Mining Text Data

Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Outline

- **Introduction**
- Document Preparation and Similarity Computation
- Specialized Clustering Methods
- Topic Modeling
- Specialized Classification Methods
- Novelty and First Story Detection
- Summary

Introduction

- Text data are copiously found in
  - Digital libraries: digitized book and paper
  - Web and Web-enabled applications: hypertext (side information), social network, Microblog, WeChat
  - Newswire services: Sina, NetEase

- Modeling of Text
  - A sequence (string)
  - A multidimensional record
    - ✓ More Popular

Multidimensional Representations

- Terminology
  - Data point: document
  - Data set: corpus
  - Feature: word, term
  - The set of features: lexicon
- Vector Space Representation
  1. Common words are removed
  2. Variations of the same word are consolidated
  3. Normalized frequencies are associated with the individual words

Specific Characteristics of Text

- Number of “Zero” Attributes (Sparsity)
  - A document may contain only a few hundred words
  - Affect many fundamental aspects of text mining, such as distance computation
- Nonnegativity
  - Frequencies are nonnegative
  - The presence of a word is statistically more significant than its absence
- Side Information
  - Hyperlinks or other Metadata
  - Friendship in social network

Outline

- Introduction
- **Document Preparation and Similarity Computation**
- Specialized Clustering Methods
- Topic Modeling
- Specialized Classification Methods
- Novelty and First Story Detection
- Summary

Feature Extraction

- Stop Word Removal
  - Words in a language that are not very discriminative for mining
  - Articles, prepositions, and conjunctions
- Stemming
  - Consolidate variations of the same
  - Singular and plural representations
  - Different tenses of the same word
- Punctuation Marks
  - Commas, semicolons, digits, hyphens

Document Normalization

- Inverse Document Frequency

$$id_i = \log(n/n_i)$$

  - n_i is the number of documents in which the ith term occurs

- Frequency Damping

$$f(x_i) = \sqrt{x_i}$$

$$f(x_i) = \log(x_i).$$

  - x_i is the frequency of the ith term

- Normalized Frequency

$$h(x_i) = f(x_i)id_i$$

Similarity Computation

- The Cosine Measure

$$\cos(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} h(x_i)h(y_i)}{\sqrt{\sum_{i=1}^{d} h(x_i)^2}\sqrt{\sum_{i=1}^{d} h(y_i)^2}}$$

- Jaccard Coefficient

$$J(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} h(x_i)h(y_i)}{\sum_{i=1}^{d} h(x_i)^2 + \sum_{i=1}^{d} h(y_i)^2 - \sum_{i=1}^{d} h(x_i)h(y_i)}$$

  - Commonly used in sparse binary data as well as sets

Specialized Preprocessing for Web Documents

- Leverage the Structure
  - Title is more important than body
  - Add anchor text to the document which it points to
- Remove Specific Parts
  - Remove tags
  - Identify the main block
    - ✓ Block labeling as a classification problem
      - Extracts visual features, label manually
    - ✓ Tree matching approach
      - Extract tag trees, determine template

Outline

- Introduction
- Document Preparation and Similarity Computation
- **Specialized Clustering Methods**
- Topic Modeling
- Specialized Classification Methods
- Novelty and First Story Detection
- Summary

Representative-Based Algorithms

- The k-Means Algorithm
  - Sum of Square Errors

$$\min_{\bar{Y}_1,\ldots,\bar{Y}_k} O = \sum_{i=1}^{n} \left[\min_j \left\| \bar{X}_i - \bar{Y}_j \right\|_2^2 \right]$$

1. Assign Step: determine clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$

$$C(\bar{X}_i) = \operatorname{argmin}_j \left\| \bar{X}_i - \bar{Y}_j \right\|_2^2$$

2. Optimize Step

$$\bar{Y}_j = \operatorname{argmin}_{\bar{Y}} \sum_{\bar{X}_i \in \mathcal{C}_j} \|\bar{X}_i - \bar{Y}\|_2^2 = \frac{1}{|\mathcal{C}_j|} \sum_{\bar{X}_i \in \mathcal{C}_j} \bar{X}_i$$

Representative-Based Algorithms

- Modifications
  - Choice of the Similarity Function
    - ✓ Cosine similarity function
  - Computation of the Cluster Centroid
    - ✓ The low-frequency words in the cluster are projected out
    - ✓ Keep a representative set of topical words for the cluster (200 to 400 words
    - ✓ Have significant effectiveness advantages

Scatter/Gather Approach

- While the k-means algorithm is more efficient $O(kn)$, it is sensitive to the choice of seeds
- While hierarchical partitioning algorithms are very robust, they typically scale worse than $\Omega(n^2)$
- A Two-phase Approach
  1. Apply either the *buckshot* or *fractionation* procedures to create a robust set of initial seeds
  2. Apply a k-means approach on the resulting set of seeds

Buckshot

1. Select a seed superset of size $\sqrt{kn}$
   - k is the number of clusters
   - n is the number of documents
2. Agglomerates them to k seeds
   - The time complexity is $O(kn)$

- Bottom-up (agglomerative) Methods
  - The individual data points are successively agglomerated into higher-level clusters

Fractionation (1)

1. Break up the corpus into n/m buckets, each of size m
2. An agglomerative algorithm is applied to each bucket to reduce them by a factor ν
3. Then, we obtain νn agglomerated documents over all buckets
   - Concatenation of the documents in a cluster
4. Repeat the above process until k agglomerated documents

Fractionation (2)

Types of Partition

1. Random partitioning

A. Sort the documents by the index of the jth most common word in the document,

B. Contiguous groups of m documents in this sort order are mapped to clusters

Time Complexity

- $O\big(nm(1 + v + v^2 + \cdots)\big) = O(nm)$

k-means algorithm

- Each document is assigned to the nearest of the k cluster centers
- The centroid of each such cluster is determined as the concatenation of the documents in that cluster
- Furthermore, the less frequent words of each centroid are removed.

