Enhancing Document Clustering Using Hybrid Models for Semantic Similarity

Ahmed K. Farahat Mohamed S. Kamel

Pattern Analysis and Machine Intelligence (PAMI) Research Group
Department of Electrical and Computer Engineering
University of Waterloo

Text Mining Workshop 2010

Outline

1. Introduction
2. Background and Related Work
   - Document Clustering
   - Document Representation
3. Hybrid Models for Semantic Analysis
4. Experiments and Results
5. Conclusion and Future Work

Outline

1. Introduction
2. Background and Related Work
   - Document Clustering
   - Document Representation
3. Hybrid Models for Semantic Analysis
4. Experiments and Results
5. Conclusion and Future Work

Introduction

Document clustering is a fundamental task in text mining that is concerned with the organization of documents into groups according to their topics.

Document clustering has been used to organize documents for end-user browsing, and to improve the effectiveness and efficiency of other text mining tasks.

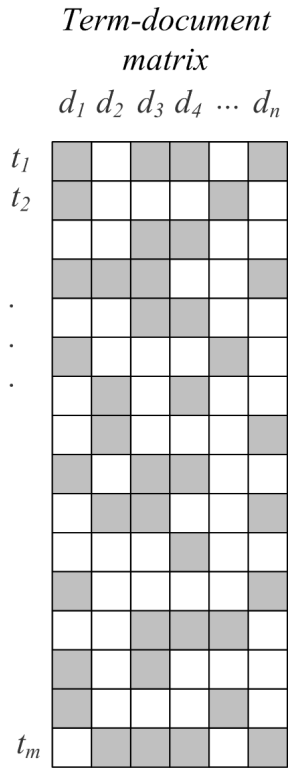
Introduction

Document clustering is a fundamental task in text mining that is concerned with the organization of documents into groups according to their topics.

Document clustering has been used to organize documents for end-user browsing, and to improve the effectiveness and efficiency of other text mining tasks.

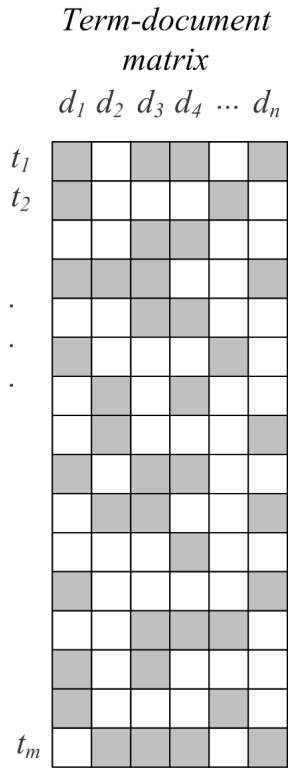
Introduction (Cont.)

- Most of document clustering algorithms use vector space model (VSM) for document representation.
- VSM represents documents as vectors in the space of terms, and measures proximity between document vectors using inner-products.



Introduction (Cont.)

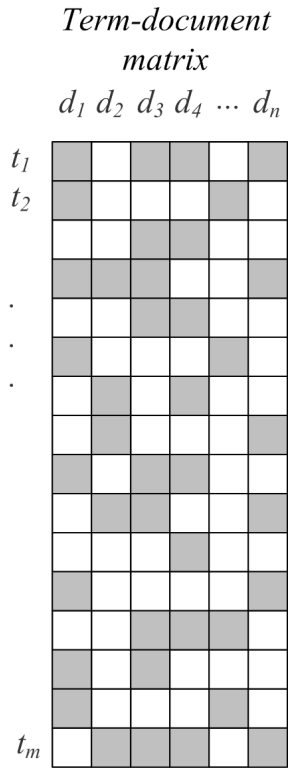
- Most of document clustering algorithms use vector space model (VSM) for document representation.
- VSM represents documents as vectors in the space of terms, and measures proximity between document vectors using inner-products.



Motivations

Problems with VSM

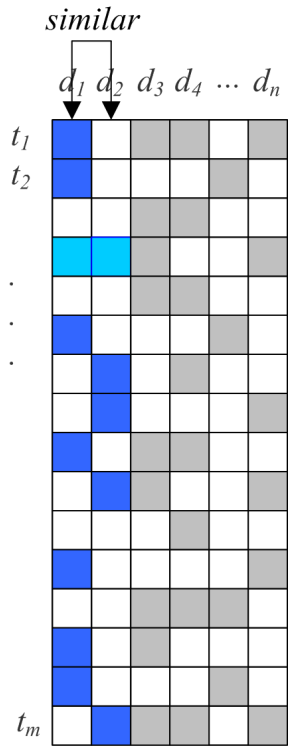
- VSM ignores semantic relations between terms.
- Proximity of document vectors in the space of terms does not reflect the true semantic similarity between documents.



Motivations

Problems with VSM

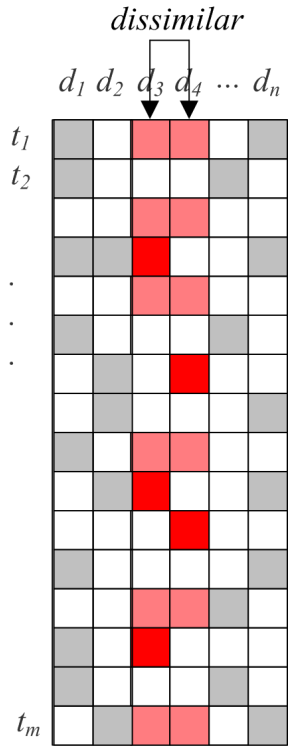
- VSM ignores semantic relations between terms.
- Proximity of document vectors in the space of terms does not reflect the true semantic similarity between documents.



