Producing Accurate Interpretable Clusters from High-Dimensional Data

Derek Greene, Pádraig Cunningham

University of Dublin, Trinity College,
Dublin 2, Ireland

Abstract. The primary goal of cluster analysis is to produce clusters that accurately reflect the natural groupings in the data. A second objective that is important for high-dimensional data 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 applied technique in many 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 the document collection. A second objective that is often overlooked is the provision of information to facilitate the human interpretation of the clustering solution.

The primary choice of representation for text mining procedures has been the *vector space model*, where each document is encoded as a vector whose dimensions correspond to the vocabulary of content-bearing terms in the corpus. 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, as documents tend to become equally similar to one another as the dimensionality increases.

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 document clustering, dimension reduction methods based on spectral analysis have been frequently applied due to their ability to uncover the latent relationships in a corpus [**?**,**?**].

The application of dimension reduction techniques to document clustering has largely focused on the algorithms' ability to accurately partition 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 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 subsequently 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 readily 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 the proposed algorithms with existing methods on a range of datasets, and discuss the generation of useful cluster descriptions.

The remainder of this paper is organised as follows. The next section provides a summary of existing methods, based on dimension reduction by matrix decomposition, that have been shown to be effective in document clustering. Section 3 describes a technique for deriving interpretable weights from the spectral co-clustering model. In Section 4 we introduce a method to further refine a set of soft clusters by employing matrix factorisation with non-negativity constraints. In Section 5 we evaluate the effectiveness of the proposed algorithms. Finally, Section 6 presents concluding remarks and suggestions for future work.

2 Matrix Decomposition Methods

In this section, we present an overview of two existing dimension reduction methods which have been previously applied to the task document clustering. In order to describe the algorithms discussed in this paper, we introduce the following notation. We let $\mathbf{A}$ denote the $m \times n$ term-document matrix of a corpus that consists of n documents represented by m-dimensional feature vectors. 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 [**?**,**?**,**?**]. In simple terms, these algorithms analyse the eigendecomposition of the matrix representation of a dataset in order to uncover its underlying structure in the form of a disjoint partition. 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. Consequently, a suitable post-processing

heuristic is required to extract the final clusters. While several authors have proposed novel techniques to produce a final partition directly from the truncated eigenbasis (e.g. [**?**]), the most common approach remains the application of the classical k-means algorithm to the points in the reduced space.

A novel approach for simultaneously clustering documents and terms was suggested by Dhillon [**?**], where the co-clustering problem was formulated as the task of approximating 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 term-document matrix $\mathbf{A_n} = \mathbf{D_1}^{-1/2}\mathbf{A}\mathbf{D_2}^{-1/2}$, where $\mathbf{D_1}$ and $\mathbf{D_2}$ are diagonal matrices such that

$$[D_1]_{ii} = \sum_{j=1}^{n} A_{ij}\,, \quad [D_2]_{jj} = \sum_{i=1}^{m} A_{ij} \tag{1}$$

Given the left and right singular vectors of $\mathbf{A_n}$, corresponding to the $\log_2 k$ largest non-trivial singular values, a reduced representation $\mathbf{Z}$ is constructed by normalised and arranging the truncated factors as follows:

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

Since the matrix $\mathbf{Z}$ can be viewed as a l-dimensional geometric embedding of the original data, the k-means algorithm may be applied in this space to produce a disjoint co-clustering of the corpus.

2.2 Non-negative Matrix Factorisation (NMF)

Non-Negative Matrix Factorisation (NMF) [**?**] 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 dataset [**?**]. Unlike spectral decomposition, NMF is constrained to produce non-negative factors. By modelling each document as the additive combination of a set of non-negative semantic variables, a readily interpretable clustering of the data can be produced without the requirement for further post-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 non-negativity of these matrices allows them to be directly interpreted as a soft k-way co-clustering. The choice of factors is determined by a given objective function that seeks to minimise the error of the reconstruction of $\mathbf{A}$ by the approximation $\mathbf{U}\mathbf{V}^\mathbf{T}$. In practice, these factors can be calculated by applying a diagonally rescaled gradient descent optimisation scheme that results in convergence to a local minimum.

3 Soft Spectral Clustering

Dimension reduction methods that transform the original dataset to a lower-dimensional space often achieve improved accuracy at the expense of interpretability. 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.

3.1 Motivation

Spectral clustering algorithms have generally focused on the production of disjoint clusters, making the assumption that the underlying structure of the data consists of k well-separated classes. However, in text corpora it is not unusual for a single document to relate to more than one topic. As stated previously, the provision of weights reflecting document-cluster associations and term relevance can be an important aid to domain users. Therefore, we propose a soft spectral clustering algorithm where each document and term may be associated with multiple clusters.

