Kernel Spectral Clustering of Scientific Journals Using a Data Integration Scheme

Carlos Alzate[a,*], Frizo Janssens[b,a], Wolfgang Glänzel[b,c],
Johan A. K. Suykens[a], Bart De Moor[a]

[a] *K.U. Leuven, SCD-SISTA, Dept. ESAT, Leuven (Belgium)*
[b] *K.U. Leuven, Centre for R&D Monitoring, Dept. MSI, Leuven (Belgium)*
[c] *Hungarian Academy of Sciences, IRPS, Budapest (Hungary)*

Abstract

An integrated approach for clustering scientific journals using spectral clustering is presented. Textual information is combined with cross-citation information in order to obtain a coherent grouping of the scientific journals and to improve over existing journal categorizations. Spectral clustering methods are powerful algorithms known to perform well in cases where classical methods such as k-means fail. This family of methods uses the eigenvectors of a Laplacian matrix derived from the data to obtain the grouping. In this work, a different approach to spectral clustering is used. The formulation is cast as a weighted form of kernel principal component analysis (PCA) where the clusters can be obtained through encoding and decoding schemes. Different data integration schemes of the two sources of information are studied including simple concatenation and integration at the kernel level. The performance of the results is assessed using internal and external validation criteria with respect to existing methods and well-known journal categorizations. Experimental results show that the proposed integration scheme

[*] Corresponding author.
Email address: carlos.alzate@esat.kuleuven.be (Carlos Alzate)

produces coherent clusters in terms of internal validation measures and compares favorably in terms of agreement with external validation criteria.
Keywords: Spectral clustering, scientific journal clustering, data integration.

1. Introduction

The idea of hybrid clustering is not new. Bibliometricians started combining co-citation analysis with the lexical approach to compensate the apparently low recall of co-citation analysis concerning the current work of specialties (Braam et al., 1991). Not much later the reverse way was suggested in France (e.g. (Zitt and Bassecoulard, 1994)) in order to allow contrasting the empirical characterization of trend with immediacy by means of co-citation analysis. Two years later, (Weiss et al., 1996) introduced the HyPursuit system, a hierarchical network search engine which exploits content-link clustering of hypertext documents for browsing and search activities. The complete link hierarchical agglomerative clustering method was used and the similarity between documents was based on the number of terms, ancestors, and descendants in common, as well as whether they point to one another. In the presented prototype, the integrated document similarity function was limited to selecting the maximal value of the text-based and hyperlink similarities, which already offered benefits to the hypertext document clustering.

(Modha and Spangler, 2000) introduced the toric k-means algorithm for clustering hypertext documents using words, out-links and in-links. The relative importance of these information sources was determined by searching the parameter space for an optimal figure-of-merit. Similarity was calculated as a weighted sum of the inner products between the individual text-based or link-based components.

(He et al., 2001, 2002) discussed Web document clustering by incorporating information from hyperlink structure, co-citation patterns, and textual contents of documents. The hyperlink structure was used as the dominant factor in the combined similarity measure, and the textual content was used to modulate the strength of each hyperlink. However, textual similarity between pages was neglected if both were not connected by a hyperlink. Integration with co-citation was achieved by a linear combination of co-citation and the weighted adjacency matrix of the graph. The resulting weighted graph was the input to a spectral clustering method.

(Joachims et al., 2001) combined kernel functions for text and co-citation in support vector machine classification of hypertext. (Wang and Kitsuregawa, 2002) evaluated a contents-link coupled clustering algorithm for retrieved web pages and studied the effect of out-links, in-links, specific terms, and their combination. Results suggested that both links and contents are important for Web page clustering and that better results are achieved with appropriate integration weights. By using a Bayesian network model, (Calado et al., 2006) combined link-based similarity measures with text-based classifiers to improve classification results for web collections . In their experiments on Web pages, the link information alone outperformed the text-only classifier, but the combination could improve results.

Finally, (Zhang et al., 2005) applied genetic programming techniques to discover the best fusion framework to integrate citation-based information and structural content in order to improve document classification. Except for the latter study, all mentioned related research was applied to hyperlink documents, while the present study is applied to scientific literature.

In a first study by authors related to the current work, the pilot study of (Glenis-

son et al., 2005b), further extended and confirmed (Glenisson et al., 2005a), full-text analysis and traditional bibliometric methods were serially combined to improve the efficiency of the individual methods. It was clear that clusters found through application of text mining provided additional information that could be used to extend and explain structures found by bibliometric methods, and vice-versa. However, the integration was still limited to serial combination. Real hybrid methodologies were therefore developed in subsequent studies and proved even more valuable tools to facilitate endeavours in mapping fields of science or technology.

In a previous study (Janssens et al., 2006), the concept structure of library and information science (LIS) was obtained by full-text mining of almost 1000 articles and notes published in the period $2002 - 2004$ in 5 representative journals with strong focus on information science. Only the pure text corpus was analyzed, excluding any bibliographic or bibliometric components. However, the authors have assessed the performance of clustering and classification algorithms using various data integration schemes on multiple data sets (Janssens, 2007). The best outcome was obtained by hybrid methods that exploited both text and citations. An integration method based on Fishers inverse chi-square and another one based on linear combination of distance matrices were among the best methods and significantly outperformed corresponding text-only and link-only methods, as well as a concatenation of vectors. In (Janssens et al., 2008), we used these methods to achieve an updated, hybrid mapping of information science by using the full-text information as well as citations, and we compared the results with the text-only clustering. The same techniques from text mining, Web mining and bibliometrics were also applied to a data set containing bioinformatics research articles (Janssens et al.,

2007).

Classical spectral clustering methods, (Fiedler, 1975; Chung, 1997; Shi and Malik, 2000) consist of relaxations of graph partitioning problems. These relaxations typically lead to eigenvalue problems of a Laplacian matrix derived from the data. The clusters can then be obtained from the eigenvectors of the Laplacian using different techniques such as recursive bipartitioning and reclustering, (Shi and Malik, 2000). Spectral clustering algorithms can solve clustering problems that are known to be difficult to standard methodologies such as k-means and hierarchical clustering. The present study compares the results obtained in another recent study, (Janssens et al., 2009), with new results based on spectral clustering. More specifically, it compares data integration at the kernel level with linear combination of similarity matrices. This is contrasted by (Janssens et al., 2008) and (Janssens et al., 2007), where a method based on Fishers inverse chi-square was used. Whereas the analyses of (Glenisson et al., 2005b,a) and (Janssens et al., 2006, 2008) were based on the full text extracted from scientific articles, the textual component of the current study and that of (Janssens et al., 2009) is based on titles, abstracts and keywords. Both analyses also differs from the previous studies in that they are situated on the journal instead of the paper level, i.e., here we cluster journals, not papers. The underlying paper set is also at least two orders of magnitude larger (over 6 million papers). Furthermore, the link component of the integration used is based on journal cross-citations, whereas bibliographic coupling was used in other studies (Janssens et al., 2008, 2007).

This paper is organized as follows. Section 2 describes the data and the textual and cross-citation information available. Section 3 contains a summary of classical spectral clustering methods. In Section 4, we describe the spectral clustering

method used in this study. In Section 5, several criteria to assess the clustering quality are described. Section 6 contains the experimental results and in Section 7, we give conclusions.