Motivations

Problems with VSM

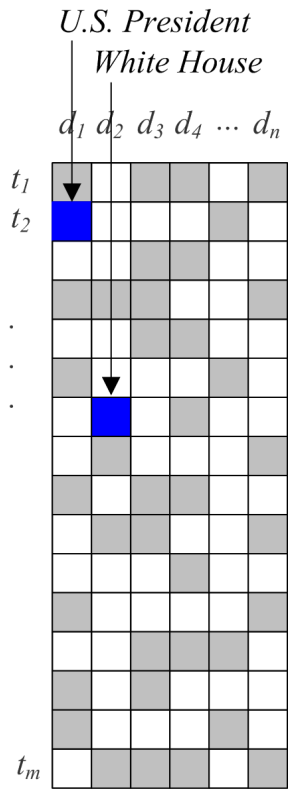
- VSM ignores semantic relations between terms.
- Proximity of document vectors in the space of terms does not reflect the true semantic similarity between documents.



Motivations

Problems with VSM

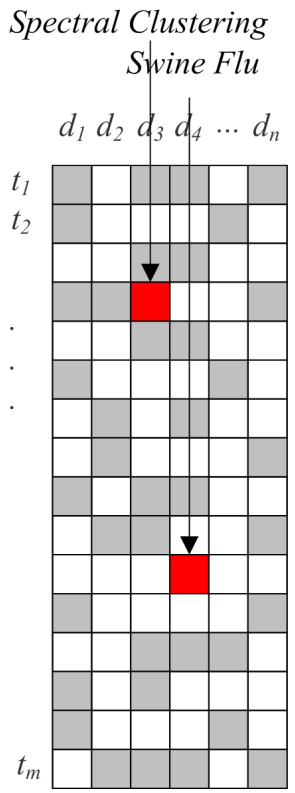
- VSM ignores semantic relations between terms.
- Proximity of document vectors in the space of terms does not reflect the true semantic similarity between documents.



Motivations

Problems with VSM

- VSM ignores semantic relations between terms.
- Proximity of document vectors in the space of terms does not reflect the true semantic similarity between documents.



Semantic Similarity Models

Explicit Models

Estimate similarity between documents based on measures of correlations between terms

- Generalized VSM (GVSM) - Wong et al. 1985 [13]

Latent Models

Use dimension reduction techniques to obtain a latent representation of concepts

- Latent semantic indexing (LSI) - Deerwester et al. 1990 [2]

Semantic Similarity Models

Explicit Models

Estimate similarity between documents based on measures of correlations between terms

- Generalized VSM (GVSM) - Wong et al. 1985 [13]

Latent Models

Use dimension reduction techniques to obtain a latent representation of concepts

- Latent semantic indexing (LSI) - Deerwester et al. 1990 [2]

This Work

Propose new hybrid models that combine explicit and latent models of semantic similarity

Use these hybrid models to enhance the effectiveness of document clustering algorithms

This Work

Propose new hybrid models that combine explicit and latent models of semantic similarity

Use these hybrid models to enhance the effectiveness of document clustering algorithms

Outline

1 Introduction

2 Background and Related Work
- Document Clustering
- Document Representation

3 Hybrid Models for Semantic Analysis

4 Experiments and Results

5 Conclusion and Future Work

Outline

1 Introduction

2 Background and Related Work
- Document Clustering
- Document Representation

3 Hybrid Models for Semantic Analysis

4 Experiments and Results

5 Conclusion and Future Work

Document Clustering

Hierarchical algorithms

Represent hierarchy of topics, locally optimal decisions
(Jain et al. 1999 [8], Zhao and Karypis 2005 [15])

Partitional algorithms

- k-means (Jain et al. 1999 [8]), spherical k-means (Dhillon and Modha, 2001 [4])
  simple iterative algorithms, local minimum
- Spectral clustering (Shi and Malik 2000 [12], Ng et al. 2002 [10])
  global minimum, computationally expensive
- Non-negative matrix factorization (NMF) (Lee and Seung 1999 [9], Xu et al. 2003 [14])
  meaningful parts, local minimum

Document Clustering

Hierarchical algorithms

Represent hierarchy of topics, locally optimal decisions
(Jain et al. 1999 [8], Zhao and Karypis 2005 [15])

Partitional algorithms

- k-means (Jain et al. 1999 [8]), spherical k-means (Dhillon and Modha, 2001 [4])
  simple iterative algorithms, local minimum
- Spectral clustering (Shi and Malik 2000 [12], Ng et al. 2002 [10])
  global minimum, computationally expensive
- Non-negative matrix factorization (NMF) (Lee and Seung 1999 [9], Xu et al. 2003 [14])
  meaningful parts, local minimum

Document Clustering

Hierarchical algorithms

Represent hierarchy of topics, locally optimal decisions
(Jain et al. 1999 [8], Zhao and Karypis 2005 [15])

Partitional algorithms

- k-means (Jain et al. 1999 [8]), spherical k-means (Dhillon and Modha, 2001 [4])
  simple iterative algorithms, local minimum
- Spectral clustering (Shi and Malik 2000 [12], Ng et al. 2002 [10])
  global minimum, computationally expensive
- Non-negative matrix factorization (NMF) (Lee and Seung 1999 [9], Xu et al. 2003 [14])
  meaningful parts, local minimum

Document Clustering

Hierarchical algorithms

Represent hierarchy of topics, locally optimal decisions
(Jain et al. 1999 [8], Zhao and Karypis 2005 [15])

Partitional algorithms

- k-means (Jain et al. 1999 [8]), spherical k-means (Dhillon and Modha, 2001 [4])
  simple iterative algorithms, local minimum
- Spectral clustering (Shi and Malik 2000 [12], Ng et al. 2002 [10])
  global minimum, computationally expensive
- Non-negative matrix factorization (NMF) (Lee and Seung 1999 [9], Xu et al. 2003 [14])
  meaningful parts, local minimum

