Semantic Smoothing of Document Models for Agglomerative Clustering

Xiaohua Zhou, Xiaodan Zhang, Xiaohua Hu
Drexel University
College of Information Science & Technology
xiaohua.zhou@drexel.edu, xzhang@ischool.drexel.edu, thu@ischool.drexel.edu

Abstract

In this paper, we argue that the agglomerative clustering with vector cosine similarity measure performs poorly due to two reasons. First, the nearest neighbors of a document belong to different classes in many cases since any pair of documents shares lots of "general" words. Second, the sparsity of class-specific "core" words leads to grouping documents with the same class labels into different clusters. Both problems can be resolved by suitable smoothing of document model and using Kullback-Leibler divergence of two smoothed models as pairwise document distances. Inspired by the recent work in information retrieval, we propose a novel context-sensitive semantic smoothing method that can automatically identifies multiword phrases in a document and then statistically map phrases to individual document terms. We evaluate the new model-based similarity measure on three datasets using complete linkage criterion for agglomerative clustering and find out it significantly improves the clustering quality over the traditional vector cosine measure.

1 Introduction

Document clustering algorithms can be categorized into *agglomerative* and *partitional* approaches according to the underlying clustering strategy (Kaufman and Rousseeuw, 1990). The agglomerative approaches initially assign each document into its own cluster and repeatedly merge pairs of most similar clusters until only one cluster is left. The partitional approaches iteratively re-estimate the cluster model (or the cluster centroid) and reassign each document into the closest cluster until no document is moved any longer. In comparison with partitional approaches, the agglomerative approach does not need initialization and gives very intuitive explanation of why a set of documents are grouped together. However, it suffers from the $O(n^2)$ clustering time and performs poorly in general in terms of cluster quality (Steinbach et al., 2000). In this paper, we will analyze the underlying reasons of its poor performance and propose a solution.

Steinbach et al. (2000) argue that 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. Besides, we think the poor performance of the agglomerative clustering can also be attributed to the sparsity of core words in a document. A document is often short and contains very few number of core words. Thus, two documents from the same class may share few core words and be falsely grouped into different clusters when using vector cosine similarity metric. To solve this problem, we should assign reasonable positive counts to "unseen" core words if its related topical terms occur in the document.

Discounting seen words and assigning reasonable counts to unseen words are two exact goals of the probabilistic language model smoothing. In this paper, we view the calculation of pairwise document similarity as a process of document model smoothing and comparison. As usual, we use the *Kullback-Leibler divergence* distance function to measure the difference of two models (i.e. word probability distributions). So the problem is reduced to obtaining a good smoothed language model for each document in the corpus. The language modeling approach to information retrieval (IR) has been received much attention in recent years due to its mathematical foundation and empirical effectiveness. In a nugget, the language modeling approach to IR is to smooth document models (Lafferty and Zhai, 2001). To the best of our knowledge, the document model smoothing has not been studied in the context of agglomerative clustering. In this paper, we adapt the existing smoothing methods used in language modeling IR to the context of agglomerative clustering and hypothesize that document model smoothing can significantly improve the quality of the agglomerative hierarchical clustering.

In IR, a simple but effective smoothing strategy is to interpolate document models with a background collection model. For example, Jelinek-Mercer, Dirichlet, Absolute discount (Zhai and Lafferty, 2001) and Two-stage smoothing (Zhai and Lafferty, 2002) are all based on this strategy. In document clustering, TF-IDF score is often used as the dimension values of document vectors. The effect of TF-IDF scheme is roughly equivalent to the background model smoothing. However, a potentially more significant and effective smoothing method is what may be referred to as semantic smoothing where context and sense information are incorporated into the model (Lafferty and Zhai, 2001). The first trial of semantic smoothing may be dated back to latent semantic indexing (LSI, Deerwester et al., 1990) which projects documents in corpus into a reduced space where document semantics becomes clear. LSI explores the structure of term co-occurrence and can solve synonymy. However, it brings noise while reducing the dimensionality because it is unable to recognize the polysemy of a same term in different contexts. In practice, it is also criticized for the lack of scalability and interpretability.

