Where do Topics Live? Learning the Structure of Information from Words and Documents

Miles Efron, Junliang Zhang, and Gary Marchionini
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599
{efrom, march}@ils.unc.edu, junliang@email.unc.edu

ASIST SIG CR Workshop Long Beach, CA October 18, 2003

Research Question: What kinds of topics can we learn by analyzing terms and documents?

Research Motivation: Supporting Information Seeking in Complex Data

To understand a webspace partition, people must first understand what is in the partition: What is the nature of the information? What is its form and extent? How is it organized? … These questions require that an interface must represent … the overall structure of the information to users.

Marchionini and Brunk (2003)

Research Motivation: Supporting Information Seeking in Complex Data

To understand a webspace partition, people must first understand what is in the partition: What is the nature of the information? What is its form and extent? How is it organized? … These questions require that an interface must represent … the overall structure of the information to users.

Marchionini and Brunk (2003)

Key implementation challenges are related to acquiring the appropriate data (slicing the data by an attribute may make good sense from a user perspective but this may entail creating customized metadata for the interface).

Goals of the Work

1. To discover a manageable and empirically valid set of topics in complex data sets and represent them meaningfully
2. To associate documents in the data with the inferred topics

Key implementation challenges are related to acquiring the appropriate data (slicing the data by an attribute may make good sense from a user perspective but this may entail creating customized metadata for the interface.

Goals of the Work

1. To discover a manageable and empirically valid set of topics in complex data sets and represent them meaningfully
2. To associate documents in the data with the inferred topics

Automatic Metadata Extraction,

cf. Han *et al.* (2003).

```
<document>
<topic1 weight="0.02" />
<topic2 weight="0.63" />
<topic3 weight="0.35" />
…
</document>
```