Document Clustering

Hierarchical algorithms

Represent hierarchy of topics, locally optimal decisions
(Jain et al. 1999 [8], Zhao and Karypis 2005 [15])

Partitional algorithms

- k-means (Jain et al. 1999 [8]), spherical k-means (Dhillon and Modha, 2001 [4])
  simple iterative algorithms, local minimum
- Spectral clustering (Shi and Malik 2000 [12], Ng et al. 2002 [10])
  global minimum, computationally expensive
- Non-negative matrix factorization (NMF) (Lee and Seung 1999 [9], Xu et al. 2003 [14])
  meaningful parts, local minimum

Document Clustering

Hierarchical algorithms

Represent hierarchy of topics, locally optimal decisions
(Jain et al. 1999 [8], Zhao and Karypis 2005 [15])

Partitional algorithms

- k-means (Jain et al. 1999 [8]), spherical k-means (Dhillon and Modha, 2001 [4])
  simple iterative algorithms, local minimum
- Spectral clustering (Shi and Malik 2000 [12], Ng et al. 2002 [10])
  global minimum, computationally expensive
- Non-negative matrix factorization (NMF) (Lee and Seung 1999 [9], Xu et al. 2003 [14])
  meaningful parts, local minimum

Outline

1 Introduction

2 Background and Related Work
- Document Clustering
- Document Representation

3 Hybrid Models for Semantic Analysis

4 Experiments and Results

5 Conclusion and Future Work

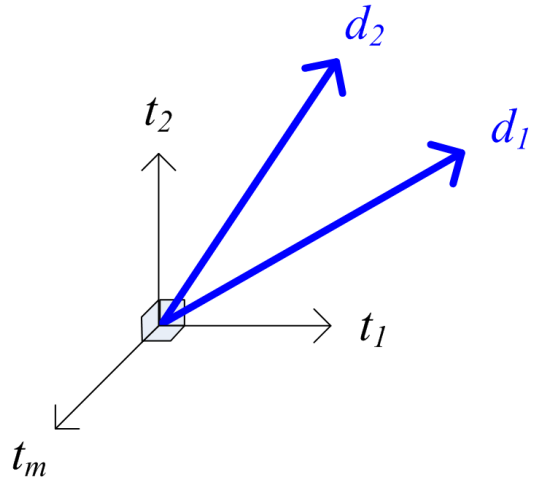
Document Representation Models

Vector space model (VSM) (Salton et al. 1975 [11])

- Documents as vectors in the space of terms: $\overrightarrow{d_j} = \overrightarrow{x_j}$
- Term vectors are orthogonal
- Similarity based on inner-product of document vectors:

$$\left\langle \overrightarrow{d_a}, \overrightarrow{d_b} \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ia} x_{jb}$$

- Ignore semantic relations between terms



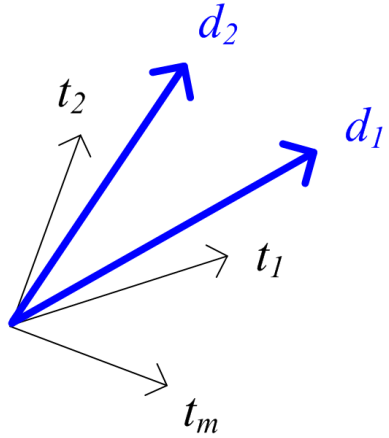
Document Representation Models (Cont.)

Generalized VSM (GVSM) (Wong et al. 1985 [13])

- Term vectors are non-orthogonal.
- Documents as linear combinations of term vectors: $\overrightarrow{d_j} = \sum_{i=1}^{m} x_{ij} \overrightarrow{t_i}$
- Similarity based on inner-product of term vectors:

$$\left\langle \overrightarrow{d_a}, \overrightarrow{d_b} \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ia} x_{jb} \left\langle \overrightarrow{t_i}, \overrightarrow{t_j} \right\rangle$$

- How to estimate inner-products between terms?



Document Representation Models (Cont.)

Latent semantic indexing (LSI) (Deerwester et al. 1990 [2])

$$\overrightarrow{d_j} = U_k^T \overrightarrow{x_j}$$

- U_k is a matrix of the first k left singular vectors of X: $X = U\Sigma V^T$.
- Preserves VSM-based similarity between documents
- Equivalent to principal component analysis (PCA) if X is column-centered
- Computationally expensive, difficult to estimate k

Document Representation Models (Cont.)

Semantic similarity using a knowledge-base

$$\overrightarrow{d_j} = P\overrightarrow{x_j}$$

- Concepts are extracted from an external knowledge-base, and a concept-term proximity matrix P is estimated.
- Lexical databases such as WordNet (Hotho et al 2003 [7]) or encyclopedias such as Wikipedia (Gabrilovich and Markovitch 2007 [6])
- Computationally expensive, noise in representation

