Co-Similarity Approach to Co-Clustering

Application to Text Mining and Bioinformatics

PhD defense of

Syed Fawad Hussain

Laboratory TIMC-IMAG

28th September, 2010

Supervisors: Mirta Gordon and Gilles Bisson

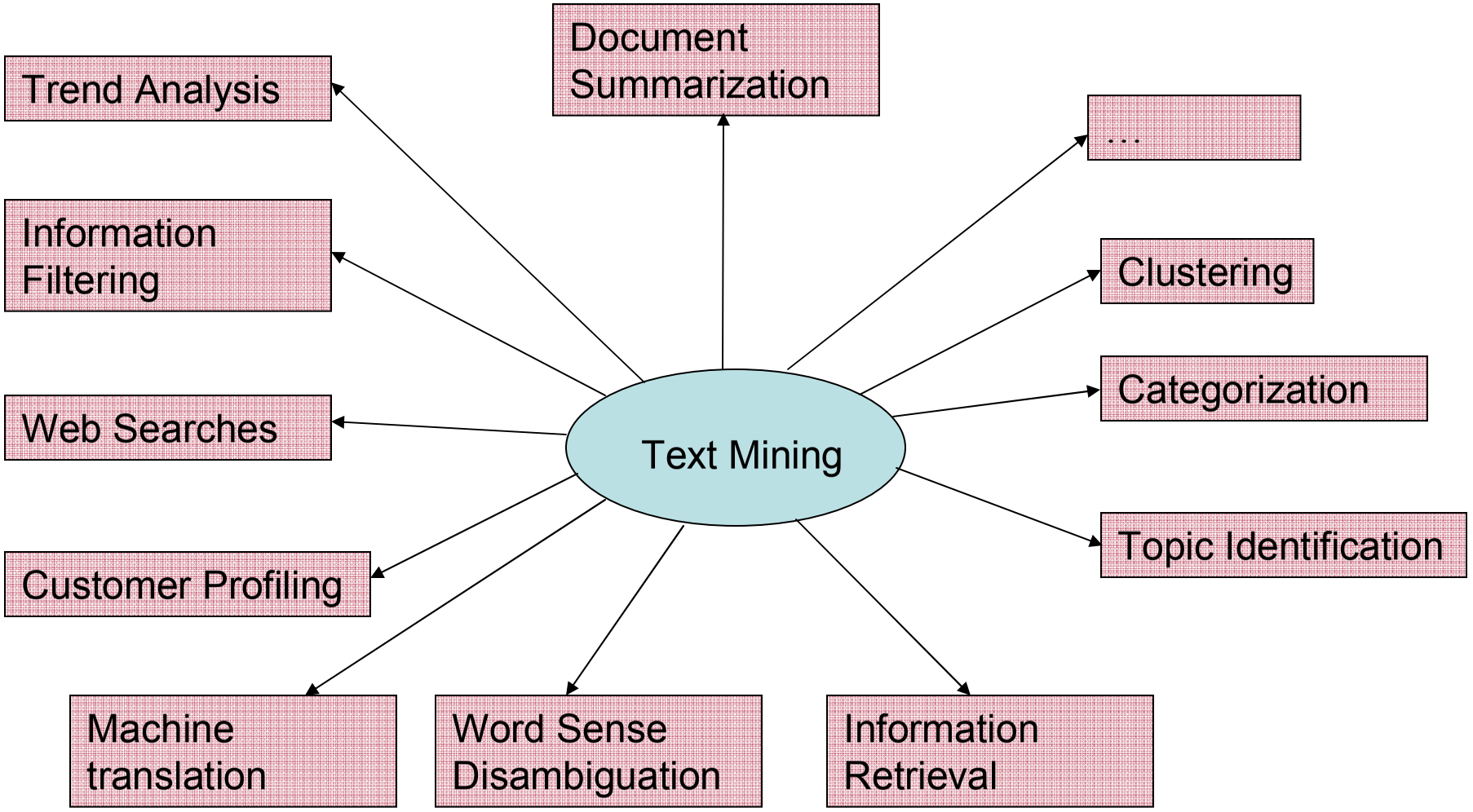
What is Text Mining?

- A huge percentage of information stored in unstructured data (mostly text)
  documents, journals, web pages, emails...

- Text Mining

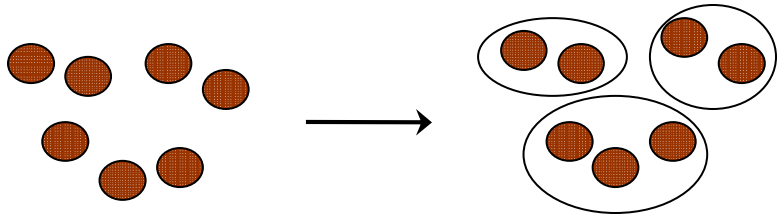
  "...extraction of implicit, previously unknown and potentially useful information from (large amounts of) textual data" [Frawley 92]

Potential Applications of Text Mining



What is Clustering?

- Division of data into groups of ‘similar objects’



- Classical clustering algorithms are based on “similarities”
  - The output of the clustering algorithm is based upon the goodness of the “similarity measure”
- Text data have certain characteristics, such as
  - High dimensionality
  - Unequal lengths of documents
  - Synonymy, polysemy, etc

Outline

1. **Introduction**
2. Our Proposed Similarity Measure (χ-Sim)
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. Experimentation
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

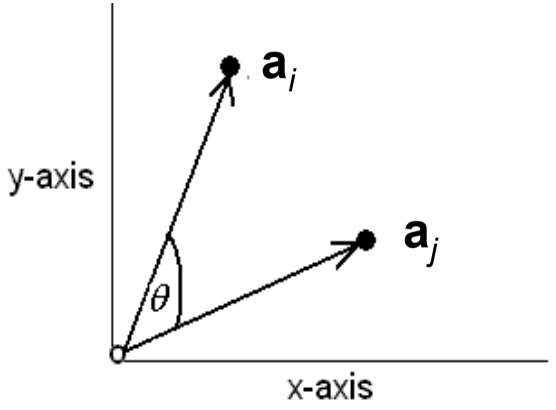
Document Representation

- Vector Space Model [Salton 70]
  - A document is a collection of words
  - $\mathbf{d}_i=[w_1,w_2,...,w_n]$, $\mathbf{w}_j=[d_1,d_2,...d_n]$
  - Matrix **A** represent a collection of *m* documents with *n* words

	$\mathbf{w_1}$	$\mathbf{w_2}$	...	$\mathbf{w_m}$
$\mathbf{d_1}$	A_{11}	A_{12}	...	A_{1m}
$\mathbf{d_2}$	A_{21}	A_{22}	...	A_{2m}
...	...	...	...	...
$\mathbf{d_m}$	A_{m1}	A_{m2}	...	A_{mn}

- Similarity between $\mathbf{d}_i$ and $\mathbf{d}_j$

$$Cosine(\mathbf{d}_i,\mathbf{d}_j)=\frac{\sum_{k=1..n} A_{ik} * A_{jk}}{\sqrt{\sum_{k=1..n}(A_{ik})^2 * \sum_{k=1..n}(A_{jk})^2}}$$