2. Data Sources

The data used in this work correspond to more than six million scientific papers indexed by the Web of Science (WoS) in the period $2002 - 2006$. The type of manuscripts considered is article, letter, note and review. Textual information has been extracted from titles, abstracts and keywords of each paper together with citation information. From these data, the resulting number of journals under consideration is $8,305$. The two resulting datasets contain textual and cross-citation information and are described as follows.

2.1. Term/Concept By Journal dataset

The textual information was processed using the term frequency - inverse document frequency (TF-IDF) weighting procedure (Baeza-Yates and Ribeiro-Neto, 1999). Terms which occur only in one document and stop words were not considered into the analysis. The Porter stemmer was applied to the remaining terms in the abstract, title and keyword fields. This processing leads to a term-by-document matrix of around six million papers and $669,860$ term dimensionality. The final journal-by-term dataset is a $8,305 \times 669,860$ matrix. Additionally, latent semantic indexing (LSI) (Deerwester et al., 1990) was performed on this dataset to reduce the term dimensionality to 200 factors. Typical values for the number of factors found in literature range from 100 (Deerwester et al., 1990) (for less than 2000 abstracts and 7000 terms) to about 300 (Berry et al., 1995) (for about 12.000 documents and 40.000 terms).

2.2. Journal cross-citation dataset

A different form of analyzing cluster information at the journal level is through a cross-citation graph. This graph contains aggregated citations between papers forming a journal-by-journal cross-citation matrix. The direction of the citations is not taken into account which leads to an undirected graph and a symmetric cross-citation matrix; intra-cluster self-citations are counted only once. At the level of journal clusters, the journal cross-citations can be further aggregated into inter-cluster citations. From the raw number of cross-citations C_{ij} between two journals (or clusters) i and j, a normalized similarity can be calculated as $\frac{C_{ij}}{\sqrt{\sum_k C_{ik} \sum_k C_{jk}}}$. For clustering, however, we calculated the similarity of two journals somewhat differently because we didnt want to ignore, for instance, that both journals could be highly cited by a third one. The similarities S_{ij} used for clustering were found by calculating the cosine of the angle between the pair of vectors containing all symmetric journal cross-citation values between the two respective journals (i and j) and all other journals (i.e., row or column of the matrix C):

$$S_{ij} = \frac{\sum_k C_{ik} C_{jk}}{\sqrt{\sum_k C_{ik}^2} \sqrt{\sum_k C_{jk}^2}} \tag{1}$$

Hence, the ultimate similarity of two journals is based on their respective similarities with all other journals. In the remainder of this article we will refer to these similarities as cosine similarities of cross-citation vectors.

3. Spectral Clustering

3.1. Classical formulation

In graph theory, a set of training data points $\mathcal{D} = \{x_i\}_{i=1}^{N}, x_i \in \mathbb{R}^d$ and a similarity (also called affinity) measure $s_{ij} \in \mathbb{R}^+$ between each pair of points x_i and x_j define an undirected graph if the similarity measure is symmetric $s_{ij} = s_{ji}$. The vertices of the graph are represented by the data points and the weight of the edges is determined by s_{ij}. A classical problem is then to split the graph into two disjoint parts by cutting edges with low weight. This problem is called graph bipartioning and it has been studied in (Fiedler, 1975; Chung, 1997). Several cut criteria have been proposed to obtain meaningful partitions of the graph. Typical criteria include the normalized cut *NCut* (Shi and Malik, 2000) and the ratio cut *RCut* (Hagen and Kahng, 1992). The multiway cut problem (also called k-way cut) consists of partitioning the graph into k disjoint sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ with $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \ldots \cap \mathcal{A}_k = \emptyset$ such that points in the same cluster are maximally similar and points in different clusters are maximally dissimilar. In general, these partitioning formulations are NP-hard due to the combinatorial nature of the problem. However, useful relaxations have been proposed in the literature (Fiedler, 1975; Chung, 1997; Shi and Malik, 2000) leading to eigenvalue problems where the clusters can be obtained from the eigenvectors.

The k-way normalized cut relaxation (Gu et al., 2001) can be defined as

$$\max_{G} \text{Tr}(G^T L G)$$
$$\text{such that } G^T G = I_k,$$

where $G \in \mathbb{R}^{N \times k}$ is a cluster indicator matrix, L is the symmetric normalized

Laplacian $L = D^{-1/2}SD^{-1/2}$ and I_k is the $k \times k$ identity matrix. This maximization problem is a special form of Fan's theorem (Fan, 1949). The optimal solution of the relaxed problem is given by $G^\star = BR_2$ where $B \in \mathbb{R}^{N\times k}$ is any orthonormal basis of the k-th principal subspace of L and $R_2 \in \mathbb{R}^{k\times k}$ is an arbitrary orthogonal matrix (Fan, 1949; Bach and Jordan, 2006). This means that the relaxed solution is contained on the eigenvectors corresponding to the k largest eigenvalues of the normalized Laplacian L but the eigenvectors are prone to arbitrary rotations.

3.2. Converting eigenvectors to clusters

Obtaining clusters from the eigenvectors is not a straightforward task due to the arbitrary rotations. In the ideal case, when the clusters are tight and well separated, the top eigenvectors of the non-symmetric normalized Laplacian $D^{-1}S$ display some structure known as the piecewise constant property (Meila and Shi, 2001a,b) for well-chosen similarity measures. If the eigenvectors are piecewise constant with respect to a partitioning, then input data points in the same cluster are mapped to a single point in the eigenspace. Thus, spectral clustering can be seen as a pre-processing step mapping to a new space where the points are easy to cluster. In this new space, a simple algorithm such as k-means can be applied to obtain the final clustering. This scheme is called reclustering (Shi and Malik, 2000) and corresponds to one of the most popular approaches to obtain clusters from eigenvectors. Reclustering using k-means works well if the clusters in the eigenspace are represented as single points or distributed spherically around the cluster prototype. Other schemes include rounding (Bach and Jordan, 2004), and the linearized cluster assignment (Ding and He, 2004).

4. Spectral Clustering via Weighted Kernel PCA

In this section, we summarize a multiway spectral clustering based on weighted kernel PCA introduced in (Alzate and Suykens, 2009a). This formulation allows the spectral clustering problem to be cast as a weighted version of kernel PCA where different clusters are obtained using encoding and decoding schemes.

4.1. Multiway Formulation

Given training data $\{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$, the multiway spectral clustering problem for k clusters can be written in the primal as:

$$\min_{w^{(l)},e^{(l)},b_l} \quad \frac{1}{2N}\sum_{l=1}^{k-1}\gamma_l e^{(l)^T} D^{-1} e^{(l)} - \frac{1}{2}\sum_{l=1}^{m} w^{(l)^T} w^{(l)} \tag{2}$$

$$\text{such that} \quad \begin{cases} e^{(1)} = \Phi w^{(1)} + b_1 1_N \\ \vdots \\ e^{(k-1)} = \Phi w^{(k-1)} + b_{k-1} 1_N \end{cases}$$

where $e^{(l)} = [e_1^{(l)}, \ldots, e_N^{(l)}]$ denotes the projected variables $e_i^{(l)} = w^{(l)^T}\varphi(x_i) + b_l$, $i = 1, \ldots, N$, the $N \times n_h$ matrix of feature maps is $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_N)^T]$, $\varphi(\cdot)$ is the mapping to a feature space of dimension n_h (which can be infinite dimensional), γ_l are regularization constants, D^{-1} is the inverse graph degree matrix, b_l are the bias terms and $l = 1, \ldots, k-1$. Each score variables vector $e^{(l)}$ represents a binary clustering where the cluster indicators can be obtained through sign$(e_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1$.