Document Similarity

- Each model for document representation maps terms and documents to some semantic space.
- Inner-products of document vectors define ***a kernel matrix*** K:
  - $K_{VSM} = X^T X$
  - $K_{GVSM} = X^T G X$
  - $K_{LSI} = X^T U_k U_k^T X$
  - $K_{KB} = X^T P P X$
- Cosine similarity between documents:

$$Sim = L^{-1/2} K L^{-1/2}$$

Document Similarity

- Each model for document representation maps terms and documents to some semantic space.
- Inner-products of document vectors define ***a kernel matrix*** K:
  - $K_{VSM} = X^T X$
  - $K_{GVSM} = X^T G X$
  - $K_{LSI} = X^T U_k U_k^T X$
  - $K_{KB} = X^T P P X$
- Cosine similarity between documents:

$$Sim = L^{-1/2} K L^{-1/2}$$

Document Similarity

- Each model for document representation maps terms and documents to some semantic space.
- Inner-products of document vectors define ***a kernel matrix*** K:
  - $K_{VSM} = X^T X$
  - $K_{GVSM} = X^T G X$
  - $K_{LSI} = X^T U_k U_k^T X$
  - $K_{KB} = X^T P P X$
- Cosine similarity between documents:

$$Sim = L^{-1/2} K L^{-1/2}$$

Outline

1 Introduction

2 Background and Related Work
- Document Clustering
- Document Representation

3 Hybrid Models for Semantic Analysis

4 Experiments and Results

5 Conclusion and Future Work

Hybrid Models

Map documents to a semantic space where similarity represents statistical correlation between terms

Apply dimension reduction techniques to document vectors in the semantic space

Apply the clustering algorithm to the low-dimension document vectors

Hybrid Models

Map documents to a semantic space where similarity represents statistical correlation between terms

Apply dimension reduction techniques to document vectors in the semantic space

Apply the clustering algorithm to the low-dimension document vectors

Hybrid Models

Map documents to a semantic space where similarity represents statistical correlation between terms

Apply dimension reduction techniques to document vectors in the semantic space

Apply the clustering algorithm to the low-dimension document vectors

Mapping Documents to Semantic Space

GVSM-based model:

$$\left\langle \vec{d_a}, \vec{d_b} \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ia} x_{jb} \left\langle \vec{t_i}, \vec{t_j} \right\rangle$$

In a matrix form:

$$K = X^T G X$$

G is a Gram matrix that represents the inner-product of term vectors in some semantic space.

How to estimate G?

Mapping Documents to Semantic Space

GVSM-based model:

$$\left\langle \overrightarrow{d_a}, \overrightarrow{d_b} \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ia} x_{jb} \left\langle \overrightarrow{t_i}, \overrightarrow{t_j} \right\rangle$$

In a matrix form:

$$K = X^T G X$$

G is a Gram matrix that represents the inner-product of term vectors in some semantic space.

How to estimate G?

Mapping Documents to Semantic Space

GVSM-based model:

$$\left\langle \vec{d_a}, \vec{d_b} \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ia} x_{jb} \left\langle \vec{t_i}, \vec{t_j} \right\rangle$$

In a matrix form:

$$K = X^T G X$$

G is a Gram matrix that represents the inner-product of term vectors in some semantic space.

How to estimate G?

How to estimate G?

- Our recent work studied different estimates (Farahat and Kamel 2009 [5])
- Estimate G from a corpus C with a term-document matrix Q:
  - Association matrix: $G_{ASSC} = QQ^T$
  - Normalized association matrix: $G_{ASSC_N} = L_Q^{-1/2} QQ^T L_Q^{-1/2}$
  - Covariance matrix: $G_{COV} = \frac{1}{n_c - 1} \widetilde{Q}\widetilde{Q}^T = \frac{1}{n_c - 1} QHQ^T$
  - Matrix of Pearson's correlation coefficients: $G_{PCORR} = \frac{1}{n_c - 1} L_{\widetilde{Q}}^{-1/2} QHQ^T L_{\widetilde{Q}}^{-1/2}$
- Different estimates of G were analyzed from a geometric perspective.

How to estimate G?

- Our recent work studied different estimates (Farahat and Kamel 2009 [5])
- Estimate G from a corpus C with a term-document matrix Q:
  - Association matrix: $G_{ASSC} = QQ^T$
  - Normalized association matrix: $G_{ASSC_N} = L_Q^{-1/2} QQ^T L_Q^{-1/2}$
  - Covariance matrix: $G_{COV} = \frac{1}{n_c - 1} \widetilde{Q}\widetilde{Q}^T = \frac{1}{n_c - 1} QHQ^T$
  - Matrix of Pearson's correlation coefficients: $G_{PCORR} = \frac{1}{n_c - 1} L_{\widetilde{Q}}^{-1/2} QHQ^T L_{\widetilde{Q}}^{-1/2}$
- Different estimates of G were analyzed from a geometric perspective.

How to estimate G?

- Our recent work studied different estimates (Farahat and Kamel 2009 [5])
- Estimate G from a corpus C with a term-document matrix Q:
  - Association matrix: $G_{ASSC} = QQ^T$
  - Normalized association matrix: $G_{ASSC_N} = L_Q^{-1/2} QQ^T L_Q^{-1/2}$
  - Covariance matrix: $G_{COV} = \frac{1}{n_c - 1} \widetilde{Q}\widetilde{Q}^T = \frac{1}{n_c - 1} QHQ^T$
  - Matrix of Pearson's correlation coefficients: $G_{PCORR} = \frac{1}{n_c - 1} L_{\widetilde{Q}}^{-1/2} QHQ^T L_{\widetilde{Q}}^{-1/2}$
- Different estimates of G were analyzed from a geometric perspective.

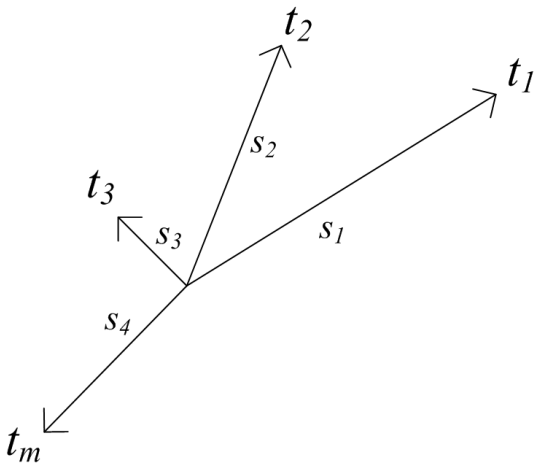
How to estimate G?

