

Using Machine Learning to Populate Dynamic Interfaces

Miles Efron
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599
efrom@ils.unc.edu

ASIST IA Summit Austin, TX February 29, 2004

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

Research Question: How can we bring machine learning techniques to bear on the problem of enabling dynamic search interfaces for complex document collections?

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).

Data: The Bureau of Labor Statistics Website

- 15,165 text/html documents
- A topical structure has been defined on 65 "high-level" documents.
- But this structure is redundant and weakly motivated.
- **Can we derive topics based on the data themselves?**



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).

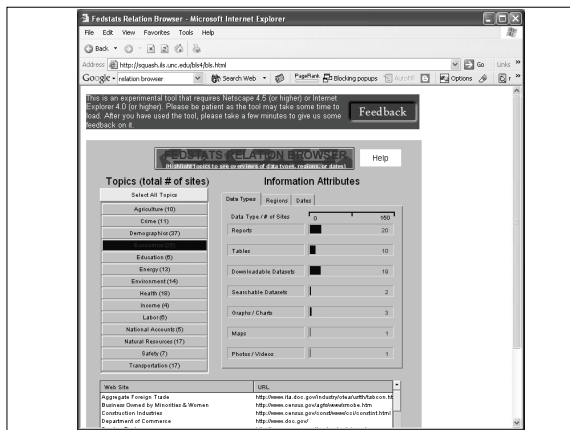
```
<document>
<topic1 weight="0.02" />
<topic2 weight="0.63" />
<topic3 weight="0.35" />
...
</document>
```

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

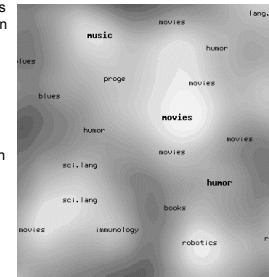
Enabling information seeking via dynamic user interfaces.
cf. Marchionini and Brunk (2003)

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



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

A three-stage process...

1. *Use domain knowledge:* Begin analysis on a focused subset of the data
2. *Model inter-document relationships Probabilistically:* Use model-based clustering to identify the dominant groupings of documents
3. *Extend the model:* Having built a model on a subset of the collection, create collection-wide metadata by applying the model to unseen documents.

Finding Structure in Information Space

A three-stage process...

1. *Use domain knowledge:* Begin analysis on a focused subset of the data
2. *Model inter-document relationships Probabilistically:* Use model-based clustering to identify the dominant groupings of documents
3. *Extend the model:* Having built a model on a subset of the collection, create collection-wide metadata by applying the model to unseen documents.

Initial learning constrained to $N=1279$ documents that are highly topical, and that contain subject metadata.

Use term distributions to inform k -means clustering, putting each document into 1 of k clusters. This is a "hard" version of probabilistic (EM) clustering.

Extend the model to remaining docs either by naive Bayes (applying the cluster model), or another classifier, such as SVM or Prind.

Comparing Document Representations for Clustering

1279 Documents from daily column *The Editor's Desk* contain:

| Title: | Body: | Keyword Metadata: |
|---|--|--|
| A succinct statement of the article's subject matter. E.G. <i>Job Gains and Losses in the 2nd Quarter of 2003.</i> | A brief exposition of a single topic in journalistic prose. Each document is approximately 200 words in length. | A list of human-assigned terms that capture the topic of the article. Median number of terms per doc: 7 E.G. <i>employment, growth, jobs</i> |

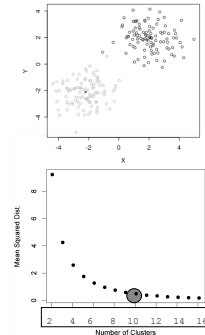
Subject Headings:

1112 documents contained controlled-vocabulary subject headings. We used the most common 20 of these (accounting for 609 docs) for cluster evaluation

Defining the Cluster Models

Clustering by *k*-means:

- full-text (FT)
- title-only (TO)
- keyword-only (KW)



Number of clusters chosen by inspection of MSD & analysis of model likelihood: **k=10**

Evaluating the Cluster Models

Subject Headings:

1112 documents contained controlled-vocabulary subject headings. We used the most common 20 of these (accounting for 609 docs) for cluster evaluation

| Model | + | - | acc |
|-------|-----|-----|------|
| FT | 392 | 217 | 0.64 |
| TO | 441 | 168 | 0.72 |
| KW | 601 | 8 | 0.98 |

$$H_0 : acc_{KW} = acc_{TO} \quad p \approx 0$$

| Subject Heading | Doc. Count |
|--|------------|
| Prices | 92 |
| Unemployment | 55 |
| Occupational safety and health | 53 |
| International comparisons, prices | 48 |
| Manufacturing, prices | 45 |
| Employment | 44 |
| Productivity | 40 |
| Consumer expenditures | 36 |
| Earnings and wages | 27 |
| Employment and unemployment | 27 |
| Compensation costs | 25 |
| Earnings and wages, metropolitan areas | 18 |
| Benefits, compensation costs | 18 |
| Earnings and wages, occupations | 17 |
| Employment, occupations | 14 |
| Benefits | 14 |
| Earnings and wages, regions | 13 |
| Work stoppages | 12 |
| Earnings and wages, industries | 11 |
| Total | 609 |

Topics Derived by Keyword Clustering

| Benefits | Costs | International | Jobs |
|-----------|--------------|---------------|------------|
| plans | compensation | import | employment |
| benefits | costs | prices | jobs |
| employees | benefits | petroleum | youth |

| Occupations | Prices | Productivity | Safety |
|-------------|-----------|--------------|--------------|
| workers | prices | productivity | safety |
| earnings | index | output | health |
| operators | inflation | nonfarm | occupational |

K=10 Clusters, labeled manually.

Each cluster shown with 3 "best" terms with respect to log-odds ratio

| Spending | Unemployment |
|--------------|--------------|
| expenditures | unemployment |
| consumer | mass |
| spending | jobless |

Generalizing via Document Classification

Having categorized 1279 documents, the task remains to extend this model to approximately 14,000 yet unseen documents.

The goal here is to estimate the strength of association between each document d_i and each class C_k .

| Method | Citation | Motivation |
|-------------|-----------------|---|
| Prind | Joachims (1997) | Estimate the proximity of doc_i to the centroid of class k |
| Naive Bayes | Mitchell (1997) | Estimate the likelihood that doc_i belongs to class k , choose the most likely, <i>a posteriori</i> |
| SVM | Joachims (1998) | Defines the maximally separating hyperplane in a high-dimensional, transformed feature space. |

Generalizing via Document Classification

Cross-validation study: 10-fold cross-validation was conducted (67/33% train/test split) to evaluate the accuracy of the automatic classifiers. N. B. This evaluation only applied to documents within the 1279-document focused subset of the collection...more evaluation remained to be done.

| Method | Av. % Accuracy | S.E. |
|-------------|----------------|------|
| Prind | 59.07 | 1.07 |
| Naive Bayes | 75.57 | 0.4 |
| SVM | 75.08 | 0.68 |

Not significantly better than SVM, but much simpler, and MUCH more scalable.

Generalizing via Document Classification

Augmenting the training set: To mitigate the small training sets, a second naive Bayes classifier (N.B. augmented) was constructed by querying Google for BLS documents about each topic. This led to a training set of 4113 documents, instead of the non-augmented model's 1279.

| Method | Av. % Accuracy | S.E. |
|-------------|----------------|------|
| Prind | 59.07 | 1.07 |
| Naive Bayes | 75.57 | 0.4 |
| SVM | 75.08 | 0.68 |
| N.B. (aug) | 58.16 | 0.32 |

Appeared significantly worse than the plain N.B. model on these data...but *not* in the next round.

Generalizing via Document Classification

Applying the Models: To gauge the ability of our models to classify documents from outside the *Editor's Desk* portion of the collection, we selected $N=100$ random documents and asked 11 volunteer judges to classify these into as many of the topics as they deemed appropriate. The two naive Bayes classifiers were then tested against the reviewers' first and second topical choices for each document.

| Human Judges 1st Choice by Majority Vote | | |
|--|------------------|------------------|
| Model | Model 1st Choice | Model 2nd Choice |
| N.B. | 24 | 1 |
| N.B. (aug) | 14 | 24 |

$N=100$ documents

| Human Judges 2nd Choice by Majority Vote | | |
|--|------------------|------------------|
| Model | Model 1st Choice | Model 2nd Choice |
| N.B. | 21 | 4 |
| N.B. (aug) | 14 | 21 |