Berger and Lafferty (1999) proposed a kind of semantic smoothing approach referred to as the statistical translation language model which statistically mapped document terms onto query terms.

$$p(q \mid d) = \sum_{w} t(q \mid w) p(w \mid d) \tag{1}$$

where $t(q \mid w)$ is the probability of translating the document term w to the query term q and $p(w \mid d)$ is the maximum likelihood estimator of the document model. With term translations, a document containing “star” may be returned for the query “movie”. Likewise, a document with the dimension of “star” but not “movie” may be merged into a cluster of “entertainment” together with a document containing “movie” but not “star”. However, like the LSI, this approach also suffers from the context-insensitivity problem, i.e., unable to incorporate contextual information into the model. Thus, the resulting translation may be fairly general and contain mixed topics. For example, “star” can be either from the class of “entertainment” (movie star) or from the class of “military” (star war).

Unlike Berger and Lafferty (1999) who estimated word translation probabilities purely based on word distributions in a corpus, Cao et al. (2005) constrained word relationships with human knowledge (i.e. relationships defined in WordNet) in order to reduce noise. They further combined linearly such a semantic-constrained translation model with a smoothed unigram document model. However, their model still did not solve the context-insensitivity problem in essence.

Compound terms often play an important role for a machine to understand the meaning of texts because they usually have constant and unambiguous meanings. Bai et al. (2005) adopted compound terms for text classification. However, compound terms are not used for smoothing purpose in their work. Instead, compound terms are directly working as features in conjunction with single-word features. In our previous work (Zhou et al., 2006), we proposed a context-sensitive semantic smoothing method for language modeling IR. The method decomposes a document into a set of weighted context-sensitive topic signatures and then statistically maps topic signatures into individual terms; a topic signature is defined as a pair of two concepts which are semantically and syntactically related to each other. The topic signature is similar to a compound term in the sense that both have constant and unambiguous meanings in most cases. For instance, if “star” and “movie” forms a topic signature, its context may be highly related to “entertainment”, but rarely to “war”. The extraction of concepts and concept pairs, however, relies on domain ontology, which is impractical for many public domains.

To overcome this limitation, we propose the use of multiword phrases (e.g. “star war”, “movie star”) as topic signatures in this paper. Same as a concept pair, a multiword phrase is often unambiguous. Furthermore, multiword phrases can be extracted from a corpus by existing statistical approaches without human knowledge. Last, documents are often full of multiword phrases; thus, it is robust to smooth a document model through statistical translation of multiword phrases in a document to individual terms.

We evaluate the new model-based document similarity metric on three datasets using agglomerative clustering with complete linkage criterion (Kaufman and Rousseeuw, 1990). The experiment results show that the KL-divergence similarity metric performs consistently better than the vector cosine metric. Moreover, the KL-divergence metric with semantic smoothing significantly outperforms with simple background smoothing. The result of the agglomerative clustering with semantic smoothing is comparable to that of the K-Means partitional clustering on three testing datasets.

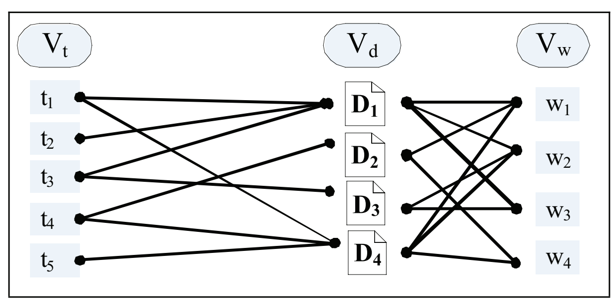

Figure 1. Illustration of document indexing. VB_{tB}, VB_{dB} and VB_{wB} are phrase set, document set and word set, respectively.

2 Document Model Smoothing

2.1 Semantic Smoothing of Document Model

Suppose we have indexed all documents in a given collection C with terms (individual words) and topic signatures (multiword phrases) as illustrated in Figure 1.