- Our recent work studied different estimates (Farahat and Kamel 2009 [5])
- Estimate G from a corpus C with a term-document matrix Q:
  - Association matrix: $G_{ASSC} = QQ^T$
  - Normalized association matrix: $G_{ASSC_N} = L_Q^{-1/2} QQ^T L_Q^{-1/2}$
  - Covariance matrix: $G_{COV} = \frac{1}{n_c - 1} \widetilde{Q}\widetilde{Q}^T = \frac{1}{n_c - 1} QHQ^T$
  - Matrix of Pearson's correlation coefficients: $G_{PCORR} = \frac{1}{n_c - 1} L_{\widetilde{Q}}^{-1/2} QHQ^T L_{\widetilde{Q}}^{-1/2}$
- Different estimates of G were analyzed from a geometric perspective.

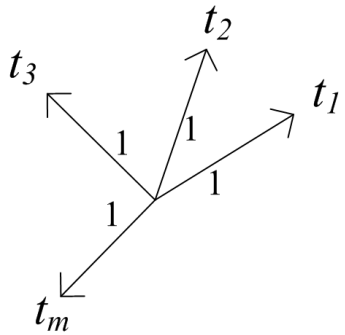
How to estimate G? (Cont.)

The length of a term vector in the semantic space

G_{ASSC}, G_{COV}



G_{ASSC_N}, G_{PCORR}

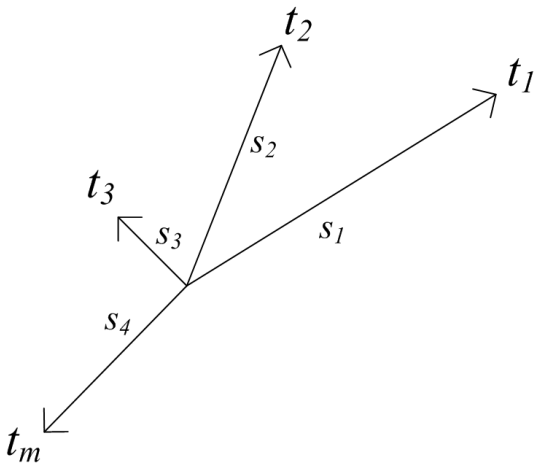


G_{COV}: length $\propto$ term variance quality (TVQ) (Dhillon et al. 2003 [3])

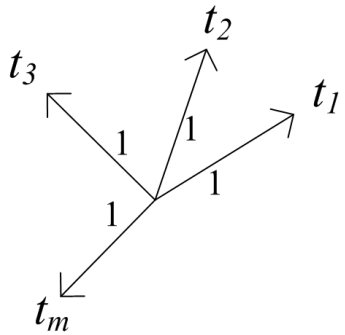
How to estimate G? (Cont.)

The length of a term vector in the semantic space

G_{ASSC}, G_{COV}



G_{ASSC_N}, G_{PCORR}



G_{COV}: length $\propto$ term variance quality (TVQ) (Dhillon et al. 2003 [3])

How to estimate G? (Cont.)

The cosine similarity between term vectors in the semantic space

t_1, t_2 positively-correlated, t_3, t_4 uncorrelated, t_5, t_6 negatively-correlated

G_{COV} and G_{PCORR} differentiate between uncorrelated and negatively-correlated terms.

How to estimate G? (Cont.)

The cosine similarity between term vectors in the semantic space

t_1, t_2 positively-correlated, t_3, t_4 uncorrelated, t_5, t_6 negatively-correlated

G_{COV} and G_{PCORR} differentiate between uncorrelated and negatively-correlated terms.

How to estimate G? (Cont.)

Semantic similarities based on covariance matrices are expected to achieve the best performance, as they:

- weight terms according to their discrimination ability, and
- differentiate between uncorrelated and negatively-correlated terms.

Mapping Documents to Semantic Space (Cont.)

- GVSM-based model with covariance matrices:

$$K = X^T G_{COV} X$$

- Use $\mathcal{D}$ to estimate G

$$G_{COV} = \frac{1}{n-1} \widetilde{X} \widetilde{X}^T$$

- K can be written as:

$$K = W^T W, \qquad W = \frac{1}{\sqrt{n-1}} \widetilde{X}^T X$$

W represents documents in the semantic space.

Mapping Documents to Semantic Space (Cont.)

- GVSM-based model with covariance matrices:

$$K = X^T G_{COV} X$$

- Use $\mathcal{D}$ to estimate G

$$G_{COV} = \frac{1}{n-1} \widetilde{X} \widetilde{X}^T$$

- K can be written as:

$$K = W^T W, \qquad W = \frac{1}{\sqrt{n-1}} \widetilde{X}^T X$$

W represents documents in the semantic space.

Mapping Documents to Semantic Space (Cont.)

- GVSM-based model with covariance matrices:

$$K = X^T G_{COV} X$$

- Use $\mathcal{D}$ to estimate G

$$G_{COV} = \frac{1}{n-1} \widetilde{X} \widetilde{X}^T$$

- K can be written as:

$$K = W^T W, \qquad W = \frac{1}{\sqrt{n-1}} \widetilde{X}^T X$$

W represents documents in the semantic space.

Mapping Documents to Semantic Space (Cont.)

- GVSM-based model with covariance matrices:

$$K = X^T G_{COV} X$$

- Use $\mathcal{D}$ to estimate G

$$G_{COV} = \frac{1}{n-1} \widetilde{X} \widetilde{X}^T$$

- K can be written as:

$$K = W^T W, \qquad W = \frac{1}{\sqrt{n-1}} \widetilde{X}^T X$$

W represents documents in the semantic space.