The Lagrangian of the constrained optimization problem (2) is:

$$\begin{aligned}\mathcal{L}(w^{(l)}, e^{(l)}, b_l; \alpha^{(l)}) =& \frac{1}{2N}\sum_{l=1}^{k-1}\gamma_l e^{(l)^T} D^{-1} e^{(l)} - \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} \\ & - \sum_{l=1}^{k-1}\alpha^{(l)^T}\left(e^{(l)} - \Phi^{(l)} w^{(l)} - b_l 1_N\right)\end{aligned} \tag{3}$$

where $\alpha^{(l)} \in \mathbb{R}^N, l = 1, \ldots, k-1$ are Lagrange multiplier vectors. Using the Karush-Kuhn-Tucker (KKT) optimality conditions $\partial\mathcal{L}/\partial w^{(l)} = 0, \partial\mathcal{L}/\partial e^{(l)} = 0, \partial\mathcal{L}/\partial b_l = 0, \partial\mathcal{L}/\partial\alpha^{(l)} = 0$, eliminating the primal variables $w^{(l)}, e^{(l)}, b_l$ and applying the kernel trick $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ leads to the following dual eigenvalue problem:

$$D^{-1}M_D\Omega\alpha^{(l)} = \lambda_l\alpha^{(l)}, \tag{4}$$

$l = 1, \ldots, k-1$, where $M_D = I_N - (1/1_N^T D^{-1} 1_N) 1_N 1_N^T D^{-1}$ is a weighted centering matrix, Ω is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j)$, $\lambda_l = N/\gamma_l$ and $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a Mercer kernel function which should be bounded. Note that, (4) is related to the random walks model (Meila and Shi, 2001b). This relationship can be interpreted as first removing a weighted mean[1] from the columns of the kernel matrix and then applying the random walks algorithm.

The projected variables $e^{(l)}$ can now be written in terms of the eigenvectors of (4):

$$e_i^{(l)} = w^{(l)^T}\varphi(x_i) + b_l = \sum_{j=1}^{N}\alpha_j^{(l)} K(x_j, x_i) + b_l$$

[1]The weights are given by $D^{-1}1_N$

where the bias terms b_l follow from the KKT conditions and become:

$$b_l = (1/1_N^T D^{-1} 1_N) 1_N^T D^{-1} \Omega \alpha^{(l)}.$$

4.2. Encoding

Once the eigenvectors $\alpha^{(l)}$ and projected variables $e^{(l)}$ have been computed, the encoding step combines the $k-1$ binary cluster indicators obtained from $e^{(l)}$ into a multiway clustering of k groups. For training data, the cluster indicators can also be obtained from the eigenvectors since $\text{sign}(\alpha_i^{(l)}) = \text{sign}(e_i^{(l)}), i = 1, \ldots, N, l = 1, \ldots, k-1$. Due to the bias terms introduced in (2), the leading eigenvector $\alpha^{(1)}$ already contains discriminatory information which is not the case in the random walks model. A codebook can then be obtained from the rows of the binarized eigenvector matrix $A = [\text{sign}(\alpha^{(1)}), \ldots, \text{sign}(\alpha^{(k-1)})]$ with $A \in \{-1, 1\}^{N \times (k-1)}$. Thus, the i-th training point x_i has an associated binary encoding represented by the i-th row of A. If the data are well-represented and the eigenvectors are piecewise constant then there will be exactly k different encodings which form the codebook. However, if the eigenvectors are approximately piecewise constant, then there will be spurious encodings. In this case, the codebook is formed by the k encodings with most occurrences.

4.3. Differences with classical spectral clustering

The kernel spectral clustering approach differs from classical spectral clustering methods with respect to the following aspects:

- **Bias terms:** Including a bias term in the primal problem (2) leads to a special centering of the data in the feature space. Another advantage induced by the bias terms follows from the KKT conditions and leads to zero mean

eigenvectors. This property avoids the determination of additional threshold values in order to binarize the eigenvectors for encoding.

- **Encoding:** The process of encoding the eigenvectors to obtain cluster indicators is inspired in multiclass support vector machine formulations (Suykens et al., 2002). Section 4.2 describes a particular binary encoding, but in general, the kernel spectral clustering framework allows the use of different encoding schemes.

- **Predictions:** The clustering model can be extended naturally to out-of-sample points by means of the projections onto the eigenvectors. This can be used for predictive purposes, model selection and generalization as shown in (Alzate and Suykens, 2009a). However in this work, we are not making use of the predictive out-of-sample extension.

- **Extendability:** Due to the primal and dual nature of the formulation, the core clustering problem can be easily extended towards advanced elements such as sparseness (Alzate and Suykens, 2008) and incorporation of prior knowledge (Alzate and Suykens, 2009b).

5. Validation and Parameter Selection

5.1. Internal Validation

Internal validation consists of assessing the cluster quality by measuring some distortion criterion on the resulting grouping. In this work, we use the following criteria:

- **Silhouette index** (Rousseeuw, 1987). This index ranges from -1 to 1 and measures how similar data points in the same clusters are to each other compared to data points in other clusters. The silhouette plot shows the value of the silhouette index for each data point as a horizontal line. The more the line is to the positive side of the plot, the more coherent this data point is to points in the same cluster. Negative values in the silhouette plot indicates that this data point should be rather assigned to a different cluster. The mean silhouette value (MSV) gives an average measure of the coherence of the groups and is defined as:

$$\text{MSV} = \frac{1}{N}\sum_{i=1}^{N}\frac{\min(f(i,\mathcal{C}_p)) - r(i)}{\max(r(i),\min(f(i,\mathcal{C}_p)))}, \tag{5}$$

  where $f(i,\mathcal{C}_p)$ is the average distance from the i-th data point to points in another cluster p and $r(i)$ is the average distance from the i-th data point to points in its own cluster.

- **Modularity** (Newman and Girvan, 2004; Newman, 2006). The modularity Q of a corresponding partition quantifies the clustering quality in terms of the number of intra-cluster connections minus the average number of connections in a graph with random connections. Well-formed clusters will have a high modularity value tending to 1. The modularity (up to a scaling factor) is defined as:

$$Q = \text{Tr}(G^T Y G) \tag{6}$$

  where G is a cluster indicator matrix and Y is the modularity matrix with

ij-entry $Y_{ij} = A_{ij} - M_{ij}$, A is the adjacency matrix

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge joining vertices } i, j, \\ 0 & \text{otherwise.} \end{cases}$$

and M_{ij} denotes the expected number of edges between vertices i and j.

5.2. Data Integration

The objective of data integration is to combine two sources of information such as text and cross-citations in order to obtain a better clustering which is coherent with the available data. We consider two types of data integration in this work: concatenation and linear combinations of the kernel (similarity) matrices.

5.2.1. Concatenation

The cross-citation dataset and the text/concept dataset are concatenated to form a joint dataset where each row represents a given journal with $669,860 + 8,305$ variables in case raw terms are used and $200 + 8,305$ variables in case the terms are preprocessed by LSI. After concatenation, the rows are re-normalized to have unitary norm.