  - Euclidean, Manhattan, Jaccard, etc

Limitation of Traditional Similarity Measures

- Consider the following sentences

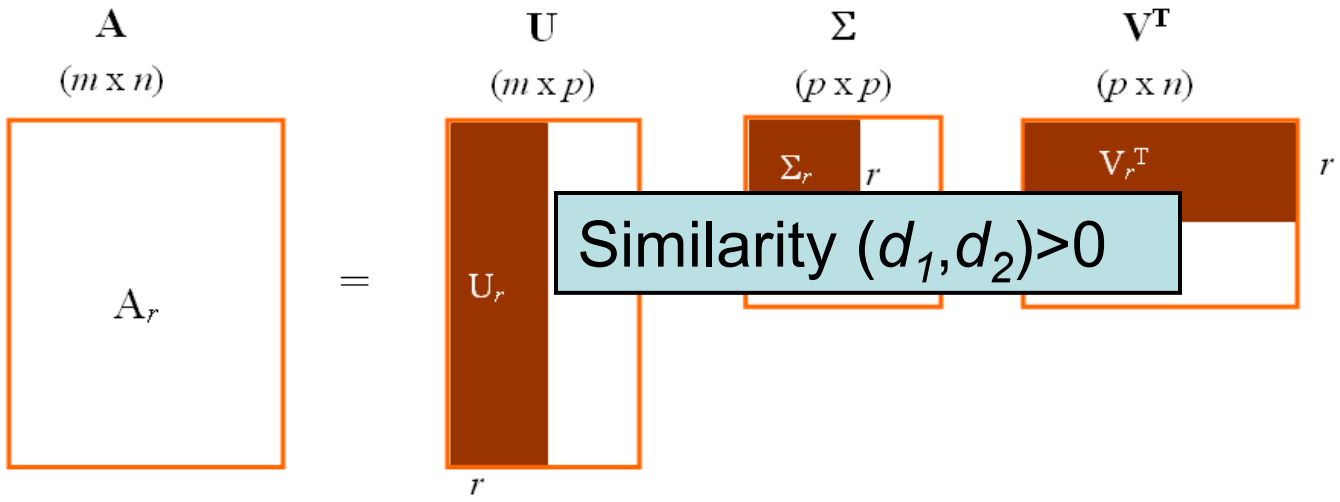
d_1	d_2
Boeing recently unveiled its new B787 aircraft dubbed the "Dreamliner".	Airbus' latest A350 is a next generation plane is due to fly in 2013

- Automated Text Clustering (using Cosine, etc)
  - No terms in common
  - Similarity $(d_1,d_2) = 0$

- Two main alternative approaches
  - Low rank approximation
  - Co-Clustering

First Approach: Low Rank Approximation

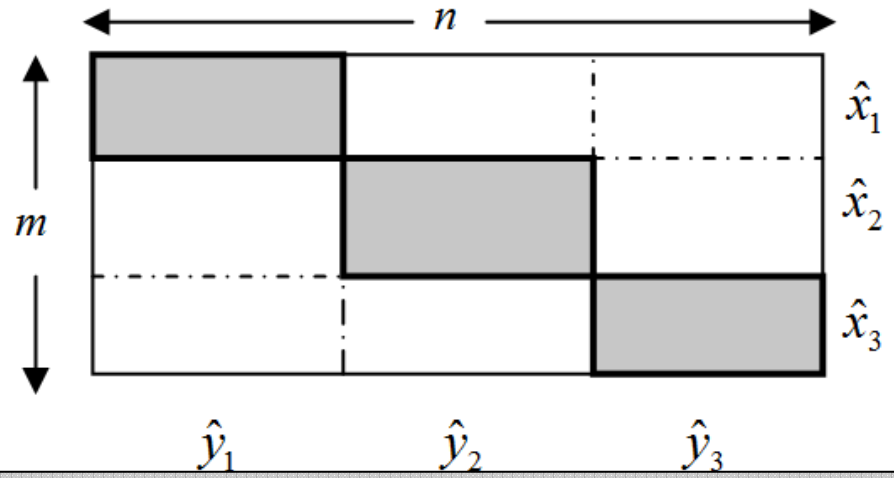
- Data Transformation
  - Low rank approximation e.g. Latent Semantic Analysis [Deerwester 89]
  - Decompose **A** into $\mathbf{U}\sum\mathbf{V}^T$
  - Trancate the lowest *r* singular values
  - Re-compute an approximate matrix A_r



Compute similarity in the "new" space

Second Approach: Co-clustering

- Simultaneously cluster documents and words
  - Start with initial partitioning of row and columns
  - Iteratively update partitioning to optimize some objective function
    - Information theoretic [Dhillion et al, 03]
    - Matrix decomposition [Long et al, 05]
    - …
  - End up in the same cluster in the "reduced" space



$$d_1, d_2 \in \hat{x}_i$$

Outline

1. Introduction
2. **Our Proposed Similarity Measure (χ-Sim)**
   1. The χ-Sim Measure
   2. Relationship to Previous Work
   3. Experimentation
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Outline

1. Introduction
2. **Our Proposed Similarity Measure (χ-Sim)**
   a. **The χ-Sim Measure**
   b. Relationship to Previous Work
   c. Experimentation
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Preliminaries

- Basic Idea
  - Two documents having no common words might still be related based on having "*similar words*".
  - Two words are similar if they occur in "*similar documents*"
  - Two documents are similar if they contain "*similar words*"
  - The dual nature of this similarity is what we refer to as "Co-Similarity"

- Notation used
  - **A** : a *document-term* matrix of size m (documents) by n (words/terms)
    - $\mathbf{d}_i = [A_{i1} \ldots A_{in}]$, a row vector corresponding to document i
    - $\mathbf{w}_j = [A_{1j} \ldots A_{mj}]$, a column vector corresponding to word j
    - $fs(A_{ij}, A_{kl})$: similarity function between elements of A
  - **R** a *document* $\times$ *document* similarity matrix, $R_{ij} \in [0,1]$
  - **C** a *word* $\times$ *word* similarity matrix, $C_{ij} \in [0,1]$

Principle of Our Approach

	$\mathbf{w_1}$	$\mathbf{w_2}$	…	$\mathbf{w_m}$
$\mathbf{d_1}$	A_{11}	A_{12}	…	A_{1m}
$\mathbf{d_2}$	A_{21}	A_{22}	…	A_{2m}
…	…	…	…	…
$\mathbf{d_m}$	A_{m1}	A_{m2}	…	A_{mn}

- Classical Similarity: Number of common terms

$$Sim(\mathbf{d}_i, \mathbf{d}_j) = f_s(A_{i1}, A_{j1}) + \ldots + f_s(A_{in}, A_{jn})$$

- Now, if we "generalize" to all pair of features

$$Sim(\mathbf{d}_i, \mathbf{d}_j) = f_s(A_{i1}, A_{j1}).C_{11} + f_s(A_{i1}, A_{j2}).C_{12} + \ldots + f_s(A_{i1}, A_{jn}).C_{1n} +$$

$Sim(\mathbf{w}_i, \mathbf{w}_j)$ can be computed similarly

$$\ldots, A_{jn}).C_{2n} +$$

$$\ldots$$

$$f_s(A_{in}, A_{j1}).C_{n1} + f_s(A_{in}, A_{j2}).C_{n2} + \ldots + f_s(A_{in}, A_{jn}).C_{nn}$$

The χ-Sim Measure

- We define χ-Sim as a recursive approach

- The algorithm is as follows
  - *Step* 1 - Given **A**, define $\mathbf{R}^{(0)}=\mathbf{I}$, $\mathbf{C}^{(0)}=\mathbf{I}$
  - *Step* 2 – for k=1 to t, do

$$R_{ij}^{(k+1)} = Sim(\mathbf{d}_i, \mathbf{d}_j) / N(\mathbf{d}_i, \mathbf{d}_j)$$

$$C_{ij}^{(k+1)} = Sim(\mathbf{w}_i, \mathbf{w}_j) / N(\mathbf{w}_i, \mathbf{w}_j)$$

    where $N(,)$ is a *normalization* factor s.t. $Sim(,) \in [0,1]$
  - *Step* 3: Output $\mathbf{R}^{(t)}$ and $\mathbf{C}^{(t)}$

