Document Routing

- Dynamic document collection
- Stable user interests
- Based on content
- Varied terminology
  - Information Filtering
  - Selective Dissemination of Information

Retrieve

Route

Document Arrivals

Interest Change
Outline

- Vector Space Model
- User Modeling
- Gaussian User Model
- Discriminant Analysis
- Implications for Future Research
Model for Text Selection

- Document Representation
  - Vector Space
- User Interest Representation
  - Gaussian User Model
- Comparison Function
  - Cosine Similarity Measure
- User Interface
  - Rank order by decreasing similarity
Vector Space Model

- Choose "terms"
  - Remove stopwords
  - Apply stemming or other morphology
  - Add phrases
- Assign term weights
  - Emphasize within-document frequency
  - Deemphasize across-document frequency
- Document representation is the sparse vector of term weights
Cosine Similarity Measure

- Cosine of angle between two vectors
  - Deemphasizes document length
  - Fast computation on sparse vectors
    - Normalized inner product
- Monotone with:
  - Arc distance on a hypersphere
  - Euclidean distance on tangent plane
Relevance Judgements

- Match documents with topics
- For tractability, assume relevance is
  - Binary
  - Stable
  - Known
- Poor intra-rater & inter-rater reliability
Effectiveness Measures

• **Fixed size sets**
  - Recall: Fraction of the relevant documents that are in a set
  - Precision: Fraction of the documents in a set that are relevant

• **Ranked output**
  - Average Precision
    - Extend cutoff to increase recall
    - Area under precision-recall curve
User Modeling Approaches

- **Information Retrieval**
  - Explicit specification
  - Relevance feedback
- **Statistical Decision Theory**
  - Discriminant analysis
- **Machine Learning**
  - Neural networks
  - Genetic algorithms
Latent Semantic Indexing

- **Sparse term-document matrix**
  - Columns represent documents
  - Rows represent terms
- **Orthogonal decomposition**
  - Singular Value Decomposition (SVD)
  - ULV decomposition
- **Rank reduction**
  - Smaller singular values result from noise
  - Noise equates to term usage variation
- **Retains "conceptual" dimensions**
LSI Representation

LSI Represents Terms and Documents in k-space

After Berry, Dumais and Letsche, 1995
Bellcore LSI Routing

- Segregate training examples by topic
- Map positive examples into LSI space
- Form mean vector to represent topic
  - Insight is that relevant documents cluster
- Use cosine similarity measure in LSI space
TREC Results

- LSI Routing outperforms explicit topic specification
- Gain from LSI is hard to characterize
  - Cornell team did well with simple relevance feedback
Gaussian User Model

- **Goal:** exploit mean and covariance
  - Distribution is unimodal but not symmetric
- **Use LSI to reduce the feature space**
  - Find sample mean with Bellcore technique
  - Estimate the covariance matrix
  - Bias similarity using covariance matrix
Covariance Estimation

- Sample covariance matrix may not have full rank
- Spectral decomposition reveals excessive eigenvalue skew
- Regularization compensates for both problems
  - Linear combination of each eigenvalue with the mean eigenvalue
  - Parameter chosen using cross-validation
Comparison Function

- Asymmetric angular distance measure
  - Emphasize directions with low variance
- Mahalanobis distance in a hyperplane
  - Normal to the mean vector
  - Tangent to the unit hypersphere
Experimental Results

- Evaluated on Cranfield collection
  - 1398 documents and 225 queries
  - Used cross-validation to find average rank
- Optimum parameter choice is 1.0 times mean eigenvalue
  - Equivalent to Bellcore routing technique!
- Using covariance information hurts routing performance
  - Suggests that training data is biased
Experimental Results

![Graph showing the relationship between Average Precision and Fraction of Mean Eigenvalue. The graph includes a line labeled "Raw_Term_Frequency." ]
Alternative Approaches

- Characterize non-relevant training documents
  - Least intrusive approach
- Explore the document space
  - Random training data
  - System-directed exploration
Discriminant Analysis

- Assume both sets are Gaussian
- Find boundary which minimizes classification error
- Rank order by Mahalanobis distance from the boundary
- Evaluated on TREC at Xerox PARC
  - Quadratic and linear boundaries
  - With and without regularization
- No better than Bellcore technique
Observations on Learning

- Users can associate positive examples with topics
  - Supervised learning works
- How to classify negative examples?
  - Disjunctive induction
    - Greedy covering
    - Exception-based learning
  - Two-pass approach
- Training data is expensive
  - Online learning
Take-Away Message

- User modeling for document routing is a machine learning task with biased noisy binary training data in which we seek to learn a rank ordering of the documents.
- In the vector space model the relevant documents have a unimodal distribution but the non-relevant document distribution is multimodal.
Conclusions

- Two-pass technique is a good choice
  - Significant computational complexity
- Unsupervised learning algorithms offer promise
  - Clustering techniques
  - Neural networks
For Further Reading

- http://www.ee.umd.edu/medlab/filter/
  - These slides
  - Bellcore papers
  - Xerox PARC papers
  - Other research groups