Producing Accurate Interpretable Clusters from High-Dimensional Data

Derek Greene and Pádraig Cunningham

University of Dublin, Trinity College,
Dublin 2, Ireland
{derek.greene, padraig.cunningham}@cs.tcd.ie

Abstract. The primary goal of cluster analysis is to produce clusters that accurately reflect the natural groupings in the data. A second objective is to identify features that are descriptive of the clusters. In addition to these requirements, we often wish to allow objects to be associated with more than one cluster. In this paper we present a technique, based on the spectral co-clustering model, that is effective in meeting these objectives. Our evaluation on a range of text clustering problems shows that the proposed method yields accuracy superior to that afforded by existing techniques, while producing cluster descriptions that are amenable to human interpretation.

1 Introduction

The unsupervised grouping of documents, a frequently performed task in information retrieval systems, can be viewed as having two fundamental goals. Firstly, we seek to identify a set of clusters that accurately reflects the topics present in a corpus. A second objective that is often overlooked is the provision of information to facilitate human interpretation of the clustering solution.

The primary choice of representation for text mining procedures has been the *vector space model*. However, corpora modelled in this way are generally characterised by their sparse high-dimensional nature. Traditional clustering algorithms are susceptible to the well-known problem of the *curse of dimensionality*, which refers to the degradation in algorithm performance as the number of features increases. Consequently, these methods will often fail to identify coherent clusters when applied to text data due to the presence of many irrelevant or redundant terms. In addition, the inherent sparseness of the data can further impair an algorithm's ability to correctly uncover the data's underlying structure.

To overcome these issues, a variety of techniques have been proposed to project high-dimensional data to a lower-dimensional representation in order to minimise the effects of sparseness and irrelevant features. In particular, dimension reduction methods based on spectral analysis have been frequently applied to improve the accuracy of document clustering algorithms, due to their ability to uncover the latent relationships in a corpus. However, from the perspective of domain users, the production of clear, unambiguous descriptions of cluster content is also highly important. A simple but effective means of achieving this

A. Jorge et al. (Eds.): PKDD 2005, LNAI 3721, pp. 486–494, 2005.

goal is to generate weights signifying the relevance of the terms in the corpus vocabulary to each cluster, from which a set of cluster labels can be derived. The provision of document weights can also help the end-user to gain an insight into a clustering solution. In particular, when a document is assigned to a cluster, one may wish to quantify the confidence of the assignment. Additionally, the use of soft clusters allows us to represent cases where a given document relates to more than one topic.

In this paper, we introduce a co-clustering technique, based on spectral analysis, that provides interpretable membership weights for both terms and documents. Furthermore, we show that by applying an iterative matrix factorisation scheme, we can produce a refined clustering that affords improved accuracy and interpretability. We compare our algorithms with existing methods on a range of datasets, and briefly discuss the generation of useful cluster descriptions. Note that an extended version of this paper is available as a technical report with the same title [1].

2 Matrix Decomposition Methods

In this section, we present a brief summary of two existing dimension reduction methods that have been previously applied to document clustering. To describe the algorithms discussed in the remainder of the paper, we let $\mathbf{A}$ denote the $m \times n$ term-document matrix of a corpus of n documents, each of which is represented by an m-dimensional feature vector. We assume that k is an input parameter indicating the desired number of clusters.

2.1 Spectral Co-clustering

Spectral clustering methods have been widely shown to provide an effective means of producing disjoint partitions across a range of domains [2,3]. In simple terms, these algorithms analyse the spectral decomposition of a matrix representing a dataset in order to uncover its underlying structure. The reduced space, constructed from the leading eigenvectors or singular vectors of the matrix, can be viewed as a set of semantic variables taking positive or negative values.

A novel approach for simultaneously clustering documents and terms was suggested by Dhillon [4], where the co-clustering problem was formulated as the approximation of the optimal normalised cut of a weighted bipartite graph. It was shown that a relaxed solution may be obtained by computing the *singular value decomposition* (SVD) of the normalised matrix $\mathbf{A_n} = \mathbf{D_1}^{-1/2}\mathbf{A}\mathbf{D_2}^{-1/2}$, where $[D_1]_{ii} = \sum_{j=1}^{n} A_{ij}$ and $[D_2]_{jj} = \sum_{i=1}^{m} A_{ij}$ are diagonal matrices. A reduced representation $\mathbf{Z}$ is then constructed from the the left and right singular vectors of $\mathbf{A_n}$, corresponding to the $\log_2 k$ largest non-trivial singular values. Viewing the matrix $\mathbf{Z}$ as a l-dimensional geometric embedding of the original data, the k-means algorithm is applied in this space to produce a disjoint co-clustering.

