An Improved Co-Similarity Measure for Document Clustering

Syed Fawad Hussain, Clément Grimal and Gilles Bisson

December 12^{th}, 2010

UNIVERSITÉ DE GRENOBLE

The text mining context

Document#1:
A contruction found in villages and in the suburbs of bigger town, used to house a family.

Document#2:
A building which main purpose is to provide accomodation to human beings.

The text mining context

Document#1:
A contruction found in villages and in the suburbs of bigger town, used to house a family.

Document#2:
A building which main purpose is to provide accomodation to human beings.

With a classical approach:

No shared terms between the two documents

$\rightarrow \text{Similarity}(Document\#1, Document\#2) = 0$

The text mining context

Document#1:
A contruction found in villages and in the suburbs of bigger town, used to house a family.

Document#2:
A building which main purpose is to provide accomodation to human beings.

With a classical approach:

No shared terms between the two documents

$\rightarrow \text{Similarity}(Document\#1, Document\#2) = 0$

Using a co-similarity approach:

Clustering of the terms

$\rightarrow \text{Similarity}(Document\#1, Document\#2) > 0$

Model

We used the classical Vector Space Model of Salton (1971):
$\mathbf{M}$: documents/words matrix of r rows and c columns

- $\mathbf{m}_{i:} = [m_{i1} \dots m_{ic}]$: row vector describing document i
- $\mathbf{m}_{:j} = [m_{1j} \dots m_{rj}]$: column vector describing word j

Model

We used the classical Vector Space Model of Salton (1971):
$\mathbf{M}$: documents/words matrix of r rows and c columns

- $\mathbf{m}_{i:} = [m_{i1} \dots m_{ic}]$: row vector describing document i
- $\mathbf{m}_{:j} = [m_{1j} \dots m_{rj}]$: column vector describing word j

We want to compute:

- $\mathbf{SR}$: square similarity matrix (documents) of size r, with $sr_{ij} \in [0, 1]$
- $\mathbf{SC}$: square similarity matrix (words) of size c, with $sc_{ij} \in [0, 1]$

Model

We used the classical Vector Space Model of Salton (1971):
$\mathbf{M}$: documents/words matrix of r rows and c columns

- $\mathbf{m}_{i:} = [m_{i1} \dots m_{ic}]$: row vector describing document i
- $\mathbf{m}_{:j} = [m_{1j} \dots m_{rj}]$: column vector describing word j

We want to compute:

- $\mathbf{SR}$: square similarity matrix (documents) of size r, with $sr_{ij} \in [0, 1]$
- $\mathbf{SC}$: square similarity matrix (words) of size c, with $sc_{ij} \in [0, 1]$

Basic Idea

- Two documents are similar if they contain similar words.
- Two words are similar if they appear in similar documents.

Model

We used the classical Vector Space Model of Salton (1971):
$\mathbf{M}$: documents/words matrix of r rows and c columns

- $\mathbf{m}_{i:} = [m_{i1} \dots m_{ic}]$: row vector describing document i
- $\mathbf{m}_{:j} = [m_{1j} \dots m_{rj}]$: column vector describing word j

We want to compute:

- $\mathbf{SR}$: square similarity matrix (documents) of size r, with $sr_{ij} \in [0, 1]$
- $\mathbf{SC}$: square similarity matrix (words) of size c, with $sc_{ij} \in [0, 1]$

Basic Idea

- Two documents are similar if they contain similar words.
- Two words are similar if they appear in similar documents.

Joint construction of the two similarity matrices $\mathbf{SR}$ and $\mathbf{SC}$.

Outline

Similarity between two documents

- Classical approach: similarity = f(shared words)

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathrm{F_s}(m_{i1}, m_{j1}) + \cdots + \mathrm{F_s}(m_{ic}, m_{jc})$$

with $\mathrm{F_s}$ a similarity function (absolute difference, product, etc.).

Similarity between two documents

- Classical approach: similarity = f(shared words)

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathrm{F_s}(m_{i1}, m_{j1}) + \cdots + \mathrm{F_s}(m_{ic}, m_{jc})$$

with $\mathrm{F_s}$ a similarity function (absolute difference, product, etc.).