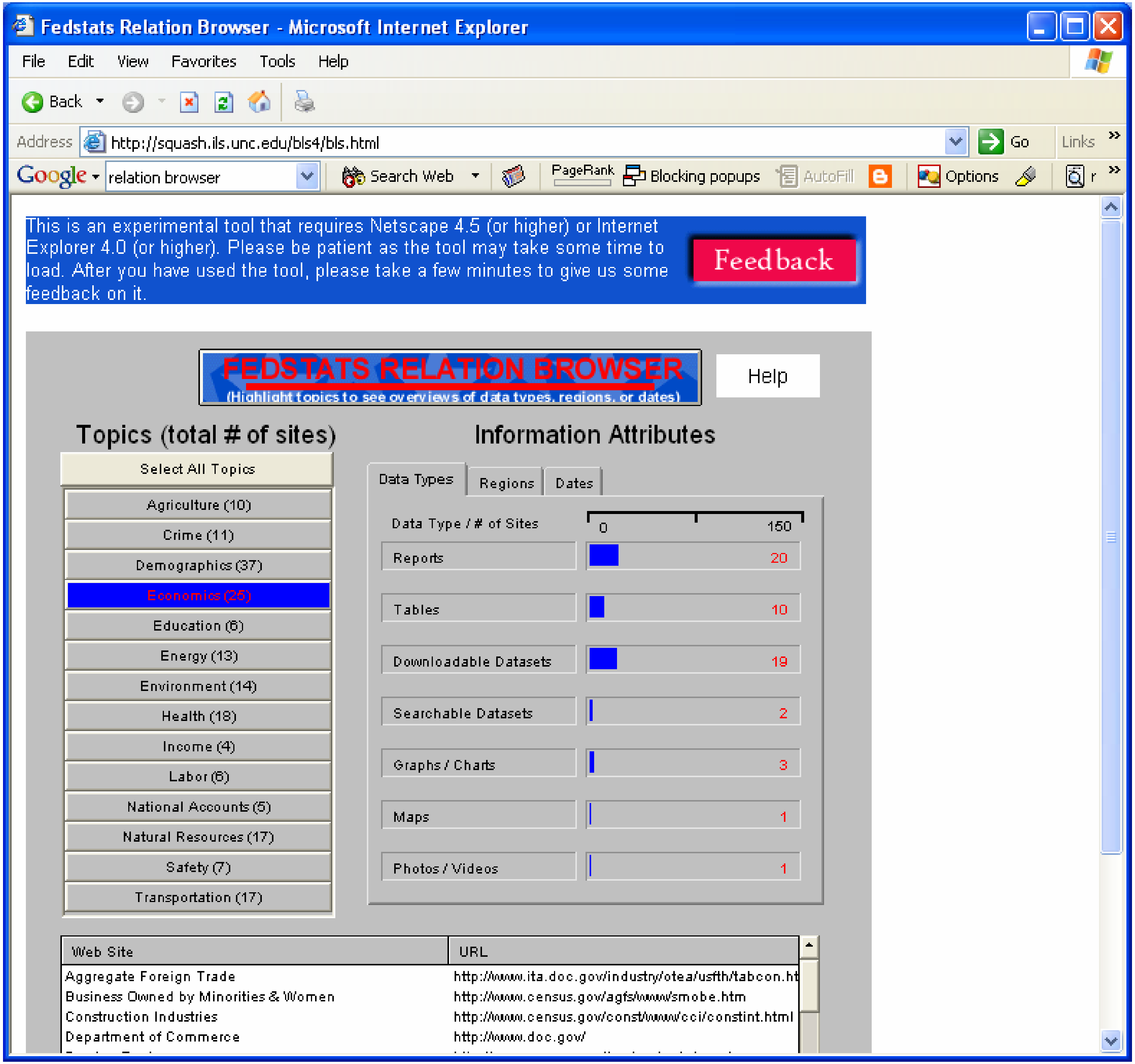
Goals of the Work

1. To discover a manageable and empirically valid set of topics in complex data sets and represent them meaningfully
2. To associate documents in the data with the inferred topics

Enabling information seeking via dynamic user interfaces.

cf. Marchionini and Brunk (2003)

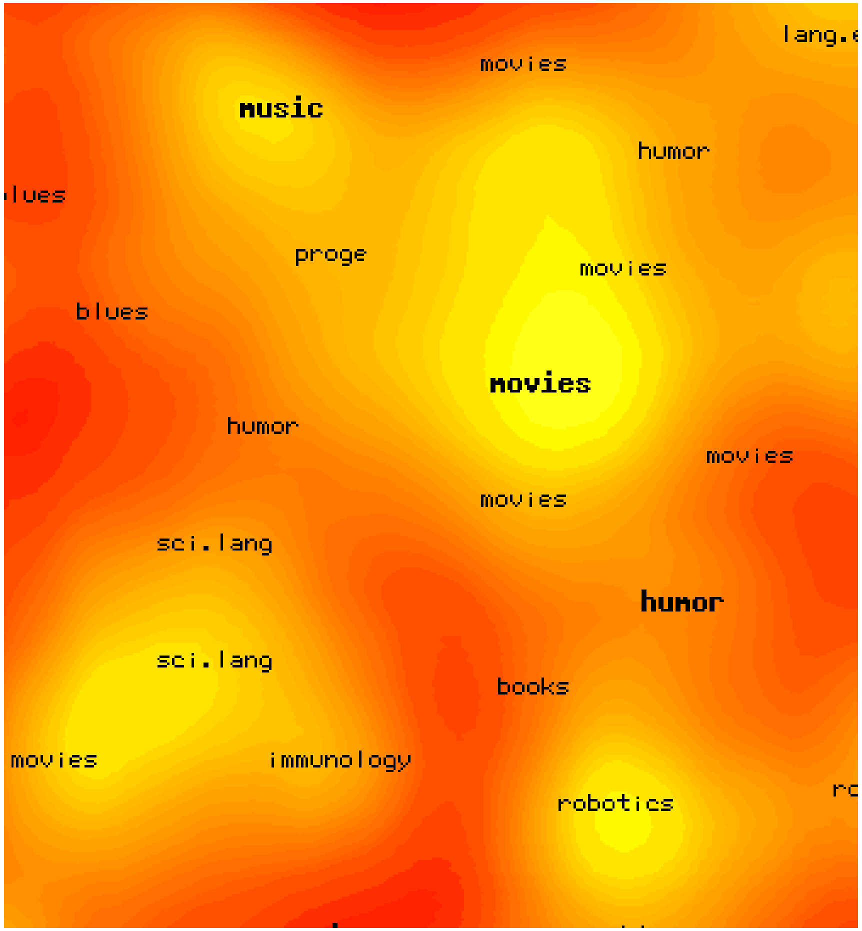
```
<document>
<topic1 weight="0.02" />
<topic2 weight="0.63" />
<topic3 weight="0.35" />
…
</document>
```


Fedstats Relation Browser - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://squash.ils.unc.edu/bls4/bls.html
Go
Links
relation browser
Search Web
PageRank
Blocking popups
AutoFill
Options
This is an experimental tool that requires Netscape 4.5 (or higher) or Internet Explorer 4.0 (or higher). Please be patient as the tool may take some time to load. After you have used the tool, please take a few minutes to give us some feedback on it.
Feedback
FEDSTATS RELATION BROWSER
(Highlight topics to see overviews of data types, regions, or dates)
Help
Topics (total # of sites)
Select All Topics
Agriculture (10)
Crime (11)
Demographics (37)
Economics (25)
Education (6)
Energy (13)
Environment (14)
Health (18)
Income (4)
Labor (6)
National Accounts (5)
Natural Resources (17)
Safety (7)
Transportation (17)
Information Attributes
Data Types
Regions
Dates
Data Type / # of Sites
0
150
Reports
20
Tables
10
Downloadable Datasets
19
Searchable Datasets
2
Graphs / Charts
3
Maps
1
Photos / Videos
1
Web Site
URL
Aggregate Foreign Trade
http://www.ita.doc.gov/industry/otea/usfth/tabcon.ht
Business Owned by Minorities & Women
http://www.census.gov/agfs/www/smobe.htm
Construction Industries
http://www.census.gov/const/www/cci/constint.html
Department of Commerce
http://www.doc.gov/

Finding Structure in Information Space

Implicit in most statistical approaches to unsupervised learning is the notion that topics can be modeled as aspects of the probability density functions that generated the data.