Enhancements

- Split Operation
  1. Identify groups that are not very coherent
     - ✓ Average similarity of the documents in a cluster to its centroid or each other
  2. Apply the buckshot procedure by using $k = 2$ and then recluster
- Join Operation
  - Merge similar clusters into a single one
    - ✓ Topical words of each cluster are computed
    - ✓ Clusters with significant overlap between the topical words

Probabilistic Algorithms

- Unsupervised Naïve Bayes
- The Generative Process
  1. Select a cluster $\mathcal{G}_m$, where $m \in \{1, \dots, k\}$
  2. Generate the document based on the term distribution of $\mathcal{G}_m$
     - ✓ Bernoulli Model or multinomial model
- Parameters
  - Prior probability $P(\mathcal{G}_m)$
  - Conditional distribution $P(w_j | \mathcal{G}_m)$

The EM Algorithm

- E-step: Estimate posterior probability of membership of documents to clusters using Bayes rule

$$P(\mathcal{G}_m|\overline{X}) \propto P(\mathcal{G}_m) \prod_{w_j \in \overline{X}} P(w_j|\mathcal{G}_m) \prod_{w_j \notin \overline{X}} (1 - P(w_j|\mathcal{G}_m))$$

- M-step: Estimate $P(w_j|\mathcal{G}_m)$ and $P(\mathcal{G}_m)$

$$P(w_j|\mathcal{G}_m) = \frac{\sum_{\overline{X}} P(\mathcal{G}_m|\overline{X}) \cdot I(\overline{X}, w_j)}{\sum_{\overline{X}} P(\mathcal{G}_m|\overline{X})}$$

$$P(\mathcal{G}_m) = \frac{\sum_{\bar{X}} P(\mathcal{G}_m|\bar{X})}{n}$$

Simultaneous Document and Word Cluster Discovery

Co-clustering

- Rearrange the rows and columns

	CHAMPION	ELECTRON	TROPHY	RELATIVITY	QUANTUM	TOURNAMENT
D_1	2	0	1	1	0	3
D_2	0	2	0	1	3	0
D_3	1	3	0	1	2	0
D_4	2	0	2	0	0	3
D_5	0	2	1	1	3	0
D_6	1	0	2	0	0	3

(a) Document-term matrix

SPORTS CO-CLUSTER

	CHAMPION	TROPHY	TOURNAMENT	ELECTRON	RELATIVITY	QUANTUM
D_1	2	1	3	0	1	0
D_4	2	2	3	0	0	0
D_6	1	2	3	0	0	0
D_2	0	0	0	2	1	3
D_3	1	0	0	3	1	2
D_5	0	1	0	2	1	3

PHYSICS CO-CLUSTER

(b) Re-arranged document-term matrix

Simultaneous Document and Word Cluster Discovery

- Co-clustering
  - the ith cluster is associated with a set of rows $\mathcal{R}_i$ (documents) and a set of columns $\mathcal{V}_i$ (words)
  - The rows $\mathcal{R}_i$ are disjoint from one another over different values of i
  - The columns $\mathcal{V}_i$ are disjoint from one another over different values of i
  - The words representing the columns of $\mathcal{V}_i$ are topical words for cluster $\mathcal{R}_i$

How can the co-clustering problem be solved?

- Minimize the weights of the nonzero entries outside these shaded blocks

	CHAMPION	TROPHY	TOURNAMENT	ELECTRON	RELATIVITY	QUANTUM
D_1	2	1	3	0	1	0
D_4	2	2	3	0	0	0
D_6	1	2	3	0	0	0
D_2	0	0	0	2	1	3
D_3	1	0	0	3	1	2
D_5	0	1	0	2	1	3

SPORTS CO-CLUSTER →

← PHYSICS CO-CLUSTER

(b) Re-arranged document-term matrix

A Bipartite Graph Partitioning Problem

- A node set N_d
  - Each node represents a document
- A node set N_w
  - Each node represents a word
- An undirected bipartite graph $G = (N_d \cup N_w, A)$
  - An edge (i, j) corresponds to a nonzero entry in the document-term matrix
  - The weight of an edge is equal to the frequency of the term in the document