- Using $\mathbf{SC}$ (usually, $sc_{ii} = 1$):

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sum_{l=1}^{c} \mathrm{F_s}(m_{il}, m_{jl}) \times sc_{ll}$$

Similarity between two documents

- Classical approach: similarity = f(shared words)

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathrm{F_s}(m_{i1}, m_{j1}) + \cdots + \mathrm{F_s}(m_{ic}, m_{jc})$$

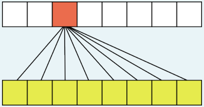
with $\mathrm{F_s}$ a similarity function (absolute difference, product, etc.).

- Using $\mathbf{SC}$ (usually, $sc_{ii} = 1$):

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sum_{l=1}^{c} \mathrm{F_s}(m_{il}, m_{jl}) \times sc_{ll}$$

- Now comparing every pair of words:

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sum_{l=1}^{c} \sum_{n=1}^{c} \mathrm{F_s}\left(m_{il}, m_{jn}\right) \times sc_{ln}$$



New approach - Pseudo-norm k

- If $\mathrm{F_s}(m_{ij}, m_{kl}) = m_{ij} \times m_{kl}$:

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathbf{m}_{i:} \times \mathbf{SC} \times \mathbf{m}_{j:}^{\mathrm{T}}$$

New approach - Pseudo-norm k

- If $\mathrm{F_s}(m_{ij}, m_{kl}) = m_{ij} \times m_{kl}$:

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathbf{m}_{i:} \times \mathbf{SC} \times \mathbf{m}_{j:}^{\mathrm{T}}$$

- We introduce a pseudo-norm k (see [Aggarwal et al.(2001)]):

$$\mathrm{Sim}^k(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sqrt[k]{(\mathbf{m}_{i:})^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k} = \langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k$$

$\rightarrow$ we have $\|\mathbf{m}_{i:}\|_{\mathbf{SC}}^k = \sqrt{\langle \mathbf{m}_{i:}, \mathbf{m}_{i:} \rangle_{\mathbf{SC}}^k}$

New approach - Pseudo-norm k

- If $\mathrm{F_s}(m_{ij}, m_{kl}) = m_{ij} \times m_{kl}$:

$$\mathrm{Sim}(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \mathbf{m}_{i:} \times \mathbf{SC} \times \mathbf{m}_{j:}^{\mathrm{T}}$$

- We introduce a pseudo-norm k (see [Aggarwal et al.(2001)]):

$$\mathrm{Sim}^k(\mathbf{m}_{i:}, \mathbf{m}_{j:}) = \sqrt[k]{(\mathbf{m}_{i:})^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k} = \langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k$$

$\rightarrow$ we have $\|\mathbf{m}_{i:}\|_{\mathbf{SC}}^k = \sqrt{\langle \mathbf{m}_{i:}, \mathbf{m}_{i:} \rangle_{\mathbf{SC}}^k}$

- Then we need to normalize this similarity:

$$sr_{ij} = \frac{\sqrt[k]{(\mathbf{m}_{i:})^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} \in [0, 1]$$

Generic form

- Now:

$$sr_{ij} = \frac{\sqrt[k]{\left(\mathbf{m}_{i:}\right)^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} = \frac{\langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})}$$

Generic form

- Now:

$$sr_{ij} = \frac{\sqrt[k]{\left(\mathbf{m}_{i:}\right)^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} = \frac{\langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})}$$

- With special values for k, $\mathbf{SC}$ and $\mathcal{N}$, we have:
  - Jaccard: $\mathbf{SC} = \mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1 - \mathbf{m}_{i:}\mathbf{m}_{j:}^{\mathrm{T}}$
  - Dice: $\mathbf{SC} = 2\mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1$
  - "Classical" χ-Sim: $k = 1$, $\mathcal{N} = |\mathbf{m}_{i:}| \times |\mathbf{m}_{j:}|$
  - Generalized Cosine: $\mathbf{SC} > 0$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_{\mathbf{SC}} \times \|\mathbf{m}_{j:}\|_{\mathbf{SC}}$

Generic form

- Now:

$$sr_{ij} = \frac{\sqrt[k]{\left(\mathbf{m}_{i:}\right)^k \times \mathbf{SC} \times \left(\mathbf{m}_{j:}^{\mathrm{T}}\right)^k}}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})} = \frac{\langle \mathbf{m}_{i:}, \mathbf{m}_{j:} \rangle_{\mathbf{SC}}^k}{\mathcal{N}(\mathbf{m}_{i:}, \mathbf{m}_{j:})}$$