5.2.2. Integration at the Kernel Level

The cross-citation and the text/concept datasets can also be integrated at the kernel level by considering the following linear combination of kernel matrices:

$$\Omega^{\text{integr}} = \rho\Omega^{\text{cross-cit}} + (1-\rho)\Omega^{\text{text}}$$

where $0 \leq \rho \leq 1$ is a user-defined integration weight which value can be obtained from internal validation measures for cluster distortion, $\Omega^{\text{cross-cit}}$ is the cross-citation kernel matrix with ij-th entry $\Omega_{ij}^{\text{cross-cit}} = K(x_i^{\text{cross-cit}}, x_j^{\text{cross-cit}})$, $x_i^{\text{cross-cit}}$ is the i-th journal represented in terms of cross-citation variables, Ω^{text} is the textual kernel matrix with ij-th entry $\Omega_{ij}^{\text{text}} = K(x_i^{\text{text}}, x_j^{\text{text}})$, x_i^{text} is the i-th journal represented in terms of textual variables and $i, j = 1, \ldots, N$.

5.3. External Validation

On the other hand, external validation consists of comparing the resulting groups with respect to a known partitioning. We use the following external validation criteria:

- **Jaccard similarity coefficient** The *Jaccard index* is the ratio of the cardinality of the intersection of two sets and the cardinality of their union. The *Jaccard similarity coefficient* is an extension of the *Jaccard* index and can be used as a measure for external cluster validation. It is used to compare a clustering result $\Delta = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k\}$ with an external partitioning $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_l\}$, where k and l represent the number of clusters and partitions, respectively. For Δ and $\mathcal{P}$ we define $n \times n$ matrices B^{Δ} and $B^{\mathcal{P}}$, where n is the total number of documents. Each Boolean value B_{ij}^{Δ} indicates whether documents i and j belong to the same cluster in Δ, while B^{P} indexes all documents that are in the same partition in $\mathcal{P}$.

  Let N_{00} represent the number of pairs of documents that do not belong to the same cluster in Δ, nor in $\mathcal{P}$, i.e., the number of elements (i, j) for which $B_{ij}^{C} = B_{ij}^{\mathcal{P}} = 0$. Likewise, N_{01} counts the number of elements for which $B_{ij}^{\Delta} = 0$ and $B_{ij}^{\mathcal{P}} = 1$. N_{10} and N_{11} are defined analogously.

The *Jaccard* coefficient is then defined as follows:

$$J(\Delta, \mathcal{P}) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}}. \tag{7}$$

The resulting value between 0 and 1 quantifies the correlation between the two binary matrices, while disregarding negative agreements (N_{00}).

- **Rand index (RI)** (Jain and Dubes, 1988). The Rand index is another external validation measure to quantify the correspondence between a clustering outcome Δ and a ground-truth categorization $\mathcal{P}$. In contrast to the Jaccard coefficient, the Rand index does take into account negative matches as follows:

$$\text{RI}(\Delta, \mathcal{P}) = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}}. \tag{8}$$

  The result is also a value between 0 and 1, with 1 indicating that Δ and $\mathcal{P}$ are identical.

- **Adjusted Rand index (ARI)** (Hubert and Arabie, 1985; Yeung and Ruzzo, 2001). A problem with the Rand index is that the expected value for the agreement between two random partitions is not a constant value and can be rather high. It assumes the generalized hypergeometric distribution as a model of randomness to compute the expected value $E\big(\text{RI}(\Delta, \mathcal{P})\big)$. The adjusted Rand index is then computed as:

$$\text{ARI}(\Delta, \mathcal{P}) = \frac{\text{RI}(\Delta, \mathcal{P}) - E\big(\text{RI}(\Delta, \mathcal{P})\big)}{\max\big(\text{RI}(\Delta, \mathcal{P})\big) - E\big(\text{RI}(\Delta, \mathcal{P})\big)}. \tag{9}$$

  The use of the adjusted Rand index for external validation has been recom-

mended in (Milligan and Cooper, 1986).

- **Normalized mutual information (NMI)** (Strehl and Ghosh, 2002). The mutual information is a symmetric measure to quantify the statistical information shared between an obtained clustering and a known partitioning used as a ground truth. Given the obtained clustering Δ and a known partitioning represented by $\mathcal{P}$, the NMI can be defined as:

$$\text{NMI} = 2\frac{I(\Delta, \mathcal{P})}{H(\Delta)H(\mathcal{P})} \tag{10}$$

  where $I(\cdot, \cdot)$ is the mutual information and $H(\cdot)$ is the entropy.

5.4. Parameter Selection

In this particular scenario, the model selection consists of determining the optimal integration weight ρ. Since we use a normalized linear kernel (which is equivalent to the cosine similarity measure) $K(x, z) = x^T z/(||x||||z||)$, there are no kernel parameters to tune. The number of clusters is considered fixed to 22 to be able to compare the proposed approach to other existing methodologies. In order to tune the integration weight ρ, we use the mean silhouette value (MSV) as an internal cluster validation criterion to select the value of ρ which gives more coherent clusters. The MSV is computed on the eigenvectors since they represent the data in the eigenspace where the cluster decisions take place.

6. Experimental Results

Figure 1 shows the silhouette curves for the 22 ESI fields in terms of textual and cross-citation information. According to the silhouette index, clusters

$2, 4, 7, 19$ and 21 seem not to be coherent enough from both a cross-citation and a textual perspective. This complies with a a former study (Janssens et al., 2009), where this existing intellectual subject-classification scheme was evaluated and compared with a hybrid text/citation-based clustering method. It was substantiated that field classification according to the ESI scheme could still be improved by automatic clustering. In this paper, we compare different hybrid clustering methods in order to find a consistent grouping of the scientific journals with an even better agreement to the available data.

Table 1 shows the 22 broad science ESI fields. Figure 2 shows the selection of the integration weight ρ. The mean silhouette value is computed on the eigenvectors of the integrated LSI+cross-citation results for $k = 22$ while varying the integration weight ρ. The optimal integration weight according to the silhouette value is $\rho = 0.5$. This means that both textual and cross-citation information are equally contributing in the integration scheme. The spectral clustering results are shown in Table 2 for several integration methods together with a hybrid hierarchical clustering algorithm (denoted as hierarchical Ward's method) using a linear combination of textual and citation information (Janssens, 2007). The proposed integrated spectral clustering methods outperform in terms of internal validation with the integrated MSV and in terms of external validation with the NMI. Ward's method wins in modularity for both internal and external criteria and the concatenation approaches outperform the rest in terms of the ARI. Table 3 shows the best 20 terms per cluster according to the proposed LSI+cross-citation algorithm displaying a coherent structure. Figure 3 shows Jaccard indices measuring similarities between the integrated spectral clustering method (LSI+cross-citation) with respect to the hybrid algorithm and to the 22 ESI fields. Figure 4 shows a multidi-

mensional scaling (MDS) plot of the cluster centroids with respect to the 22 ESI fields. Some similarities between the two clusterings are clearly visible from the plot, e.g. cluster 6 with the ESI field psychology/psychiatry, cluster 20 with the ESI field chemistry and clusters $2, 8$ with the ESI field economics/business. However, cluster 3 appears clearly isolated in the MDS plot. The best terms in cluster 3 reflects a more *nutrition*-oriented cluster. Silhouette plots based on integrated data are shown in Figure 5. This plot shows that the proposed method outperforms the 22 ESI fields (see Figure 1) in terms of the silhouette index with respect to the combined textual and cross-citation information. With respect to computation times, the proposed kernel spectral clustering methods take in average 17 seconds in a Dual Opteron 250 with 4 GB of RAM using GNU Linux, MATLAB 7.9 and the Lanczos method for eigenvalue decomposition. Ward's method takes 9 seconds to produce clustering results on the same machine.