- Principal Component Analysis (*cf.* Jolliffe (1986)).
- Self Organizing Maps (*cf.* Kohonen (1997) and Lin *et al*. (2003)).
- Latent Semantic Indexing (*cf.* Deerwester *et al.* (1990)).



http://websom.hut.fi/websom/milliondemo/html/root.html

Finding Structure in Information Space

Implicit in most statistical approaches to unsupervised learning is the notion that topics can be modeled as aspects of the probability density functions that generated the data.

$$\underset{n\times p}{\mathbf{A}} = \underset{n\times p}{\mathbf{T}} \; \underset{p\times p}{\Sigma} \; \underset{p\times p}{\mathbf{D}'}$$

Given a document collection with *n* documents in *p* terms, we define the document-term matrix **A**, and its singular value decomposition:

Finding Structure in Information Space

Implicit in most statistical approaches to unsupervised learning is the notion that topics can be modeled as aspects of the probability density functions that generated the data.

$$\underset{n\times p}{\mathbf{A}} = \underset{n\times p}{\mathbf{T}} \; \underset{p\times p}{\Sigma} \; \underset{p\times p}{\mathbf{D}'}$$

Given a document collection with *n* documents in *p* terms, we define the document-term matrix **A**, and its singular value decomposition:

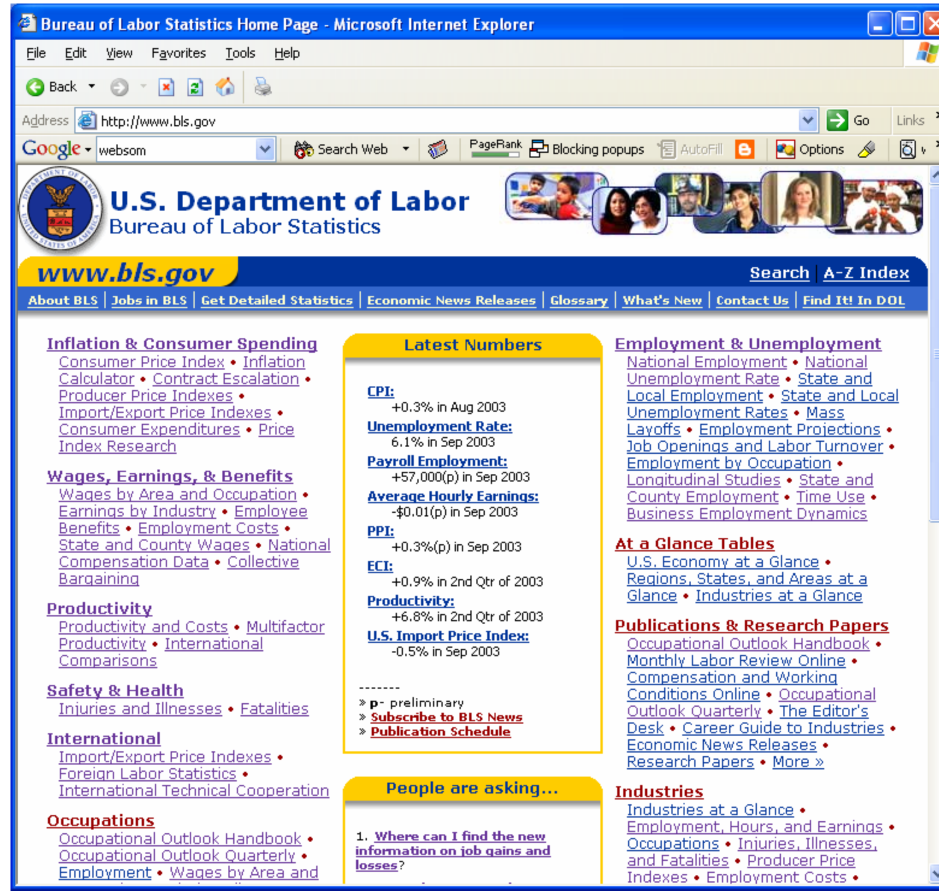
Diagonal matrix Sigma contains the singular values of **A**. These are the square roots of the eigenvalues of matrices **A'A** and **AA'**. Thus the principal components of term-space and document-space are identically descriptive.