- With special values for k, $\mathbf{SC}$ and $\mathcal{N}$, we have:
  - Jaccard: $\mathbf{SC} = \mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1 - \mathbf{m}_{i:}\mathbf{m}_{j:}^{\mathrm{T}}$
  - Dice: $\mathbf{SC} = 2\mathbf{I}$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_1 + \|\mathbf{m}_{j:}\|_1$
  - "Classical" χ-SIM: $k = 1$, $\mathcal{N} = |\mathbf{m}_{i:}| \times |\mathbf{m}_{j:}|$
  - Generalized Cosine: $\mathbf{SC} > 0$, $k = 1$, $\mathcal{N} = \|\mathbf{m}_{i:}\|_{\mathbf{SC}} \times \|\mathbf{m}_{j:}\|_{\mathbf{SC}}$
  - χ-SIM^k : $\mathcal{N} = \|\mathbf{m}_{i:}\|_{\mathbf{SC}}^k \times \|\mathbf{m}_{j:}\|_{\mathbf{SC}}^k$

Pruning parameter p

In such a corpus...

Many words are not specific enough, and creates a lot of irrelevant similarities. These similarities can be considered as noise.

Example: Astronomy / Mythology

The word *Hercules* can appear once in an astronomy document, and “link” it to all mythology documents dealing with greek heroes...

Pruning parameter p

In such a corpus...

Many words are not specific enough, and creates a lot of irrelevant similarities. These similarities can be considered as noise.

Example: Astronomy / Mythology

The word *Hercules* can appear once in an astronomy document, and “link” it to all mythology documents dealing with greek heroes...

How to deal with it?

Hypothetis: these irrelevant similarities are small.
$\rightarrow$ At each iteration, we remove the smallest $p\%$ of the similarity matrices.

Algorithm for χ-Sim_p^k

1. $\mathbf{SR}^{(0)}$ and $\mathbf{SC}^{(0)}$ are initialized with the identity matrix.

Algorithm for χ-$\textsc{Sim}_p^k$

1. $\mathbf{SR}^{(0)}$ and $\mathbf{SC}^{(0)}$ are initialized with the identity matrix.
2. At each iteration t, we update both similarity matrices :
    3. Update $\mathbf{SR}^{(t)}$ using $\mathbf{SC}^{(t-1)}$
    4. Prune $\mathbf{SR}^{(t)}$

    5. Update $\mathbf{SC}^{(t)}$ using $\mathbf{SR}^{(t-1)}$
    6. Prune $\mathbf{SC}^{(t)}$

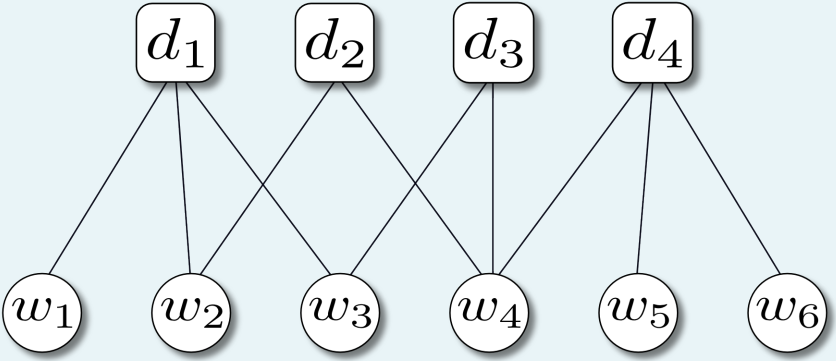
Algorithm for $\chi\text{-}\mathrm{Sim}_p^k$

1. $\mathbf{SR}^{(0)}$ and $\mathbf{SC}^{(0)}$ are initialized with the identity matrix.
2. At each iteration t, we update both similarity matrices :
   3. Update $\mathbf{SR}^{(t)}$ using $\mathbf{SC}^{(t-1)}$
   4. Prune $\mathbf{SR}^{(t)}$
   5. Update $\mathbf{SC}^{(t)}$ using $\mathbf{SR}^{(t-1)}$
   6. Prune $\mathbf{SC}^{(t)}$

Usually, $t = 4$ is enough.