7. Conclusions

An application of kernel spectral clustering towards finding groups of similar scientific journals is presented. Different data integration schemes such as simple concatenation and integration at the kernel level are studied. The method used in this work corresponds to a primal and dual approach to kernel spectral clustering where the cluster information is contained in the eigenvectors of a Laplacian matrix derived from the data. The proposed methodologies are compared in terms of several cluster quality criteria with respect to well-known journal categorizations. The experimental results indicate that the proposed integrated approaches produce clusters that are more coherent with respect to the available data.

Acknowledgments

This work was supported by grants and projects for the Research Council K.U.Leuven (GOA-Mefisto 666, GOA-Ambiorics, several PhD / Postdocs & fellow grants), the Flemish Government Centre for R&D Monitoring (ECOOM), FWO: PhD / Postdocs grants, projects G.0240.99, G.0211.05, G.0407.02, G.0197.02, G.0080.01, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0226.06, G.0302.07, ICCoS, ANMMM; AWI;IWT:PhD grants, GBOU (McKnow) Soft4s, the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/01/40), the EU(FP5-Quprodis, ERNSI, Eureka 2063-Impact;Eureka 2419-FLiTE) and Contracts Research/Agreements (ISMC/IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard). Johan Suykens is a professor at K.U.Leuven, Belgium. Bart De Moor is a full professor at K.U.Leuven, Belgium. The scientific responsibility is assumed by its authors.

References

Alzate, C., Suykens, J. A. K., 2008. Sparse kernel models for spectral clustering using the incomplete Cholesky decomposition. In: Proc. of the 2008 International Joint Conference on Neural Networks (IJCNN 2008). pp. 3555–3562.

Alzate, C., Suykens, J. A. K., 2009a. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. IEEE Transactions on Pattern Analysis and Machine Intelligence. In Press.

Alzate, C., Suykens, J. A. K., 2009b. A regularized formulation for spectral clustering with pairwise constraints. In: Proc. of the 2009 International Joint Conference on Neural Networks (IJCNN 2009). pp. 141–148.

Bach, F. R., Jordan, M. I., 2004. Learning spectral clustering. In: Thrun, S., Saul, L. K., Schölkopf, B. (Eds.), Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA.

Bach, F. R., Jordan, M. I., 2006. Learning spectral clustering, with application to speech separation. Journal of Machine Learning Research 7, 1963–2001.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley.

Berry, M. W., Dumais, S. T., Obrien, G. W., 1995. Using linear algebra for intelligent information retrieval. SIAM Review 37 (4), 573–595.

Braam, R., Moed, H., van Raan, A., 1991. Mapping of science by combined co-citation and word analysis. i. structural aspects. Journal of the American Society of Information Science 42 (4), 233–251.

Calado, P., Cristo, M., Gonçalves, M. A., de Moura, E. S., Ribeiro-Neto, B., Ziviani, N., 2006. Link-based similarity measures for the classification of web documents. Journal of the American Society for Information Science and Technology 57 (2), 208–221.

Chung, F. R. K., 1997. Spectral Graph Theory. American Mathematical Society.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A., 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 (6), 391–407.

Ding, C., He, X., 2004. Linearized cluster assignment via spectral ordering. In: ICML 2004: Proceedings of the twenty-first International Conference on Machine Learning. ACM Press, New York, NY, USA, p. 30.

Fan, K., 1949. On a theorem of Weyl concerning eigenvalues of linear transformations. Proc. Natl. Acad. Sci. USA 35 (11), 652–655.

Fiedler, M., 1975. A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory. Czech. Math. J. 25 (100), 619–633.

Glenisson, P., Glänzel, W., Janssens, F., De Moor, B., 2005a. Combining full text and bibliometric information in mapping scientific disciplines. Information Processing & Management 41 (6), 1548–1572.

Glenisson, P., Glänzel, W., Persson, O., 2005b. Combining full-text analysis and bibliometric indicators. A pilot study. Scientometrics 63 (1), 163–180.

Gu, M., Zha, H., Ding, C., He, X., Simon, H., 2001. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering. Technical report, Penn. State Univ, Computer Science and Engineering.

Hagen, L., Kahng, A., 1992. New spectral methods for ratio cut partitioning algorithms. IEEE Trans. Computer-Aided Design 11 (9), 1074–1085.

He, X., Ding, C. H. Q., Zha, H., Simon, H. D., 2001. Automatic topic identification using webpage clustering. In: ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, pp. 195–202.

He, X., Zha, H., Ding, C., Simon, H. D., 2002. Web document clustering using hyperlink structures. Computational Statistics & Data Analysis 41 (1), 19–45.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification, 193–218.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice Hall.

Janssens, F., 2007. Clustering of scientific fields by integrating text mining and bibliometrics. Ph.D. thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium.

Janssens, F., Glänzel, W., De Moor, B., 2007. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007). ACM, New York, NY, USA, pp. 360–369.

Janssens, F., Glänzel, W., De Moor, B., 2008. A hybrid mapping of information science. Scientometrics 75 (3), 607–631.

Janssens, F., Leta, J., Glänzel, W., De Moor, B., 2006. Towards mapping library and information science. Information Processing & Management 42 (6), 1614–1642.

Janssens, F., Zhang, L., De Moor, B., Glänzel, W., 2009. Hybrid clustering for validation and improvement of subject-classification schemes. Information Processing & Management 45 (6), 683 – 702.

Joachims, T., Cristianini, N., Shawe-Taylor, J., 2001. Composite kernels for hypertext categorisation. In: ICML ’01: Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 250–257.

Meila, M., Shi, J., 2001a. Learning segmentation by random walks. In: Leen, T. K., Dietterich, T. G., Tresp, V. (Eds.), Advances in Neural Information Processing Systems 13. MIT Press.

Meila, M., Shi, J., 2001b. A random walks view of spectral segmentation. In: Artificial Intelligence and Statistics AISTATS.

Milligan, G. W., Cooper, M. C., 1986. A study of the comparability of external criteria for hierarchical cluster-analysis. Multivariate Behavioral Research 21 (4), 441–458.

Modha, D. S., Spangler, W. S., 2000. Clustering hypertext with applications to web searching. In: HYPERTEXT ’00: Proceedings of the eleventh ACM on Hypertext and Hypermedia. ACM Press, New York, NY, USA, pp. 143–152.

Newman, M. E. J., 2006. Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA 103 (23), 8577–8582.

Newman, M. E. J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E 69 (2).

Rousseeuw, P. J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20 (1), 53–65.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 22 (8), 888–905.

Strehl, A., Ghosh, J., 2002. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific, Singapore.

Wang, Y., Kitsuregawa, M., 2002. Evaluating contents-link coupled web page clustering for web search results. In: CIKM '02: Proceedings of the eleventh international Conference on Information and Knowledge Management. ACM Press, New York, NY, USA, pp. 499–506.

Weiss, R., Vélez, B., Sheldon, M. A., 1996. Hypursuit: a hierarchical network search engine that exploits content-link hypertext clustering. In: HYPERTEXT '96: Proceedings of the the seventh ACM conference on Hypertext. ACM, New York, NY, USA, pp. 180–193.