The translation probabilities from a topic signature t_k to any individual term w, denoted as $p(w|t_k)$, are also given. Then we can easily obtain a document model below:

$$p_t(w \mid d) = \sum_k p(w \mid t_k) p_{ml}(t_k \mid d) \qquad (2)$$

The likelihood of a given document generating the topic signature t_k can be estimated with

$$p_{ml}(t_k \mid d) = \frac{c(t_k, d)}{\sum_i c(t_i, d)} \qquad (3)$$

where $c(t_i, d)$ is the frequency of the topic signature t_i in a given document d.

We refer to the above model as *translation model* after Berger and Lafferty's work (1999). As we discussed in the introduction, the translation from multiword phrase to individual term would be very specific. Thus, the translation model not only weakens the effect of "general" words, but also relieves the sparsity of class-specific "core" words. However, not all topics in a document can be expressed by topic signatures (i.e., multiword phrases). If only translation model is used, there will be serious information loss. A natural extension is to interpolate the translation model with a unigram language model below:

$$p_b(w \mid d) = (1 - \alpha) p_{ml}(w \mid d) + \alpha p(w \mid C) \qquad (4)$$

Here _ is a coefficient accounting for the background collection model $p(w \mid C)$ and $p_{ml}(w \mid d)$ is a maximum likelihood estimator. In the experiment, _ is set to 0.5. We refer to this unigram model as *simple language model* or *baseline language model*. We use Jelinek-Mercer smoothing on the purpose of further discounting "general" words.

The final document model for clustering use is described in equation (5). It is a mixture model with two components: *a simple language model* and *a translation model*.

$$p_{bt}(w \mid d) = (1 - \lambda) p_b(w \mid d) + \lambda p_t(w \mid d) \qquad (5)$$

The translation coefficient (_) is to control the influence of two components in the mixture model. With training data, the translation coefficient can be trained by optimizing the clustering quality.

2.2 Topic Signature Extraction and Translation

Zhou et al (2006) implemented topic signatures as concept pairs and developed an ontology-based approach to extract concepts and concept pairs from documents. However, for many domains, ontology is not available. For this reason, we propose the use of multiword phrases as topic signatures and employ *Xtract* (Smadja, 1993) to identify phrases in documents. *Xtract* is a kind of statistical extraction tool with some syntactic constraints. It is able to extract noun phrases frequently occurring in the corpus without any external knowledge. *Xtract* uses four parameters, strength (k_0), peak z-score (k_1), spread (U_0), and percentage frequency (T), to control the quantity and quality of the extracted phrases. In the experiment, the four parameters are set to 1, 1, 4, and 0.75, respectively.

Table 1. Examples of phrase-word translations. The three phrases are automatically extracted from the collection of 20-newsgroup by *Xtract*. We list the top 20 topical words for each phrase.

Arab Country		**Nuclear Power**		**Gay People**	
Term	Prob.	Term	Prob.	Term	Prob.
Arab	0.061	nuclear	0.084	gay	0.078
country	0.048	power	0.046	homosexual	0.052
Israel	0.046	plant	0.039	sexual	0.046
Jew	0.037	technology	0.030	church	0.027
Israeli	0.032	air	0.026	persecute	0.023
Jewish	0.024	fuel	0.025	friend	0.023
Palestine	0.020	fossil	0.025	Abolitionist	0.019
1948	0.018	reactor	0.024	parent	0.019
Syria	0.018	steam	0.024	Society	0.019
expel	0.016	contaminate	0.024	lesbian	0.019
terror	0.015	water	0.024	themselves	0.019
Iraq	0.015	cold	0.023	lover	0.018
Davidsson	0.014	Cool	0.022	lifestyle	0.018
War	0.013	Tower	0.022	emotion	0.018
homeland	0.013	industry	0.022	thier	0.018
Egypt	0.013	radioactive	0.020	repress	0.018
Zionist	0.013	boil	0.020	affirm	0.018
legitimism	0.012	site	0.019	Ministry	0.017
Kaufman	0.012	built	0.019	straight	0.017
rejoinder	0.012	temperature	0.018	preach	0.017

For each phrase t_k, we have a set of documents (D_k) containing that phrase. Intuitively, we can use this document set D_k to estimate the translation probabilities for t_k, i.e., determining the probability of translating the given phrase t_k to terms in the vocabulary. If all terms appearing in the document set center on the sub-topic represented by t_k, we can simply use the maximum likelihood estimator and the problem is as simple as term frequency counting. However, some terms address the issue of other sub-topics while some are background terms of the whole collection. We then use a mixture language model to remove noise. Assuming the set of documents containing t_k is generated by a mixture language model (i.e., all terms in the document set are either translated by the given topic signature model $p(w \mid \theta_{t_k})$ or generated by the background collection model $p(w \mid C)$), we have:

$$p(w \mid \theta_{t_k}, C) = (1 - \beta) p(w \mid \theta_{t_k}) + \beta p(w \mid C) \qquad (6)$$

where _ is a coefficient accounting for the background noise and θ_{t_k} denotes parameter set of translation probabilities for t_k. Under this mixture language model, the log likelihood of generating the document set D_k is:

$$\log p(D_k \mid \theta_{t_k}, C) = \sum_w c(w, D_k) \log p(w \mid \theta_{t_k}, C) \qquad (7)$$

where $c(w, D_k)$ is the document frequency of term w in D_k., i.e., the cooccurrence count of w and t_k in the whole collection. The translation model can be estimated using the EM algorithm (Dempster et al., 1977). The EM update formulas are:

$$\hat{p}^{(n)}(w) = \frac{(1-\beta)p^{(n)}(w \mid \theta_{t_k})}{(1-\beta)p^{(n)}(w \mid \theta_{t_k}) + \beta p(w \mid C)} \quad (8)$$

$$p^{(n+1)}(w \mid \theta_{t_k}) = \frac{c(w, D_k)\hat{p}^{(n)}(w)}{\sum_i c(w_i, D_k)\hat{p}^{(n)}(w_i)} \quad (9)$$

In the experiment, we set the background coefficient _=0.5. We also truncate terms with extremely small translation probabilities for two purposes. First, with smaller number of translation space, the document smoothing will be much more efficient. Second, we assume terms with extremely small probability are noise (i.e. not semantically related to the given topic signature). In detail, we disregard all terms with translation probability less than 0.001 and renormalize the translation probabilities of the remaining terms.

2.3 The KL-Divergence Distance Metric

After estimating a language model for each document in the corpus with context-sensitive semantic smoothing, we use the Kullback-Leibler divergence of two language models as the distance measure of the corresponding two documents. Given two probabilistic document models $p(w|d_1)$ and $p(w|d_2)$, the KL-divergence distance of $p(w|d_1)$ to $p(w|d_2)$ is defined as:

$$\Delta(d_1, d_2) = \sum_{w \in V} p(w \mid d_1) \log \frac{p(w \mid d_1)}{p(w \mid d_2)} \quad (10)$$

where V is the vocabulary of the corpus. KL-divergence distance will be a non-negative score. It gets the zero value if and only if two document models are exactly same. However, KL-divergence is not a symmetric metric. Thus, we define the distance of two documents as the minimum of two KL-divergence distances. That is,

$$dist(d_1, d_2) = \min\{\Delta(d_1, d_2),\ \Delta(d_2, d_1)\} \quad (11)$$

The calculation of KL-divergence involves scanning the vocabulary, which makes the solution computationally inefficient. To solve this problem, we truncate terms with its distribution probability less than 0.001 while estimating document model using the equation (5) and renormalize the probabilities of remaining terms. Because we keep terms with high probability values in document models, it makes almost no difference in clustering results.

