Statistical Translation, Heat Kernels, and Expected Distances

Joshua Dillon
School of Elec. and Computer Engineering
Purdue University - West Lafayette, IN
jvdillon@ecn.purdue.edu

Yi Mao
School of Elec. and Computer Engineering
Purdue University - West Lafayette, IN
ymao@ecn.purdue.edu

Guy Lebanon
Department of Statistics, and
School of Elec. and Computer Engineering
Purdue University - West Lafayette, IN
lebanon@stat.purdue.edu

Jian Zhang
Department of Statistics
Purdue University - West Lafayette, IN
jianzhan@stat.purdue.edu

Abstract

High dimensional structured data such as text and images is often poorly understood and misrepresented in statistical modeling. The standard histogram representation suffers from high variance and performs poorly in general. We explore novel connections between statistical translation, heat kernels on manifolds and graphs, and expected distances. These connections provide a new framework for unsupervised metric learning for text documents. Experiments indicate that the resulting distances are generally superior to their more standard counterparts.

1 Introduction

Modeling text documents is an essential part in a wide variety of applications such as classification, segmentation, visualization and retrieval of text. Most approaches start by representing a document by its word count or histogram. They then proceed to fit statistical models or compute distances based on that word histogram content. Representing a document by its word frequency may be motivated by the assumption that the appearance of words in documents follows a multinomial distribution. The word histogram is then the maximum likelihood estimator of the multinomial parameter and may be used in lieu of the word sequence.

A statistical analysis of word frequency representation of documents requires the assumption of a geometry on the simplex - the closure of the space of all document histogram representations (below, and elsewhere we assume that the vocabulary $V = \{1, \ldots, m+1\}$)

$$\overline{\mathbb{P}}_m = \left\{ \theta \in \mathbb{R}^{m+1} : \forall i \ \theta_i \geq 0, \sum_j \theta_j = 1 \right\}.$$

The geometric assumption may be explicit, as is the case in nearest neighbor classifiers. In other cases, such as logistic regression and boosting, it is made implicitly.

Many geometries have been suggested for the simplex for use in modeling word histogram document representation. Two canonical examples are the Euclidean distance and the Fisher geodesic distance

$$d(\theta, \eta) = \arccos\left(\sum_{i=1}^{m+1} \sqrt{\theta_i \eta_i} \right) \qquad \theta, \eta \in \overline{\mathbb{P}}_m \tag{1}$$

which is based on the Fisher information Riemannian metric [12, 1]. Both distance functions, as well as other similarity measures (such as tfidf cosine similarity), suffer from the slow convergence rate of high dimensional histograms to their expectations. In other words, the word histogram of a document containing several dozen words serves as a poor estimate for the multinomial parameter whose dimensionality is much higher (corresponding to the vocabulary size). As a result, standard distances are not well suited for use in the analysis of word frequency representation of documents.

Following a discussion of related work in Section 2 we present the translation model and its related expected geometry. We conclude with some experimental results and a discussion.

2 Related Work

Distributional clustering [18] was first introduced to cluster words (such as nouns) according to their distributions in syntactic contexts (such as verbs). The model is estimated by minimizing the free energy, and is used to address the data sparseness problem in language modeling. It also serves as an aggressive feature selection method in [2] to improve document classification accuracy. Unlike previous work, we use word contextual information for constructing a word translation model rather than clustering words.

Our method of using word translation to compute document similarity is also closely related to query expansion in information retrieval. Early work [20] used word clusters from a word similarity matrix for query expansion. A random walk model on a bipartite graph of query words and documents was introduced in [15], and was later generalized to a more flexible family of random walk models [9]. Noisy channels were originally used in communication for data transmission, but served as a platform for considerable research in statistical machine translation. An interesting work by Berger and Lafferty [4] formulated a probabilistic approach to information retrieval based upon the ideas and methods of statistical machine translation.

