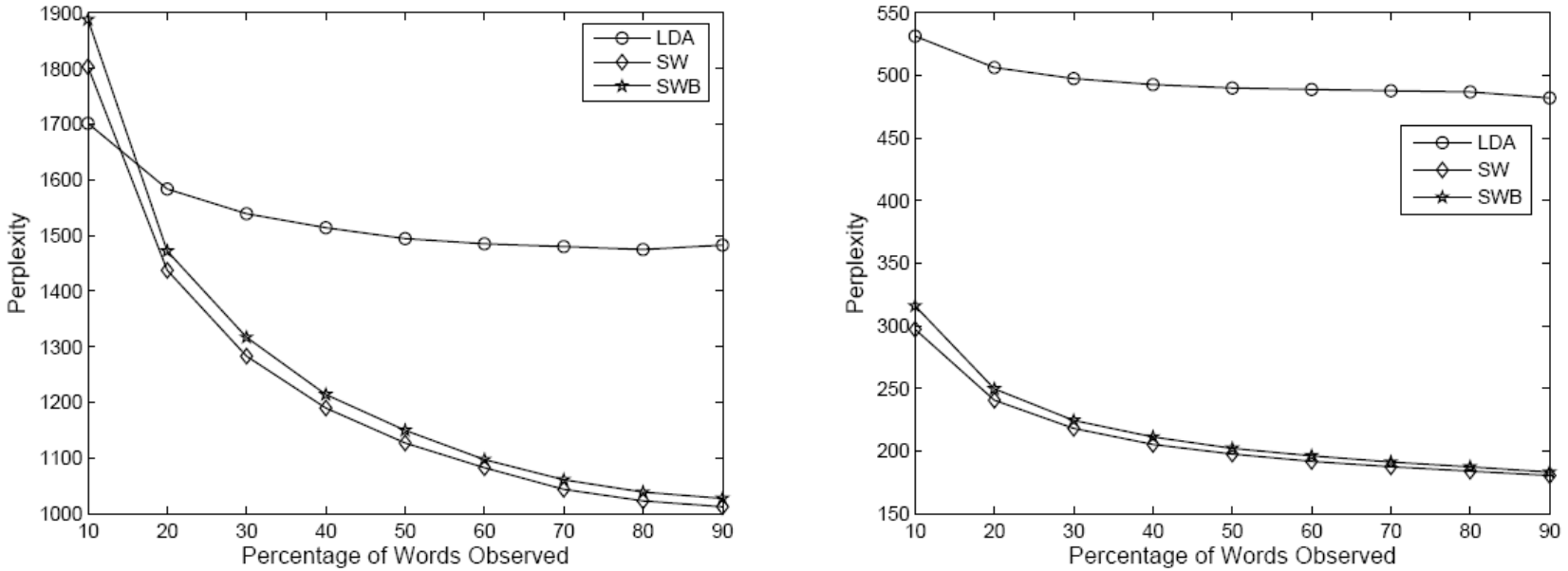


Figure 4: Average perplexity of the two special words models and the standard topics model as a function of the percentage of words observed in test documents on the NIPS data set (left) and the PATENTS data set (right).

Information Retrieval Results

- 500 queries of length 3-5
- Using randomly selected low-frequency words
- Summarizing top k-ranked documents (1, 10, 50, 100)

Method	1 Ret Doc	10 Ret Docs	50 Ret Docs	100 Ret Docs
TF-IDF	100.0	100.0	100.0	100.0
LSI	97.6	82.7	64.6	54.3
LDA	90.0	80.6	67.0	58.7
SW	99.2	97.1	79.1	67.3
SWB	99.4	96.6	78.7	67.2

Table 2: Percentage of retrieved documents containing at least one query word (NIPS corpus).

Information Retrieval Precision

- The score for a document d relative to a query q: p(q|d)

$$p(q|d) \approx \prod_{w \in q} [p(x=0|d) \sum_{t=1}^{T} p(w|z=t)p(z=t|d) \; + \; p(x=1|d)p'(w|d) \; + \; p(x=2|d)p''(w)]$$

MAP:
mean average precision

Pr@10d:
the precision for the top 10 docs retrieved

AP

MAP				Pr@10d			
Method	Title	Desc	Concepts	Method	Title	Desc	Concepts
TF-IDF	.353	.358	.498	TF-IDF	.406	.434	.549
LSI	.286	.387	.459	LSI	.455	.469	.523
LDA	.424	.394	.498	LDA	.478	.463	.556
SW	**.466***	**.430***	**.550***	SW	**.524***	**.509***	.599*
SWB	.460*	.417	.549*	SWB	.513*	.495	**.603***

FR

MAP				Pr@10d			
Method	Title	Desc	Concepts	Method	Title	Desc	Concepts
TF-IDF	.268	.272	.391	TF-IDF	.300	.287	.483
LSI	.329	.295	.399	LSI	.366	.327	.487
LDA	.344	.271	.396	LDA	.428	.340	.487
SW	.371	.323*	**.448***	SW	**.469**	.407*	**.550***
SWB	**.373**	**.328***	.435	SWB	.462	**.423***	.523

*=sig difference wrt LDA

Conclusions

- The new proposed SWB model accounts for both general and specific aspects of documents.
- This models allows documents to be modeled as a mixture of words generated either by general topics or in a manner specific to that document.
- Experimental results on information retrieval show that SWB model performs significantly better than the generalization techniques, such as LSI and LDA, when faced with very specific query words.

Questions ?