Practicalities of Learning Concepts

- Documents have extra-linguistic information that may inform topic modeling.
  - Link structure
  - Prior Classifications
  - Transaction Logs
- Terms also have information not captured by the matrix **A**. However, methods for utilizing this information for statistical learning is non-trivial.

The Bureau of Labor Statistics Website

- 25,530 documents
- 26,772 terms (after stemming, and filtering terms through WordNet)
- Part of an implicit network of statistical information websites (e.g. Census, EIA, etc.)



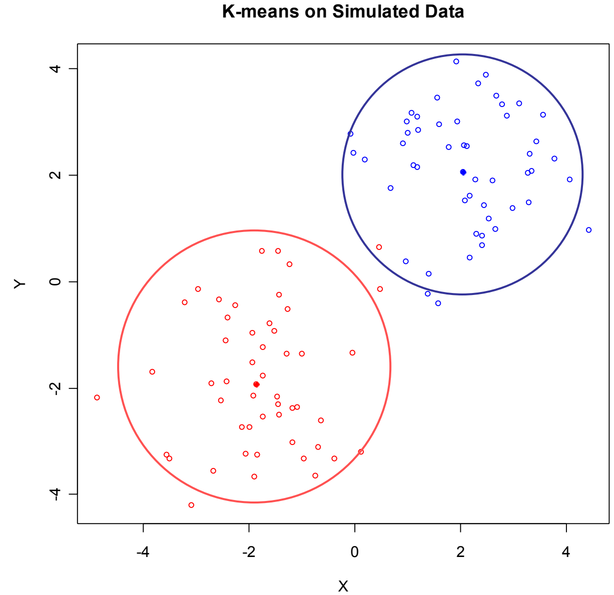
This research is part of the GovStat Project (http://www.ils.unc.edu/govstat)

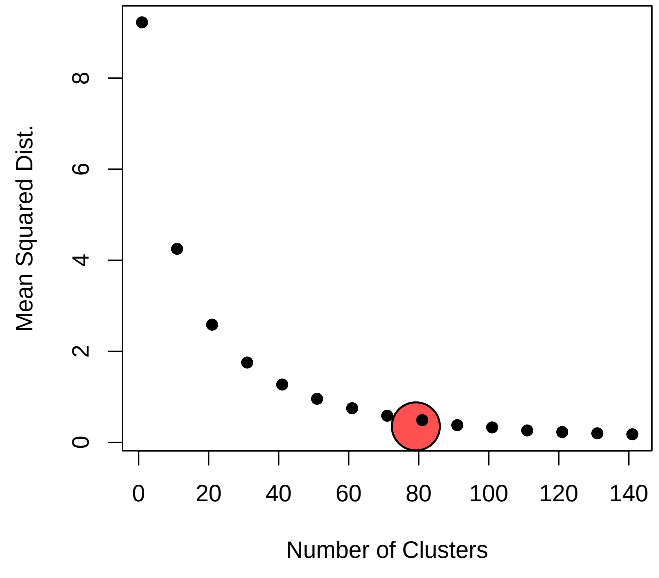
Points of Comparison

- **Model Specification:** What type of learning algorithm is definable on a given space?
- **Feature Selection:** How does the specified model enable us to reduce the dimensionality of the search space?
- **Knowledge Representation:** After analysis, how are the inferred topics represented? How useful/meaningful is this representation?
- **Quality of Learned Topics:** Does the learned structure accurately describe the information space? How to measure this remains an open question.

Learning from Terms

- *Model specification:* *k*-means clustering
- *Feature selection:* Salton's term discrimination model
- *Knowledge representation:* mutually exclusive classification of terms into clusters





Learning from Terms

- *Model specification:* *k*-means clustering
- *Feature selection:* Salton's term discrimination model
- *Knowledge representation:* mutually exclusive classification of terms into clusters

Salton (1975) argues that the best discriminators have document frequency on the interval $\left[\frac{n}{100}, \frac{n}{10}\right]$ where *n* is # of docs.

Using Salton's model led us to represent each document in 1882-space.

Learning from Terms

- *Model specification:* *k*-means clustering
- *Feature selection:* Salton's term discrimination model
- *Knowledge representation:* mutually exclusive classification of terms into clusters

Pilot Study: Sufficiency of Salton's Term Discrimination Model

- Created a 2nd clustering, adding the 100 most frequently occurring terms (after stoplist application) to the representation.
- 9 participants chose 1 term for each cluster that best exemplified that cluster's topical domain.
- The term discrimination model appeared to miss some important terms…

Learning from Terms

- *Model specification:* *k*-means clustering
- *Feature selection:* Salton's term discrimination model
- *Knowledge representation:* mutually exclusive classification of terms into clusters

Pilot Study: Sufficiency of Salton's Term Discrimination Model

Terms omitted by TD Model

annual	detail
office	service
area	median
publish	**transportation**
code	metropolitan
question	**wages**
construction	number
research	**workers**