Yeung, K. Y., Ruzzo, W. L., 2001. Principal component analysis for clustering gene expression data. Bioinformatics 17 (9), 763–774.

Zhang, B., Chen, Y., Fan, W., Fox, E. A., Goncalves, M. A., Cristo, M., Calado, P., 2005. Intelligent fusion of structural and citation-based evidence for text classification. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York, NY, USA, pp. 667–668.

Zitt, M., Bassecoulard, E., 1994. Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis. Scientometrics 30 (1), 333–351.

Figure Captions

- Figure 1: Silhouette plots of textual and cross-citation data with respect to the 22 ESI fields clustering. Note that, journals in clusters $2, 4, 7, 19$ and 21 have a lot of negative values which means they are inconsistent with the data according to the silhouette index.

- Figure 2: Parameter selection - Tuning ρ for the integrated results (LSI+cross-citation). The optimal integration weight ρ is 0.5 according to the mean silhouette value measured on the eigenvectors.

- Figure 3: Visual Jaccard Index. Agreement of the proposed integrated scheme with respect to the 22 ESI clustering and the Ward's method with linearly combined distances. **Top:** Clusters $2, 4, 8, 12, 20$ of the proposed approach are very similar to clusters $21, 11, 6, 12, 3$ in the 22 ESI indicators respectively. **Bottom:** Clusters $6, 15, 16, 17, 19, 20$ of the proposed clustering agree with clusters $21, 18, 11, 19, 20, 4$ of Ward's method respectively.

- Figure 4: Multidimensional scaling (MDS) map of the cluster centroids of the 22 ESI fields with respect to the proposed clustering scheme (LSI+cross-citations). A cluster centroid is defined as the linear combination of all documents in the cluster. The crosses indicate the cluster centroids for the proposed method. Similarities between the different clusterings are e.g. cluster 6 with the ESI field psychology/psychiatry, cluster 20 with the ESI field chemistry and clusters $2, 8$ with the ESI field economics/business.

- Figure 5: Kernel spectral clustering results - Silhouette plots. **Top:** Silhouette index with respect to textual information (LSI). The mean silhouette value is 0.145. **Center:** Silhouette index with respect to cross-citations. The mean silhouette value is 0.062. **Bottom:** Silhouette index with respect to integrated information. Mean

silhouette value is 0.197. Note the improvements compared to Figure 1 in terms of cluster quality according to the silhouette index.

Field	Name	Field	Name
1	Agricultural sciences	12	Mathematics
2	Biology and biochemistry	13	Microbiology
3	Chemistry	14	Molecular biology & genetics
4	Clinical medicine	15	Multidisciplinary
5	Computer science	16	Neuroscience & behavior
6	Economics and business	17	Pharmacology & toxicology
7	Engineering	18	Physics
8	Environment/Ecology	19	Plant & animal science
9	Geosciences	20	Psychology / Psychiatry
10	Immunology	21	Social sciences
11	Materials sciences	22	Space science

Table 1: The 22 science fields according to the essential science indicators (ESI)

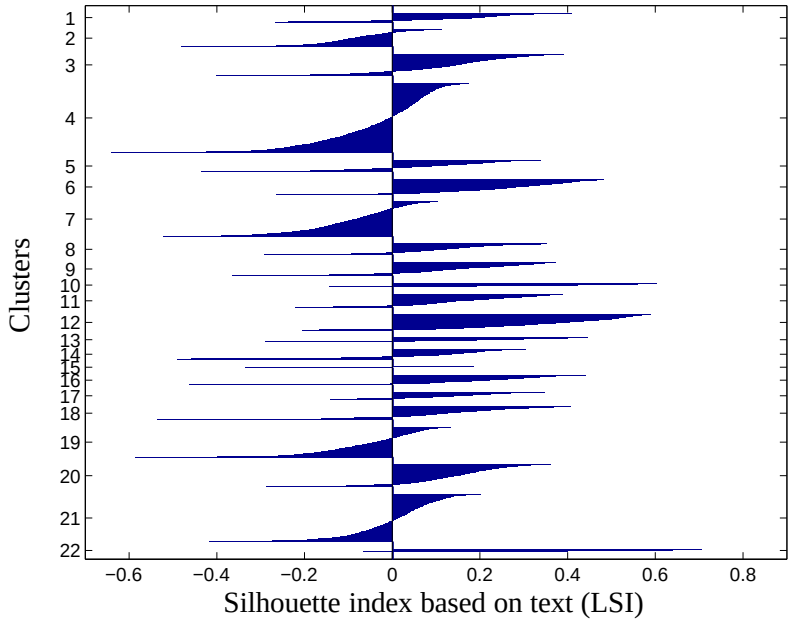
	Internal validation				External validation		
	MSV textual	MSV cross-cit.	MSV integrated	Modularity cross-cit.	Modularity ISI 254	ARI 22 ESI	NMI 22 ESI
22 ESI fields	0.057	0.016	0.063	0.475	0.526*	1.000	1.000
Cross-citations	0.093	0.057	0.189	**0.547**	0.442	0.278	0.516
Textual (LSI)	0.118	0.035	0.130	0.505	0.451	0.273	0.516
Hierarch. Ward's method $\rho = 0.5$	0.121	0.055	0.190	**0.547**	**0.488**	0.285	0.540
Concat. Terms+Cross-citations	**0.153**	0.055	0.191	0.528	0.457	*0.325*	0.556
Concat. LSI+Cross-citations	**0.153**	0.049	0.185	0.520	0.446	**0.329**	0.552
Integr. Terms+Cross-citations $\rho = 0.5$	0.138	**0.064**	**0.201**	*0.533*	*0.465*	0.294	*0.557*
Integr. LSI+Cross-citations $\rho = 0.5$	*0.145*	*0.062*	*0.197*	0.527	*0.465*	0.308	**0.560**

Table 2: Spectral clustering results of several integration methods in terms of mean Silhouette value (MSV), modularity, adjusted Rand index (ARI) and normalized mutual information (NMI). The first four rows correspond to existing clustering results used for comparison. The last four rows correspond to the proposed spectral clustering algorithms. For external validation, the clustering results are compared with respect to the 22 ESI fields and the ISI 254 subject categories. The highest value per column is indicated in bold while the second highest value appears in italic. For MSV, a standard t-test for the difference in means revealed that differences between highest and second highest values are statistically significant at the 1% significance level (p-value $< 10^8$). The selected method for further comparisons is the integrated LSI+Cross-citations approach since it wins in external validation with one highest value (NMI) and one second highest value (Modularity).