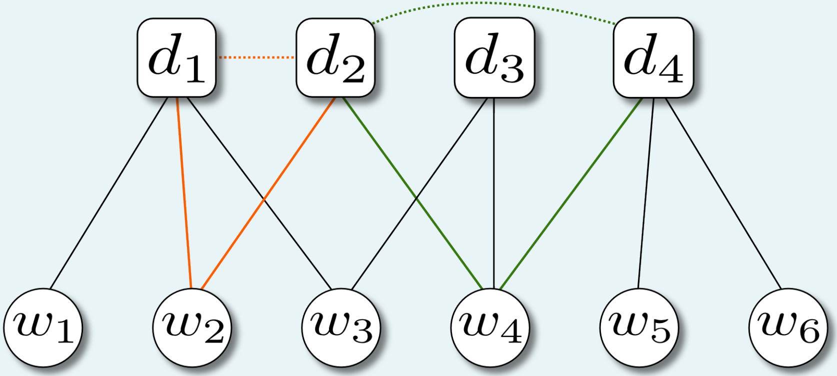
Meaning of an iteration

Bi-partite graph representing a simple corpus



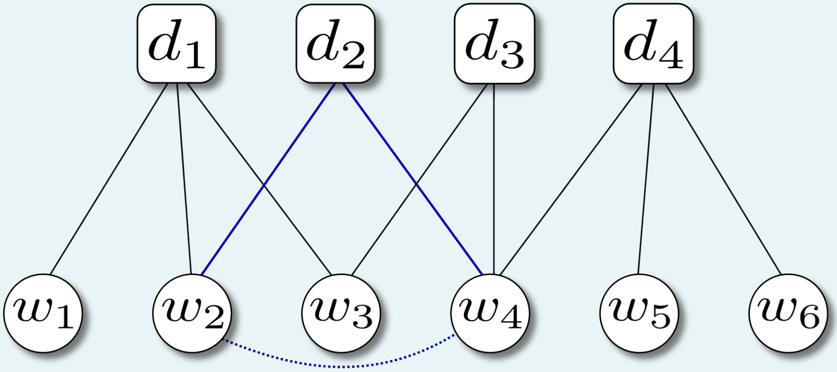
Meaning of an iteration

First iteration: $sr_{12} > 0$ and $sr_{24} > 0$, but $sr_{14} = 0$



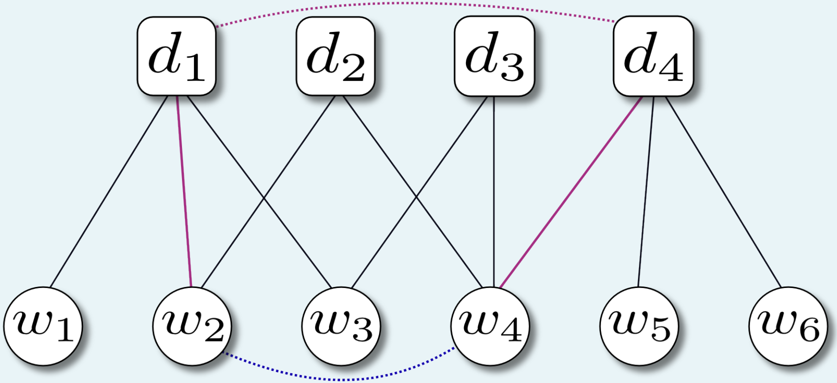
Meaning of an iteration

Second part of the first iteration: $sc_{24} > 0$



Meaning of an iteration

Second iteration: through sc_{24}, now $sr_{14} > 0$



Outline

Methods

- Five similarity measures
  - Cosine
  - χ-Sim (with or without k and p) [Hussain et al.(2010)]
  - LSA (Latent Semantic Analysis) [Deerwester et al.(1990)]
  - SNOS (Similarity in Non-Orthogonal Space) [Liu et al.(2004)]
  - CTK (Commute Time Kernel) [Yen et al.(2009)]

  \+ Ascendant Hierarchical Clustering, with Ward's index

- Three co-clustering methods
  - ITCC (Information Theoric Co-Clustering) [Dhillon et al.(2003)]
  - *BVD (Block Value Decomposition) [Long et al.(2005)]*
  - *RSN (k-partite graph partioning algorithm) [Long et al.(2006)]*

Methodology and Data

Methodology

- We randomly select subsets of documents already labeled
- We measure the quality of the clusters using the micro-averaged precision

The subsets:

Name	Newsgroups included	#clusters.	#docs.
M2	talk.politics.mideast, talk.politics.misc	2	500
M5	comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast	5	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	10	500
NG1	rec.sports.baseball, rec.sports.hockey	2	400
NG2	comp.os.ms-windows.misc, comp.windows.x, rec.motorcycles, sci.crypt, sci.space	5	1000
NG3	comp.os.ms-windows.misc, comp.windows.x, misc.forsale, rec.motorcycles, sci.crypt, sci.space, talk.politics.mideast, talk.religion.misc	8	1600