The idea of diffusion or heat kernel $\exp(-t\mathcal{L})$ based on the normalized graph Laplacian $\mathcal{L}$ [8] has been studied for discrete input space such as graphs [13], and applied to classification problems with kernel-based learning methods and semi-supervised learning [22, 3]. It has also been formally connected to regularization operators on graphs [19], and can be thought as a smoothing operator over graphs. In our case the diffusion kernel is used to generate a stochastic matrix which is then used to define a translation model between words. This has the effect of translating one word to its semantic neighbor connected by similar contextual information.

Several methods have been proposed to learn a better metric for classification. In [21] a new metric is learned using side-information such as which pairs of examples are similar and dissimilar, and the task is formulated as a convex optimization problem. In [11] a quadratic Gaussian metric is learned based on the idea that examples in the same class should be collapsed, and the solution can be found by convex optimization techniques. In [16] a new metric is estimated in an unsupervised manner based on the geometric notion of volume elements. A similar line of research develops new similarity measures and word clustering [5, 10] based on word co-occurrence information.

In most methods, a linear transformation of the original Euclidean metric is learned using labeled data with criteria such as better separability or prediction accuracy. Unlike those methods, our approach is based on word translation and expected distances and is totally unsupervised.

3 Translation Model

Given a word in the vocabulary $v \in V = \{1, \ldots, m+1\}$, we define its contextual distribution $q_v \in \overline{\mathbb{P}}_m$ to be $q_v(w) = p(w \in d | v \in d)$, where p is a generative model for the documents d. In other words, assuming that v occurs in the document, $q_v(w)$ is the probability that w also occurs in the document. Note that in particular, $q_v(v)$ measures the probability of a word v re-appearing a second time after its first appearance.

In general, we do not know the precise generative model and have to resort to an estimate such as

$$\hat{q}_v(w) \propto \sum_{d: v \in d} \mathrm{tf}(w, d)$$

where $\mathrm{tf}(w, d)$ is the relative (or normalized) frequency of occurrences of word w in document d. Note that the above estimate requires only unlabeled data and can leverage large archival databases to produce accurate estimates.

As pointed out by several researchers, the contextual distributions q_w, q_v convey important information concerning the words w, v. For example, similar distributions indicate a semantic similarity between the words. In this paper, we explore the geometric structure of the simplicial points $\{q_w : w \in V\}$ in order to define a statistical translation model that produces a better estimate of the document multinomial parameter. We describe below the translation model and conclude this section with a more formal motivation in terms of a translation based generative model.

3.1 Diffusion Kernel on $\{q_w : w \in V\}$

The key idea behind the translation model is that replacing occurring words with non-occurring but similar words in the document is likely to improve the original histogram estimate of the document's multinomial parameter. For example, we may stochastically translate the word `policeman` appearing in a certain document to the word `cop`. Despite the fact that the word `cop` was not generated initially, it is probably relevant to the document and the multinomial parameter $\theta_{\texttt{cop}}$ corresponding to it should be non-negligible. As a result of the above observation we wish to stochastically translate a document y into a new document z and represent the document as a histogram of z rather than of y. However, since the translation of y to z is probabilistic, we need to consider the standard geometric quantities such as distance as random variables leading to the notion of expected geometry.

We approximate the probability of translating word u into word v by the similarity between their contextual distributions q_u and q_v. A natural way to measure the similarity is through the heat or diffusion kernel on $\overline{\mathbb{P}}_m$. The particular choice of the Fisher geometry on $\overline{\mathbb{P}}_m$ is axiomatically motivated [7, 6] and the heat kernel has several unique properties characterizing it in a favorable way [14]. However, the Riemannian heat kernel is defined for the entire simplex $\overline{\mathbb{P}}_m$ which includes distributions that are irrelevant for modeling translations. The contextual distributions $\{q_w : w \in V\}$ which corresponds to the vocabulary words are our main objects of interest can be viewed as a graph embedded in the simplex. The natural restriction of the Riemannian heat kernel to the graph $G = (V, E)$ is the heat kernel on the graph whose edge weight $e(q_u, q_v) \in E$ is defined by the Riemannian heat kernel $K_t(q_u, q_v)$ on $\overline{\mathbb{P}}_m$. We elaborate on this below.

We construct an undirected weighted graph whose vertices correspond to word contextual distributions $\{q_w : w \in V\}$. Since the graph nodes are embedded in $\overline{\mathbb{P}}_m$, we define the graph edge weight connecting the two nodes u and v as the corresponding approximated Riemannian heat flow on $\overline{\mathbb{P}}_m$ [14]:

$$e(u, v) = \exp\left(-\frac{1}{\sigma^2}\arccos^2\left(\sum_w \sqrt{q_u(w)q_v(w)}\right)\right).$$

The graph heat kernel can then be computed via the matrix exponential of the normalized graph Laplacian [8]

$$\mathcal{L} = D^{-1/2}(D - E)D^{-1/2}$$

where D is a diagonal matrix with $D_{ii} = \sum_j e_{ij}$. Specifically the matrix exponential $T = \exp(-t\mathcal{L})$ where t describes the time of heat flow is viewed as the graph analog of the Riemannian heat kernel, and models the flow of heat across the graph. The normalized matrix corresponding to T is thus a stochastic matrix whose rows represent translation probabilities $p(w_i \to w_j)$ that correspond to flow of heat from q_{w_i} to q_{w_j} based on the geometry of the graph $\{q_w : w \in V\}$ and the Fisher geometry of the simplex $\overline{\mathbb{P}}_m$.

The time parameter t in $T = \exp(-t\mathcal{L})$ controls the amount of translation. Small t would yield $T \approx I$ while large t would yield an approximately uniform T. It is well known that the diffusion kernel is closely related to lazy random walks. In particular, the matrix T corresponds to lazy random walk after n steps with $n \to \infty$ [19]. As a result, the translation matrix T combines information from multiple paths of various lengths between any pair of contextual distributions.

3.2 A Generative Model

The above motivation may be expressed more formally in two ways. The first interpretation is to assume a multinomial model p_{θ_d} that generates each document. The representation problem becomes that of obtaining a good estimate for θ_d. The word histogram, while being unbiased of θ_d results in poor estimation performance in high dimension due to high variance. A biased estimator such as the translation model can improve performance by drastically reducing variance using an external data source such as an unlabeled corpus. This is in direct analogy with methods such as lasso and ridge in linear regression and in general the notion of regularization. One key difference between our approach and standard regularization techniques is that in our case the regularization is data dependent.

A second way to interpret the framework is to assume the following generative model. The observed documents are actually noisy versions of some original document where the noise is modeled via the heat flow on the graph $\{q_v : v \in V\}$ embedded in $\overline{\mathbb{P}}_m$. The task is to recover the original representation before the translation procedure "corrupted" the initial form. Here, the translation corresponds to denoising or filtering under complex non-iid noise.

In both cases, since the statistical translation stochastically converts a document y into a document z, we should consider quantities of interest $f(z)$ to be random objects motivating the use of expectation $E_{p(z|y)}f(z)$ or $E_{p(z|y)p(w|x)}f(z,w)$. This leads us to the notion of expected distances on word histogram where the expectation is taken with respect to the translation model. Alternatively, we can consider the expected distances as the solution of a metric learning problem. The expected distances based on the heat kernel form a new metric which is fitted to the data based on a large unsupervised corpus.

4 Expected Distances

As mentioned in the previous section, we are interested in computing expected distances instead of distances based on the observed word histogram. Below, we denote the histogram of a document $y = \langle y_1, \ldots, y_N \rangle$ as $\gamma(y)$ where $[\gamma(y)]_k = N^{-1} \sum_{i=1}^{N} \delta_{k,y_i}$.

As we show below the expected distance

$$d(\gamma(x), \gamma(w)) \stackrel{\text{def}}{=} E_{p(y|x)p(z|w)} d'(\gamma(y), \gamma(z))$$

has a closed form expression for $d'(p,q) = \|p-q\|_2^2$. In this case

$$\begin{aligned} d(\gamma(x), \gamma(w)) &= E_{p(y|x)p(z|w)} \|\gamma(y) - \gamma(z))\|_2^2 \\ &= E_{p(y|x)} \big\langle \gamma(y), \gamma(y) \big\rangle + E_{p(z|w)} \big\langle \gamma(z), \gamma(z) \big\rangle - 2\, E_{p(y|x)p(z|w)} \big\langle \gamma(y), \gamma(z) \big\rangle. \quad (2) \end{aligned}$$

The closed form expression for the expected distance can be obtained by substituting the expectations (below, T represents the heat kernel-based stochastic word to word translation matrix)

$$\begin{aligned} E_{p(y|x)p(z|w)} \big\langle \gamma(y), \gamma(z) \big\rangle &= N_1^{-1} N_2^{-1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{m+1} E_{p(y|x)p(z|w)} \delta_{k,y_i} \delta_{k,z_j} \\ &= N_1^{-1} N_2^{-1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{m+1} T_{x_i,k} T_{w_j,k} = N_1^{-1} N_2^{-2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} (TT^\top)_{x_i,w_j} \\ E_{p(y|x)} \big\langle \gamma(y), \gamma(y) \big\rangle &= N_1^{-2} \sum_{k=1}^{m+1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_1} E_{p(y|x)} \delta_{k,y_i} \delta_{k,y_j} \\ &= N_1^{-2} \sum_{i=1}^{N_1} \sum_{j \in \{1,\ldots,N_1\} \setminus \{i\}} (TT^\top)_{x_i,x_j} + N_1^{-2} \sum_{i=1}^{N_1} \sum_{k=1}^{m+1} T_{x_i,k} \\ &= N_1^{-2} \sum_{i=1}^{N_1} \sum_{j \in \{1,\ldots,N_1\} \setminus \{i\}} (TT^\top)_{x_i,x_j} + N_1^{-1}, \end{aligned}$$

in (2) to obtain

$$d(\gamma(x),\gamma(w)) = N_1^{-2}\sum_{i=1}^{N_1}\sum_{j\in\{1,\ldots,N_1\}\setminus\{i\}}(TT^\top)_{x_i,x_j} + N_2^{-2}\sum_{i=1}^{N_2}\sum_{j\in\{1,\ldots,N_2\}\setminus\{i\}}(TT^\top)_{w_i,w_j}$$
$$- 2N_1^{-1}N_2^{-2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(TT^\top)_{x_i,w_j} + N_1^{-1} + N_2^{-1}.$$

It is worth mentioning several facts concerning the above expression. If $T = I$ the above expected distance reduces to the standard Euclidean distance between the histogram representations. While the above equation is expressed using the sequential document contents, the expected distance remains the same under permutation of the words within a document since it is a pure bag of words construct. Finally, it is possible to pre-compute $TT^\top$ in order to speed up the distance computation. The next section contains some experimental results demonstrating the use of the expected distances in text classification.

5 Experimental Results

We experimented with the effect of using the expected L_2 distance vs. the L_2 distance in the context of nearest neighbor text classification using the Reuters RCV1 corpus. RCV1 has a considerable label hierarchy which generates many possible binary classification tasks. The effect of replacing L_2 with its expected version was positive in general, but not on every single labeling task. The top 1500 words (words that appear in most documents) were excluded from participating in the translation. This exclusion follows from the motivation of obtaining a more accurate estimate of low frequency terms. The most common terms already appear often and there is no need to translate from them to other words and vice versa. It is important to realize that this exclusion does not eliminate or downweight the frequent words like the tfidf representation in any way. It merely limits the translations between these words and other words.

Figure 1 (right) shows the test set error rate as a function of training set size for one specific labeling task, `C18` vs. `C31`. Figure 1 (left) shows the difference in error rate between the two methods as a function of which labeling task is being examined. In this case the labeling tasks on the x axis are ordered so the curve indicates that 70% of the labeling tasks gained improvement by using the expected distance. Note also how the amount of improvement is significantly higher than the amount of potential damage.

Figure 2 demonstrates the binary classification mean error rate of all-pairs of sub-categories. The sub-categories were taken directly from the RCV1 topic hierarchy, with the exception of `E01`, `C01`, `C02`, `G01`, and `G02`. These sub-categories were newly created parents for all leaves not currently members of any sub-category. The assignment of leaves to `C01`/`C02` and `G01`/`G02` was arbitrary. All binary classification experiments were averaged over 40 cross validations.