Learning from Terms

- *Model specification:* *k*-means clustering
- *Feature selection:* Salton's term discrimination model
- *Knowledge representation:* mutually exclusive classification of terms into *small* clusters

flight
fly
pilot
airline
aircraft

mass
unemployment
layoff

logging
conservation
forest

pension
plan
retirement
benefits
contribution
coverage
definition
employment
employee

Learning from Terms

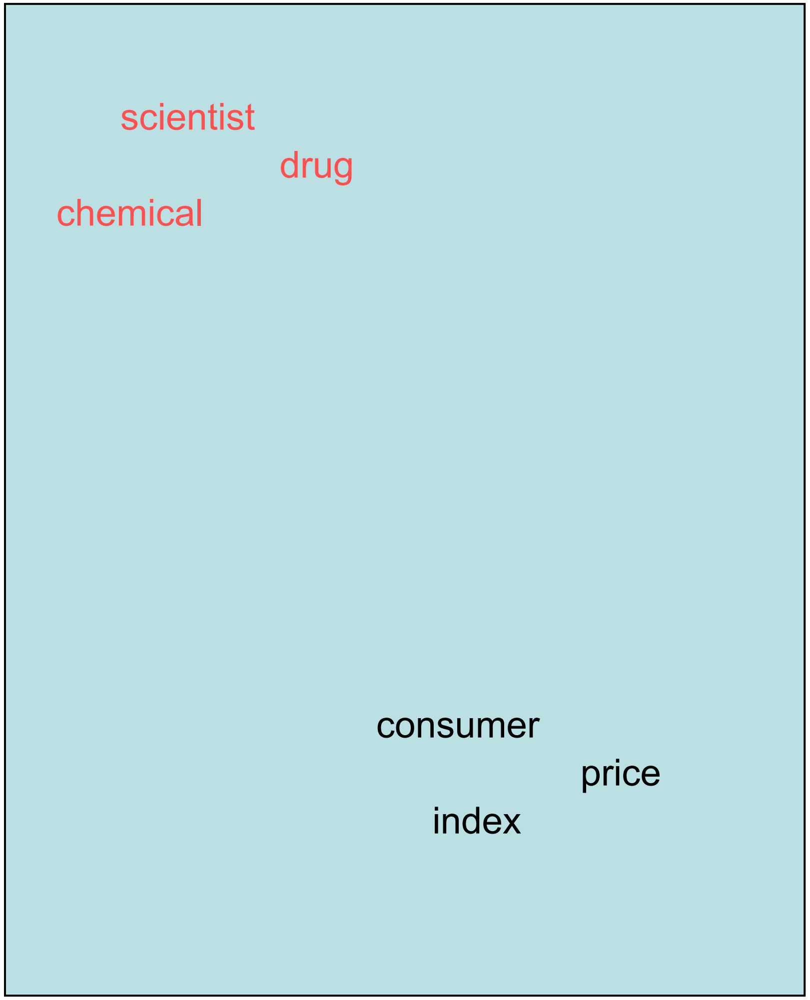
- **Quality of learned topics**
  - **Most of the clusters were intuitively coherent**
  - **But several were not, and their problems fell into a variety of types**

lower	
multiple	
region	
tract	
trunk	
upper	
chest	
foot	
internal	
arm	
location	

louisiana	
maryland	
mexico	
michigan	
midwest	
mississippi	
minnesota	
…	
iowa	
kansas	
kentucky	