3.2 Related Techniques

For the task of generating feature weights from a hard clustering, a common approach is to derive values from each cluster's centroid vector, which can be viewed as providing a summary of the content of each cluster [**?**]. In the *spherical k-means* algorithm [**?**], term weights are extracted from the unit normalised centroid of each cluster. However, an analogous technique for spectral clustering is not feasible due to the presence of negative values in centroid vectors formed in the reduced space. Another possible approach is to consider the membership weights of a given document as being a function of the similarity between the document and each cluster centroid [**?**]. Documents that are highly similar to a particular cluster centroid will be assigned a high membership weight for that cluster, whereas documents that bear little similarity to the centroid will have a low weight. In the case of a co-clustering, we can also derive term membership weights in the same manner.

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 [**?**]. 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.

As an alternative to inducing soft weights from a hard partition, one may advocate the application of existing fuzzy clustering techniques. An analogous approach to spectral k-means clustering would involve the application of the fuzzy

c-means algorithm [**?**]. However, its effectiveness as a post-processing method for spectral document clustering is limited due to its reliance on a squared-norm metric to measure similarity and its inability to deal with outliers. The membership values produced by a fuzzy co-clustering of a spectral embedding will also be subject to the effects of the distortion described above. Taking these factors into account, we choose to focus on the derivation of soft clusters from a disjoint partition.

3.3 Inducing Soft Clusters

As a starting point, we construct a reduced space based on the spectral co-clustering model described in Section **??**. 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 [**?**]. 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, and we denote the k centroids of the clustering by $\{c_1, \ldots, c_k\}$.

As the spectral co-clustering strategy is based on the principle of the duality of clustering documents and terms [**?**], 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 $\mathbf{P_1}$ indicates the assignments of terms to clusters and $\mathbf{P_2}$ indicates the assignment of documents. An intuitive approach to producing term weights is to apply the transformation $\mathbf{A}\hat{\mathbf{P}}_\mathbf{2}$, where $\hat{\mathbf{P}}_\mathbf{2}$ denotes the matrix $\mathbf{P}_2$ 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 $\mathbf{A}^T\hat{\mathbf{P}}_\mathbf{1}$, thereby projecting the embedded term cluster centroids to the original data.

However, we observe that due to the "winner takes all" nature of the k-means algorithm, membership weights derived using 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:

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

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

3.4 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 $(\mathbf{U}, \mathbf{V})$, where $\mathbf{U}$ represents the term-cluster membership function and $\mathbf{V}$ represents the document-cluster membership function.

As an appropriate 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 consisting of overlapping classes.

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 [**?**]. Our choice is also motivated by the observation that the binary indicators in $\hat{\mathbf{P}}_\mathbf{2}$ result in sparse discriminative weight vectors, whereas the projection based on $\mathbf{S_2}$ leads to term weights where 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 $\mathbf{V} = \mathbf{A}^T\mathbf{S_1}$.

While the traditional approach to initialise the k-means algorithm is to randomly select data points as cluster centroids, we generate an initial clustering solution in a manner similar to that proposed in [**?**] to avoid instability. In this initialisation strategy, 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 approach to document clustering by dimension reduction that builds upon the co-clustering techniques described in Section **??** to produce a refined clustering that affords improved accuracy while retaining the interpretability of the clusters.

4.1 Motivation

The dimensions of the reduced space produced by spectral decomposition are constrained to be orthogonal. However, as text corpora will typically contain documents that pertain 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 reduction techniques such as NMF, where each document may be represented as an additive combination of the topics [**?**]. However, the standard approach of initialising the factors $(\mathbf{U}, \mathbf{V})$ with random non-zero values followed by the application of multiplicative update rules can lead to convergence to a range of solutions of varying quality. To address this issue, a novel strategy for seeding NMF was introduced in [**?**], where the centroids resulting from the *spherical k-means* algorithm were used to provide a good initialisation. The authors demonstrate that, by providing the factorisation algorithm with initial basis vectors that are as linearly independent as possible, the factorisation algorithm will consistently converge to an improved solution.

We argue that initial factors, produced using the soft cluster induction techniques discussed previously, can provide a good set of well-separated "core" clusters. By subsequently applying matrix factorisation with non-negativity constraints to the corresponding membership matrices, we can effectively uncover overlaps between clusters. The combination of the global information available to spectral techniques with the local nature of iterative matrix factorisation methods can yield accuracy superior to that achieved by either of the individual approaches. In addition, the sparseness of the factors produced by NMF improves our ability to identify outlying documents and eliminate irrelevant terms.