	Best 20 terms
Cluster 1	diabet therapi hospit arteri coronari physician renal hypertens mortal syndrom cardiac nurs chronic infect pain cardiovascular symptom serum cancer pulmonari
Cluster 2	polit war court reform parti legal gender urban democraci democrat civil capit feder discours economi justic privat liber union welfar
Cluster 3	diet milk fat intak cow dietari fed meat nutrit fatti chees vitamin ferment fish dry fruit antioxid breed pig egg
Cluster 4	alloi steel crack coat corros fiber concret microstructur thermal weld film deform ceram fatigu shear powder specimen grain fractur glass
Cluster 5	infect hiv vaccin viru immun dog antibodi antigen pathogen il pcr parasit viral bacteri dna therapi mice bacteria cat assai
Cluster 6	psycholog cognit mental adolesc emot symptom child anxieti student sexual interview school abus psychiatr gender attitud mother alcohol item disabl
Cluster 7	text music polit literari philosophi narr english moral book essai write discours philosoph fiction ethic poetri linguist german christian religi
Cluster 8	firm price busi trade economi invest capit tax wage financi compani incom custom sector bank organiz corpor stock employ strateg
Cluster 9	nonlinear finit asymptot veloc motion stochast elast nois turbul ltd vibrat iter crack vehicl infin singular shear polynomi mesh fuzzi
Cluster 10	soil seed forest crop leaf cultivar seedl ha shoot fruit wheat fertil veget germin rice flower season irrig dry weed
Cluster 11	soil sediment river sea climat land lake pollut wast fuel wind ocean atmospher ic emiss reactor season forest urban basin

	Best 20 terms
Cluster 12	algebra theorem manifold let finit infin polynomi invari omega singular inequ compact lambda graph conjectur convex proof asymptot bar phi
Cluster 13	pain surgeri injuri lesion muscl bone brain ey surgic nerv mri ct syndrom fractur motor implant arteri knee spinal stroke
Cluster 14	rock basin fault sediment miner ma tecton isotop mantl volcan metamorph seismic sea magma faci earthquak ocean cretac crust sedimentari
Cluster 15	web graph fuzzi logic queri schedul semant robot machin video wireless neural node internet traffic processor retriev execut fault packet
Cluster 16	student school teacher teach classroom instruct skill academ curriculum literaci learner colleg write profession disabl faculti english cognit peer gender
Cluster 17	habitat genu fish sp forest predat egg nest larva reproduct taxa bird season prei nov ecolog island breed mate genera
Cluster 18	star galaxi solar quantum neutrino orbit quark gravit cosmolog decai nucleon emiss radio nuclei relativist neutron cosmic gaug telescop hole
Cluster 19	film laser crystal quantum atom ion beam si nm dope thermal spin silicon glass scatter dielectr voltag excit diffract spectra
Cluster 20	polym catalyst ion bond crystal solvent ligand hydrogen nmr molecul atom polymer poli aqueou adsorpt methyl film spectroscopi electrod bi
Cluster 21	receptor rat dna neuron mice enzym genom transcript brain mutat peptid kinas inhibitor metabol cancer mrna muscl ca2 vitro chromosom
Cluster 22	cancer tumor carcinoma breast therapi prostat malign chemotherapi tumour surgeri lesion lymphoma pancreat recurr resect surgic liver lung gastric node

Table 3: Best 20 terms per cluster according to the integrated results (LSI+cross-citation) with $\rho = 0.5$. The terms found display a coherent structure in the clusters.


Clusters
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
Silhouette index based on text (LSI)

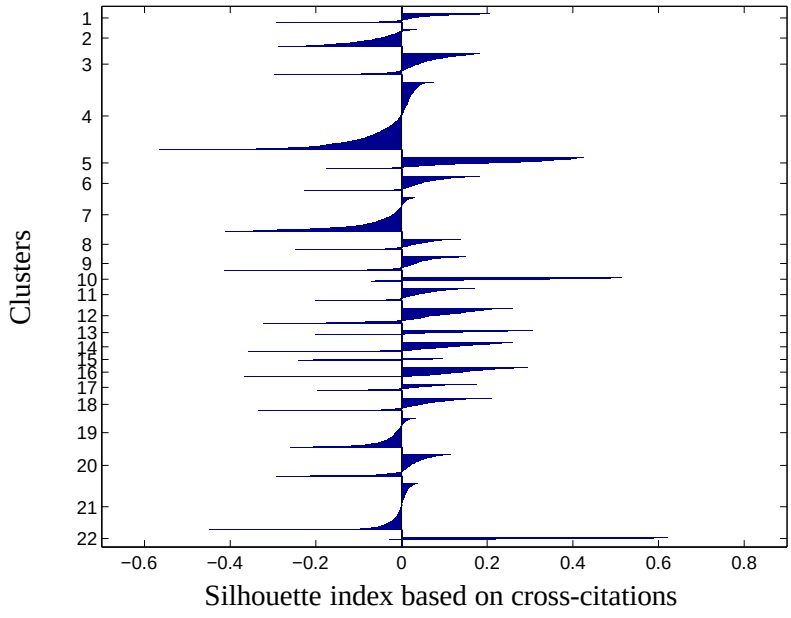

Clusters
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
Silhouette index based on cross-citations

Figure 1:

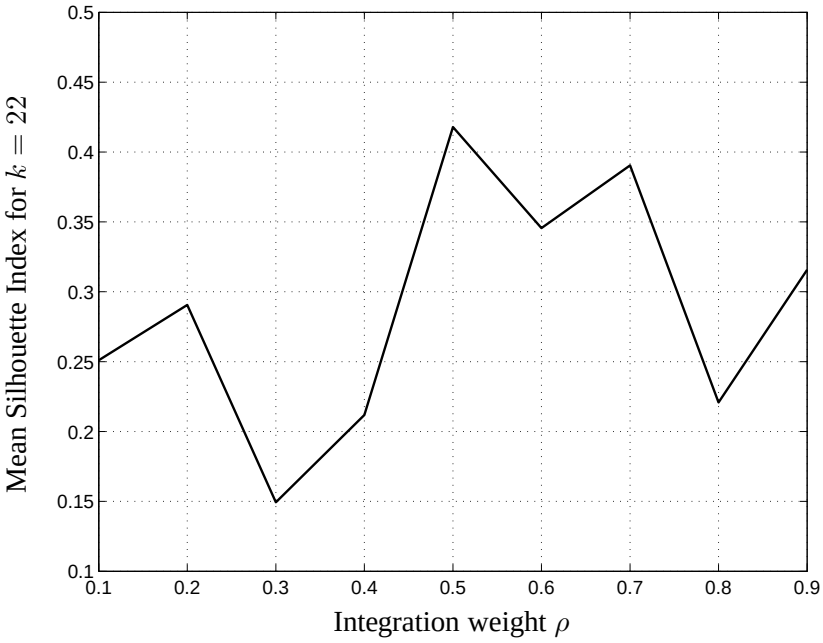
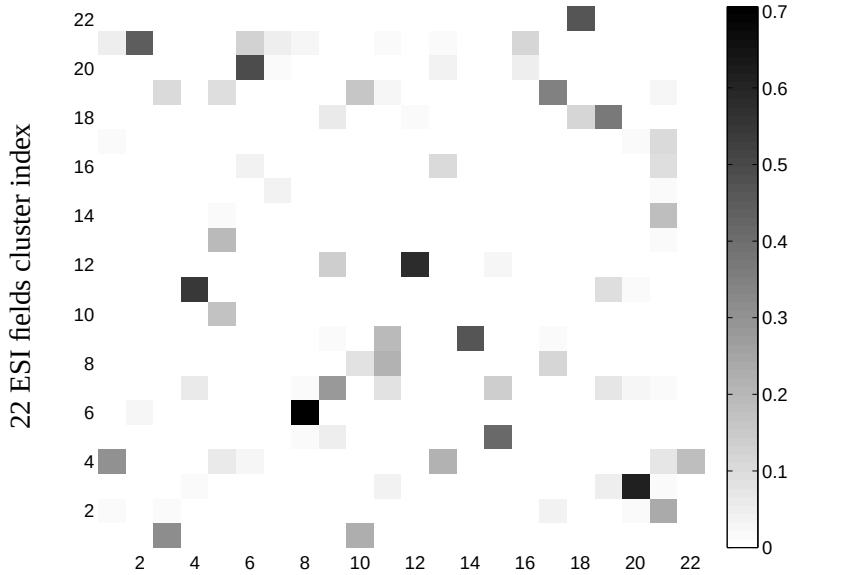
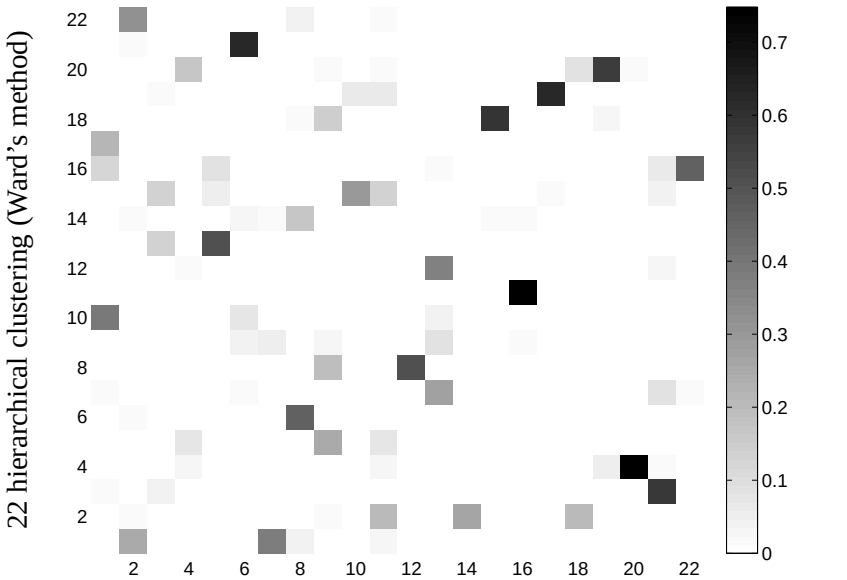