A Undirected Bipartite Graph

- 2-way Cut

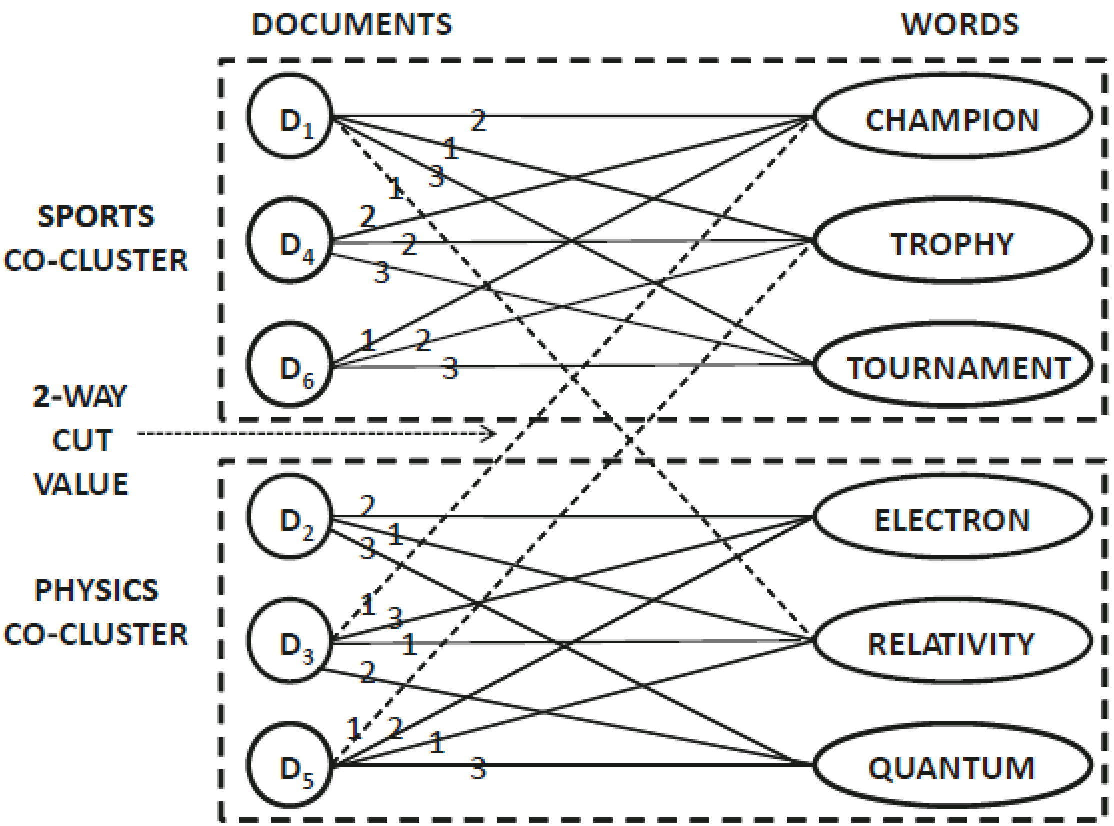


Figure 13.2: Graph partitioning for co-clustering

Each partition contains a set of documents and a corresponding set of words

A Undirected Bipartite Graph

- 2-way Cut

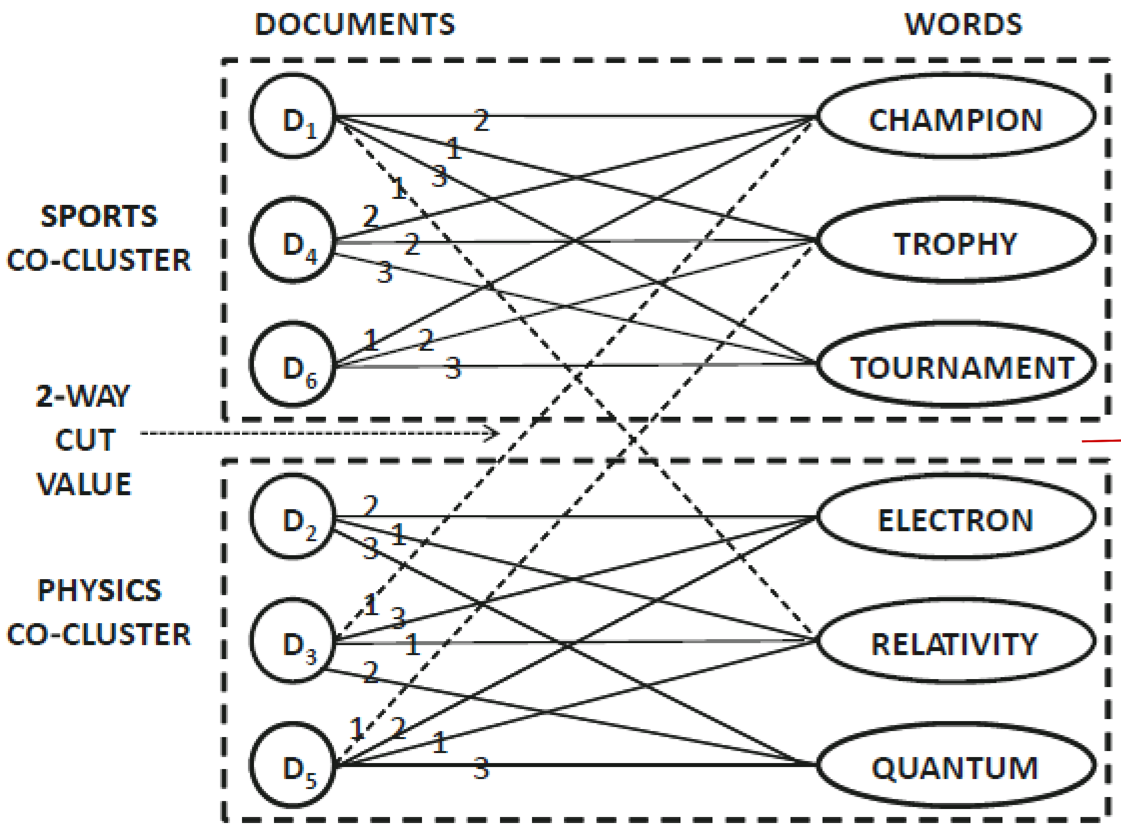


Figure 13.2: Graph partitioning for co-clustering

Edges across the partition correspond to nonzero entries in the nonshaded regions

The General Procedure

□ A k-way Co-clustering Problem

1. Create a graph $G = (N_d \cup N_w, A)$ with nodes in N_d representing documents, nodes in N_w representing words, and edges in A with weights representing nonzero entries in matrix D.

2. Use a k-way graph partitioning algorithm to partition the nodes in $N_d \cup N_w$ into k groups.

3. Report row–column pairs $(\mathcal{R}_i \mathcal{V}_i)$ for $i \in \{1 \dots k\}$. Here, $\mathcal{R}_i$ represents the rows corresponding to nodes in N_d for the ith cluster, and $\mathcal{V}_i$ represents the columns corresponding to the nodes in N_w for the ith cluster.