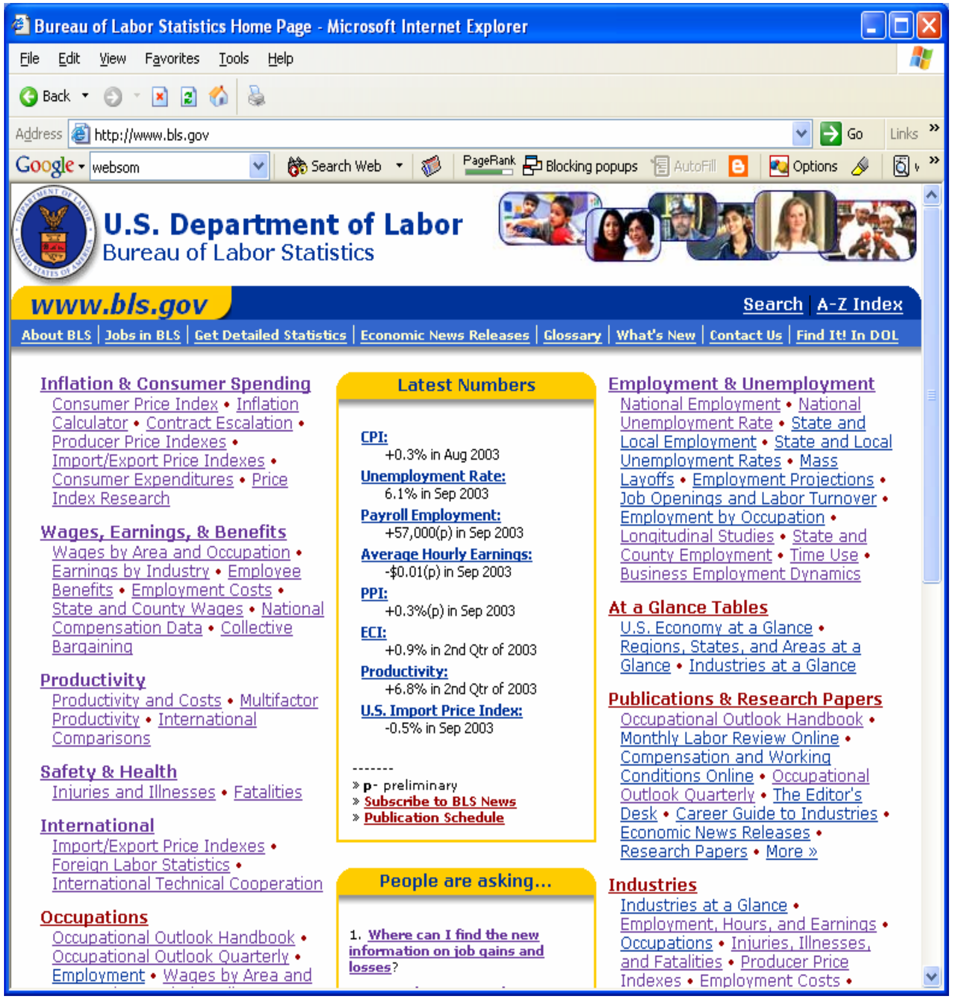
Learning from Terms

- **Quality of learned topics**
  - **Perhaps most problematic: insufficient detail in coverage of the topic space**



Learning from Documents

- Work with a privileged subset of n=107 “top-level” documents
- BLS has assigned each of these documents to 1 or more of 15 top-level classes
- Use the BLS classification implicitly to inform our own analysis



Learning from Documents

BLS' 15 top-level document classes

Inflation	Occupations	Tabular data
Wages	Demographics	Publications
Productivity	Other sites	Industries
Safety	BLS offices	Business costs
International	Employment/un-employment	Geography

Each of the 107 pages linked to from www.bls.gov is associated with 1 or more of these topics.

Learning from Documents

- *Model specification:* naïve Bayes
- *Feature selection:* information gain
- *Knowledge representation:* a probability that a given document is about each of 15 topics

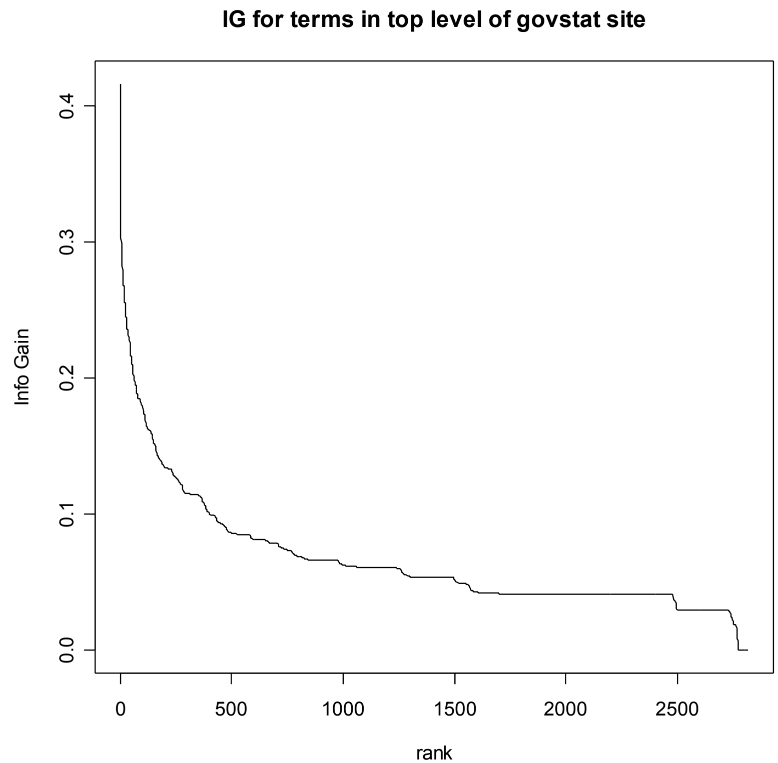
Given document d_i and 15 classes, $C_1 \ldots C_{15}$. model the document as a 15-vector of probabilities:

$$\mathbf{d}_i = \begin{bmatrix} P(C_1 \mid d_i) = \dfrac{P(d_i \mid C_1)P(C_1)}{P(d_i)} \\ P(C_2 \mid d_i) = \dfrac{P(d_i \mid C_2)P(C_2)}{P(d_i)} \\ \vdots \\ P(C_{15} \mid d_i) = \dfrac{P(d_i \mid C_{15})P(C_{15})}{P(d_i)} \end{bmatrix}$$