3 Experiment Settings and Result Analysis

3.1 Evaluation Methodology

Cluster quality is evaluated by three extrinsic measures, *purity* (Zhao and Karypis, 2001), *entropy* (Steinbach et al., 2000) and *normalized mutual information* (NMI, Banerjee and Ghosh, 2002). Due to the space limit, we only list the result of NMI, an increasingly popular measure of cluster quality. The other two measures are consistent with NMI on all runs. NMI is defined as the mutual information between the cluster assignments and a pre-existing labeling of the dataset normalized by the arithmetic mean of the maximum possible entropies of the empirical marginals, i.e.,

$$NMI(X,Y) = \frac{I(X;Y)}{(\log k + \log c)/2} \quad (12)$$

where X is a random variable for cluster assignments, Y is a random variable for the pre-existing labels on the same data, k is the number of clusters, and c is the number of pre-existing classes. Regarding the details of computing I(X; Y), please refer to (Banerjee and Ghosh, 2002). NMI ranges from 0 to 1. The bigger the NMI is the higher quality the clustering is. NMI is better than other common extrinsic measures such as purity and entropy in the sense that it does not necessarily increase when the number of clusters increases.

We take complete linkage criterion for agglomerative hierarchical clustering. The two document similarity metrics are the traditional vector cosine and the Kullback-Leibler divergence proposed in this paper. For cosine similarity, we try three different vector representations: term frequency (tf), normalized term frequency (i.e., tf divided by the vector length), and TF-IDF. For KL-divergence metric, we use document models with semantic smoothing as described in equation (5) and test 11 translation coefficients (_) ranging from 0 to 1. When _=0, it actually uses simple background smoothing.

In order to compare with the partitional approach, we also implement a basic K-Means using cosine similarity metric on three vector representations (TF, NTF, and TF-IDF). The calculation of the cluster centroid uses the following formula:

$$centroid = \frac{1}{|C|}\sum_{d \in C} d \quad (13)$$

where C is the corpus. Since the result of K-Means clustering varies with the initialization. We run ten times with random initialization and average the results. For various vector representations, each run has the same initialization.

3.2 Datasets

We conduct clustering experiments on three datasets: TDT2, LA Times (from TREC), and 20-newsgroups (20NG). The TDT2 corpus has 100 document classes, each of which reports a major news event. LA Times news are labeled with 21 unique section names, e.g., Financial, Entertainment, Sports, etc. 20-Newsgroups dataset is collected from 20 different Usenet newsgroups, 1,000 articles from each.

We index 9,638 documents in TDT2 that have a unique class label, 21,623 documents from top ten sections of LA Times, and all 19,997 documents in 20-newsgroups. For each document, we index its title and body content with both multiword phrases and individual words, and ignore other sections including Meta data. A list of 390 stop words is used. In the testing stage, 100 documents are randomly picked from each class of a given dataset and merged into a big pool for clustering. For each dataset, we create five such random pools and average the experimental results. The ten classes selected from TDT2 are *20001, 20015, 20002, 20013, 20070, 20044, 20076, 20071, 20012, and 20023*. The ten sections selected from LA Times are *Entertainment, Financial, Foreign, Late Final, Letters, Metro, National, Sports, Calendar, and View*. All 20 classes of 20NG are selected for testing.

Table 2. Statistics of three datasets

Dataset Name	TDT2	LA Times	20NG
# of indexed docs	9,638	21,623	19,997
# of words	37,994	63,510	140,269
# of phrases	8,256	9,517	10,902
Avg. doc length (word)	240	103	193
Avg. doc length (phrase)	21	10	10
# of classes	10	10	20

3.3 Experiment Results and Analysis

The *Dragon Toolkit* [Zhou et al., 2006] is used to conduct clustering experiments. The translation coefficient (_) in equation (5) is trained over TDT2 dataset by maximizing the NMI of clustering. The optimal value _=0.8 is then applied to two other datasets. The NMI result of the agglomerative hierarchical clustering with complete linkage criterion is listed in Table 3. When the vector cosine measure is used as pairwise document similarity, the TF-IDF scheme performs slightly better than the TF scheme. As we discussed before, the heuristic TF-IDF weighting scheme can discount "general" words and strengthen "specific" words in a document vector. Thus, it can improve the agglomerative clustering quality. The KL-divergence similarity measure with background smoothing of document models (i.e., _=0) consistently outperforms the cosine measure on both TF and TF-IDF schemes. As expected, the KL-divergence measure with context-sensitive semantic smoothing significantly improves the quality of the agglomerative clustering on all three datasets. After semantic smoothing, the class-independent general words will be dramatically weakened and the class-specific "core" words will be strengthened even if it does not appear in the document at all. Thus, the distance of intra-class documents will be decreased while the distance of inter-documents will be increased, and hence improve the clustering quality.

Table 3. NMI results of the agglomerative hierarchical clustering with complete linkage criterion

Dataset	Cosine		KL-Divergence	
	TF/NTF	TF-IDF	background	Semantic
TDT2	0.369	0.484	0.550	**0.743**
LA Times	0.059	0.054	0.104	**0.202**
20NG	0.135	0.135	0.155	**0.216**

To see the robustness of the semantic smoothing method, we show the performance curve in Figure 2. Except for the point of _=1, the semantic smoothing always improve the cluster quality over the simple background smoothing. In general, NMI will increase with the increase of translation coefficient till the peak point (around 0.8 in our case) and then go downward. In our experiment, we only consider phrases appearing in more than 10 documents as topic signatures in order to obtain a good estimate of translation probabilities. Moreover, not all topics in a document can be expressed by multiword phrases. Thus, the phrase-based semantic smoothing will cause information loss. We interpolate the translation model with a unigram language model to make up the loss. Now it is easy to understand why the NMI goes downward when the influence of the semantic smoothing is too high. Actually, LSI also causes information loss when the dimensionality reduction is too aggressive; but there is no mechanism to recover the loss. In this sense, the semantic smoothing approach is more flexible than LSI.

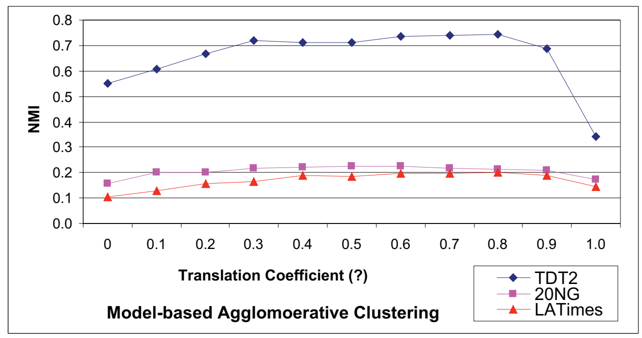

Figure 2. The variance of the cluster quality with the translation coefficient (_) which controls the influence of semantic smoothing

Steinbach et al. (2000) reported that K-Means performed as good as or better than agglomerative approaches. Our experiment also repeated this finding. Using vector cosine similarity measure, the complete linkage algorithm performs significantly worse than the K-Means (see table 3 and 4). However, with semantic smoothing of document models, the result of complete linkage clustering is comparable to that of K-Means on three representation schemes (TF, Norm TF, and TF-IDF). This is also a kind of indication that semantic smoothing of document models is very effective in improving agglomerative clustering approaches.

Table4. NMI results of the regular K-Means clustering. K is the number of true classes listed in table 2.

Dataset	TF	NTF	TF-IDF
TDT2	0.792	**0.805**	0.790
LA Times	0.194	**0.197**	0.166
20NG	0.197	0.161	**0.374**

4 Conclusions and Future Work

The quality of agglomerative hierarchical clustering often highly depends on pairwise document similarity measures. The density of class-independent "general" words and the sparsity of class-specific "core" words in documents make the traditional vector cosine a poor similarity measure for agglomerative clustering. To solve this problem, we develop a context-sensitive semantic smoothing method to “smooth” document models, i.e. discounting seen “general” words and assigning reasonable positive counts to unseen “core” words in a document, and further use Kullback-Leibler divergence of smoothed probabilistic models as the document similarity measure for clustering. The clustering experiments on three different datasets show that the combination of semantic smoothing and the KL-divergence similarity measure can significantly improve agglomerative hierarchical clustering.