- Graph partitioning is addressed in Sect. 19.3 of Chap. 19
- Actually, spectral clustering can be applied

Outline

- Introduction
- Document Preparation and Similarity Computation
- Specialized Clustering Methods
- **Topic Modeling**
- Specialized Classification Methods
- Novelty and First Story Detection
- Summary

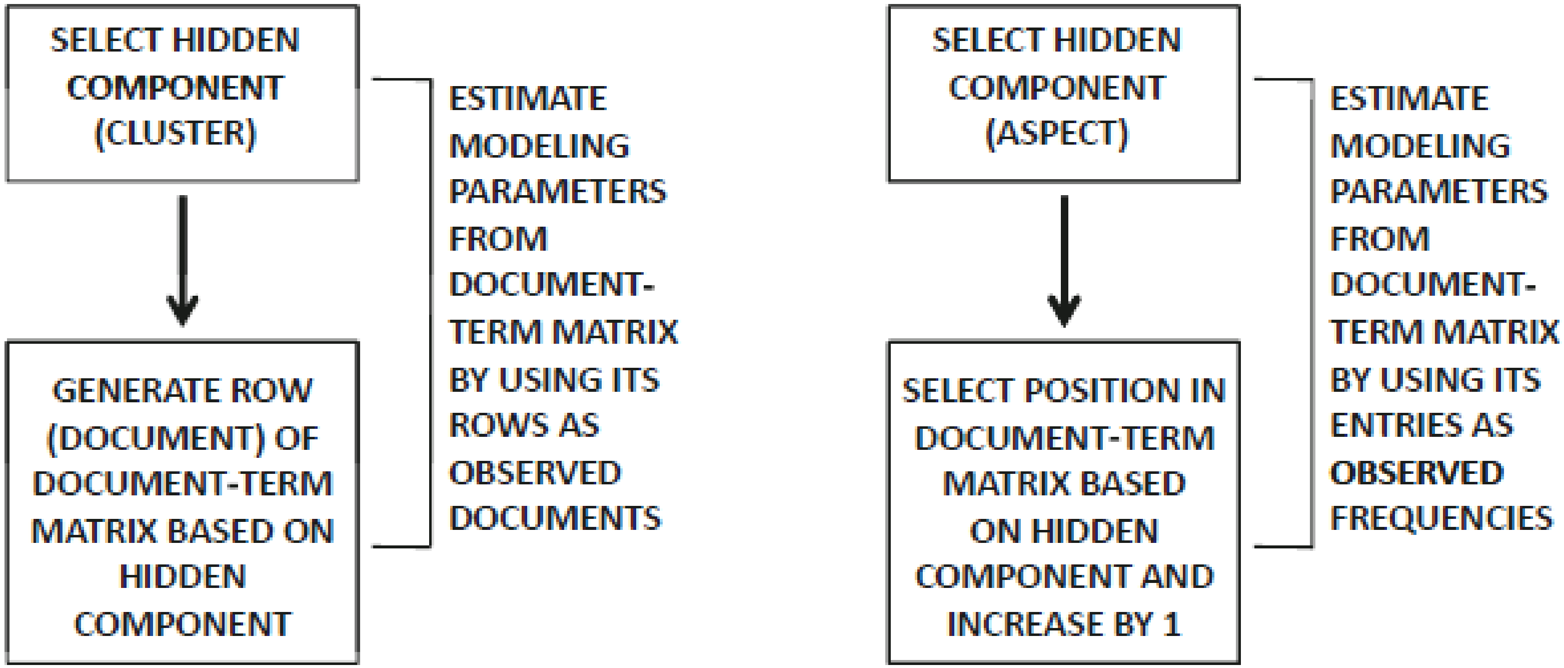
Probabilistic Latent Semantic Analysis (PLSA)

- A Probabilistic Variant of LSA (SVD)
- An Expectation Maximization-based Mixture Modeling Algorithm
  - Designed for dimensionality reduction rather than clustering

1. Select a latent component $\mathcal{G}_m$, where $m \in \{1, \dots, k\}$
2. Generate the indices (i, j) or $(\bar{X}_i, w_j)$ with probabilities $P(\bar{X}_i|\mathcal{G}_m)$ and $P(w_j|\mathcal{G}_m)$
   - ✓ The frequency of entry (i, j) in the document-term matrix is increased by 1

EM-clustering v.s. PLSA (1)

- Row v.s. Entry



(a) EM-clustering (section 13.3.2) (b) *PLSA*

EM-clustering v.s. PLSA (2)

- The clustering model generates a document from a unique hidden component (cluster)
  - The final soft clustering is due to uncertainty in estimation from observed data
- In PLSA, different parts of the same document may be generated by different aspects, even at the generative modeling level
  - Documents are generated by a combination of mixture components

The EM Algorithm (1)

- (E-step) Estimate posterior Probability $P(\mathcal{G}_m|\overline{X}_i, w_j)$ for each entry
  - The Bayes rule

$$P(\mathcal{G}_m|\overline{X_i}, w_j) = \frac{P(\mathcal{G}_m) \cdot P(\overline{X_i}, w_j|\mathcal{G}_m)}{P(\overline{X_i}, w_j)}$$

  - Conditionally independent assumption

$$P(\overline{X_i}, w_j|\mathcal{G}_m) = P(\overline{X_i}|\mathcal{G}_m) \cdot P(w_j|\mathcal{G}_m)$$

  - Law of total probability

$$P(\overline{X_i}, w_j) = \sum_{m=1}^{k} P(\mathcal{G}_m) \cdot P(\overline{X_i}, w_j|\mathcal{G}_m) = \sum_{m=1}^{k} P(\mathcal{G}_m) \cdot P(\overline{X_i}|\mathcal{G}_m) \cdot P(w_j|\mathcal{G}_m)$$

The EM Algorithm (2)

