Semantically rich spaces for document clustering

Roberto Basili, Paolo Marocco and Daniele Milizia
Department of Computer Science, Systems and Production
University of Rome Tor Vergata
00133 Rome (Italy)

Abstract

Dimensionality reduction techniques address a relevant problem of Vector Space Models that is the size of involved dictionaries. Certain geometrical transformations applied over the original feature space, like the Latent Semantic Analysis (LSA), aim at preserving and discovering semantic relations between documents within small dimensional spaces. In this paper, a linear transformation method, named Locality Preserving Projections (LPP), is evaluated with respect to a document clustering task and results are compared with LSA. LPP is here applied directly on the original space, through an efficient ***C****-based implementation, and different parameterizations are investigated. Experimental results suggest that LPP is an effective technique able to account for the availability of a priori knowledge within an unsupervised learning framework.*

1. Introduction

Lexical meaning is at the basis of most tasks in Information Retrieval, such as ad hoc retrieval, document clustering or summarization. Although the formalization of word meaning is an old topic in AI and Philosophy, IR approaches have traditionally faced this huge problem by relying on simple meaning surrogates, i.e. the words themselves, with an extraordinary success, in terms of accuracy and scalability, given the shallow nature of the adopted representation. When using lexicalized features (such as the words occurring in a text to express the latter's semantics), several advantages arise. First, all the observations of the proper features are objectives and errors in the data interpretation are avoided. The well known distributional hypothesis suggests that word meaning can be acquired through a wittgensteinian "*language in use*" perspective and the growing availability of collections of digital documents allows to explore it on a large scale. Discrete, although large scale, feature sets can thus be naturally mapped into possibly high-dimensional vector space representations, where geometrical metrics supply principled real-valued functions as models of semantic similarity. Finally, analytical methods for manipulating the derived space can be inherited from the huge tradition of linear algebra and optimization theory. The analysis of different vector space models (VSM) for capturing word meaning is a relevant research line for problems such as the large scale acquisition of lexical knowledge, document management as well as social network analysis. Studies on learning methods for pattern recognition and automatic classification tasks have outlined the role of geometric transformations for dimensionality reduction ([8, 7]). These aim at capturing the subset of significant information implicit in the data distribution itself, and representing this source information by means of the minimal number of dimensions. Dimensionality reduction (DR) can be very helpful in clustering as the latter can be directly applied over the lower dimensional representation obtained. Dimensionality reduction usually leads up to a significant reduction of time and memory consumption in a clustering process, although it does not impact on its theoretical complexity. Several dimension reduction techniques are known: Independent Component Analysis (ICA), Latent Semantic Analysis (LSA), Random Projection (RP). These are analyzed and compared in [10]. Our aim is studying advanced geometrical transformations as rich models for lexical semantics, that are meaning preserving and reusable for a large number of text processing tasks: classification, clustering or word semantic disambiguation. We will discuss here methods that emphasize *neighborhood* information in the source data distribution to provide suitable dimensionality reductions, namely Locality Preserving Projections (LPP) ([5]), and compare them with previous works (i.e. LSA, [1]). In [2] LPP is firstly used in a document clustering task. The authors compare LPP with other dimension reduction techniques: LSA, NMF (Nonnegative Matrix Factorization) and LDA (Linear Discriminant Analysis). The comparisons indicate that the *locality model* is the key factor that characterizes LPP with respect to other techniques.

In this work an extensive experimental comparison is

carried out between LPP and LSA on a document clustering task. Clustering here is seen as a way to measure the impact of geometrical transformations on the learnability of text topicality. Moreover, as for the duality between documents and terms and the availability of semantic dictionaries, LPP is also promising for automatic learning of lexical taxonomies. In Section 2 we will introduce the two DR technologies. The experimental investigation is then presented in Section 3 with results discussed in Section 4.

2 Neighborhood information in VSMs

Vector space models in IR capture contextual information by expressing the distribution of words across text collections: individual texts are thus represented via linear combinations of (usually orthonormal) vectors corresponding to their component words. Similarity in the space can be captured by distances like euclidean norms or the cosine measure. These models very elegantly map documents and words in vector spaces (there are as many dimensions as words in the dictionary) and individual collections into distributions of data-points. Every distribution implicitly expresses:

- *global properties*, as the idf scores computed for terms across the entire collection and irrespectively from their word senses
- *local regularities*, as for example, the existence of subsets of the dictionary that tend to appear only in some documents. These tend to be also closer in the space

Meaning representation is certainly sensitive to both dimensions but geometrical transformation methods have been devised that are quite differently related to the two sources of information.

2.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is an algorithm presented by Deerwester et al. in [4], and afterwards diffused by Landauer [8]: it can be seen as a variant on the Principal Component Analysis idea. LSA aims to find the best subspace approximation to the original document space, in the sense of minimizing the global reconstruction error projecting data along the directions of maximal variance. It captures term (semantic) dependencies by applying a matrix decomposition process called Singular Value Decomposition (SVD). The original term-by-document matrix M, that describes traditional term-based document space, is transformed into the product of three new matrices: U, S, and V such that $M = USV^T$. Matrix M is approximated by $M_k = U_k S_k V_k^T$ in which only first k columns of U and V are used, and only first k greatest singular values are considered. This approximation supplies a way to project term vectors into the k-dimensional space using $Y_{terms} = U_k S_k^{1/2}$ and document vectors using $Y_{docs} = S_k^{1/2} V_k^T$. Notice that the SVD process accounts for the eigenvectors of the entire original distribution (matrix M). LSA is thus an example of a decomposition process tightly dependent on a global property. The original statistical information about M is captured by the new k-dimensional space which preserves the global structure. Each dimension (i.e. an induced LSA feature) may be thought of as an artificial concept and represents emerging meaning components from many different words and documents [8].

2.2 The LPP Algorithm

An alternative to LSA, much tighter to local properties of data, is the Locality Preserving Projections (LPP), a linear approximation of the nonlinear Laplacian Eigenmap algorithm, recently introduced by Xiaofei and Niyogi [5]. LPP is a linear dimensionality reduction method whose goal is, given a set $x_1, x_2, .., x_m$ in R^n, to find a transformation matrix A that maps these m points into a set of points $y_1, y_2, .., y_m$ in R^k ($k \ll n$). LPP achieves this result through a cascade of processing steps described hereafter.

(1) Construction of an Adjacency graph

Let **G** denote a graph with m nodes. Nodes *i* and *j* have got a weighted connection if vectors x_i and x_j are "close" according to an arbitrary measure of similarity. There are many ways to build an adjacency graph. In this paper we explore two possibilities:

- the *cosine* graph with cosine weighting scheme: given two vectors x_i and x_j, the weight w_{ij} between them is set by $w_{ij} = max\{0, \frac{cos(x_i,x_j)-\tau}{|cos(x_i,x_j)-\tau|} \cdot cos(x_i, x_j)\}$. here a cosine threshold τ is necessary.
- the *ϵ-neighborhoods* graph with Gauss Kernel weighting scheme: given two vectors x_i and x_j, the weight between them is set by $w_{ij} = max\{0, \frac{\epsilon - ||x_i - x_j||^2}{|\epsilon - ||x_i - x_j||^2|} \cdot GK(i,j,t)\}$, with $GK(i,j,t) = e^{-\frac{||x_i - x_j||^2}{t}}$. It is necessary to select here a threshold ϵ.

The adjacency graph can be represented using a symmetric $m \times m$ adjacency matrix, named W, whose element W_{ij} contains the weight between nodes *i* and *j*.

(2) Solve an Eigenmap problem

Compute the eigenvectors and eigenvalues for the generalized eigenvector problem:
(1) $XLX^T\mathbf{a} = \lambda XDX^T\mathbf{a}$
where X is a $n \times m$ matrix whose columns are the original m vectors in R^n, D is a diagonal $m \times m$ matrix whose entries are column (or row) sums of W, $D_{ii} = \sum_j W_{ij}$ and

$L = D - W$ is the Laplacian matrix. The solution of problem (1) is the set of eigenvectors $\mathbf{a}_0, \mathbf{a}_1, .., \mathbf{a}_{n-1}$, ordered according to their eigenvalues $\lambda_0 < \lambda_1 < .. < \lambda_{n-1}$. LPP projection matrix A is obtained by selecting the k eigenvectors corresponding to the k smallest eigenvalues: therefore it is a $n \times k$ matrix whose columns are the selected n-dimensional k eigenvectors. Final projection of original vectors into R^k can be performed by $Y = A^T X$. This transformation provides a valid kernel that can be efficiently embedded in kernel-based classifiers.

While LSA finds a projection according to the global properties of the space, LPP tries to preserve the local structures of the data. LPP exploits the adjacency graph to represent neighborhood information. It computes a transformation matrix which maps data points into a lower dimensional subspace. This transformation preserves optimally the local neighborhood information expressed by the graph. It is well possible to combine the LSA and the LPP model, using the former as pre-processing step to reduce the feature space dimensions, and the latter to emphasize the emergence of local semantic features, as also done in [2]. The embedded model adopted in our experiments is discussed in Section 3.

3 Local and global properties in clustering

In order to investigate the two different DR methodologies, Latent Semantic Analysis (LSA) and Locality Preserving Projections (LPP), we will analyse the impact of several of their possible settings in a document clustering task. The main distinguishing aspect of the LPP technology is the adjacency graph: it is possible to express neighborhood information inside the graph using internal or external similarity metrics. Internal metrics suggest connections by maximizing (or minimizing) similarity (or distance), as computed directly on the source data distribution. However, external metrics can be used as well. They can be inspired by independent sources such as semantic dictionaries, used to superimpose a connection when synonyms are found in both documents.

In this work, we focused on internal metrics. However, we also applied an "*ideal*" adjacency graph that makes direct use of the topic association. The so-called "*topic*" graph is built such that two documents have a connection only if they belong to the same corpus category, and the cosine between them is used as a weight. The application of LPP fed by such a graph has been then used to trigger unsupervised clustering against a gold standard (e.g. the Reuters collection). The agreement between clusters and the Reuters categories is reported in Table 1 (the adopted settings and metrics are discussed later in Section 4). Interestingly, LPP seems to converge towards a very expressive space as clustering almost fully reproduce the targeted classification scheme. It seems that the use of *a priori* knowledge about the target task is perfectly captured by the LPP transformation. While LSA induces semantic information through a purely statistical procedure, LPP allows the integration of external semantic information through a suitable adjacency graph. The adoption of the proper a priori knowledge about the target task can be thus seen as a promising research direction.

Corpus Pre-Processing. Document clustering was used in order to compare LSA and LPP. We employed two extensively used document collections in our experiments: Reuters-21578[1] and 20Newsgroups[2]. These allowed us to have a comparative evaluation in line with [2]. LSA and LPP reduction algorithms were applied first, and then an extended version of the k-Means algorithm ([6]) was run on document representations obtained after the reduction step. The Reuters data corpus contains 21,578 documents and 135 topics created manually, and each document in the corpus has been assigned one or more topics (categories). The 20Newsgroups corpus is made by 19,997 documents, from the Usenet newsgroups collection, grouped into 20 topics. As an hard clustering approach is used, in order to consistently compare results of different models, we discarded multi-topic documents in both collections. The largest 30 categories in Reuters were selected (as also done in [2]). For both collections, we filtered stopwords, and punctuation tokens. Rare terms in the collections were then removed in order to limit the source dimensionality. Thresholds of 1 and 9 occurrences were applied to Reuters and 20Newsgroups respectively in order to provide tractable feature sets: matrixes of 9,675 columns (document vectors) and 18,349 rows (vocabulary size), for Reuters, and 18,828 columns and 21,500 rows for 20Newsgroups were finally obtained. The TFIDF weighting was applied to both cases.

Dimensionality Reduction. We used LSA and LPP as dimensionality reduction techniques. We also applied LPP combined with the LSA algorithm (we call this representation LSA+LPP), as follows. When X is the original space matrix, the equation of the LSA can be expressed by $\mathbf{Y} = A_{LSA}^T \mathbf{X}$ space, while $\mathbf{Y} = A_{LPP}^T \mathbf{X}$ stands for the LPP transformation. The LSA+LPP model can thus be written as $\mathbf{Z} = A_{LSA} A_{LPP} \mathbf{X}$ where the transformation $A_{LSA+LPP} = A_{LSA} A_{LPP}$ is implicitly adopted. Substantially we first apply the LSA transformation over X matrix using a final space dimension k, and we obtain a new matrix Y. Then we apply LPP on the Y matrix, choosing a lower final dimension k_1 such that $k_1 < k$.

Document Clustering. An extended version of the well-known k-Means algorithm ([6]) was used to cluster documents in different representation spaces. k-Means is a partitional (hard) clustering algorithm that starts with an initial (typically random) setting of a specified number k of

[1] http://kdd.ics.uci.edu/databases/reuters21578/

[2] http://www.ai.mit.edu/people/jrennie/20Newsgroups/

centroids and adjusts them iteratively. The extended algorithm used during the experiments offers some parameters that make k-Means more flexible ([6]). *Infra_cluster thresholding*: the allowed similarity between members and a cluster centroid is maintained over a threshold. The higher the threshold, the higher is the number of derived clusters. *Inter_cluster thresholding*: the similarity allowed between two centroids is kept below a threshold. If the cosine between two centroids is higher than the threshold they will be merged, resulting in a larger cluster. The *Maximum size* corresponds to maximal cardinality allowed for a cluster.

4 Experimental Results

The objective of the experiments is the investigation about the combined effects of several design choices on the adopted DR techniques:

- reduction factor (k) for LSA, LPP and LSA+LPP.
- types of adjacency graph for LPP, based on different measures and thresholds ϵ and τ.
- clustering parameters, i.e. *Infra_cluster thresholding* and *Inter_cluster thresholding*

Given a representation space (among VSM, LSA, LPP and LSA+LPP), a dimension for the representation space, and the selection of the desired graph and threshold for LPP, the experimental workflow proceeds by executing the clustering according to a combination of the parameters, and measuring the clustering quality through external metrics against the gold standard.

External clustering quality was evaluated using *Accuracy* and *Normalized Mutual Information* metrics ([3, 2]). Clustering Accuracy is evaluated by mapping every cluster to a topic label: a majority voting function was used here to label every cluster with one of the corpus topics. Given the i-th document, let A_i be the topic label assigned to the cluster where i-th document has been placed by the method, while O_i expresses the expected topic label as defined by the oracle. *Accuracy* is obtained by

$$AC = \frac{\sum_{i=1}^{n} \delta(A_i, O_i)}{N} \quad (1)$$

where N is the total number of documents and $\delta(A_i, O_i)$ is 1 only if $A_i = O_i$ and 0 otherwise.

Given the set T of corpus topics (i.e. the collection categories) and the set C of the generated clusters, the *Normalized mutual information* is defined as follows:

$$NMI(T, C) = \frac{\sum_{t \in T, c \in C} p(t, c) log_2 \frac{p(t,c)}{p(t) \cdot p(c)}}{min(H(T), H(C))} \quad (2)$$

where where $p(t)$ and $p(c)$ are the probabilities that a document arbitrarily selected from the corpus belongs to topic t and cluster c, respectively; $p(t, c)$ is the joint probability that the arbitrarily selected document belongs both to topic t and cluster c; $H(T)$ and $H(C)$ are the entropies[3] of topics in T and clusters in C, respectively. While an extensive experimentation of all parameters settings has been carried out and discussed in [9] we will report hereafter the most significant results obtained. Notice how all training documents and all classes (O_i) are here employed, in contrast with [2], where variable data sets, ranging from 10% to 90% of the original data, are employed.

4.1 Comparing LSA and LPP

The comparison between LSA and LPP for different choices of final space dimension is illustrated in Figure 1. We used the *cosine* adjacency graph with threshold $\tau = 0.9$ for LPP: this kind of graph turned out to be the best among the ones investigated in [9]. These results show that the LPP

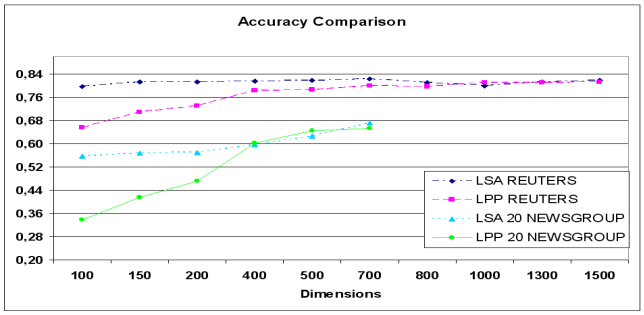

Figure 1. Accuracy of LSA and LPP

method which uses an internal metric of similarity can't succeed in outperform LSA's results globally. The best result is obtained using LSA, while LPP improves on LSA only when some favourable dimensions are used. Results differ from those in [2] as different training and testing conditions are here adopted, e.g. all classes and all documents are employed for each measure. A more interesting result was obtained using the "*topic*" graph. Table 1 confirms our intuition that bringing topic knowledge inside LPP adjacency graph can succeed in gaining a greater discriminating power inside the reduced representation space. We point out that the "*topic*" graph could not be used in a realistic clustering or categorization experiment as it is not available for unseen data. However, this experiment shows that the availability of perfect "a priori" evidence about the data is fully captured by LPP.