Dimension Reduction

- Apply DR techniques (LSI or PCA) on W

$$W = U\Sigma V^T$$

- Compute a low-rank approximation of W as:

$$W_d = U_d^T W = \Sigma_d V_d^T$$

- Compute K_d and Sim_d as:

$$K_d = W_d^T W_d, \qquad Sim = L_d^{-1/2} K_d L_d^{-1/2}$$

Dimension Reduction

- Apply DR techniques (LSI or PCA) on W

$$W = U\Sigma V^T$$

- Compute a low-rank approximation of W as:

$$W_d = U_d^T W = \Sigma_d V_d^T$$

- Compute K_d and Sim_d as:

$$K_d = W_d^T W_d, \qquad Sim = L_d^{-1/2} K_d L_d^{-1/2}$$

Dimension Reduction

- Apply DR techniques (LSI or PCA) on W

$$W = U\Sigma V^T$$

- Compute a low-rank approximation of W as:

$$W_d = U_d^T W = \Sigma_d V_d^T$$

- Compute K_d and Sim_d as:

$$K_d = W_d^T W_d, \qquad Sim = L_d^{-1/2} K_d L_d^{-1/2}$$

Clustering in the Latent Semantic Space

Similarity and kernel based algorithms

- Hierarchical algorithms, spectral clustering, kernel k-means
- Apply algorithm directly to Sim_d or K_d

Vector-based algorithms

- k-means, spherical k-means
- Apply algorithm to W_d

Clustering in the Latent Semantic Space

Similarity and kernel based algorithms

- Hierarchical algorithms, spectral clustering, kernel k-means
- Apply algorithm directly to Sim_d or K_d

Vector-based algorithms

- k-means, spherical k-means
- Apply algorithm to W_d

Clustering in the Latent Semantic Space

Similarity and kernel based algorithms

- Hierarchical algorithms, spectral clustering, kernel k-means
- Apply algorithm directly to Sim_d or K_d

Vector-based algorithms

- k-means, spherical k-means
- Apply algorithm to W_d

Spherical k-means in Latent Semantic Space

Inputs: X, t_{max}, k, d, **Outputs:** $\Pi = \{\pi_1, ..., \pi_k\}$

1. $W = \frac{1}{\sqrt{n-1}} \widetilde{X}^T X$
2. $[U, \Sigma, V] = svd(W)$
3. $W_d = U_d^T W = \Sigma_d V_d^T$
4. Initialize: $\Pi_0 = \{\pi_1, ..., \pi_k\}$, $t = 1$
5. $\overrightarrow{\mu_j} = \frac{\sum_{x_i \in \pi_j} \overrightarrow{w_{di}}}{\left\| \sum_{x_i \in \pi_j} \overrightarrow{w_{di}} \right\|}$, $j = 1..k$
6. $y_i = arg\ max_j\ cos\left(\overrightarrow{w_{di}}, \overrightarrow{\mu_j}\right)$, $i = 1..n$
7. $\pi_j = \{x_i : \ y_i = j\}$, $j = 1..k$
8. $\Pi_t = \{\pi_1, ..., \pi_k\}$
9. If $(\Pi_t \neq \Pi_{t-1}$ & $t < t_{max})$ $t = t + 1$, Go to step 3.
   Else Return Π_t

Hybrid vs. Latent Models

Latent models (like LSI and PCA) essentially preserve VSM-based similarity between documents.

Hybrid models preserve semantic similarity based on term-term correlations.

In hybrid models, dimension reduction removes noise or irrelevant information in the original data matrix and in calculating term-term correlations.

Hybrid vs. Latent Models

Latent models (like LSI and PCA) essentially preserve VSM-based similarity between documents.

Hybrid models preserve semantic similarity based on term-term correlations.

In hybrid models, dimension reduction removes noise or irrelevant information in the original data matrix and in calculating term-term correlations.

Hybrid vs. Latent Models

Latent models (like LSI and PCA) essentially preserve VSM-based similarity between documents.

Hybrid models preserve semantic similarity based on term-term correlations.

In hybrid models, dimension reduction removes noise or irrelevant information in the original data matrix and in calculating term-term correlations.

Hybrid vs. Latent Models

Latent models (like LSI and PCA) essentially preserve VSM-based similarity between documents.

Hybrid models preserve semantic similarity based on term-term correlations.

In hybrid models, dimension reduction removes noise or irrelevant information in the original data matrix and in calculating term-term correlations.

Hybrid vs. Latent Models

Latent models (like LSI and PCA) essentially preserve VSM-based similarity between documents.

Hybrid models preserve semantic similarity based on term-term correlations.

In hybrid models, dimension reduction removes noise or irrelevant information in the original data matrix and in calculating term-term correlations.

Outline

1. Introduction
2. Background and Related Work
   - Document Clustering
   - Document Representation
3. Hybrid Models for Semantic Analysis
4. Experiments and Results
5. Conclusion and Future Work

Experiments

- Evaluate the effectiveness of document clustering using hybrid models compared to well-known document representation models
- Three clustering algorithms:
  - Spherical k-means
  - HAC with complete-link
  - HAC with average-link
- Thirteen benchmark data sets

Experiments

- Evaluate the effectiveness of document clustering using hybrid models compared to well-known document representation models
- Three clustering algorithms:
  - Spherical k-means
  - HAC with complete-link
  - HAC with average-link
- Thirteen benchmark data sets

Experiments

- Evaluate the effectiveness of document clustering using hybrid models compared to well-known document representation models
- Three clustering algorithms:
  - Spherical k-means
  - HAC with complete-link
  - HAC with average-link
- Thirteen benchmark data sets