- (M-step) Estimate $P(\mathcal{G}_m)$, $P(\overline{X_i}|\mathcal{G}_m)$ and $P\left(w_j|\mathcal{G}_m\right)$

$$P(\overline{X_i}|\mathcal{G}_m) \propto \sum_{w_j} f(\overline{X_i}, w_j) \cdot P(\mathcal{G}_m|\overline{X_i}, w_j) \quad \forall i \in \{1 \dots n\}, m \in \{1 \dots k\}$$

$$P(w_j|\mathcal{G}_m) \propto \sum_{\overline{X_i}} f(\overline{X_i}, w_j) \cdot P(\mathcal{G}_m|\overline{X_i}, w_j) \quad \forall j \in \{1 \dots d\}, m \in \{1 \dots k\}$$

$$P(\mathcal{G}_m) \propto \sum_{\overline{X_i}} \sum_{w_j} f(\overline{X_i}, w_j) \cdot P(\mathcal{G}_m|\overline{X_i}, w_j) \quad \forall m \in \{1 \dots k\}.$$

  - $f(\overline{X}_i, w_j)$ represent the observed frequency of the occurrence of word w_j in document $\overline{X}_i$

PLSA for Dimensionality Reduction (1)

- We have the following relation

$$P(\overline{X_i}, w_j) = \sum_{m=1}^{k} P(\mathcal{G}_m) \cdot P(\overline{X_i}|\mathcal{G}_m) \cdot P(w_j|\mathcal{G}_m)$$

  - $D_k \in \mathbb{R}^{n\times d}$ be a matrix with $[D_k]_{ij} = P(\bar{X}_i, w_j)$
  - $Q_k \in \mathbb{R}^{n\times k}$ be a matrix with $[Q_k]_{im} = P(\bar{X}_i|\mathcal{G}_m)$
  - $P_k \in \mathbb{R}^{d\times k}$ be a matrix with $[P_k]_{jm} = P(w_j|\mathcal{G}_m)$
  - $\Sigma_k \in \mathbb{R}^{k\times k}$ be a diagonal matrix with $[\Sigma_k]_{mm} = P(\mathcal{G}_m)$

$$D_k = Q_k \Sigma_k P_k^T$$

PLSA for Dimensionality Reduction (2)

- Let D be the Scaled data matrix
  - The summation of entries in D is 1

$$D \approx D_k = Q_k \Sigma_k P_k^\top$$

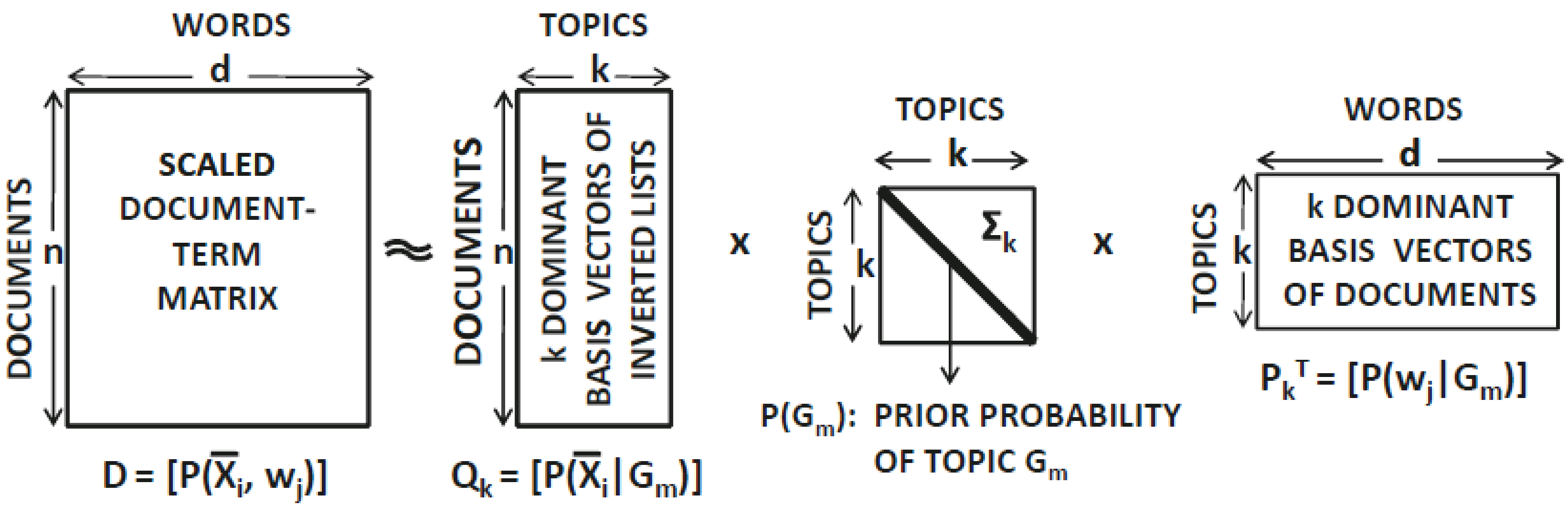


Figure 13.4: Matrix factorization of *PLSA*

PLSA for Dimensionality Reduction (3)

- Let D be the Scaled data matrix
  - The summation of entries in D is 1

$$D \approx D_k = Q_k \Sigma_k P_k^\top$$

  - $Q_k \Sigma_k \in \mathbb{R}^{n \times k}$ provide k-dimensional representations of documents
  - $\Sigma_k P_k^\top \in \mathbb{R}^{k \times d}$ provide k-dimensional representations of terms

PLSA v.s. LSA v.s. NMF

- PLSA
  - Nonnegative and have clear probabilistic interpretability (topical words of aspects)
  - Out-of-sample extension is difficult
- LSA (SVD)
  - The columns of Q_k/P_k are orthonormal
  - Out-of-sample extension is straightforward
- NMF
  - Nonnegative (but a different objective)
  - Out-of-sample extension is difficult