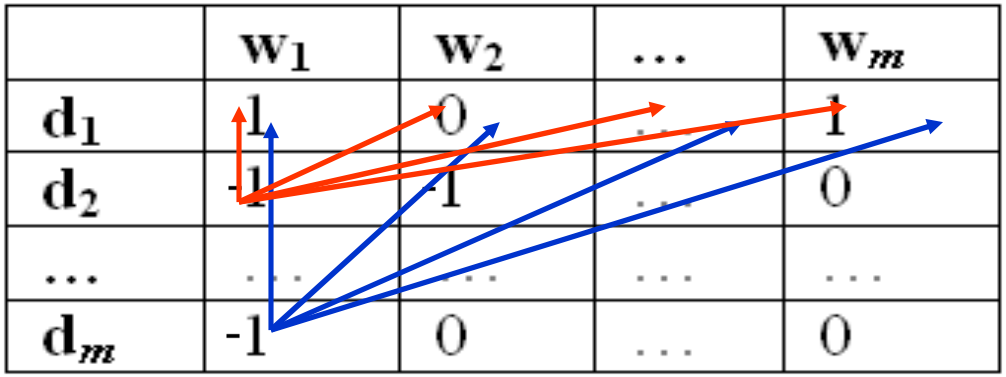
Complexity Analysis

	$\mathbf{w_1}$	$\mathbf{w_2}$	...	$\mathbf{w_m}$
$\mathbf{d_1}$	A_{11}	A_{12}	...	A_{1m}
$\mathbf{d_2}$	A_{21}	A_{22}	...	A_{2m}
...	...	...	...	...
$\mathbf{d_m}$	A_{m1}	A_{m2}	...	A_{mn}

- Total number of operations
  - **R** requires m^2 operations
    - Each $Sim(\mathbf{d}_i,\mathbf{d}_j)$ requires n^2 operations.
  - **C** requires n^2 operations
    - Each $Sim(\mathbf{w}_i,\mathbf{w}_j)$ requires m^2 operations.
  - Each of **R** and **C** is calculated recursively t times
- Complexity $O(t.m^2.n^2)$…which is clearly tooo much!
- We discuss two approaches to reduce the complexity

Complexity Reduction (1/2)

- When values of A_{ij} is of the enumerated type E.
  - $A_{ij} \in \{E\}$ (for instance A_{ij} could one of 3 states : -1,0,1)
  - |E| is small
- Many computations are repetitive



	w_1	w_2	...	w_m
d_1	1	0	...	1
d_2	-1	-1	...	0
...	...	...	...	...
d_m	-1	0	...	0

- Pre-compute similarity values $f_s(A_{ij},A_{kl})$
- Complexity is $O(\max(|E|.n.m^2, |E|.n^2.m).t)$

Complexity Reduction (2/2)

- Define $f_s(,)$ as a dot product
  - $f_s(A_{ik},A_{jl}) = A_{ik} * A_{jl}$
  - similar to Cosine, Euclidean, etc

- Similarity values can now be expressed

$$R_{ij}^{(k)} = Sim(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \mathbf{C}^{(k-1)} \mathbf{d}_j^{\mathrm{T}}}{N(\mathbf{d}_i, \mathbf{d}_j)}$$

$$C_{ij}^{(k)} = Sim(\mathbf{w}_i, \mathbf{w}_j) = \frac{\mathbf{w}_i \mathbf{R}^{(k-1)} \mathbf{w}_j^{\mathrm{T}}}{N(\mathbf{w}_i, \mathbf{w}_j)}$$

- Complexity is $O(\max(m^2.n, m.n^2).t)$

Outline

1. Introduction
2. **Our Proposed Similarity Measure (χ-Sim)**
   a. **The χ-Sim Measure**
      i. **Normalization**
      ii. **Number of iterations**
      iii. **A parameter called *pruning***
   b. Relationship to Previous Work
   c. Experimentation
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Normalization used in χ-Sim

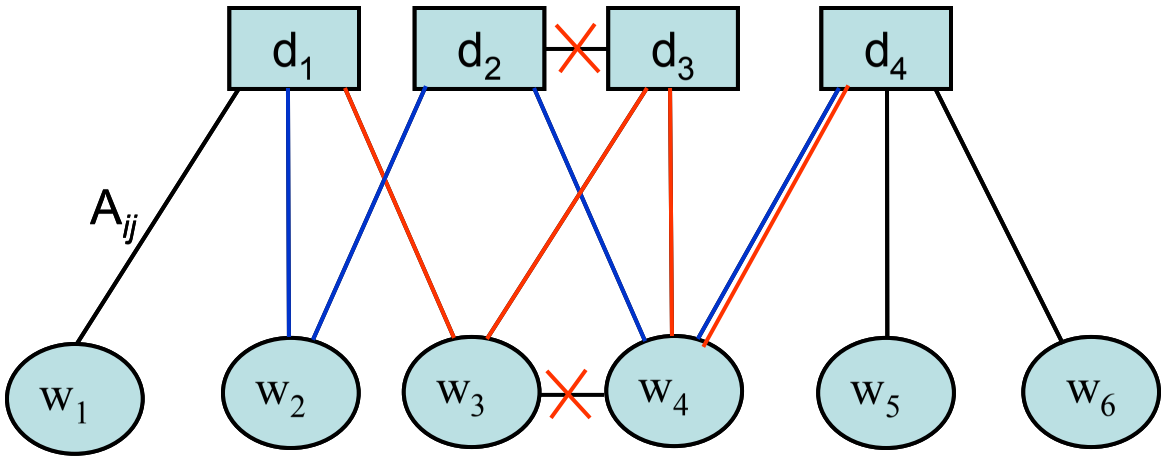
	$\mathbf{w_1}$	$\mathbf{w_2}$	...	$\mathbf{w_m}$
$\mathbf{d_1}$	A_{11}	A_{12}	...	A_{1m}
$\mathbf{d_2}$	A_{21}	A_{22}	...	A_{2m}
...	...	...	...	...
$\mathbf{d_m}$	A_{m1}	A_{m2}	...	A_{mn}

- The normalization factor renders similarity values in [0,1]
- L1 Normalization
  - Therefore maximum value of $\mathbf{a}_i\mathbf{C}(\mathbf{a}_j)^T$ is

$$N(\mathbf{d}_i,\mathbf{d}_j)=\sum_k A_{ik}\sum_k A_{jk} \qquad N(\mathbf{w}_i,\mathbf{w}_j)=\sum_k A_{ki}\sum_k A_{ki}$$

  - Takes unequal document (and word) length into account
  - Does not guarantee maximum self similarity
  - Set $R_{ii}=1$, $C_{jj}=1$
- Other possible Normalizations (e.g. L2)
  - Similar to Cosine in the first iteration
  - Does not guarantee $\mathbf{R}^{(t)}_{ij}$ or $\mathbf{C}^{(t)}_{ij} \in [0,1]$ for $t>1$

Meaning of an iteration



- Bipartite Graph
  - G=(V1,V2,E)
  - V1={d_1,d_2,…,d_m}
  - V2={w_1,w_2,…,w_n}
  - E =A_{ij}, $i \in$V1, $j \in$V2

- Iteration 1:
  - $\mathbf{R}^{(1)}$: Sim(d_1,d_2), Sim(d_1,d_3), …
  - $\mathbf{C}^{(1)}$: Sim(w_1,w_2), Sim(w_1,w_3), …
- Iteration 2:
  - $\mathbf{R}^{(2)}$: Sim(d_1,d_4) via C_{24} and C_{34} …
  - …

Successive iterations means paths of increasing length

Practically 4 iterations are enough

Pruning Parameter

- Intuition
  - Many similarity values may result from random co-occurrences
  - These values can be considered as “noise” in the data
- Low similarities may not be semantically meaningful
- Prune the similarity matrices
  - Set the least ρ% values in **R** and **C** to zero (at each iteration)
  - ρ is user defined

Outline

1. Introduction
2. **Our Proposed Similarity Measure (χ-Sim)**
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. Experimentation
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Comparison with Non-Recursive Approaches

$$Similarity(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i(\mathbf{C})\mathbf{d}_j^{\mathrm{T}}}{N(\mathbf{d}_i, \mathbf{d}_j)}$$