Learning from Documents

- *Model specification:* naïve Bayes
- *Feature selection:* information gain
- *Knowledge representation:* a probability that a given document is about each of 16 topics

Naïve Bayes by definition assumes that our terms are independent. To reduce the error incurred by this assumption, we may limit the vocabulary to the best *k* terms, where "best" is understood in the information theoretic sense.



information gain for BLS terms

Learning from Documents

- *Quality of the Results?*
- *Utility:*
  - classifying documents for use in dynamic interfaces
  - adding subject metadata to documents to inform search

Document: http://www.bls.gov/bls/demographics.htm

•demographics	0.9999377197
•geography	3.189869598e-05
•employment	1.977756158e-05
•inflation	3.003283918e-06
•occupations	2.00910276e-06
•wages	1.127429308e-06
•international	1.055563966e-06
•businessCosts	1.046589011e-06
•publications	8.456974726e-07
•productivity	4.396949932e-07
•industry	4.368836362e-07
•offices	2.543445862e-07
•safety	2.169423434e-07
•other	9.22091744e-08
•tables	7.629056542e-08

Learning from Documents

- *Quality of the Results?*
- *Utility:*
  - enriching queries with topical information

Query: “race ethnicity population”

•demographics	0.5816571154
•employment	0.3501218041
•occupations	0.01350592469
•geography	0.01069743082
•publications	0.00781181311 4
•inflation	0.00526569437 5
•wages	0.00523916415 9
•industry	0.0050356233 9
•other	0.00364026788 9
•safety	0.00315404866 8
•productivity	0.00302724427 3
•international	0.00284990903 4
•businessCosts	0.00278081441 6
•offices	0.002645050 34
•tables	0.00256809534 8

Learning from Documents

- *Quality of the Results?*
- *Utility:*
  - enriching queries with topical information

Query: “geography demographics employment”

•geography	0.2949892732
•demographics	0.1887337873
•occupations	0.1877464733
•employment	0.07008388182
•wages	0.05535215515
•other	0.0512377791
•publications	0.03792001448
•industry	0.03767006428
•offices	0.01715965884
•businessCosts	0.01690363933
•productivity	0.01545018715
•inflation	0.01260101009
•international	0.00610982941 5
•safety	0.00535487401 8
•tables	0.00268737251 3

Learning from Terms and Documents: Problems

Terms:

- No well-motivated rationale for feature selection
- Clusters are at a level of granularity that is too fine for providing global overviews of the information space
- No means of validating or weighting clusters

Documents:

- Under the current model we don’t learn any topics that BLS didn’t already posit
- We lack sufficient training data to create robust models

Learning from Terms and Documents: Problems

Terms:

- No well-motivated rationale for feature selection
- Clusters are at a level of granularity that is too fine for providing global overviews of the information space
- No means of validating or weighting clusters

Documents:

- Under the current model we don't learn any topics that BLS didn't already posit
- We lack sufficient training data to create robust models

Consider the idea of ***employment***…

Learning from Terms and Documents: Problems

Term Clusters

mass
unemployment
layoff

pension
plan
retirement
benefits
contribution
coverage
definition
employment
employee

Related BLS Categories

- Employment and Unemployment
- Wages
- Occupations

Learning from Terms and Documents: Problems

Term Clusters

mass
unemployment
layoff

pension
plan
retirement
benefits
contribution
coverage
definition
employment
employee

Query: “mass layoff unemployment”

employment	0.91262575
geography	0.04582751
offices	0.00855301
occupations	0.00569907
demographics	0.00437878
wages	0.00417550
businessCosts	0.00277032
tables	0.00252967
International	0.00248472
publications	0.00240270
other	0.00237859
industry	0.00218753
inflation	0.00153552
productivity	0.00125203
safety	0.00119923

Learning from Terms and Documents: Problems

Term Clusters

mass
unemployment
layoff

pension
plan
retirement
benefits
contribution
coverage
definition
employment
employee

Query: “pension plan retirement … employee”