Synonymy

$$
\begin{array}{ll}
 & \begin{array}{cccccc} \text{LION} & \text{TIGER} & \text{CHEETAH} & \text{JAGUAR} & \text{PORSCHE} & \text{FERRARI} \end{array} \\
\begin{array}{r} \text{CATS}\left\{\begin{array}{r} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \end{array}\right. \\ \text{BOTH}\left[\,\overline{X}_4\right. \\ \text{CARS}\left\{\begin{array}{r} \overline{X}_5 \\ \overline{X}_6 \end{array}\right. \end{array} &
\left[\begin{array}{cccccc}
2/41 & 2/41 & 1/41 & 2/41 & 0 & 0 \\
2/41 & 3/41 & 3/41 & 3/41 & 0 & 0 \\
1/41 & 1/41 & 1/41 & 1/41 & 0 & 0 \\
2/41 & 2/41 & 2/41 & 3/41 & 1/41 & 1/41 \\
0 & 0 & 0 & 1/41 & 1/41 & 1/41 \\
0 & 0 & 0 & 2/41 & 1/41 & 2/41
\end{array}\right]
\end{array}
\approx
\begin{array}{ll}
 & \begin{array}{cc} \text{CATS} & \text{CARS} \end{array} \\
\begin{array}{r} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \end{array} &
\left[\begin{array}{cc}
2/8 & 0 \\
3/8 & 0 \\
1/8 & 0 \\
2/8 & 1/4 \\
0 & 1/4 \\
0 & 2/4
\end{array}\right]
\end{array}
\times
\left[\begin{array}{cc}
8/11 & 0 \\
0 & 3/11
\end{array}\right]
\times
\begin{array}{ll}
 & \begin{array}{cccccc} \text{LION} & \text{TIGER} & \text{CHEETAH} & \text{JAGUAR} & \text{PORSCHE} & \text{FERRARI} \end{array} \\
\begin{array}{r} \text{CATS} \\ \text{CARS} \end{array} &
\left[\begin{array}{cccccc}
1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\
0 & 0 & 0 & 1/3 & 1/3 & 1/3
\end{array}\right]
\end{array}
$$

$$D \approx Q_k \times \Sigma_k \times P_k^T$$

- Two documents containing “cat” and “kitten” have positive values of the transformed coordinate for aspect “cats”

Polysemy

		LION	TIGER	CHEETAH	JAGUAR	PORSCHE	FERRARI
CATS	$\overline{X}_1$	2/41	2/41	1/41	2/41	0	0
	$\overline{X}_2$	2/41	3/41	3/41	3/41	0	0
	$\overline{X}_3$	1/41	1/41	1/41	1/41	0	0
BOTH	$\overline{X}_4$	2/41	2/41	2/41	3/41	1/41	1/41
CARS	$\overline{X}_5$	0	0	0	1/41	1/41	1/41
	$\overline{X}_6$	0	0	0	2/41	1/41	2/41

D

$\approx$

	CATS	CARS
$\overline{X}_1$	2/8	0
$\overline{X}_2$	3/8	0
$\overline{X}_3$	1/8	0
$\overline{X}_4$	2/8	1/4
$\overline{X}_5$	0	1/4
$\overline{X}_6$	0	2/4

Q_k

X

8/11	0
0	3/11

Σ_k

X

	LION	TIGER	CHEETAH	JAGUAR	PORSCHE	FERRARI
CATS	1/4	1/4	1/4	1/4	0	0
CARS	0	0	0	1/3	1/3	1/3

P_k^T

- A word with multiple meanings may have positive components in different aspects
- Other words in the document will reinforce one of these two aspects

PLSA for Clustering

- The 1st Way
  - Although it is designed for dimensionality reduction, it can also be applied to clustering by calculating

$$P(\mathcal{G}_m|\overline{X_i}) = \frac{P(\mathcal{G}_m) \cdot P(\overline{X_i}|\mathcal{G}_m)}{\sum_{r=1}^{k} P(\mathcal{G}_r) \cdot P(\overline{X_i}|\mathcal{G}_r)}$$

- The 2nd Way
  - Apply clustering algorithm, such as k-means, to $Q_k \Sigma_k \in \mathbb{R}^{n \times k}$

Limitations of PLSA

- Overfitting
  - Too many parameters $(n + d + 1)k$
- Out-of-sample extension is difficult
  - Cannot assign probabilities to unseen documents
- Latent Dirichlet Allocation (LDA)
  - Use Dirichlet priors on the topics
  - Generalizes easily to new documents

Outline

- Introduction
- Document Preparation and Similarity Computation
- Specialized Clustering Methods
- Topic Modeling
- **Specialized Classification Methods**
- Novelty and First Story Detection
- Summary

Instance-Based Classifiers

- k-nearest Neighbor Classifier
  - Find the top-k nearest neighbors with the cosine similarity
  - Return the dominant class label
    - ✓ Weight the vote with the cosine similarity

- Due to sparsity and high-dimensionality, it can be modified in two ways
  - Leverage Latent Semantic Analysis
  - Use fine-grained clustering

Leveraging Latent Semantic Analysis

- Synonymy and Polysemy lead to noise in cosine similarity
  - The significance of a word can be understood only in the context of other words in the document
- LSA ($X = U\Sigma V^{\top}$)
  - The removal of the dimensions with small eigenvalues typically leads to a reduction in the noise effects
  - $100{,}000 \rightarrow 300$
- PLSA can also be used

Centroid-Based Classification

- A fast alternative to k-nearest neighbor classifiers
  - Partition documents of each class into clusters
    - ✓ The number of clusters of each class is proportional to the number of documents in that class
  - Retaining most frequent words in centroid, which is referred to as a cluster digest
  - The k-nearest neighbor classification is performed with a smaller number of centroids

Advantages

- Efficient since the number of centroids is small
- Effective by addressing the issues of synonymy and polysemy indirectly

1. *Business schools:* business (35), management (31), school (22), university (11), campus (15), presentation (12), student (17), market (11), . . .
2. *Law schools:* law (22), university (11), school (13), examination (15), justice (17), campus (10), courts (15), prosecutor (22), student (15), . . .

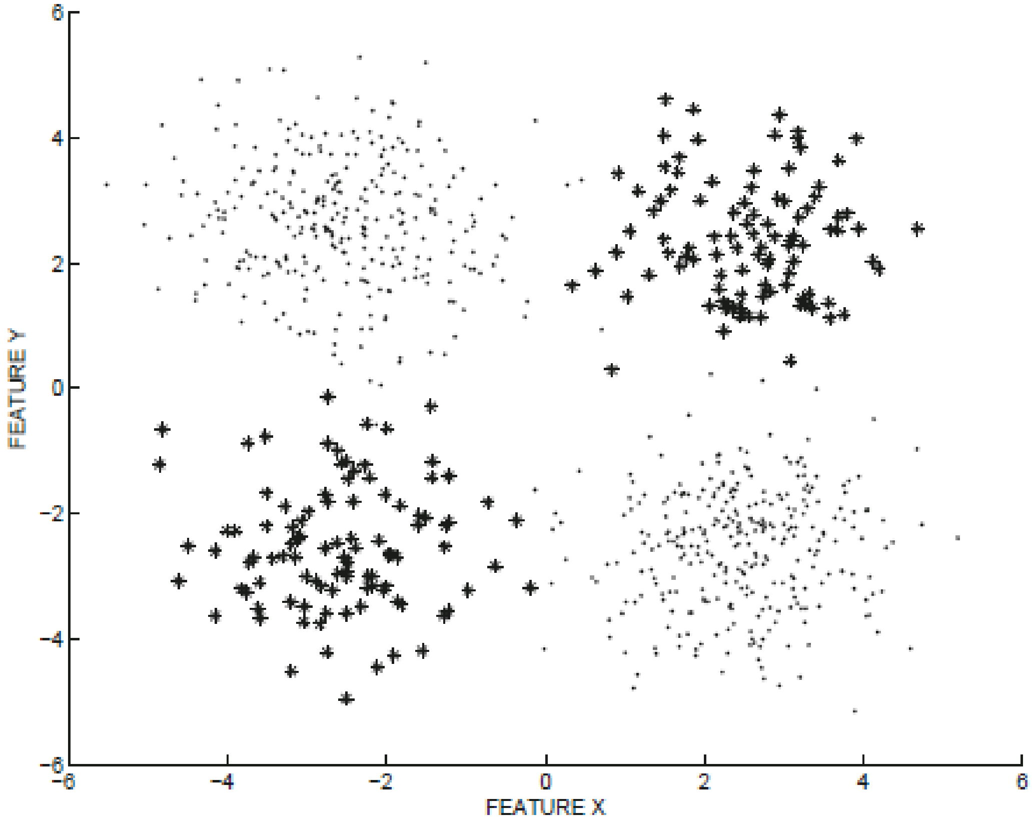
  - Similar words are represented in the same centroid
  - Words with multiple meanings can be represented in different centroids

A Special Case—Rocchio Classification

- All documents belonging to the same class are aggregated into a single centroid
  - Extremely fast
- The Class-contiguity Assumption
  - Documents in the same class form a contiguous region, and regions of different classes do not overlap

A Bad Case of Rocchio Classification

- Documents of the same class were separated into distinct clusters



Bayes Classifiers

- Bernoulli Bayes Model
  - The model for generating term is the Bernoulli model
  - Each term takes on the value of either 0 or 1
  - Does not account for the frequencies of the words in the documents
- Multinomial Bayes Model
  - The model for generating term is the Multinomial model

Bernoulli Bayes Model (1)

- The goal is to predict

$$P(C = c|x_1 = a_1, \ldots x_d = a_d)$$

- Bayes Rule

$$\begin{aligned} P(C = c|x_1 = a_1, \ldots x_d = a_d) &= \frac{P(C = c)P(x_1 = a_1, \ldots x_d = a_d|C = c)}{P(x_1 = a_1, \ldots x_d = a_d)} \\ &\propto P(C = c)P(x_1 = a_1, \ldots x_d = a_d|C = c). \end{aligned}$$

- Bernoulli Model

$$P(x_1 = a_1, \ldots x_d = a_d|C = c) = \prod_{j=1}^{d} P(x_j = a_j|C = c)$$

- The Final Probability

$$P(C = c|x_1 = a_1, \ldots x_d = a_d) \propto P(C = c) \prod_{j=1}^{d} P(x_j = a_j|C = c)$$

Bernoulli Bayes Model (2)

- Estimation of $P(x_i = a_i | C = c)$
  - Let $p(i, c)$ be the fraction of the documents in class c containing word i

$$P(x_i = 1 | C = c) = p(i, c)$$

$$P(x_i = 0 | C = c) = 1 - p(i, c)$$

- Limitations

$$P(C = c | x_1 = a_1, \ldots x_d = a_d) \propto P(C = c) \prod_{j=1}^{d} P(x_j = a_j | C = c)$$

  - Explicitly penalizes nonoccurrence of words in documents
  - Frequencies of words are ignored

Multinomial Bayes Model (1)

- Terms in a document are samples from a multinomial distribution

- The Generative Model of a Document $d = (a_1, \ldots, a_d)$
  - Sample a class c with a class-specific prior probability
  - Sample $L = \sum_{i=1}^{d} a_i$ terms with replacement from the term distribution of the chosen class c
    - ✓ which is a multinomial model

Multinomial Bayes Model (2)

- The number of possible ways to sample the different terms to result in $d = (a_1, \dots, a_d)$
$$\frac{L!}{\prod_{i:a_i>0} a_i!}$$