2.2 Non-negative Matrix Factorisation (NMF)

Non-Negative Matrix Factorisation (NMF) [5] has recently been identified as a practical approach for reducing the dimensionality of non-negative matrices such as the term-document representation of a text corpus [6]. Unlike spectral decomposition, NMF is constrained to produce non-negative factors, providing an interpretable clustering without the requirement for further processing.

Given the term-document matrix $\mathbf{A}$, NMF generates a rank-k approximation of the corpus in the form of the product of two non-negative matrices $\mathbf{A} \approx \mathbf{U}\mathbf{V}^T$. The factor $\mathbf{U}$ is a $m \times k$ matrix consisting of k basis vectors, which can be viewed as a set of semantic variables corresponding to the topics in the data, while $\mathbf{V}$ is a $n \times k$ matrix of coefficients describing the contribution of the documents to each topic. The factors are determined by a given objective function that seeks to minimise the error of the reconstruction of $\mathbf{A}$ by the approximation $\mathbf{U}\mathbf{V}^\mathbf{T}$.

3 Soft Spectral Clustering

In this section, we discuss the problem of inducing membership weights from a disjoint partition, and we propose an intuitive method to produce soft clusters based on the spectral co-clustering model.

For the task of generating feature weights from a hard clustering, a common approach is to derive values from each cluster's centroid vector [7]. However, an analogous technique for spectral clustering is not useful due to the presence of negative values in centroid vectors formed in the reduced space. Another possibility is to formulate a document's membership weights as a function of the document's similarity to each cluster centroid [8].

The success of spectral clustering methods has been attributed to the truncation of the eigenbasis, which has the effect of amplifying the association between points that are highly similar, while simultaneously attenuating the association of points that are dissimilar [3]. However, while this process has been shown to improve the ability of a post-processing algorithm to identify cohesive clusters, the truncation of the decomposition of $\mathbf{A}$ to $k \ll m$ singular vectors introduces a distortion that makes the extraction of natural membership weights problematic. As a consequence, we observe that directly employing embedded term-centroid similarity values as membership weights will not provide intuitive cluster labels.

3.1 Inducing Soft Clusters

As a starting point, we construct a reduced space based on the spectral co-clustering model described in Section 2.1. However, we choose to form the embedding $\mathbf{Z}$ from the leading k singular vectors, as truncating the eigenbasis to a smaller number of dimensions may lead to an inaccurate clustering [2]. By applying the classical k-means algorithm using the cosine similarity measure, we generate k disjoint subsets of the points in the embedded geometric space. We represent this clustering as the $(m + n) \times k$ partition matrix $\mathbf{P} = [P_1, \ldots, P_k]$,

where P_i is a binary membership indicator for the i-th cluster. We denote the k centroids of the clustering by $\{\mu_1, \ldots, \mu_k\}$.

As the spectral co-clustering strategy is based on the principle of the duality of clustering documents and terms [4], we argue that we can induce a soft clustering of terms from the partition of documents in $\mathbf{Z}$ and a soft clustering of documents from the partition of terms. Note that the matrix $\mathbf{P}$ has the structure:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P_1} \\ \mathbf{P_2} \end{bmatrix}$$

where $\mathbf{P_1}$ and $\mathbf{P_2}$ indicate the assignment of terms and documents respectively. An intuitive approach to producing term weights is to apply the transformation $\mathbf{A}^T\hat{\mathbf{P}}_\mathbf{1}$, where $\hat{\mathbf{P}}_\mathbf{1}$ denotes the matrix $\mathbf{P}_1$ with columns normalised to unit length. This effectively projects the centroids of the partition of documents in $\mathbf{Z}$ to the original feature space. Similarly, to derive document-cluster association weights $\mathbf{V}$, we can apply the transformation $\mathbf{A}\hat{\mathbf{P}}_\mathbf{2}$, thereby projecting the embedded term cluster centroids to the original data.

However, the above approach will not reflect the existence of boundary points lying between multiple clusters or outlying points that may be equally distant from all centroids. To overcome this problem, we propose the projection of the centroid-similarity values from the embedded clustering to the original data. Due to the presence of negative values in $\mathbf{Z}$, these similarities will lie in the range $[-1, 1]$. We rescale the values to the interval $[0, 1]$ and normalise the k columns to unit length, representing them by the matrix $\mathbf{S}$ as defined by:

$$S_{ij} = \frac{1 + cos(z_i, \mu_j)}{2}, \quad S_{ij} \leftarrow \frac{S_{ij}}{\sum_l S_{lj}} \tag{1}$$

As with the partition matrix of the embedded clustering, one may divide $\mathbf{S}$ into two submatrices, where $\mathbf{S_1}$ corresponds to the $m \times k$ term-centroid similarity matrix and $\mathbf{S_2}$ corresponds to the $n \times k$ document-centroid similarity matrix:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S_1} \\ \mathbf{S_2} \end{bmatrix}$$

By applying the projections $\mathbf{A}^T\mathbf{S_1}$ and $\mathbf{AS_2}$, we generate membership weights that capture both the affinity between points in the embedded space and the raw term-frequency values of the original dataset.

3.2 Soft Spectral Co-clustering (SSC) Algorithm

Motivated by the duality of the co-clustering model, we now present a spectral clustering algorithm with soft assignment of terms and documents that employs a combination of the transformation methods described in the previous section. We formulate the output of the algorithm as a pair of matrices $(\mathbf{U}, \mathbf{V})$, representing the term-cluster and document-cluster membership functions respectively. As a document membership function, we select the projection $\mathbf{A}^T\mathbf{S_1}$, on the basis

that the use of similarity values extracts more information from the embedded clustering than purely considering the binary values in $\mathbf{P}$. We observe that this generally leads to a more accurate clustering, particularly on datasets where the natural groups overlap.

The requirements for a term membership function differ considerably from those of a document membership function, where accuracy is the primary consideration. As the production of useful cluster descriptions is a central objective of our work, we seek to generate a set of weights that results in the assignment of high values to relevant features and low values to irrelevant features. Consequently, we select the projection $\mathbf{A}\hat{\mathbf{P}}_\mathbf{2}$ as previous work has shown that centroid vectors can provide a summarisation of the important concepts present in a cluster [7]. Our choice is also motivated by the observation that the binary indicators in $\hat{\mathbf{P}}_\mathbf{1}$ result in sparse discriminative weight vectors, whereas the projection based on $\mathbf{S_2}$ leads to term weights such that the highest ranking words tend to be highly similar across all clusters. We now summarise the complete procedure, which we refer to as the *Soft Spectral Co-clustering* (SSC) algorithm:

1. Compute the k largest singular vectors of $\mathbf{A_n}$ to produce the truncated factors $\mathbf{U_k} = (u_1, \ldots, u_k)$ and $\mathbf{V_k} = (v_1, \ldots, v_k)$.
2. Construct the embedded space $\mathbf{Z}$ by scaling and stacking $\mathbf{U_k}$ and $\mathbf{V_k}$:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{D_1}^{-1/2}\mathbf{U_k} \\ \mathbf{D_2}^{-1/2}\mathbf{V_k} \end{bmatrix}$$

3. Apply the k-means algorithm with cosine similarity to $\mathbf{Z}$ to produce a disjoint co-clustering, from which the matrices $\mathbf{S_1}$ and $\hat{\mathbf{P}}_\mathbf{1}$ are computed.
4. Form soft clusters by applying the projections $\mathbf{U} = \mathbf{A}\hat{\mathbf{P}}_\mathbf{2}$ and $\mathbf{V} = \mathbf{A}^T\mathbf{S_1}$.

We employ a cluster initialisation strategy similar to that proposed in [2], where each centroid is chosen to be as close as possible to 90^o from the previously selected centroids. However, rather than nominating the first centroid at random, we suggest that accurate deterministic results may be produced by selecting the most centrally located data point in the embedded space.

4 Refined Soft Spectral Clustering

We now present a novel technique for document clustering by dimension reduction that builds upon the co-clustering techniques described in Section 3.

The dimensions of the space produced by spectral decomposition are constrained to be orthogonal. However, as text corpora will typically contain documents relating to multiple topics, the underlying semantic variables in the data will rarely be orthogonal. The limitations of spectral techniques to effectively identify overlapping clusters has motivated other techniques such as NMF, where each document may be represented as an additive combination of topics [6]. However, the standard approach of initialising the factors $(\mathbf{U}, \mathbf{V})$ with random values can lead to convergence to a range of solutions of varying quality. We argue that

initial factors, produced using the soft cluster induction techniques discussed previously, can provide a set of well-separated "core" clusters. By subsequently applying matrix factorisation with non-negativity constraints to the membership matrices, we can effectively uncover overlaps between clusters.

4.1 Refined Soft Spectral Co-clustering (RSSC) Algorithm

In the SSC algorithm described in Section 3.1, our choice of projection for the construction of the term membership matrix was motivated by the goal of producing human-interpretable weights. However, the projection $\mathbf{AS_2}$ retains additional information from the embedded clustering, while simultaneously considering the original term frequencies in $\mathbf{A}$. Consequently, we apply soft spectral co-clustering as described previously, but we select $\mathbf{U} = \mathbf{A}^T\mathbf{S_1}$ and $\mathbf{V} = \mathbf{AS_2}$ as our initial pair of factors.

We refine the weights in $\mathbf{U}$ and $\mathbf{V}$ by iteratively updating these factors in order to minimise the divergence or entropy between the original term-document matrix $\mathbf{A}$ and the approximation $\mathbf{UV}^T$ as expressed by

$$D(\mathbf{A}||\mathbf{UV}^T) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij}\log\frac{A_{ij}}{[\mathbf{UV}^T]_{ij}} - A_{ij} + [\mathbf{UV}^T]_{ij}\right) \tag{2}$$

This function can be shown to reduce to the Kullback-Leibler divergence measure when both $\mathbf{A}$ and $\mathbf{UV}^T$ sum to 1. To compute the factors a diagonally scaled gradient descent optimisation scheme is applied in the form of a pair of multiplicative update rules that converge to a local minimum. We summarise the Refined Soft Spectral Co-clustering (RSSC) algorithm as follows:

1. Decompose $\mathbf{A_n}$ and construct the embedded space $\mathbf{Z}$ as described previously.
2. Apply k-means to the rows of $\mathbf{Z}$ to produce a disjoint clustering, from which $\mathbf{S_1}$ and $\mathbf{S_2}$ are constructed.
3. Generate the initial factors $\mathbf{U} = \mathbf{A}^T\mathbf{S_1}$ and $\mathbf{V} = \mathbf{AS_2}$.
4. Update $\mathbf{V}$ using the rule

$$v_{ij} \leftarrow v_{ij}\left[\left(\frac{A_{ij}}{[\mathbf{UV}^T]_{ij}}\right)^T\mathbf{U}\right]_{ij} \tag{3}$$

5. Update $\mathbf{U}$ using the rule

$$u_{ij} \leftarrow u_{ij}\left[\frac{A_{ij}}{[\mathbf{UV}^T]_{ij}}\mathbf{V}\right]_{ij}, \quad u_{ij} \leftarrow u_{ij}\frac{U_{ij}}{\sum_{l=1}^{m}U_{lj}} \tag{4}$$

6. Repeat from step 4 until convergence.

To provide a clearer insight into the basis vectors, we subsequently apply a normalisation so that the Euclidean length of each column of $\mathbf{U}$ is of unit length and we scale the factor $\mathbf{V}$ accordingly as suggested in [6].

5 Experimental Evaluation

In our experiments we compared the accuracy of the SSC and RSSC algorithms to that of spectral co-clustering (CC) based on k singular vectors and NMF using the divergence objective function given in (2) and random initialisation. Choosing the number of clusters k is a difficult model-selection problem which lies beyond the scope of this paper. For the purpose of our experiments we set k to correspond to the number of annotated classes in the data.

The experimental evaluation was conducted on a diverse selection of datasets, which differ in their dimensions, complexity and degree of cluster overlap. For a full discussion of the datasets used in our experiments, consult [1]. To pre-process these datasets, we applied standard stop-word removal and stemming techniques. We subsequently excluded terms occurring in less than three documents. No further feature selection or term normalisation was performed.