6 Discussion

The experimental results demonstrate how overall expected distances prove useful for text classification. It is likely that expected distances may be used in other areas of text analysis such as query expansion in information retrieval. Another interesting application that is worth exploring is using the expected distances in sequential visualization of documents, for example based on the lowbow framework [17].

The theoretic motivation behind the translation model as a probabilistic biased estimate of the document multinomial parameter should be further explored. A solid connection, if found, between translation based expected distances and variance reduction would be an interesting theoretical result. It would link theoretical results in estimation in high dimensions to practical information retrieval methods such as query expansion.

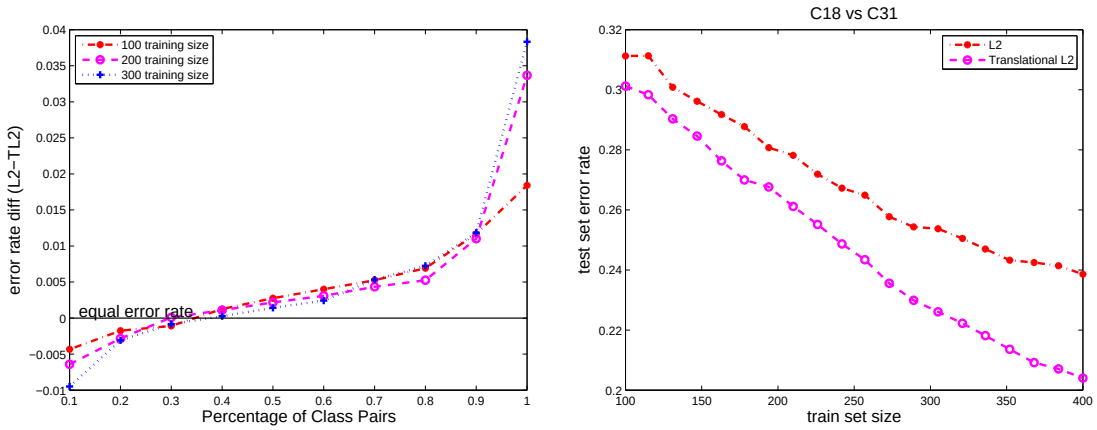

Figure 1: Left: Improvement using expected L_2 distance over L_2 distance for nearest neighbor classifier on RCV1 all pairs task as a function of the rank order of the labeling task. Right: Test set error rate as a function of train set size for one specific RCV1 labeling task `C18` vs. `C31`