- The probability of each of these sequences
$$\prod_{i:a_i>0} p(i,c)^{a_i}$$

  - $p(i,c) = \frac{n(i,c)}{\sum_i n(i,c)}$ is estimated as the fractional number of occurrences of word i in class c including repetitions

Multinomial Bayes Model (3)

- The Class Conditional Feature Distribution

$$P(x_1 = a_1, \ldots x_d = a_d | C = c) \approx \frac{L!}{\prod_{i:a_i>0} a_i!} \prod_{i:a_i>0} p(i,c)^{a_i}$$

- The Posterior Probability

$$\begin{aligned} P(C = c | x_1 = a_1, \ldots x_d = a_d) &\propto P(C = c) \cdot P(x_1 = a_1, \ldots x_d = a_d | C = c) \\ &\approx P(C = c) \cdot \frac{L!}{\prod_{i:a_i>0} a_i!} \prod_{i:a_i>0} p(i,c)^{a_i} \\ &\propto P(C = c) \cdot \prod_{i:a_i>0} p(i,c)^{a_i}. \end{aligned}$$

  - Nonoccurrence of words is ignored

SVM Classifiers

- Linear classifiers tend to work well
  - Linear SVM without intercept

$$\text{(OP1): Minimize } \frac{||\overline{W}||^2}{2} + C\frac{\sum_{i=1}^{n}\xi_i}{n}$$
$$\text{subject to: } y_i\overline{W}\cdot\overline{X_i} \geq 1-\xi_i \quad \forall i$$
$$\xi_i \geq 0 \quad \forall i.$$

- SVMPerf method

$$\text{(OP2): Minimize } \frac{||\overline{W}||^2}{2} + C\xi$$
$$\text{subject to: } \frac{1}{n}\sum_{i=1}^{n}u_i y_i\overline{W}\cdot\overline{X_i} \geq \frac{\sum_{i=1}^{n}u_i}{n} - \xi \quad \forall\, \overline{U} \in \{0,1\}^n$$
$$\xi \geq 0.$$

2^n constraints

SVM Classifiers

- Linear classifiers tend to work well
  - Linear SVM without intercept

$$\text{(OP1): Minimize } \frac{||\overline{W}||^2}{2} + C\frac{\sum_{i=1}^n \xi_i}{n}$$

$$\text{subject to: } y_i \overline{W} \cdot \overline{X_i} \geq 1 - \xi_i \quad \forall i$$

Lemma 13.5.1 *A one-to-one correspondence exists between solutions of (OP1) and (OP2), with equal values of* $\overline{W} = \overline{W^*}$ *in both models, and* $\xi^* = \frac{\sum_{i=1}^n \xi_i^*}{n}$.

$$\text{(OP2): Minimize } \frac{||\overline{W}||^2}{2} + C\xi$$

$$\text{subject to: } \frac{1}{n}\sum_{i=1}^{n} u_i y_i \overline{W} \cdot \overline{X_i} \geq \frac{\sum_{i=1}^n u_i}{n} - \xi \quad \forall \overline{U} \in \{0,1\}^n$$

$$\xi \geq 0.$$

Why (OP2) is a better formulation than (OP1)?

- A Single Slack Variable
  - Although the number of constraints is exponential
- Never use all the constraints explicitly

1. Determine optimal solution $(\overline{W}, \xi)$ for objective function of (OP2) using only constraints in the working set $\mathcal{WS}$.

2. Determine most violated constraint among the 2^n constraints of (OP2) by setting u_1 to 1 if $y_i \overline{W} \cdot \overline{X_i} < 1$, and 0 otherwise.

3. Add the most violated constraint to $\mathcal{WS}$.

  - For a constant size working set $\mathcal{WS}$, the time complexity is $O(ns)$
  - Terminates in a small number of iterations

Outline

- Introduction
- Document Preparation and Similarity Computation
- Specialized Clustering Methods
- Topic Modeling
- Specialized Classification Methods
- **Novelty and First Story Detection**
- Summary

Novelty and First Story Detection

- In the context of streams of news
  - A first story on a new topic needs to be reported as soon as possible
- The problem of first story detection
  - Determine novelties from the underlying text stream based on the history
- A simple approach
  - Compute the maximum similarity of the current document with all previous ones
  - Report the documents with very low maximum similarity values as novelties

Novelty and First Story Detection

- In the context of streams of news
  - A first story on a new topic needs to be reported as soon
- The problem of f
  - Determine nove
    text stream bas

> - High Computational Cost
>   - ✓ Reservoir sampling
> - Pairwise similarity is unstable
>   - ✓ Synonymy and Polysemy

- A simple approach
  - Compute the maximum similarity of the current document with all previous ones
  - Report the documents with very low maximum similarity values as novelties

Micro-clustering Method

- Simultaneously determines the clusters and novelties
  - Maintains k different cluster centroids
  - For an incoming document, its similarity to all the centroids is computed
    - ✓ If this similarity is larger than a user-defined threshold, then the document is added to the cluster and update the centroid
    - ✓ Otherwise, the incoming document is reported as a novelty, create a new cluster and delete one old cluster

Outline

- Introduction
- Document Preparation and Similarity Computation
- Specialized Clustering Methods
- Topic Modeling
- Specialized Classification Methods
- Novelty and First Story Detection
- **Summary**

Summary

- Document Preparation and Similarity Computation
  - TF, IDF, Cosine measure
- Specialized Clustering Methods
  - Representative-based algorithms, Probabilistic algorithms, Co-clustering
- Topic Modeling
  - PLSA, Dimensionality reduction, clustering
- Specialized Classification Methods
  - Instance-based classifiers, Bayes classifiers, SVM classifiers
- Novelty and First Story Detection
  - Micro-clustering method