[3] For a discrete distribution $X = \{x1, ..., x_n\}$ the entropy $H(X)$ is given by $H(X) = \sum_i -p(x_i) log_2(p(x_i))$.

METHOD	REUTERS	
	ACC	NMI
LSA	0.82	0.79
LPP	**0.94**	**0.99**

Table 1. Best LSA vs. upper bound LPP results based on the "*topic*" graph on Reuters.

THR	LSA (700)			
	ACC	NMI	CLUSTERS	β_{CV}
-1	0.72	0.61	30	2268.3
0.2	0.82	0.79	507	20.7
0.4	**0.86**	**0.84**	1298	4.2
THR	LSA+LPP (LSA 700, LPP 680, ϵ=0.05)			
	ACC	NMI	CLUSTERS	β_{CV}
-1	0.77	0.66	30	2775.3
0.2	0.81	0.78	491	21.6
0.4	**0.86**	**0.84**	1253	4.3

Table 2. Performances on Reuters

4.2 Evaluating the LSA+LPP model

We also applied LPP on LSA representation of textual data (LSA+LPP algorithm). We performed different experiments varying final LSA and LPP dimensions and adjacency graphs for LPP ([9]): here we report only the most remarkable results. We applied LSA using 700 dimensions on Reuters corpus and 500 on 20Newsgroups and then applied LPP on LSA representation, using a final dimension of 680 for Reuters and 480 for 20Newsgroups. Results in Table 2 and 3 show how LSA+LPP remarkably outperforms LSA when the cluster number is exactly 30, although this is not true in general.

5 Summary and Conclusions

In this paper we investigated and compared two different dimensionality reduction methodologies, Latent Semantic Analysis and Locality Preserving Projections, with respect to a document clustering task. We focused our analysis on LPP, a linear algorithm derived from non linear Locally Linear Embedding. Results obtained on the Reuters and 20Newsgroups corpora showed that internal neighborhood metrics improve over LSA only when certain space dimensions are used. The LPP projection seems to be more sensitive to dimension changes. Moreover, we found that expressing topic knowledge inside the adjacency graph is a key factor to produce an "*ideal*" representation space. In this space, categories are in fact mapped into low variance clusters, geometrically and semantically well separated. These results can be successfully exploited in a variety of tasks. In Text Categorization, in fact, LPP could be used as the kernel function of a Support Vector Machine. In the automatic learning of lexical taxonomies LPP is also useful for modeling *a priori* knowledge. These applications are justified by the empirical evidence reported in this work.

THR	LSA (500)			
	ACC	NMI	CLUSTERS	β_{CV}
-1	0.58	0.57	20	9726.92
0.2	0.59	0.59	430	66.26
0.3	**0.63**	**0.64**	720	28.99
THR	LSA+LPP (LSA 500, LPP 480, ϵ=0.05)			
	ACC	NMI	CLUSTERS	β_{CV}
-1	0.54	0.55	20	8210,61
0.2	0.59	0.60	438	62.89
0.3	**0.62**	**0.64**	724	28.74

Table 3. Performances on 20Newsgroups

References

[1] M. Berry, S. Dumais, and G. O'Brien. Using linear algebra for intelligent information retrieval, 1995.

[2] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12), 2005.

[3] P. A. Daniel Crabtree, Xiaoying Gao. Standardized evaluation method for web clustering results. In *IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI'05)*, Compiegne, France, 2005.

[4] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proc. of SIGIR '88*, New York, USA, 1988.

[5] X. He and P. Niyogi. Locality preserving projections. In *Proceedings of NIPS '03*, Vancouver, Canada, 2003.

[6] L. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, (9):1106–1115, 1999.

[7] T. J, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[8] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[9] D. Milizia. Text clustering models based on latent semantic analysis and locality preserving projections transformations. *Laurea Degree thesis*. University of Rome, Tor Vergata, 2007.

[10] B. Tang, M. Shepherd, E. Milios, and M. Heywood. Comparing and combining dimension reduction techniques for efficient text clustering. In *SIAM Int. Workshop on* Feature Selection for Data Mining, Newport Beach, CA, 2005.