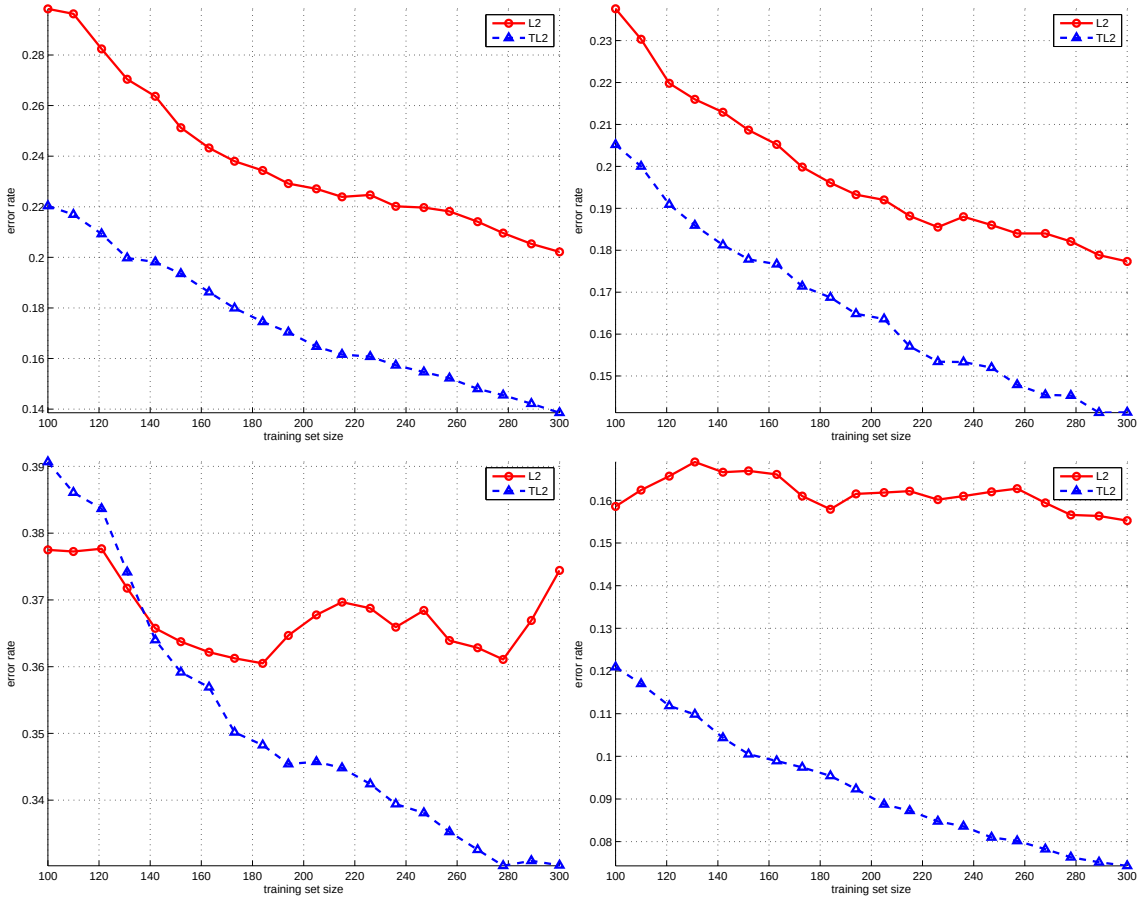

Figure 2: Line plots depicting the average L_2 test set error rate and the expected L_2 test set error rate for a nearest neighbor classifier. The figures correspond to RCV1 labeling tasks within a certain hierarchy class. Top figures demonstrate all pair classification within $\{M01, M13, M14\}$ (left) and $\{E12, E13, E14, E21, E31, E41, E51, E01\}$ (right). Bottom figures correspond to labeling population of all pairs within $\{C15, C17, C18, C31, C33, C41, C01, C02\}$ (left) and $\{G01, G02, G15\}$ (right).

References

[1] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000.

[2] L. Douglas Baker and Andrew K. McCallum. Distributional clustering of words for text classification. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, 1998.

[3] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7, 2006.

[4] Adam Berger and John D. Lafferty. Information retrieval as statistical translation. In *Proc. of ACM SIGIR Conference*, 1999.

[5] J. Byrnes and R. Rowher. Text modeling for real-time document categorization. In *Proc. of IEEE Conference on Aerospace*, 2005.

[6] L. L. Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

[7] N. N. Čencov. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, 1982.

[8] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[9] Kevyn Collins-Thompson and Jamie Callan. Query expansion using random walk models. In *Conference on Information and Knowledge Management*, 2005.

[10] Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. New experiments in distributional representations of synonymy. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, 2005.

[11] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems, 18*, 2005.

[12] R. E. Kass. The geometry of asymptotic inference. *Statistical Science*, 4(3):188–234, 1989.

[13] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th Intl. Conf. on Machine Learning (ICML)*, 2002.

[14] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 2005.

[15] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proc. of the ACM-SIGIR conference*, 2001.

[16] Guy Lebanon. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 2006.

[17] Guy Lebanon. Sequential document representations and simplicial curves. In *Proc. of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

[18] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 32st Annual Meeting of the Association for Computational Linguistics*, 1993.

[19] Alex Smola and Risi Kondor. Kernels and regularization on graphs. In *Proc. of Computational Learning Theory (COLT)*, 2003.

[20] K. Sparck-Jones and E.O. Barber. What makes an automatic keyword classification effective? *Journal of the American Society for Information Science*, 22, 1971.

[21] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russel. Distance metric learning with applications to clustering with side information. In *Advances in Neural Information Processing Systems, 15*, 2003.

[22] Xiaojin Zhu. *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005.