5.1 Results

To compare algorithm accuracy, we apply the *normalised mutual information* (NMI) external validation measure proposed in [9]. We elected to evaluate hard clusterings due to the disjoint nature of the annotated classes for the datasets under consideration, and to provide a means of comparing the non-probabilistic document weights generated by our techniques with the output of the spectral co-clustering algorithm. Therefore, we induce a hard clustering from $\mathbf{V}$ by assigning the i-th document to the j-th cluster if $j = \arg\max_j(v_{ij})$.

Table 1 summarises the experimental results for all datasets as averaged across 20 trials. In general, the quality of the clusters produced by the SSC algorithm was at least comparable to that afforded by the spectral co-clustering method described in [4]. By virtue of their ability to perform well in the presence of overlapping clusters, both the NMF and RSSC methods generally produced clusterings that were superior to those generated using only spectral analysis. However, the RSSC algorithm's use of spectral information to seed well-separated "core clusters" for subsequent refinement leads to a higher level of accuracy on most datasets. When applied to larger datasets, we observe that the NMF and CC methods exhibit considerable variance in the quality of the clusters that they

Table 1. Performance comparison based on NMI

Dataset	CC	NMF	SSC	RSSC
bbc	0.78	0.80	0.82	**0.86**
bbcsport	0.64	0.69	0.65	**0.70**
classic2	0.29	0.34	0.46	**0.79**
classic3	0.92	**0.93**	0.92	**0.93**
classic	0.63	0.70	0.62	**0.87**
ng17-19	0.39	0.36	0.45	**0.50**

Dataset	CC	NMF	SSC	RSSC
ng3	0.68	0.78	0.70	**0.84**
re0	0.33	0.39	0.35	**0.40**
re1	0.39	0.42	0.41	**0.43**
reviews	0.34	0.53	0.40	**0.57**
tr31	0.38	0.54	0.51	**0.65**
tr41	0.58	0.60	**0.67**	**0.67**

produce, whereas the deterministic nature of the initialisation strategy employed by the newly proposed algorithms leads to stable solutions.

5.2 Cluster Labels

Given the term membership weights produced by the SSC and RSSC algorithms, a natural approach to generating a set of labels for each cluster is to select the terms with the highest values from each column of the matrix **U**. Due to space restrictions, we only provide a sample of the labels selected for clusters produced by the RSSC algorithm on the *bbc* dataset in Table 2, where the natural categories are: business, politics, sport, entertainment and technology.

Table 2. Labels produced by RSSC algorithm for *bbc* dataset

Cluster	Top 7 Terms
C1	company, market, firm, bank, sales, prices, economy
C2	government, labour, party, election, election, people, minister
C3	game, play, win, players, england, club, match
C4	film, best, awards, music, star, show, actor
C5	people, technology, mobile, phone, game, service, users

6 Concluding Remarks

In this paper, we described a method based on spectral analysis that can yield stable interpretable clusters in sparse high-dimensional spaces. Subsequently, we introduced a novel approach to achieve a more accurate clustering by applying a constrained matrix factorisation scheme to refine an initial solution produced using spectral techniques. Evaluations conducted on a variety of text corpora demonstrate that this method can lead to the improved identification of overlapping clusters, while simultaneously producing document and term weights that are amenable to human interpretation.

References

1. Greene, D., Cunningham, P.: Producing accurate interpretable clusters from high-dimensional data. Technical Report CS-2005-42, Trinity College Dublin (2005)
2. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proc. Advances in Neural Information Processing. (2001)
3. Brand, M., Huang, K.: A unifying theorem for spectral embedding and clustering. In: Proc. 9th Int. Workshop on AI and Statistics. (2003)
4. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Knowledge Discovery and Data Mining. (2001) 269–274
5. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–91
6. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proc. 26th Int. ACM SIGIR. (2003) 267–273

7. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Machine Learning **42** (2001) 143–175
8. Zhao, Y., Karypis, G.: Soft clustering criterion functions for partitional document clustering: a summary of results. In: Proc. 13th ACM Conf. on Information and Knowledge Management. (2004) 246–247
9. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. JMLR **3** (2002) 583–617