Data Sets

Data set	Source	n	m	m_{doc}	k	n_{class}
20ng	20 Newsgroups	2000	28839	23.3 ± 49.1	20	100.0 ± 0.0
classic	Different Abstracts	7094	41681	6.2 ± 7.7	4	1773.5 ± 971.4
fbis	TREC	2463	2000	68.5 ± 88.7	17	144.9 ± 139.3
hitech	TREC	2301	126321	37.9 ± 27.9	6	383.5 ± 189.9
reviews	TREC	4069	126354	43.3 ± 34.8	5	813.8 ± 520.9
la12	TREC	6279	31472	43.5 ± 38.0	6	1046.5 ± 526.5
tr31	TREC	927	10128	111.9 ± 248.3	7	132.4 ± 124.0
tr41	TREC	878	7454	66.5 ± 100.5	10	87.8 ± 80.1

n: # docs, m: # terms, k: # classes, m_{doc}: # terms per docs, n_{class}:: # docs per class

Data Sets (Cont.)

Data set	Source	n	m	m_{doc}	k	n_{class}
re0	Reuters-21578	1504	2886	15.0 ± 14.5	13	115.7 ± 173.8
re1	Reuters-21578	1657	3758	15.4 ± 12.3	25	66.3 ± 91.8
k1a	WebACE	2340	21839	44.5 ± 20.8	20	117.0 ± 117.5
k1b	WebACE	2340	21839	44.5 ± 20.8	6	390.0 ± 513.3
wap	WebACE	1560	8460	43.2 ± 20.5	20	78.0 ± 81.1

n: # docs, m: # terms, k: # classes, m_{doc}: # terms per docs, n_{class}:: # docs per class

Evaluation of Cluster Quality

- Compare to ground-truth partitioning of documents assigned by humans

Quality Measures:

- **F-measure**: combines precision and recall.
- **Entropy**: measures the homogeneity of clusters with respect to classes.
- **Purity**: measures the average precision of clusters relative to their best matching classes.

Evaluation of Cluster Quality

- Compare to ground-truth partitioning of documents assigned by humans

Quality Measures:

- **F-measure**: combines precision and recall.
- **Entropy**: measures the homogeneity of clusters with respect to classes.
- **Purity**: measures the average precision of clusters relative to their best matching classes.

Evaluation of Cluster Quality

- Compare to ground-truth partitioning of documents assigned by humans

Quality Measures:

- **F-measure**: combines precision and recall.
- **Entropy**: measures the homogeneity of clusters with respect to classes.
- **Purity**: measures the average precision of clusters relative to their best matching classes.

Experimental Setup

- Six document representation models: VSM, GVSM (using covariance matrix), LSI, PCA, LSI-C (LSI with GVSM based on covariance matrix), and PCA-C (PCA with GVSM based on covariance matrix)
- Change k from 5 to 100, and calculate quality measures for each k. Evaluate each method based on the average and standard deviation of the best 10 values for the quality measure (Cai et al. [1])

Experimental Setup

- Six document representation models: VSM, GVSM (using covariance matrix), LSI, PCA, LSI-C (LSI with GVSM based on covariance matrix), and PCA-C (PCA with GVSM based on covariance matrix)
- Change k from 5 to 100, and calculate quality measures for each k. Evaluate each method based on the average and standard deviation of the best 10 values for the quality measure (Cai et al. [1])

Experimental Setup

- Six document representation models: VSM, GVSM (using covariance matrix), LSI, PCA, LSI-C (LSI with GVSM based on covariance matrix), and PCA-C (PCA with GVSM based on covariance matrix)
- Change k from 5 to 100, and calculate quality measures for each k. Evaluate each method based on the average and standard deviation of the best 10 values for the quality measure (Cai et al. [1])

Experimental Setup (Cont.)

- To average for different data sets, calculate quality measures relative to the best value for each method (Zhao and Karypis [16])

$$F_r = \frac{F}{max(F)}, \; E_r = \frac{min(E)}{E}, \; P_r = \frac{P}{max(P)}$$

- Compare pairs of methods (M_1, M_2) over all data sets using paired t-test with 95% confidence interval.

Experimental Setup (Cont.)

- To average for different data sets, calculate quality measures relative to the best value for each method (Zhao and Karypis [16])

$$F_r = \frac{F}{max(F)}, \; E_r = \frac{min(E)}{E}, \; P_r = \frac{P}{max(P)}$$

- Compare pairs of methods (M_1, M_2) over all data sets using paired t-test with 95% confidence interval.

Results - Relative F-Measure

Methods	1.VSM	2.GVSM	3.LSI	4.PCA	5.LSI-C	6.PCA-C
SKM	0.9186	0.9668	0.9750	0.9640	0.9780	0.9676
HAC-C	0.6633	0.9113	0.8503	0.8943	0.9828	0.9742
HAC-A	0.7109	0.8605	0.9160	0.9712	0.9662	0.9557

M1	1					2				3			4		5
M2	2	3	4	5	6	3	4	5	6	4	5	6	5	6	6
SKM	$<$	$<$	$=$	$<$	$<$	$=$	$=$	$=$	$=$	$=$	$=$	$=$	$=$	$=$	$=$
HAC-C	$<$	$<$	$<$	$<$	$<$	$>$	$=$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$=$
HAC-A	$<$	$<$	$<$	$<$	$<$	$=$	$<$	$<$	$=$	$=$	$<$	$=$	$=$	$=$	$=$

Results - Relative Entropy

Methods	1.VSM	2.GVSM	3.LSI	4.PCA	5.LSI-C	6.PCA-C
SKM	0.9000	0.9720	0.9706	0.9619	0.9529	0.9383
HAC-C	0.5941	0.8897	0.8187	0.8879	0.9732	0.9616
HAC-A	0.6066	0.7887	0.9002	0.9824	0.9462	0.9493