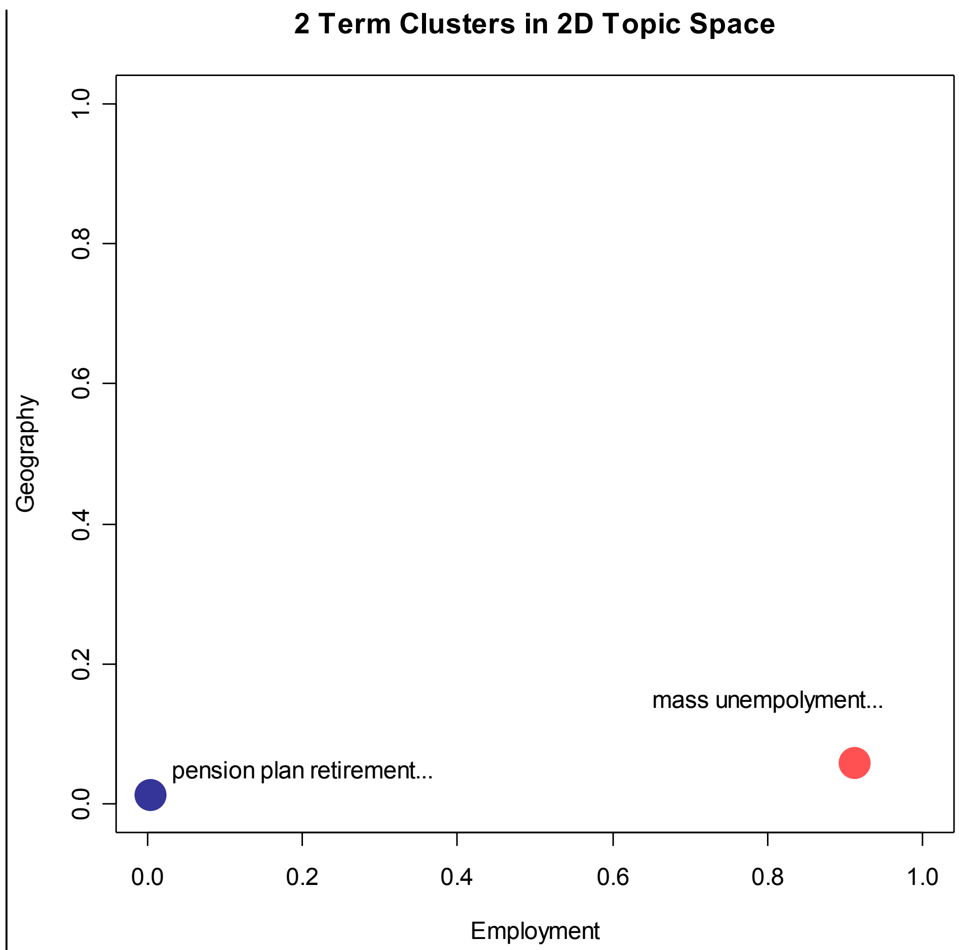
wages	0.9431528
occupations	0.03636805
businessCosts	0.004555304
productivity	0.003115273
publications	0.002744467
employment	0.002074062
industry	0.001504748
safety	0.001371867
tables	0.001307003
International	0.0008155906
geography	0.0008146394
inflation	0.0006476636
demographics	0.0005400136
offices	0.0005083938
other	0.0004800395

Learning from Terms and Documents: Problems

Term Clusters

mass
unemployment
layoff

pension
plan
retirement
benefits
contribution
coverage
definition
employment
employee



Open Questions and Future Directions

- BLS is part of a larger network of statistical information resources. When it comes time to enable search across agencies, what will be the advantage of each of the models addressed here?
- Does the term-space in fact have more structure than we've argued here? i.e. If we admitted syntax- and discourse-level analysis into our partitioning of term-space, might we be able to address the limitations discussed here?
- Evaluation: How can we decide which mapping of information space is superior? How well are we doing?

Open Questions and Future Directions

- Development of a metric for assessing a given page's quality *vis a vis* topic discovery.
- Pursuit of a middle-ground, using semi-supervised learning as described in Blum and Mitchell (1998).
- Application of NLP techniques to improve our analysis of the term-space.

References

1. Blum, A. and T. Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the 11th Annual Conference on Computational Learning Theory,* Madison, WI, 1998.
2. Deerwester, S., *et al*. Indexing by latent semantic analysis. *JASIS*, 41(6), 1990, 391-407.
3. Han, H. *et al.* Automatic Document Metadata Extraction Using Support Vector Machines. *Joint Conference on Digital Libraries. JCDL '03.* 2003.
4. Jolliffe, I. T. *Principal Component Analysis*. Springer, 1996.
5. Kohonen, T. Self-Organizing Maps. Information Sciences. Springer, second edition, 1997.
6. Lin, X.; White, H. D.; & Buzydlowski, J. (2003). Real-time author co-citation mapping for online searching. *International Journal of Information Processing & Management*, 39(5), pp. 689-706.
7. Salton, G. and C. T. Yu. A Theory of Term Importance in Automatic Text Analysis. *JASIS,* 26(1), 1975. pp. 33-44.
8. Marchionini, G. and B. Brunk. Toward a General Relation Browser: a GUI for Information Acrhitects. *Journal of Digital information*, 4(1), 2003.