- Generalized Cosine Measure
  - $\mathcal{M}(\mathbf{a}_i,\mathbf{a}_j) = \|\mathbf{a}_i\|_2 * \|\mathbf{a}_j\|_2$, $\mathbf{C}=\mathbf{I} \rightarrow$ Cosine
  - $\mathcal{M}(\mathbf{a}_i,\mathbf{a}_j) = |\mathbf{a}_i| + |\mathbf{a}_i| - \mathbf{a}_i\mathbf{a}_i^{\mathrm{T}}$, $\mathbf{C}=\mathbf{I} \rightarrow$ Jaccard
  - $\mathcal{M}(\mathbf{a}_i,\mathbf{a}_j) = |\mathbf{a}_i| + |\mathbf{a}_i|$, $\mathbf{C}=2\mathbf{I} \rightarrow$ Dice
- For χ-Sim
  - $C_{ij} > 0$ for $i \neq j$ (after the first iteration)
  - Matrix **C** depends on **R** (and vice versa)
  - **R** and **C** are computed Iteratively

Comparison with Recursive Approaches

- SimRank [Jeh et al, 02]
  - $R_{ii}=1, C_{jj}=1$
  - C_1 is "Confidence Level"
  - Does not take weights into account

$$R_{ij} = \frac{C_1}{\left|\mathbf{d}_i\right|\left|\mathbf{d}_j\right|} \sum_{k \in O(\mathbf{d}_i)} \sum_{l \in O(\mathbf{d}_j)} C_{kl}$$

where $O(\mathbf{d}_i)$ denote the indices s.t. $A_{ik}>0$

- Similarity in Non-Orthogonal Space [Liu et al, 04]
  - Similar to χ-Sim
  - Does not take document/word vector length into account

$$R_{ij}^{(k+1)} = (\lambda_1 \mathbf{A}\mathbf{C}^{(k)}\mathbf{A}^{\mathrm{T}})_{ij}$$

- The Method of Blondel [Blondel et al, 04]
  - Based on an adjacency matrix $\mathbf{L}$ between objects of the same kind

$$\mathbf{S}^{(k+1)} = \frac{\mathbf{L}\mathbf{S}^{(k)}\mathbf{L}^{\mathrm{T}} + \mathbf{L}^{\mathrm{T}}\mathbf{S}^{(k)}\mathbf{L}}{\left\|\mathbf{L}\mathbf{S}^{(k)}\mathbf{L}^{\mathrm{T}} + \mathbf{L}^{\mathrm{T}}\mathbf{S}^{(k)}\mathbf{L}\right\|_F}$$

Outline

1. Introduction
2. **Our Proposed Similarity Measure (χ-Sim)**
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. **Experimentation**
      i. **Overview**
      ii. Effect of the parameters – number of iterations and pruning
      iii. Comparison with other approaches
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Datasets

- 20 Newsgroup Dataset
- Classic 3 Dataset

Subset	Newsgroups/Abstracts included	docs/cluster	# of docs
M2	Talk.politics.mideast, talk.politics.misc	250	500
M5	comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space , talk.politics.mideast.	100	500
M10	alt.atheism, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.gun.	50	500
NG1	rec.sports.baseball, rec.sports.hockey	200	400
NG2	comp.os.ms-windows.misc, comp.windows.x,rec.motorcycles,sci.crypt,sci.space	200	1000
NG3	comp.os.ms-windows.misc, comp.windows.x, misc.forsale, rec.motorcycles, sci.crypt, sci.space, talk.politics.mideast, talk.religion.misc	200	1600
Classic3	MEDLINE, CRANFIELD, CISI	1033,1460,1400	3893

- Preprocessing
  - Knowledge intensive approach
    - Mutual Information of words with document category (SMI)
  - Unsupervised approach
    - Mutual Information of words with documents (UMI)
    - Partitioning Around Mediods (PAM)

Document Clustering

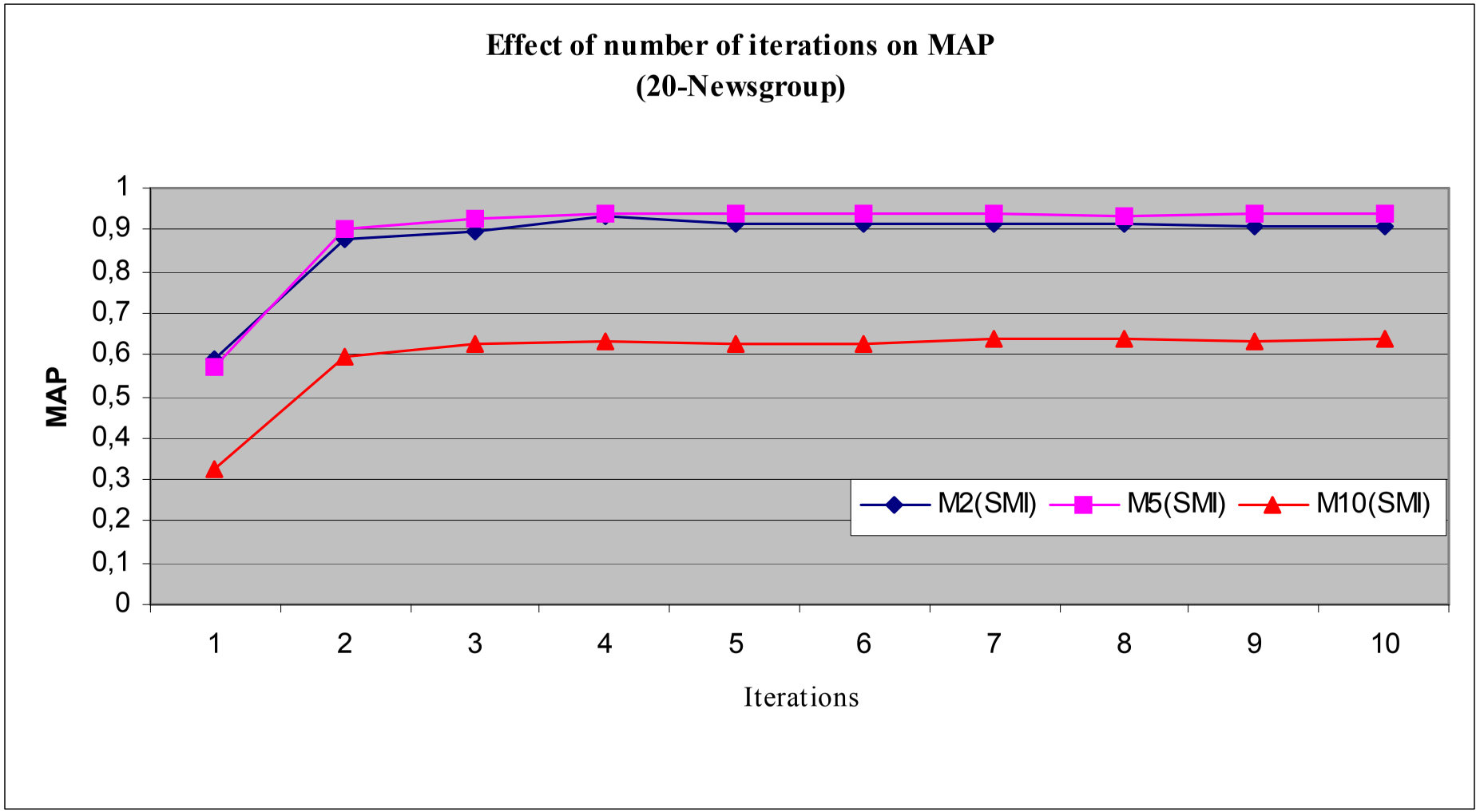
- Methodology
  - Generate document similarity matrix
  - Convert to distance matrix (by subtracting from max. value)
  - Cluster using Hierarchical Clustering (using Ward's linkage)
  - Number of clusters were fixed to actual clusters