4.2 Refined Soft Spectral Co-clustering (RSSC) Algorithm

We now describe a procedure to refine the output of methods based on the soft spectral co-clustering model. In the SSC algorithm described in Section **??**,

our choice of projection for the construction of the term membership matrix was motivated by the desire to produce natural human-interpretable weights. However, the projection $\mathbf{AS_2}$ retains additional information from the embedded clustering in the form of the set of $n \times k$ normalised similarity values, while simultaneously considering the actual term frequencies in $\mathbf{A}$. Consequently, we apply soft spectral co-clustering as described previously, but select $\mathbf{V} = \mathbf{A}^T\mathbf{S_1}$ and $\mathbf{U} = \mathbf{AS_2}$ 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

$$D(\mathbf{A}||\mathbf{UV}^T) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij}\log\frac{A_{ij}}{[\mathbf{UV}^T]_{ij}} - A_{ij} + [\mathbf{UV}^T]_{ij}\right) \tag{3}$$

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. Compute the decomposition of $\mathbf{A}$ and apply the k-means algorithm to $\mathbf{Z}$ to produce a disjoint clustering, from which $\mathbf{S_1}$ and $\mathbf{S_2}$ are constructed.
2. Generate the initial factors $\mathbf{U} = \mathbf{AS_2}$ and $\mathbf{V} = \mathbf{A}^T\mathbf{S_1}$.
3. Update $\mathbf{V}$ using the rule

$$v_{ij} \leftarrow v_{ij}\left[\left(\frac{A_{ij}}{[\mathbf{UV}^T]_{ij}}\right)^T\mathbf{U}\right]_{ij} \tag{4}$$

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

$$u_{ij} \leftarrow u_{ij}\left[\frac{A_{ij}}{[\mathbf{UV}^T]_{ij}}\mathbf{V}\right]_{ij}, \quad u_{ij} \leftarrow u_{ij}\frac{U_{ij}}{\sum_{l=1}^{m}U_{lj}} \tag{5}$$

5. Repeat from step 3 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 [**?**,**?**].

$$V_{ij} \leftarrow V_{ij}\sqrt{\sum_{i=1}^{m}U_{ij}^2}, \quad U_{ij} \leftarrow \frac{U_{ij}}{\sqrt{\sum_{i=1}^{m}U_{ij}^2}} \tag{6}$$

5 Experimental Evaluation

In this section we compare the newly proposed algorithms to the existing clustering methods described in Section **??**. The objective of this evaluation is to demonstrate that the our techniques succeed in producing clusters that are both accurate and interpretable.

5.1 Experimental Setup

In our experiments we compared the accuracy of the SSC and RSSC algorithms with that of spectral co-clustering (CC) based on k singular vectors and NMF with the divergence objective function given in (**??**) 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 make use of a value for k corresponding to the number of "natural clusters" 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. The CLASSIC2, CLASSIC3 and CLASSIC datasets are collections of technical abstracts taken from Cornell's SMART repository[1], which have been widely used in information retrieval. We constructed the *bbc* corpus from 2225 complete news articles from the BBC corresponding to stories in five topical areas from 2004-2005. The *bbcsport* corpus consists of 737 sports news articles from the same source and time period[2]. The NG17-19 dataset is a subset of the well-known 20-Newsgroups corpus, consisting of three groups relating to politics that exhibit considerable overlap. The NG3 dataset is another subset derived from the same corpus, composed of three relatively well-separated groups pertaining to astronomy, politics and computer graphics. The remaining datasets are taken from the CLUTO toolkit[3]. For further details on these corpora consult [**?**].

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 was performed and no term normalisation function was used when constructing the term-document matrix $\mathbf{A}$.

Dataset	Description	Documents	Terms	Classes
bbc	News articles from BBC	2225	9635	5
bbcsport	Sports news articles	737	4613	5
classic2	CACM/CISI	4664	4932	2
classic3	CISI/CRAN/MED	3893	6733	3
classic	CACM/CISI/CRAN/MED	7097	8276	4
hitech	Technical news articles (TREC)	2301	10002	6
ng17-19	Overlapping newsgroups	2625	11841	3
ng3	Approximately disjoint newsgroups	2900	12875	3
re0	Subset of Reuters-21578	1504	2837	13
re1	Subset of Reuters-21578	1657	3704	25
reviews	Entertainment news articles (TREC)	4069	18391	5
tr31	Derived from TREC collections	927	10041	7
tr41	Derived from TREC collections	878	7373	10

Table 1. Experimental datasets.