Results

	M2	M5	M10	NG1	NG2	NG3
Cosine	0.61 ± 0.04	0.54 ± 0.08	0.39 ± 0.03	0.52 ± 0.01	0.60 ± 0.05	0.49 ± 0.02
LSA	0.79 ± 0.09	0.66 ± 0.05	0.44 ± 0.04	0.56 ± 0.05	0.61 ± 0.06	0.52 ± 0.03
ITCC	0.70 ± 0.05	0.54 ± 0.05	0.29 ± 0.05	0.61 ± 0.06	0.44 ± 0.08	0.49 ± 0.07
SNOS	0.51 ± 0.01	0.26 ± 0.04	0.20 ± 0.02	0.51 ± 0.00	0.24 ± 0.01	0.22 ± 0.02
CTK	0.75 ± 0.10	0.78 ± 0.04	0.54 ± 0.05	0.72 ± 0.14	0.66 ± 0.06	0.58 ± 0.02

Results

	M2	M5	M10	NG1	NG2	NG3
Cosine	0.61 ± 0.04	0.54 ± 0.08	0.39 ± 0.03	0.52 ± 0.01	0.60 ± 0.05	0.49 ± 0.02
LSA	0.79 ± 0.09	0.66 ± 0.05	0.44 ± 0.04	0.56 ± 0.05	0.61 ± 0.06	0.52 ± 0.03
ITCC	0.70 ± 0.05	0.54 ± 0.05	0.29 ± 0.05	0.61 ± 0.06	0.44 ± 0.08	0.49 ± 0.07
SNOS	0.51 ± 0.01	0.26 ± 0.04	0.20 ± 0.02	0.51 ± 0.00	0.24 ± 0.01	0.22 ± 0.02
CTK	0.75 ± 0.10	0.78 ± 0.04	0.54 ± 0.05	0.72 ± 0.14	0.66 ± 0.06	0.58 ± 0.02
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
χ-SIM$_p$	0.65 ± 0.09	0.68 ± 0.06	0.47 ± 0.04	0.62 ± 0.12	0.63 ± 0.14	0.57 ± 0.04

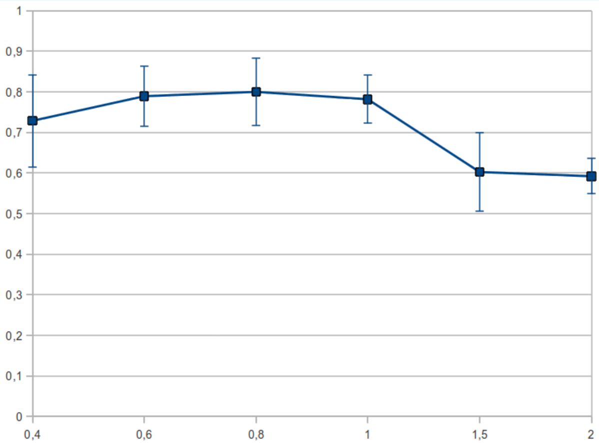
Results

	M2	M5	M10	NG1	NG2	NG3
Cosine	0.61 ± 0.04	0.54 ± 0.08	0.39 ± 0.03	0.52 ± 0.01	0.60 ± 0.05	0.49 ± 0.02
LSA	0.79 ± 0.09	0.66 ± 0.05	0.44 ± 0.04	0.56 ± 0.05	0.61 ± 0.06	0.52 ± 0.03
ITCC	0.70 ± 0.05	0.54 ± 0.05	0.29 ± 0.05	0.61 ± 0.06	0.44 ± 0.08	0.49 ± 0.07
SNOS	0.51 ± 0.01	0.26 ± 0.04	0.20 ± 0.02	0.51 ± 0.00	0.24 ± 0.01	0.22 ± 0.02
CTK	0.75 ± 0.10	0.78 ± 0.04	0.54 ± 0.05	0.72 ± 0.14	0.66 ± 0.06	0.58 ± 0.02
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
χ-SIM$_p$	0.65 ± 0.09	0.68 ± 0.06	0.47 ± 0.04	0.62 ± 0.12	0.63 ± 0.14	0.57 ± 0.04
χ-SIM$_p^{0.8}$	$\mathbf{0.81} \pm 0.10$	$\mathbf{0.79} \pm 0.05$	$\mathbf{0.55} \pm 0.04$	$\mathbf{0.81} \pm 0.02$	$\mathbf{0.72} \pm 0.02$	$\mathbf{0.64} \pm 0.04$

More results in the paper...

Influence of k

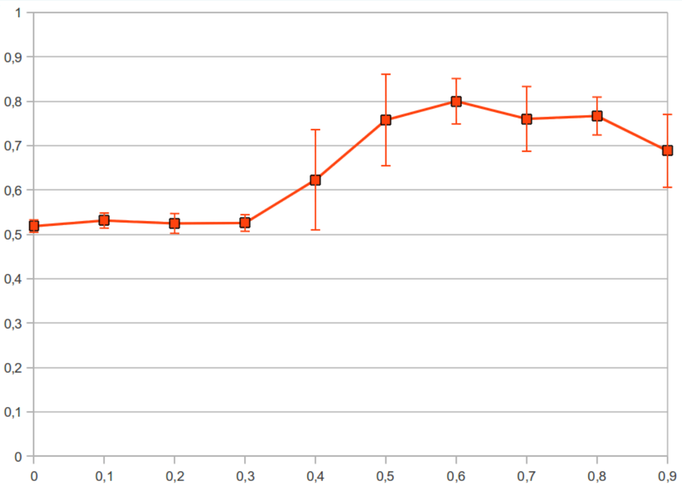
Tests on NG1, displaying the standard deviation over the 10 folds as error bars.



See [Aggarwal et al.(2001)].

Influence of p

Tests on NG1, displaying the standard deviation over the 10 folds as error bars.



Conclusion & Perspectives

Improvements of χ-Sim

- Exploration of different normed spaces (k)
- Pruning of the similarity matrices (p)
- Very good experimental results

Conclusion & Perspectives

Improvements of χ-Sim

- Exploration of different normed spaces (k)
- Pruning of the similarity matrices (p)
- Very good experimental results

Perspectives

- Use a damping factor to decrease the weight of higher order co-occurences
- Automatically find the best values for k and p
- Results are good, but a better theoritical understanding is needed
- Use the χ-Sim similarity matrices as input for the kernel-based algorithms used with CTK

Thank you very much!

Any questions?

References

S. Deerwester, S. T. Dumais, G. W. Furnas, Thomas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD*, pages 89–98, 2003.

S. F. Hussain, C. Grimal, and G. Bisson. An improved co-similarity measure for document clustering. In *Proceedings of the 9th ICMLA*, 2010.

N. Liu, B. Zhang, J. Yan, Q. Yang, S. Yan, Z. Chen, F. Bai, and W. ying Ma. Learning similarity measures in non-orthogonal space. In *Proceedings of the 13th ACM CIKM*, pages 334–341. ACM Press, 2004.

B. Long, Z. M. Zhang, and P. S. Yu. Co-clustering by block value decomposition. In *Proceedings of the 11th ACM SIGKDD*, pages 635–640, New York, NY, USA, 2005. ACM.

B. Long, Z. M. Zhang, X. Wú, and P. S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd ICML*, pages 585–592, New York, NY, USA, 2006. ACM.

L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens. Graph nodes clustering with the sigmoid commute-time kernel: A comparative study. *Data Knowl. Eng.*, 68(3):338–361, 2009.

C. C. Aggarwal and A. Hinneburg, and D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Lecture Notes in Computer Science*, 420–434, 2001.

Parameter k

The generalized χ-Sim

$$\forall i,j \in 1..r,\ \ sr_{ij} = \frac{\mathrm{Sim}^k(\mathbf{m}_{i:},\mathbf{m}_{j:})}{\sqrt{\mathrm{Sim}^k(\mathbf{m}_{i:},\mathbf{m}_{i:})} \times \sqrt{\mathrm{Sim}^k(\mathbf{m}_{j:},\mathbf{m}_{j:})}}$$

$$\forall i,j \in 1..c,\ \ sc_{ij} = \frac{\mathrm{Sim}^k(\mathbf{m}_{:i},\mathbf{m}_{:j})}{\sqrt{\mathrm{Sim}^k(\mathbf{m}_{:i},\mathbf{m}_{:i})} \times \sqrt{\mathrm{Sim}^k(\mathbf{m}_{:j},\mathbf{m}_{:j})}}$$

For $k = 1$, $\mathbf{SR}$ and $\mathbf{SC}$ are not positive semi-definite...
We are not defining an inner product so it is not a generalized cosine measure...