- Evaluation
  - Is an entire research field
  - We used two popular measures
    - Micro-Average precision (MAP)
    - Normalized Mutual Information (NMI)
  - Range between 0 and 1 (*the higher the better*)

Outline

1. Introduction
2. **Our Proposed Similarity Measure (χ-Sim)**
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. **Experimentation**
      i. Overview
      ii. **Effect of the parameters – number of iterations and pruning**
      iii. Comparison with other approaches
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Effect of Number of Iterations



Effect of Pruning (1/2)

Precision (MAP) values when using knowledge intensive approach

		Pruning Percentage											Max Gain
		0	10	20	30	40	50	60	70	80	90	100	
Datasets	$M2_{SMI}$	0.91	**0.94**	0.93	**0.94**	0.93	0.93	0.88	0.68	0.62	0.61	0.51	0.03
	$M5_{SMI}$	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.91	0.61	0.21	0.00
	$M10_{SMI}$	0.69	0.69	0.69	0.70	**0.73**	0.72	**0.73**	0.71	0.66	0.57	0.13	0.04
	$NG1_{SMI}$	0.96	0.96	**0.97**	0.96	0.96	0.96	0.96	0.96	0.92	0.65	0.51	0.01
	$NG2_{SMI}$	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.91	0.86	0.61	0.21	0.00
	$NG3_{SMI}$	0.79	0.78	0.78	0.79	0.79	0.80	0.81	**0.84**	**0.84**	0.67	0.14	0.05
Avg Gain		0.000	0.001	0.005	0.007	0.01	0.01	0.005	-0.028	-0.073	-0.251	-0.587	

Effect of Pruning (2/2)

Precision (MAP) values when using unsupervised approach

		Pruning Percentage											Max Gain
		0	10	20	30	40	50	60	70	80	90	100	
Datasets	$M2_{PAM}$	0.58	0.59	0.62	0.62	0.62	**0.65**	**0.65**	0.63	0.57	0.55	0.51	0.07
	$M5_{PAM}$	0.62	0.67	0.65	**0.68**	0.64	0.58	0.56	0.51	0.48	0.39	0.22	0.06
	$M10_{PAM}$	0.43	0.42	0.43	0.45	**0.47**	**0.47**	**0.47**	**0.47**	0.44	0.34	0.11	0.04
	$NG1_{PAM}$	0.54	0.54	0.54	0.55	0.60	0.61	**0.62**	0.59	0.60	0.57	0.51	0.07
	$NG2_{PAM}$	0.60	**0.63**	0.59	0.60	0.62	0.60	0.58	0.58	0.59	0.53	0.50	0.03
	$NG3_{PAM}$	0.47	0.48	0.52	**0.57**	0.56	0.54	0.51	0.47	0.43	0.35	0.13	0.10
Avg Gain		0.000	0.015	0.018	0.038	0.043	0.035	0.025	0.005	-0.02	-0.085	-0.212	

Outline

1. Introduction
2. **Our Proposed Similarity Measure (χ-Sim)**
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. Experimentation
      i. Overview
      ii. Effect of the parameters – number of iterations and pruning
      iii. Comparison with other approaches
3. Extension for Supervised Classification
4. χ-Sim for the biclustering task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Comparison with other Methods

- 4 similarity measure based approaches
  - Cosine Similarity
  - Latent semantic Analysis (LSA) [Deerwester et al, 90]
  - Similarity in Non-Orthogonal Spaces (SNOS) [Liu et al, 04]
  - Similarity Refinement based Co-clustering (SRCC) [Zhang, 07]
- 3 co-clustering based approaches
  - Information theoretic co-clustering (ITCC) [Dhillon et al, 03]
  - Relation Summary Networks (RSN) [Long et al, 06]
  - Block Value Decomposition (BVD) [Long et al, 05]

Comparison with other Methods

- 4 similarity measure based approaches
  - **Cosine Similarity**
  - **Latent semantic Analysis (LSA) [Deerwester et al, 90]**
  - **Similarity in Non-Orthogonal Spaces (SNOS) [Liu et al, 04]**
  - Similarity Refinement based Co-clustering (SRCC) [Zhang, 07]
- 3 co-clustering based approaches
  - **Information theoretic co-clustering (ITCC) [Dhillon et al, 03]**
  - Relation Summary Networks (RSN) [Long et al, 06]
  - Block Value Decomposition (BVD) [Long et al, 05]

Comparison with other methods

Micro-Average Precision values using Knowledge intensive approach

	χ-Sim$_\rho$	Cosine	LSA	SNOS	ITCC
$M2_{SMI}$	**0.93 ± 0.01**	0.62 ± 0.04	0.78 ± 0.15	0.55± 0.02	0.71 ± 0.15
$M5_{SMI}$	**0.94 ± 0.04**	0.6 ± 0.08	0.82 ± 0.09	0.25± 0.02	0.48 ± 0.06
$M10_{SMI}$	**0.75 ± 0.04**	0.45 ± 0.07	0.54 ± 0.09	0.24± 0.06	0.28 ± 0.04
$NG1_{SMI}$	**1 ± 0**	0.90 ± 0.11	0.95 ± 0.02	0.51± 0.01	0.67 ± 0.12
$NG2_{SMI}$	**0.93 ± 0.01**	0.60 ± 0.08	0.82 ± 0.03	0.24± 0.02	0.57 ± 0.08
$NG3_{SMI}$	**0.85 ± 0.06**	0.60 ± 0.04	0.72 ± 0.05	0.22± 0.05	0.55 ± 0.06

Micro-Average Precision values using unsupervised approach

	χ-Sim$_\rho$	Cosine	LSA	SNOS	ITCC
$M2_{UMI}$	**0.61 ± 0.01**	0.58 ± 0.01	**0.61** ± 0.04	0.51 ± 0.02	0.58 ± 0.02
$M5_{UMI}$	**0.50 ± 0.04**	0.36 ± 0.05	0.45 ± 0.04	0.22 ± 0.03	0.32 ± 0.02
$M10_{UMI}$	**0.37 ± 0.04**	0.27 ± 0.03	0.34 ± 0.03	0.16 ± 0.03	0.22 ± 0.01
$NG1_{UMI}$	**0.63 ± 0.01**	0.58 ± 0.05	**0.63 ± 0.08**	0.50 ± 0.00	0.60 ± 0.05
$NG2_{UMI}$	**0.37 ± 0.04**	0.29 ± 0.04	0.33 ± 0.03	0.23 ± 0.05	0.28 ± 0.01
$NG3_{UMI}$	**0.24 ± 0.03**	0.18 ± 0.01	0.23 ± 0.02	0.14 ± 0.01	0.16 ± 0.03

Outline

1. Introduction
2. Our Proposed Similarity Measure (χ-Sim)
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. Experimentation
3. **Extension for Text Categorization**
4. Evaluation on Biclustering Task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Text Categorization

- Labeling new documents given an existing grouping
- Using the χ-Sim measure
  - Given a document $\mathbf{d}_{train}$ whose class label is known, and a document $\mathbf{d}_{test}$ whose class label is to be determined,

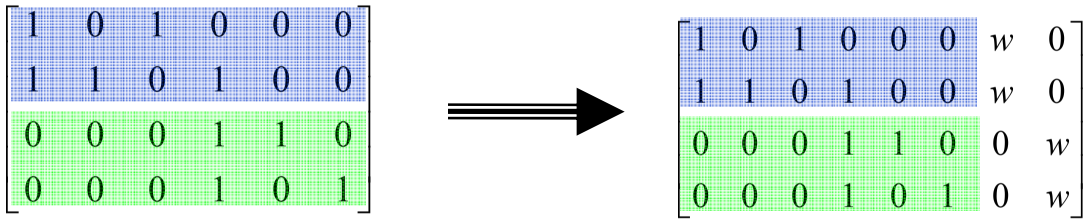
$$Sim(\mathbf{d}_{test}, \mathbf{d}_{train}) = \frac{\mathbf{d}_{test}(\mathbf{C})\mathbf{d}_{train}^{\mathrm{T}}}{\sum d_{test} \sum d_{train}}$$