[1] Available from ftp://ftp.cs.cornell.edu/pub/smart
[2] Both available from http://www.cs.tcd.ie/Derek.Greene/research/datasets.html
[3] Available from http://www.cs.umn.edu/∼karypis/cluto

5.2 Cluster Validation

When comparing the accuracy of the proposed clustering algorithms with existing techniques, we use external class information to assess cluster quality. A common external validation method is to employ mutual information to provide a robust indication of the shared information between a clustering and a set of natural classes. To produce a value in the range $[0, 1]$, we apply the *normalised mutual information* (NMI) measure proposed by Strehl and Ghosh [**?**]. Since this measure evaluates disjoint clusters, we produce a hard clustering from $\mathbf{V}$ by assigning the i-th document to cluster c_j if $j = \arg\max_j(v_{ij})$.

5.3 Discussion

Table 2 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 [**?**]. 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. Only in the case of the *hitech* dataset did the existing NMF method marginally out-perform our proposed algorithm, where the spectral-based initial factors lead the iterative optimisation procedure to converge to a poorer local minimum. When applied to larger datasets, we observe that the NMF and CC methods exhibit considerable variance in the quality of the clusters that they produce, whereas the deterministic nature of the initialisation strategy employed by the newly proposed algorithms leads to stable solutions.

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**
hitech	0.20	**0.25**	0.21	0.24
ng17-19	0.39	0.36	0.45	**0.50**
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**

Table 2. Performance comparison based on NMI.

In our implementation we take advantage of the sparse nature of text data to improve computational performance. We employ the Implicitly Restarted Lanc-

zos Method implemented by ARPACK [**?**], which allows the efficient computation of the leading singular vectors of a sparse matrix. However, due to the number of matrix multiplication operations involved in applying the iterative update rules, both the NMF and RSSC algorithms are computationally expensive. While we address these concerns somewhat by implementing NMF using sparse matrix multiplication techniques, we suggest that for larger datasets the SSC method provides reasonable accuracy coupled with sufficient interpretability to direct further processing.

5.4 Cluster Labels

Given the term membership weights produced by the SSC and RSSC algorithms, a natural approach to generating a set of human-readable 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 3. The natural categories are: technology, entertainment, sport, politics and business - the identification of which cluster corresponds to which topic is left to the reader.

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

Table 3. Labels produced by RSSC algorithm for *bbc* dataset.

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.

An issue that remains to be addressed is the selection of a suitable value for the number of clusters k. For practical applications we plan to consider several recently proposed techniques (e.g. [**?**]), where a value for k is estimated by examining the magnitude of the eigenvalues produced by spectral decomposition. In addition, as matrix factorisation has been previously used in gene expression analysis, we aim to investigate the application of our proposed methods to microarray data.

References

1. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press (1961)
2. Berry, M.W., Dumais, S.T., O'Brien, G.W.: Using linear algebra for intelligent information retrieval. SIAM Rev. **37** (1995) 573–595
3. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Knowledge Discovery and Data Mining. (2001) 269–274
4. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proc. Advances in Neural Information Processing. (2001)
5. Brand, M., Huang, K.: A unifying theorem for spectral embedding and clustering. In: Proc. 9th Int. Workshop on AI and Statistics. (2003)
6. Yu, S.X., Shi, J.: Multiclass spectral clustering. In: Proc. 9th IEEE Int. Conference on Computer Vision. (2003) 313
7. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–91
8. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proc. 26th Int. ACM SIGIR. (2003) 267–273
9. Karypis, G., Han, E.H.: Fast supervised dimensionality reduction algorithm with applications to document categorization and retrieval. In: Proc. 9th ACM Int. Conf. on Information and Knowledge Management, ACM Press (2000) 12–19
10. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Machine Learning **42** (2001) 143–175
11. 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, ACM Press (2004) 246–247
12. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
13. Wild, S., Curry, J., Dougherty, A.: Improving non-negative matrix factorizations through structured initialization. Pattern Recognition **37** (2004) 2217–2232
14. Osinski, S.: Dimensionality reduction techniques for search results clustering. Master's thesis, Department of Computer Science, University of Sheffield, UK (2004)
15. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis. Technical Report 01-040, University of Minnesota (2001)
16. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. JMLR **3** (2002) 583–617
17. Lehoucq, R.B., Sorensen, D.C., Yang, C.: ARPACK Users' Guide. Solution of large eigenvalue problems with implicitly restarted Arnoldi methods. SIAM (1997)
18. Li, W., Ng, W.K., Ong, K.L., Lim, E.P.: A spectroscopy of texts for effective clustering. In: Proc. 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy (2004) 301–312