M1	1					2				3			4		5
M2	2	3	4	5	6	3	4	5	6	4	5	6	5	6	6
SKM	<	<	<	=	=	=	=	=	=	=	=	=	=	=	=
HAC-C	<	<	<	<	<	=	=	<	<	<	<	<	<	<	=
HAC-A	<	<	<	<	<	<	<	<	<	<	=	=	=	=	=

Results - Relative Purity

Methods	1.VSM	2.GVSM	3.LSI	4.PCA	5.LSI-C	6.PCA-C
SKM	0.9408	0.9826	0.9850	0.9887	0.9857	0.9797
HAC-C	0.6583	0.9136	0.8997	0.9382	0.9869	0.9809
HAC-A	0.6775	0.8296	0.9219	0.9873	0.9470	0.9517

M1	1					2				3			4		5
M2	2	3	4	5	6	3	4	5	6	4	5	6	5	6	6
SKM	<	<	<	<	=	=	=	=	=	=	=	=	=	=	=
HAC-C	<	<	<	<	<	=	=	<	<	<	<	<	<	<	=
HAC-A	<	<	<	<	<	=	<	<	<	=	=	=	=	=	=

Outline

1. Introduction
2. Background and Related Work
   - Document Clustering
   - Document Representation
3. Hybrid Models for Semantic Analysis
4. Experiments and Results
5. Conclusion and Future Work

Conclusion

This work proposes hybrid models that combine explicit and latent semantics.

Hybrid models apply dimension reduction in a semantic space which captures statistical correlation between terms.

In the document clustering task, hybrid models are either statistically significantly better or equivalent to explicit/latent models.

Conclusion

This work proposes hybrid models that combine explicit and latent semantics.

Hybrid models apply dimension reduction in a semantic space which captures statistical correlation between terms.

In the document clustering task, hybrid models are either statistically significantly better or equivalent to explicit/latent models.

Conclusion

This work proposes hybrid models that combine explicit and latent semantics.

Hybrid models apply dimension reduction in a semantic space which captures statistical correlation between terms.

In the document clustering task, hybrid models are either statistically significantly better or equivalent to explicit/latent models.

Future Work

Reducing the computational complexity and memory requirement of semantic mapping and dimension reduction by selecting a subset of documents to estimate G

Studying of the problem of determining the intrinsic dimensionality of the documents

Future Work

Reducing the computational complexity and memory requirement of semantic mapping and dimension reduction by selecting a subset of documents to estimate G

Studying of the problem of determining the intrinsic dimensionality of the documents

Future Work

Reducing the computational complexity and memory requirement of semantic mapping and dimension reduction by selecting a subset of documents to estimate G

Studying of the problem of determining the intrinsic dimensionality of the documents

Thank you !

References I

[1] Deng Cai, Xiaofei He, and Jiawei Han.
Document clustering using locality preserving indexing.
IEEE Transactions on Knowledge and Data Engineering, 17(12):1624–1637, 2005.

[2] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman.
Indexing by latent semantic analysis.
J. Am. Soc. Inf. Sci. Technol., 41(6):391–407, 1990.

[3] I. Dhillon, J. Kogan, and C. Nicholas.
Feature selection and document clustering, chapter 4, pages 73–100.
Springer-Verlag New York Inc, 2003.

References II

[4] Inderjit S. Dhillon and Dharmendra S. Modha.
Concept decompositions for large sparse text data using clustering.
Mach. Learn., 42(1/2):143–175, 2001.

[5] Ahmed K. Farahat and Mohamed S. Kamel.
Document clustering using semantic kernels based on term-term correlations.
In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, 2009.

References III

[6] E. Gabrilovich and S. Markovitch.
Computing semantic relatedness using wikipedia-based explicit semantic analysis.
In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[7] A. Hotho, S. Staab, and G. Stumme.
Wordnet improves text document clustering.
In *Proceedings of the SIGIR 2003 Semantic Web Workshop*, pages 541–544, New York, NY, USA, 2003. ACM.

[8] A. K. Jain, M. N. Murty, and P. J. Flynn.
Data clustering: a review.
ACM Comput. Surv., 31(3):264–323, 1999.

References IV

[9] D.D. Lee and H.S. Seung.
Learning the parts of objects by non-negative matrix factorization.
Nature, 401:788–791, 1999.

[10] A. Ng, M. Jordan, and Y. Weiss.
On spectral clustering: Analysis and an algorithm.
In *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 [sic] Conference*, page 849. MIT Press, 2002.

[11] G. Salton, A. Wong, and C. S. Yang.
A vector space model for automatic indexing.
Commun. ACM, 18(11):613–620, 1975.

References V

[12] J. Shi and J. Malik.
Normalized cuts and image segmentation.
IEEE Transactions on pattern analysis and machine intelligence, 22(8):888–905, 2000.

[13] S. K. M. Wong, Wojciech Ziarko, and Patrick C. N. Wong.
Generalized vector spaces model in information retrieval.
In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 18–25, New York, NY, USA, 1985. ACM.

References VI

[14] W. Xu, X. Liu, and Y. Gong.
Document clustering based on non-negative matrix factorization.
In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, New York, NY, USA, 2003. ACM.

[15] Y. Zhao and G. Karypis.
Hierarchical clustering algorithms for document datasets.
Data Mining and Knowledge Discovery, 10(2):141–168, 2005.

References VII

[16] Ying Zhao and George Karypis.
Empirical and theoretical comparisons of selected criterion functions for document clustering.
Mach. Learn., 55(3):311–331, 2004.