- To Decide the category
  - k-NN
  - Nearest Class (NC) prototype vector (similar to Rocchio) [Joachims 97]

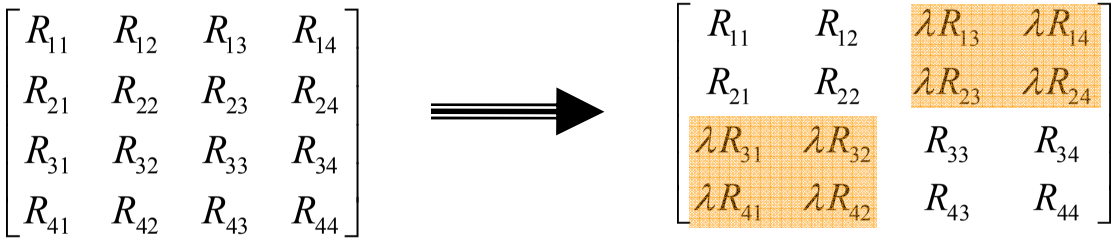
How to incorporate class label from training data?

Categorization Framework

- Increase *intra-class* similarities
  - By adding a class discriminating variables in **A** [Chakraborti, 2007]



- Decrease *inter-class* similarities
  - By penalizing inter-class similarities in **R**



Results and Comparison

- Supervised LSI
  - SLSI using sprinkling [Chakraborti et al, 06]
  - SLSI using Adaptive sprinkling [Chakraborti et al, 07]
- Higher-order co-occurrences using case based retrieval network [Wiratunga et al., 06]
  - Compute increasing levels of higher-order co-occurrences
  - Perform a weighted combination of these
- *k*-nearest neighbor (*k*-NN)
  - With cosine measure
  - Best of 'majority'
- Support Vector Machine (SVM)
  - Using linear Kernel [Joachims, 98]

Results

	20-Newsgroup			Reuters	SPAM
	9 classes	2 classes	2 classes	3-classes	2-classes
	HIERARCHICAL	HARDWARE	RELPOL	ORTHOGONAL	LINGSPAM
χ-Sim (w=2,λ= 1) + k-NN	0.66 ± 0.01	0.85 ± 0.01	0.98 ± 0.00	0.95 ± 0.00	0.88 ± 0.06
χ-Sim (w=2,λ= 1) + NC	0.77 ± 0.01	0.86 ± 0.02	0.98 ± 0.00	0.93 ± 0.00	0.88 ± 0.07
χ-Sim (w=0,λ= 0) + k-NN	0.49 ± 0.03	0.75 ± 0.02	0.96 ± 0.01	0.94 ± 0.01	0.96 ± 0.02
χ-Sim (w=0,λ= 0)+ NC	0.76 ± 0.01	0.87 ± 0.01	0.97 ± 0.01	0.95 ± 0.01	0.96 ± 0.03
χ-Sim (w=2,λ= 0) + k-NN	0.73 ± 0.01	0.86 ± 0.01	**0.99 ± 0.00**	0.96 ± 0.00	0.93 ± 0.06
χ-Sim (w=2,λ= 0) + NC	**0.77 ± 0.01**	0.87 ± 0.01	0.98 ± 0.01	0.94 ± 0.01	0.98 ± 0.01
k-NN + Cosine	0.65 ± 0.01	0.84 ± 0.01	0.98 ± 0.01	0.95 ± 0.01	0.97 ± 0.01
SLSI	-	0.80	0.94	-	**0.98**
ASLSI	0.60	-	-	0.95	-
HOCRN	-	0.80	0.94	-	-
SVM	**0.77 ± 0.01**	**0.87 ± 0.07**	0.97 ± 0.01	**0.96 ± 0.01**	0.95 ± 0.02

Outline

1. Introduction
2. Our Proposed Similarity Measure (χ-Sim)
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. Experimentation
3. Extension for Text Categorization
4. Evaluation on Biclustering Task
5. Improvement to χ-Sim
6. Conclusion + Perspectives

Gene Expression Data

- Data Matrix **A**,
  - *m* rows corresponding to genes
  - *n* columns corresponding to conditions
  - A_{ij} corresponds to intensity of gene expression measured under the given condition

- Goal
  - Find "*co-clusters*" (also referred to as biclusters) that corresponds to a subset of genes showing similar intensities over a subset of conditions.

- A "good" co-cluster is one
  - Whose genes exhibit similar patterns in the co-cluster
  - That shows significant variation under different conditions

Adaptation of χ-Sim

- Genes express different levels of intensities
  - Centering and reduction are necessary
    - Row Scaling

$$A^r_{ij} = \frac{A_{ij} - \mu_i}{\sigma_i}$$

    - Column Scaling

$$A^c_{ij} = \frac{A_{ij} - \mu_j}{\sigma_j}$$

- Use $\mathbf{A}^r$ when calculating $Sim(\text{gene}_i, \text{gene}_j)$

- Use $\mathbf{A}^c$ when calculating $Sim(\text{condition}_i, \text{condition}_j)$

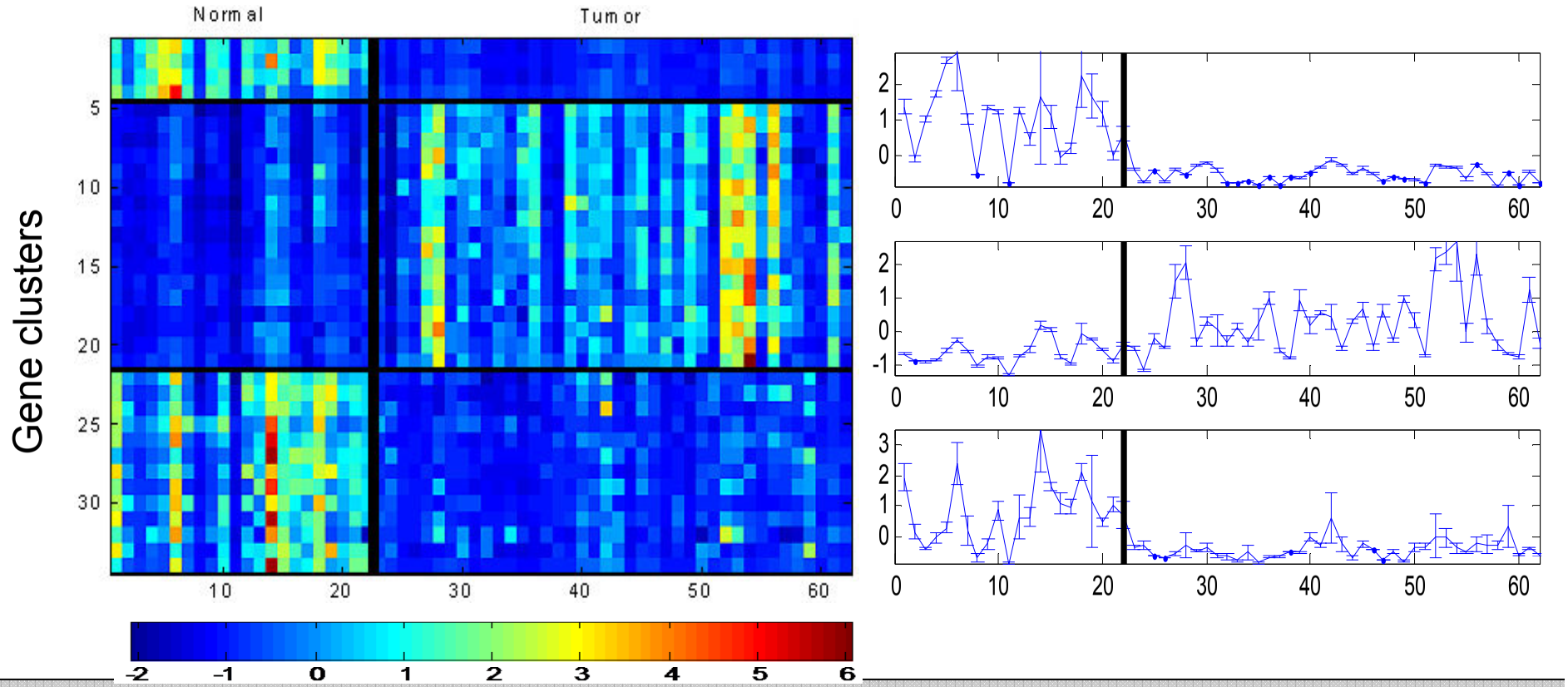
Extracting co-clusters

- Steps
  - Compute the matrices **R** and **C** as previous
  - Cluster **R** and **C** independently into k and l clusters (user given)
  - Select top k most discriminating co-clusters as follows

		Condition Clusters					
		$\hat{y}_1$	$\hat{y}_2$	…	$\hat{y}_l$	Max-Min	Index
Gene Clusters	$\hat{x}_1$	$\hat{x}_1\hat{y}_1$	$\hat{x}_1\hat{y}_2$	…	$\hat{x}_1\hat{y}_l$	$\max(\hat{x}_1\hat{y}_i) - \min(\hat{x}_1\hat{y}_j)$	$Ind_1 \in [1..l]$
	$\hat{x}_2$	$\hat{x}_2\hat{y}_1$	$\hat{x}_2\hat{y}_2$	…	$\hat{x}_2\hat{y}_l$	$\max(\hat{x}_2\hat{y}_i) - \min(\hat{x}_2\hat{y}_j)$	$Ind_2 \in [1..l]$
	…	…	…	…	…	…	…
	$\hat{x}_k$	$\hat{x}_k\hat{y}_1$	$\hat{x}_k\hat{y}_2$	…	$\hat{x}_k\hat{y}_l$	$\max(\hat{x}_k\hat{y}_i) - \min(\hat{x}_k\hat{y}_j)$	$Ind_k \in [1..l]$

Top 3 co-cluster on colon cancer dataset

- Colon Cancer dataset
  - 1096 genes
  - 62 tissues — Normal (42) + Tumor (20)



Outline

1. Introduction
2. Our Proposed Similarity Measure (χ-Sim)
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. Experimentation
3. Extension for Text Categorization
4. Evaluation on Biclustering Task
5. **Improvement to χ-Sim**
6. Conclusion + Perspectives

Improvement to χ-Sim [ICMLA 10]

- The L_p Norm has been shown to be sensitive to p when dealing with high dimensional data [Aggarwal, 01]
  - Lower values of p are preferred since

    "for higher dimensionality, the relative contrast provided by a norm with smaller parameter p is more likely to dominate another with a larger parameter".

- We introduce two modification to the original χ-Sim Measure
  1. Motivated by their results of we introduce a measure p in χ-Sim
  2. We also introduce an L_2 Normalization

Improvement to χ-Sim [ICMLA 10]

- Let $\mathbf{A}^{\circ p}$ denote the matrix $\mathbf{A}$ with $(A_{ij})^p \ \forall i,j$

  then

$$(R^{(k+1)})_{ij}^p = \frac{\sqrt[p]{(\mathbf{A}^{\circ p} \times \mathbf{C}^{(k)} \times (\mathbf{A}^{\circ p})^{\mathrm{T}})_{ij}}}{\sqrt[2p]{(R^{(k+1)})_{ii}^p \times (R^{(k+1)})_{jj}^p}}$$

  - The value for $\mathbf{C}$ can be calculated similarly
  - $R_{ii} = 1, R_{ij} \notin [0,1]$

Precision (MAP) results using Unsupervised feature selection (PAM)

	M2	M5	M10	NG1	NG2	NG3
χ-Sim	0.58 ± 0.07	0.62 ± 0.12	0.43 ± 0.04	0.54 ± 0.03	0.60 ± 0.12	0.47 ± 0.05
$\chi\text{-Sim}_{\rho}$	0.65 ± 0.09	0.68 ± 0.06	0.47 ± 0.04	0.62 ± 0.12	0.63 ± 0.14	0.57 ± 0.04
$\chi\text{-Sim}^{1}$	0.54 ± 0.06	0.62 ± 0.13	0.36 ± 0.04	0.53 ± 0.02	0.35 ± 0.09	0.30 ± 0.05
$\chi\text{-Sim}^{1}_{\rho}$	0.80 ± 0.13	0.77 ± 0.08	0.53 ± 0.05	0.75 ± 0.07	$\mathbf{0.73} \pm 0.06$	0.61 ± 0.03
$\chi\text{-Sim}^{0.8}$	0.54 ± 0.05	0.66 ± 0.07	0.37 ± 0.06	0.52 ± 0.02	0.38 ± 0.08	0.36 ± 0.04
$\chi\text{-Sim}^{0.8}_{\rho}$	$\mathbf{0.81} \pm 0.10$	$\mathbf{0.79} \pm 0.05$	$\mathbf{0.55} \pm 0.04$	$\mathbf{0.81} \pm 0.02$	0.72 ± 0.02	$\mathbf{0.64} \pm 0.04$

Outline

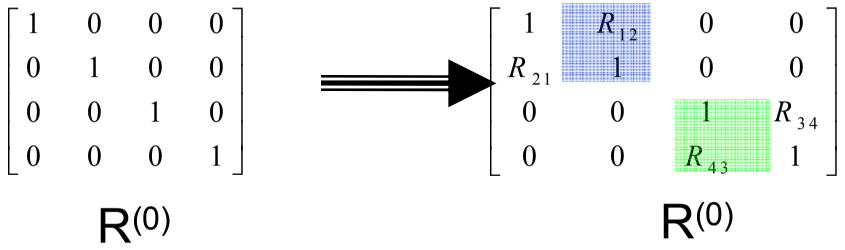
1. Introduction
2. Our Proposed Similarity Measure (χ-Sim)
   a. The χ-Sim Measure
   b. Relationship to Previous Work
   c. Experimentation
3. Extension for Text Categorization
4. Evaluation on Biclustering Task
5. Improvement to χ-Sim
6. **Conclusion + Perspectives**

Conclusion

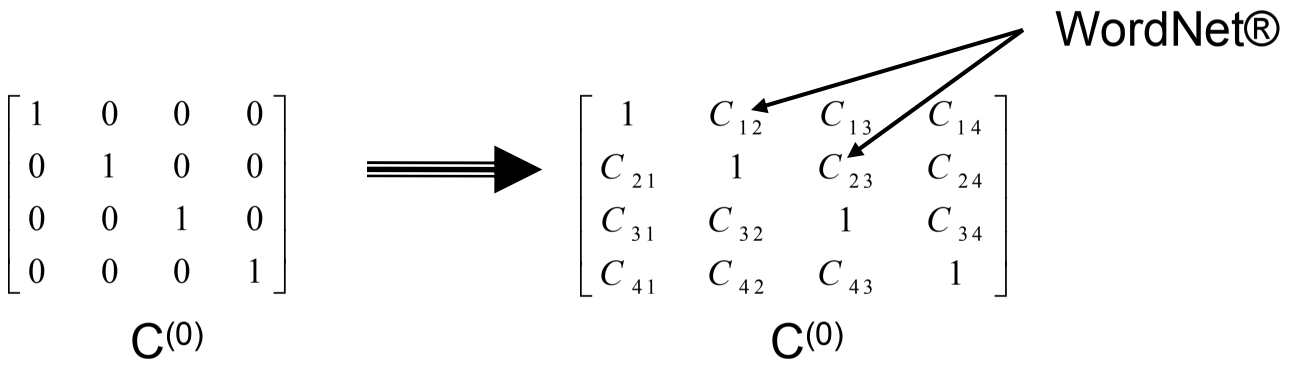
- New approach to use co-similarity for co-clustering
  - Takes higher order co-occurrences into account
  - Performs implicit co-clustering
  - Allows for both one-way clustering and co-clustering
- Allows to use all the clustering methods based on a similarity
  - Partitioning, Hierarchical,...
- Using co-similarity reduces the effects of sparsity and improves accuracy
- We propose an extension of χ-Sim for the supervised case by
  - Increasing intra-class similarity values
  - Decreasing inter-class similarity values
- Evaluated χ-Sim for biclustering gene expression data

