Context Sensitive Semantic Smoothing for Model-based Document Clustering*

Xiaodan Zhang and Xiaohua Zhou
College of Information Science & Technology, Drexel University

Abstract

A document is often full of class-independent “general” words and short of class-specific “core” words, which leads to the difficulty of document clustering. We argue that both problems will be relieved after suitable smoothing of document models in agglomerative approaches and of cluster models in partitional approaches, and hence improve clustering quality. To the best of our knowledge, most model-based clustering approaches use Laplacian smoothing to prevent zero probability while most similarity-based approaches employ the heuristic TF*IDF scheme to discount the effect of “general” words. Inspired by a series of statistical translation language model for text retrieval, we propose in this paper a novel smoothing method referred to as context-sensitive semantic smoothing for document clustering purpose. The comparative experiment on three datasets shows that model-based clustering approaches with semantic smoothing is effective in improving cluster quality.

Background: Why Semantic Smoothing for DC?

A Scenario

How do you judge the similarity of two documents which do not share topical words?

- Doc1
  - I am looking for any information about the **space program**. This includes NASA, the shuttles, history, anything! I would like to know if anyone could suggest books, periodicals, even ftp sites for a novice who is interested in the space program.

- Doc2
  - the Phobos **mission** did return some useful data including images of Phobos itselfBy the way, the new book entitled "**Mars**" (Kieffer et al, 1992, University of Arizona Press) has a great chapter on **spacecraft** exploration of the planet. The chapter is co-authored by V.I. Moroz of the Space Research Institute

- Doc3:
  - **ROCKET LAUNCH** OBSERVED! A bright light phenomenon was observed in the Eastern Finland on April 21. I don't know if there were **satellite launches** in Plesetsk Cosmodrome near Arkhangelsk, but this may be a **rocket** experiment too.

Model-based Clustering

Model-based Agglomerative clustering:
The key of agglomerative clustering is to measure the distance of two clusters, which is further reduced to the calculation of pairwise document distance (KL-Divergence).

Model-based Partitional Clustering:
It assumes that there are k parameterized models, one for each cluster. Basically, the algorithm iterates between a model re-estimation step and a sample re-assignment step.

Problem to Solve

The agglomerative hierarchical clustering perform poorly because the nearest neighbors of a document belong to different classes in many cases. According to their examination on the data, each class has a “core” vocabulary of words and remaining “general” words may have similar distributions on different classes. Thus, two documents from different classes may share many general words (e.g. stop words) and will be viewed similar in terms of vector cosine similarity. To solve this problem, we should “discount” general words and “emphasize” more importance on core words in a vector.

Problem with Existing Solution

Laplacian Smoothing: treat term equally

$$p(w|c_j)=\frac{1+c(w,c_j)}{|V|+\sum_w c(w,c_j)}$$

TF*IDF: can not discount general term properly for small document sets.

Solution: Context-sensitive Semantic Smoothing

Solution Highlight:

- ♦ How to Define Context?
- ♦ How to Extract Contextual Information?
- ♦ How to Estimate Context-sensitive Translation Probability?
- ♦ How to Incorporate Context-sensitive Semantic Smoothing?

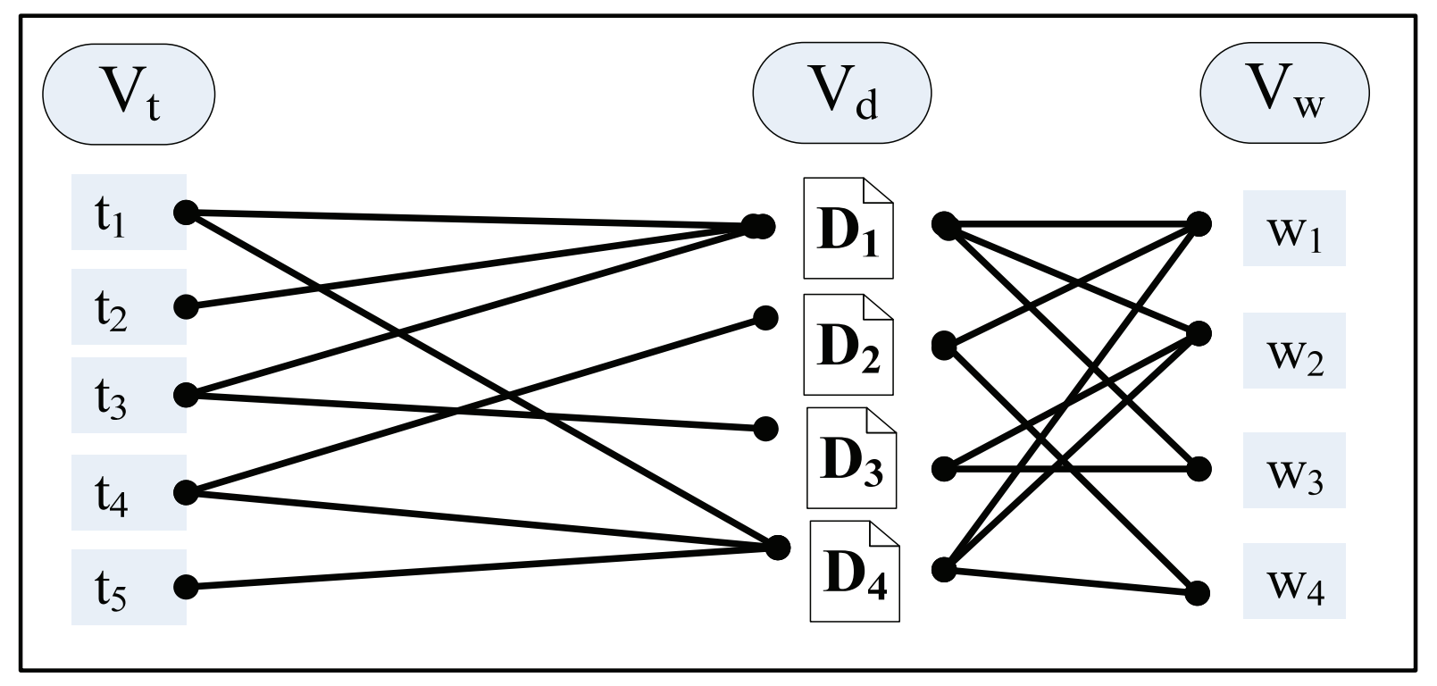


Illustration of document indexing. V_t, V_d, and V_w are phrase set , document set and word set, respectively .

Phrase as Context

The semantics of a phrase is clear and explicit since a multiword phrase is unambiguous in most cases. Thus, the translation of phrases to individual terms will be very specific.

Context Sensitive Translation Probability Estimates

All terms in the document set are either translated by the given topic signature model or generated by the background collection model, we have:

$$p(w|\theta_{t_k},C)=(1-\beta)p(w|\theta_{t_k})+\beta p(w|C)$$

The probability of translating a topic signature t_k to a concept w can be estimated by the Expectation Maximum (EM) algorithm with the following update formulas:

$$\hat{p}^{(n)}(w)=\frac{(1-\alpha)p^{(n)}(w|\theta_{t_k})}{(1-\alpha)p^{(n)}(w|\theta_{t_k})+\alpha p(w|C)}$$

$$p^{(n+1)}(w|\theta_{t_k})=\frac{c(w,D_k)\hat{p}^{(n)}(w)}{\sum_i c(w_i,D_k)\hat{p}^{(n)}(w_i)}$$

Where: D_k is the set of documents containing phrase t_k; $c(w, D_k)$ is the frequency count of word w in D_k; A is the background noise coefficient; C denotes the background model.

Translation probability of phrase “space program” to words

Space:
**Space 0.245; shuttle 0.057; launch 0.053; flight 0.042; air 0.035;
Program 0.031; center 0.030; administration 0.026; develop 0.025;
Like 0.023; look 0.022; world 0.020; director 0.020; plan 0.018;
Release 0.017; problem 0.017; work 0.016; place 0.016; mile 0.015
Base 0.014;**

Program:
**Program 0.193; Washington 0.026; congress 0.026;
Administration 0.024; need 0.024; billion 0.023; develop 0.023;
Bush 0.020; plan 0.020;money 0.020; problem 0.020;
Provide 0.020; writer 0.018; d 0.018; help 0.018; work 0.017;
President 0.017; house .017; million 0.016; increase 0.016;**

Space Program
Space 0.101; program 0.071; NASA 0.048; shuttle 0.043; astronaut 0.041; launch 0.040; mission 0.038; flight 0.037; earth 0.037; moon 0.035; orbit 0.032; satellite 0.031; Mar 0.030; explorer 0.028; station 0.028; rocket 0.027; technology 0.026; project 0.025; science 0.023; budget 0.023;

Document Model Smoothing

Agglomerative Clustering

♦ **Document model estimation**

$$p_{bt}(w|d)=(1-\lambda)p_b(w|d)+\lambda p_t(w|d)$$

♦ **KL-divergence distance**

$$\Delta(d_1,d_2)\equiv\sum_{w\in V}p(w|d_1)\log\frac{p(w|d_1)}{p(w|d_2)}$$

♦ **Simple language model**

$$p_b(w|d)=(1-\alpha)p_{ml}(w|d)+\alpha p(w|C)$$

♦ **The translation model smoothes document models by statistically mapping context-sensitive phrase into individual terms**

$$p_{ml}(t_k|d)=\frac{c(t_k,d)}{\sum_i c(t_i,d)} \qquad p_t(w|d)=\sum_k p(w|t_k)p_{ml}(t_k|d)$$

Partitional Clustering

♦ **k parameterized models, one for each cluster**

$$p(w|c_j)=(1-\lambda)p_b(w|c_j)+\lambda p_t(w|c_j)$$

$$p_b(w|c_j)=(1-\alpha)p_{ml}(w|c_j)+\alpha p(w|C)$$

$$p_t(w|c_j)=\sum_k p(w|t_k)p(t_k|c_j)$$

Evaluation

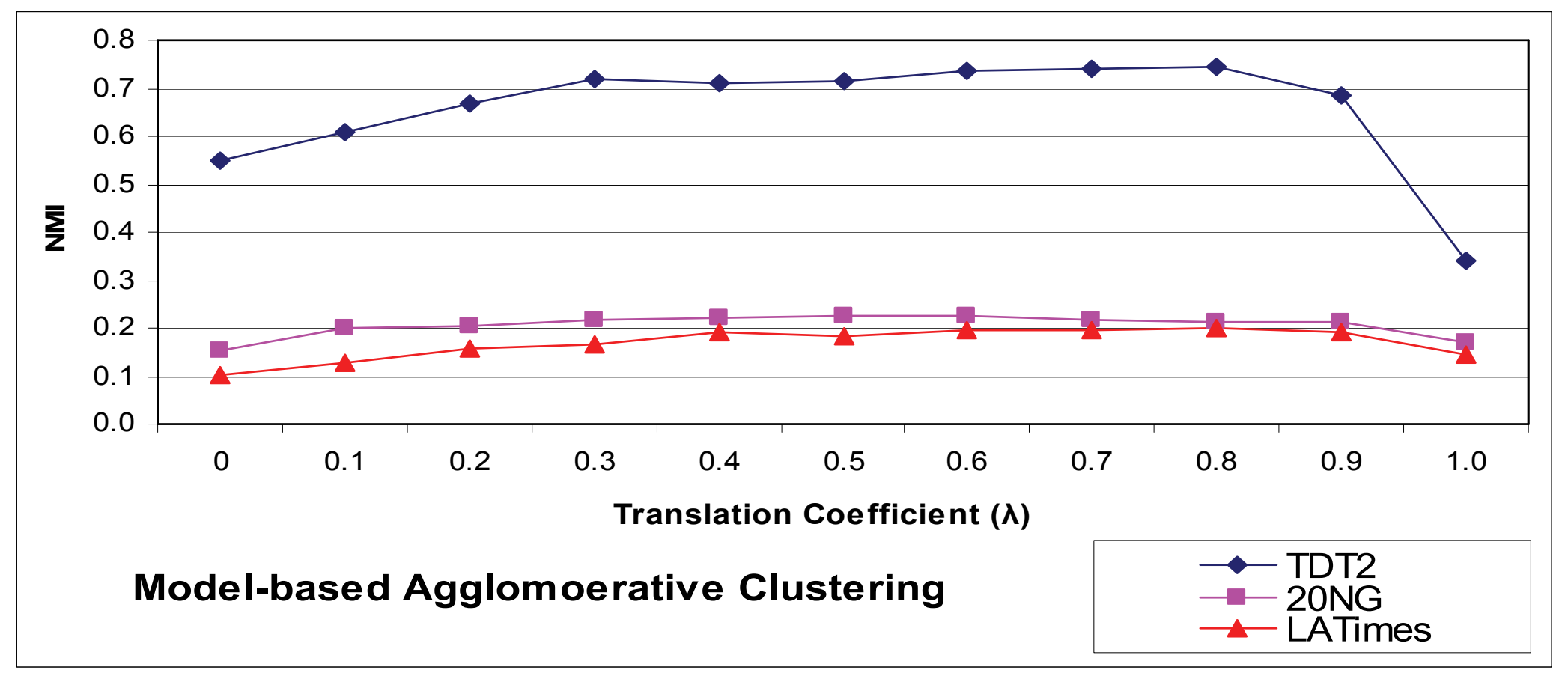
Evaluation Highlight:

- ♦ Testing Collections? 20NG, TDT2, LA Times
- ♦ Evaluation Measure?
  Normalized Mutual Information (NMI), Purity
- ♦ Evaluation Logic?
  Context-sensitive Model vs. Baseline Language Model
  Context-sensitive Model vs. Vector Cosine Model

Agglomerative clustering (small dataset). The comparison of the baseline language model(Bkg) to document smoothing model (Semantic) and vector cosine model(TF,TF-IDF). The λ parameter control the percentage of semantic smoothing.

Dataset (NMI)	Vector Cosine		KL-Divergence	
	TF	TF-IDF	Bkg	Semantic
TDT2	0.369	0.484	0.550	**0.743(λ=0.8)**
20NG	0.135	0.135	0.155	**0.227(λ=0.6)**
LA Times	0.059	0.054	0.104	**0.202(λ=0.8)**

Dataset (Purity)	Vector Cosine		KL-Divergence	
	TF	TF-IDF	Bkg	Semantic
TDT2	0.446	0.54	0.606	**0.826(λ=0.8)**
20NG	0.095	0.103	0.158	**0.236(λ=0.5)**
LA Times	0.128	0.127	0.227	**0.333(λ=0.8)**

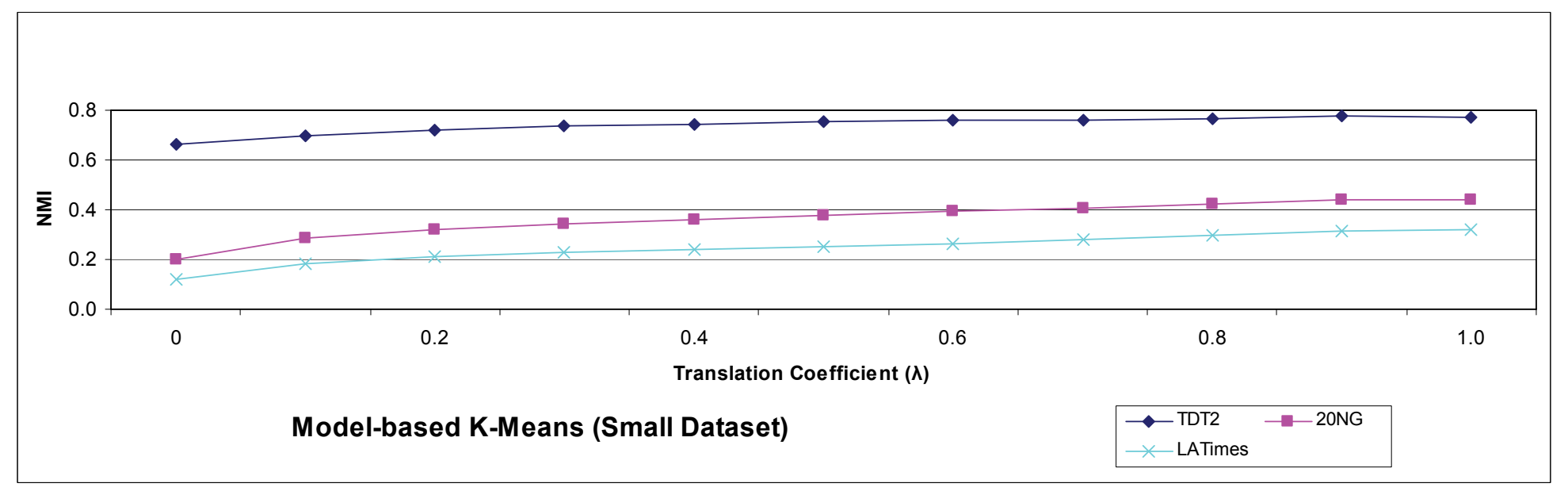


Model-based Agglomoerative Clustering

Partitional clustering (small dataset)

Lap: Laplacian

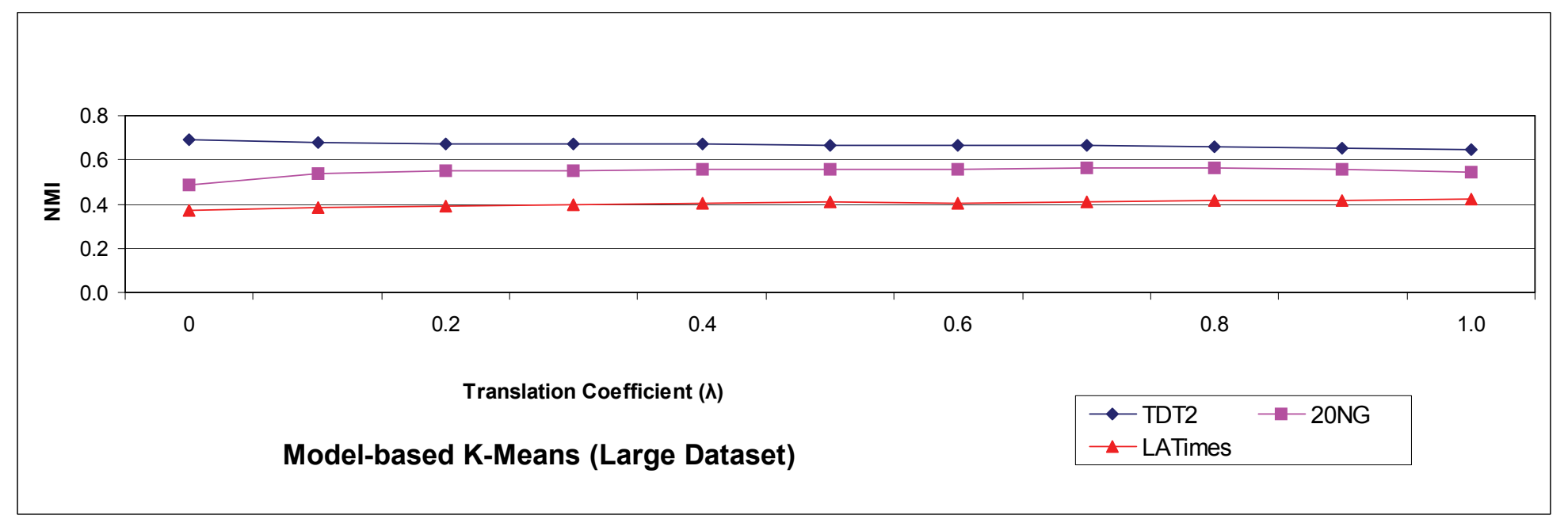
Dataset(NMI)	NTF	TF-IDF	Lap	Bkg	Semantic	λ
TDT2	0.791	**0.794**	0.651	0.665	0.774	1.0
20NG	0.176	0.391	0.240	0.201	**0.441**	1.0
LATimes	0.200	0.185	0.145	0.122	**0.322**	1.0



Model-based K-Means (Small Dataset)

Partitional clustering (large dataset)

Dataset	NTF	TF-IDF	Lap	Bkg	Max. Semantic	Min. Semantic
TDT2	0.702	**0.715**	0.684	0.689	0.678	0.649
20NG	0.192	0.506	0.493	0.489	**0.564**	0.536
LATimes	0.201	0.349	0.382	0.371	**0.420**	0.383



Model-based K-Means (Large Dataset)

Conclusions

Findings From the Experiment:

- ♦ **Semantic smoothing is much more effective than other schemes on agglomerative clustering where data sparsity is the major problem.**
- ♦ **The effectiveness of semantic smoothing with partitional clustering depends on the size of the dataset. When dataset is small and data sparsity is the major problem, semantic smoothing is very effective; otherwise, it equals to background smoothing.**

Contributions of This Paper:

- ♦ **Propose a new framework studying the relationship between model smoothing methods and clustering quality and successfully incorporate context-sensitive semantic smoothing to model based document clustering.**
- ♦ **Empirically prove the effectiveness of context-sensitive semantic smoothing for document clustering.**

References

*** Two papers based on this research project have been published in two top AI conferences, IJCAI 2007 and ICDM 2006, respectively**

[1]Berger, A. and Lafferty J., “Information Retrieval as Statistical Translation”, SIGIR’99, 222-229.
[2]Lafferty, J. and Zhai, C., “Document Language Models, Query Models, and Risk Minimization for Information Retrieval”, SIGIR’01, 111-119
[3]Xiaodan Zhang, Xiaohua Zhou, Xiaohua Hu: Semantic Smoothing for Model-based Document Clustering. ICDM 2006: 1193-1198
[4]Xiaohua Zhou, Xiaodan Zhang, Xiaohua Hu: Semantic Smoothing of Document Models for Agglomerative Clustering. IJCAI 2007: 2928-2933