Generalizing via Document Classification

Applying the Models: To gauge the ability of our models to classify documents from outside the *Editor's Desk* portion of the collection, we selected $N=100$ random documents and asked 11 volunteer judges to classify these into as many of the topics as they deemed appropriate. The two naive Bayes classifiers were then tested against the reviewers' first and second (for 73 multi-class docs) topical choices for each document.

| Human Judges 1st Choice by Majority Vote | | |
|--|------------------|------------------|
| Model | Model 1st Choice | Model 2nd Choice |
| N.B. | 24 | 1 |
| N.B. (aug) | 14 | 24 |

The non-augmented model is superior if we only care about a hard classification.

| Human Judges 2nd Choice by Majority Vote | | |
|--|------------------|------------------|
| Model | Model 1st Choice | Model 2nd Choice |
| N.B. | 21 | 4 |
| N.B. (aug) | 14 | 21 |

$acc(N.B.) = 0.45$
 $acc(N.B. aug) = 0.28$

Generalizing via Document Classification

Applying the Models: To gauge the ability of our models to classify documents from outside the *Editor's Desk* portion of the collection, we selected $N=100$ random documents and asked 11 volunteer judges to classify these into as many of the topics as they deemed appropriate. The two naive Bayes classifiers were then tested against the reviewers' first and second topical choices for each document.

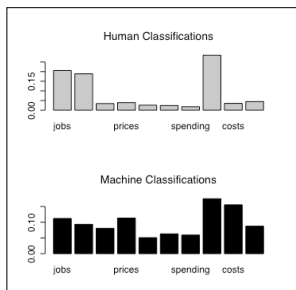
| Human Judges 1st Choice by Majority Vote | | |
|--|------------------|------------------|
| Model | Model 1st Choice | Model 2nd Choice |
| N.B. | 24 | 1 |
| N.B. (aug) | 14 | 24 |

But in the context of the relation browser, a correct topic in 2nd place is good. Under this lens, augmenting the training set appears to improve generalization.

| Human Judges 2nd Choice by Majority Vote | | |
|--|------------------|------------------|
| Model | Model 1st Choice | Model 2nd Choice |
| N.B. | 21 | 4 |
| N.B. (aug) | 14 | 21 |

$acc(N.B.) = 0.5$
 $acc(N.B. aug) = 0.73$

Room for Improvement



Are the 10 topics we've learned by document clustering independent? Human judges suggest not.

The classifiers distributed documents across topics more uniformly.

Current Work and Future Directions

- Relaxing k -means, we are using the EM (expectation maximization) algorithm (Mitchell, 1997) to derive a "soft clustering" of all documents, eliminating the need for a separate classification step
- To enable this, however, we need high-quality document features. We are exploring hypertext-based indicators of document relationships to supplement internal semantic evidence.
- Given the complexity of such large-scale clustering, can we reduce the dimensionality? Eigenvalue/eigenvector methods may help, but we are more enthusiastic about document projections along their independent components.

Open Questions

- Given the noise in text data, how can we constrain unsupervised learning algorithms to derive concepts that are useful for information architecture purposes?
- How can we take better advantage of current IA in a collection to improve the learning process? i.e. can we use pre-existing structures?
- Non-topical documents traits (time, location, audience) often inform IA activity. This approach can't speak to these concerns as it stands. Are there other approaches that could?

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
- Application of NLP techniques to improve our analysis of the term-space.

References

- Deerwester, S., et al. Indexing by latent semantic analysis. *JASIS*, 41(6), 1990, 391-407.
- Han, H. et al. Automatic document metadata extraction using support vector machines. *Joint Conference on Digital Libraries. JCDL '03*. 2003.
- Joachims, T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text classification. In D. E. Fisher, Ed., *Proceedings of ICML-97 14th International Conference on Machine Learning*. 143-154, 1997.
- Joachims, T. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*. 137-142, 1998.
- Kohonen, T. *Self-Organizing Maps*. Information Sciences. Springer, second edition, 1997.
- 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.
- Marchionini, G. and B. Brunk. Toward a general relation browser: a GUI for information architects. *Journal of Digital information*, 4(1), 2003.
- Mitchell, T. *Machine Learning*. Mc-Graw Hill, 1997.