Perspective

- Explore different ways to incorporate external knowledge

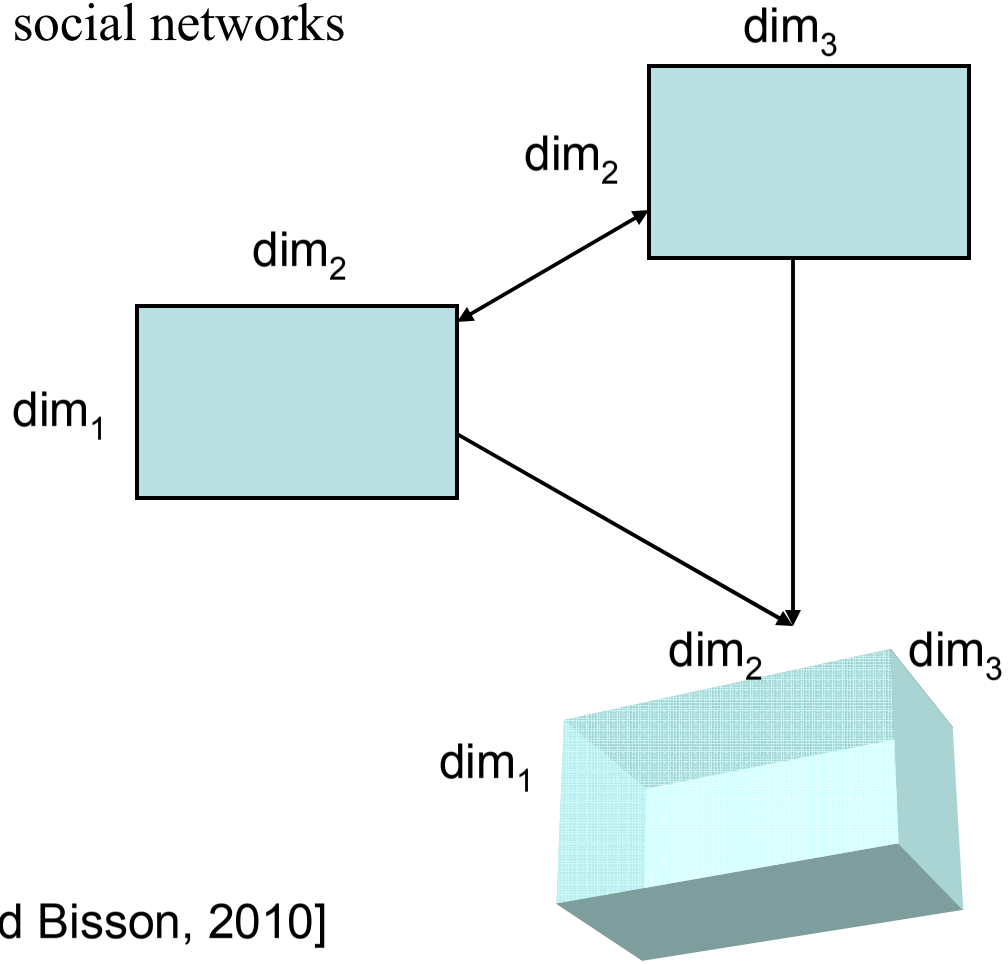


- Use external sources to incorporate word similarities



Perspective

- Extension to *k*-partite graphs
  - e.g. In social networks



  - [Grimal and Bisson, 2010]

References

- Blondel, V. D, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren. 2004. “A measure of similarity between graph vertices: Applications to synonym extraction and web searching.” *Siam Review* 46:647–666.
- Chakraborti, S., R. Lothian, N. Wiratunga, and S. Watt. 2006. “Sprinkling: supervised latent semantic Indexing.” *Advances in Information Retrieval* 3936:510-514.
- Chakraborti, S., R. Mukras, et al. 2007. “Supervised latent semantic indexing using adaptive sprinkling.” Pp. 1582–1587 in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*.
- Deerwester, S., S. T Dumais, G. W Furnas, T. K Landauer, and R. Harshman. 1990. “Indexing by latent semantic analysis.” *Journal of the American society for information science* 41:391–407.
- Dhillon, I. S, S. Mallela, and D. S Modha. 2003. “Information-theoretic co-clustering.” Pp. 89–98 *in Proceedings of the 9th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*
- Frawley, W. J, G. Piatetsky-Shapiro, and C. J Matheus. 1992. “Knowledge discovery in databases: An overview.” *Ai Magazine* 13:57.
- Grimal, Clément, and G. Bisson. 2010. “Classification à partir d'une collection de matrices.” in *REcherche et REcommandation d'information dans les RESeaux sOciaux (REiSO)*.
- Jeh, G., and J. Widom. 2002. “SimRank: A measure of structural-context similarity ” *in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

References

- Joachims, T. 1998. “Text categorization with support vector machines: Learning with many relevant features.” in *Machine Learning: ECML-98 10th European Conference on Machine Learning, Chemnitz, Germany*.
- Liu, N. et al. 2004. “Learning similarity measures in non-orthogonal space.” Pp. 334–341 *in Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM)*
- Long, B., X. Wu, Z. M Zhang, and P. S Yu. 2006. “Unsupervised learning on k-partite graphs.” Pp. 317–326 in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Long, B., Z. M Zhang, and P. S Yu. 2005. “Co-clustering by block value decomposition.” Pp. 635–640 *in Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge discovery in data mining*
- N. Wiratunga, Chakraborti, S., R. Lothian, and S. Watt. 2007. “Acquiring Word Similarities with Higher Order Association Mining.” *Case Based Resoning, Research and Development* 4626:61-76.
- Salton, G., A. Wong, and C. S. Yang. 1975. “A vector space model for automatic indexing.” *Communications of the ACM* 18:620.
- Zhang, J. 2007. “Co-clustering by similarity refinement.” Pp. 381–386 *in Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA)*
- Aggarwal, C. C., Hinerburg, A., Keim, D.A., 2001. “On the suprising behavior of distance metrics in high dimensional space”. *In Proceedings of International Conference on Database theory (ICDT)*

Publications

- International Conferences
  - *Hussain F, Grimal C., Bisson G. : "An improved Co-Similarity Measure for Document Clustering",* **9th IEEE International Conference on Machine Learning and Applications (ICMLA)**, 12-14th Dec. 2010, Washington, United States. **[*To Appear*]**
  - *Hussain F. and Bisson G. : "Text Categorization using Word Similarities Based on Higher Order Co-Occurrences",* ***Society for Industrial and Applied Mathematics International Conference on Data Mining (SDM 2010)****, Columbus, Ohio, April 29-May 1, 2010.*
  - Bisson G., Hussain F. : "*X-sim: A new similarity measure for the co-clustering task*", **7th IEEE International Conference on Machine Learning and Applications (ICMLA)**, 11-13th Dec. 2008, San Diego, United States.

- National Conferences
  - CAP 2010
  - CAP 2009

Thank You

fawadsyed@gmail.com