Our semantic smoothing approach uses unambiguous multiword phrases as topic signature and statistically maps phrases onto individual terms. Conceptually, a phrase here corresponds to a latent dimension in latent semantic indexing (LSI). However, the semantics of phrase is explicit and clear. Because multiword phrases are unambiguous in most cases, its translation to individual terms is very specific whereas LSI brings noise when exploring latent semantic structures due to term polysemy. LSI also causes information loss during dimensionality reduction. Our approach can recover the information loss by interpolating the phrase translation model with a smoothed unigram language model. But how to obtain optimal weights for each component in the mixture model will be an open problem. In this paper, we empirically tuned a fixed translation coefficient to optimize the clustering results. Ideally, this coefficient should be optimized over each document. In addition, we use natural language processing techniques to identify sub-topics (phrases) from texts. This is somehow ad hoc nature and could be improved in future.

Recent advances in document clustering have shown that model-based partitional approaches are more efficient and effective than similarity-based approaches in general (Zhong and Ghosh, 2005). However, most generative models simply use Laplacian smoothing to smooth the cluster models on the purpose of avoiding zero probability. For future work, we will apply context-sensitive semantic smoothing to model-based partitional approaches, which may further improve their clustering quality.

Acknowledgment

This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667). We also thank three anonymous reviewers for their instructive comments on the paper.

References

[Bai et al., 2005] Bai, J., Nie, J.-Y., and Cao, G. Integrating Compound Terms in Bayesian Text Classification, *Web Intelligence 2005*, France.

[Banerjee and Ghosh, 2002] Banerjee, A. and Ghosh, J. Frequency sensitive competitive learning for clustering on high-dimensional hperspheres. *Proc. IEEE Int. Joint Conference on Neural Networks*, pp. 1590-1595.

[Berger and Lafferty, 1999] Berger, A. and Lafferty J. Information Retrieval as Statistical Translation. *ACM SIGIR 1999*, pp. 222-229.

[Cao et al., 2005] Cao, G., Nie, J.-Y., and Bai, J. Integrating Word Relationships into Language Models, *ACM SIGIR 2005*, pp. 298-305.

[Deerwester et al., 1990] Deerwester, S., Dumais, T.S., Furnas, W.G., Landauer, K.T., and Harshman, R. Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science*, 1990, 41(6): 391- 407

[Dempster et al., 1977] Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977, 39: 1-38.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseuw, P.J. *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley and Sons, 1990.

[Lafferty and Zhai, 2001] Lafferty, J. and Zhai, C. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of the 24th ACM SIGIR Conference on Research and* Development in IR, 2001, pp.111-119.

[Smadja, 1993] Smadja, F. Retrieving collocations from text: Xtract. *Computational Linguistics*, 1993, 19(1), pp. 143--177.

[Steinbach et al., 2000] Steinbach, M., Karypis, G., and Kumar, V. *A Comparison of document clustering techniques*. Technical Report #00-034, Department of Computer Science and Engineering, University of Minnesota, 2000.

[Zhai and Lafferty, 2002] Zhai, C. and Lafferty, J. Two-Stage Language Models for Information Retrieval. *ACM SIGIR 2002*, pp. 49-56

[Zhai and Lafferty, 2001] Zhai, C. and Lafferty, J. A Study of Smoothing Methods for Language Models Applied to Ad hoc Information Retrieval. *ACM SIGIR 2001*, pp. 334-342.

[Zhao and Karypis, 2001] Zhao, Y. and Karypis, G. *Criterion functions for document clustering: experiments and analysis*, Technical Report, Department of Computer Science, University of Minnesota, 2001.

[Zhong and Ghosh, 2005] Zhong, S. and Ghosh, J. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3): 374-384, 2005.

[Zhou et al., 2006] Zhou, X., Hu, X., Zhang, X., Lin, X., and Song, I.-Y.. Context-Sensitive Semantic Smoothing for the Language Modeling Approach to Genomic IR. *ACM SIGIR 2006*, Aug 6-11, 2006, pp. 170-177.

[Zhou et al., 2006] Zhou, X., Zhang, X., and Hu, X. *The Dragon Toolkit*, Data Mining and Bioinformatics Lab (DMBio), iSchool at Drexel University, PA 19104, USA http://www.ischool.drexel.edu/dmbio/dragontool