Mean Silhouette Index for $k = 22$
Integration weight ρ
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure 2:



Integrated clustering (LSI+Cross-citations) with $\rho = 0.5$ and $k = 22$



Integrated clustering (LSI+Cross-citations) with $\rho = 0.5$ and $k = 22$

Figure 3:

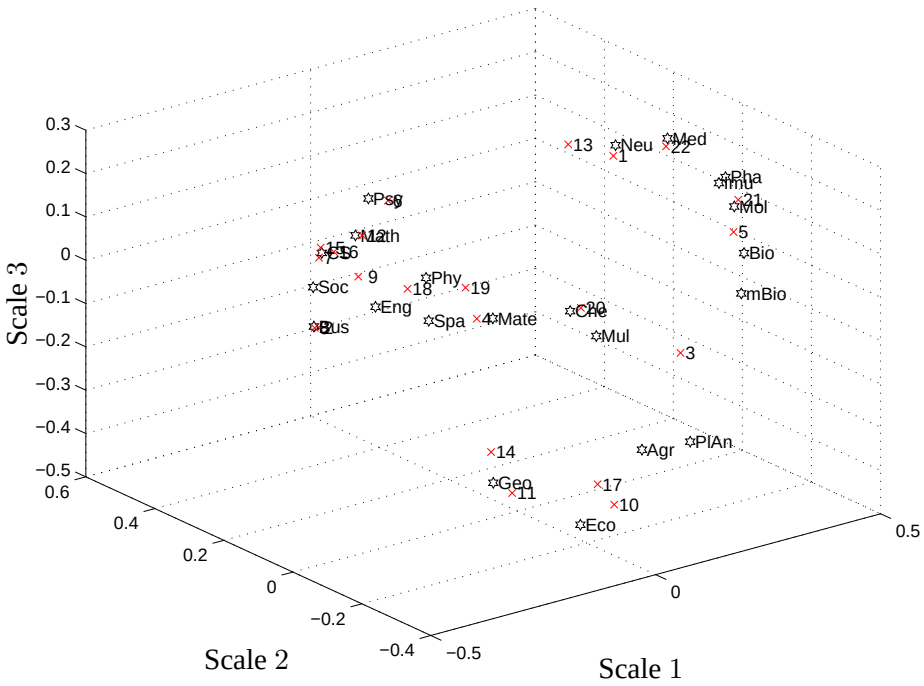

Scale 3
Scale 2
Scale 1
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
0.6
0.4
0.2
0
−0.2
−0.4
−0.5
0
0.5
13
Neu
Med
22
1
Pha
Imu
21
Mol
5
Bio
mBio
Psy
Math
9
Phy
18
19
Soc
Eng
Spa
4
Mate
Che
20
Mul
3
14
Agr
PlAn
Geo
11
17
10
Eco

Figure 4:

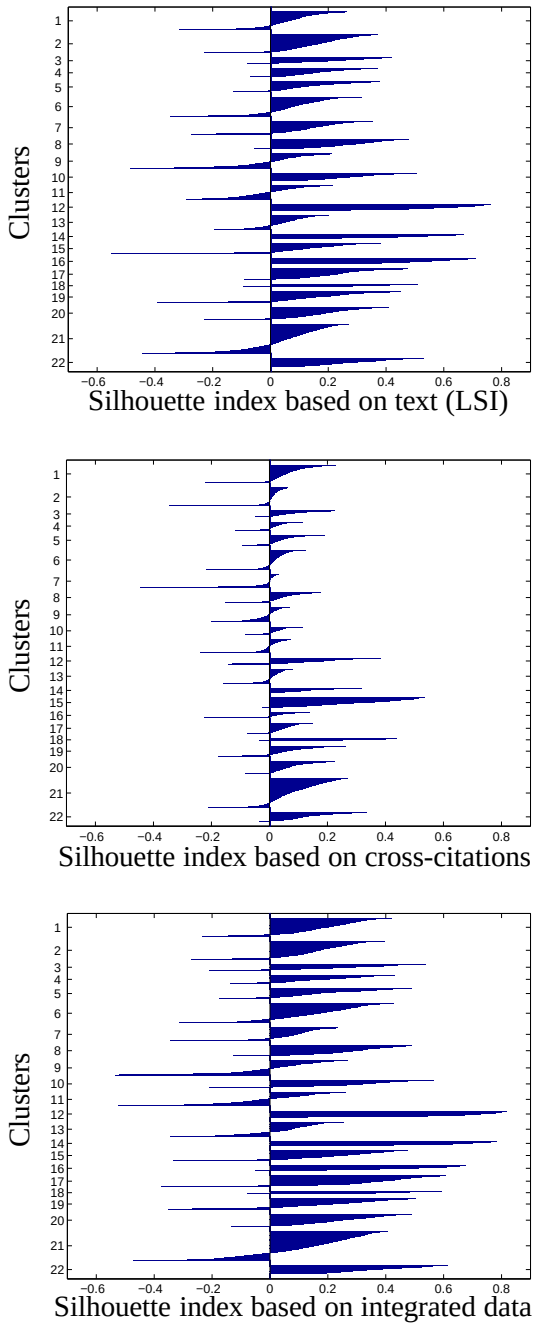

Clusters
Silhouette index based on text (LSI)
Clusters
Silhouette index based on cross-citations
Clusters
Silhouette index based on integrated data

Figure 5: