Scientometrics (2011) 88:819–839
DOI 10.1007/s11192-011-0348-3

Hybrid clustering of multi-view data via Tucker-2 model and its application

Xinhai Liu · Wolfgang Glänzel · Bart De Moor

Received: 15 January 2011 / Accepted: 27 January 2011 / Published online: 3 June 2011

Abstract With the modern technology fast developing, most of entities can be observed by different perspectives. These multiple view information allows us to find a better pattern as long as we integrate them in an appropriate way. So clustering by integrating multi-view representations that describe the same class of entities has become a crucial issue for knowledge discovering. We integrate multi-view data by a tensor model and present a hybrid clustering method based on Tucker-2 model, which can be regarded as an extension of spectral clustering. We apply our hybrid clustering method to scientific publication analysis by integrating citation-link and lexical content. Clustering experiments are conducted on a large-scale journal set retrieved from the Web of Science (WoS) database. Several relevant hybrid clustering methods are cross compared with our method. The analysis of clustering results demonstrate the effectiveness of the proposed algorithm. Furthermore, we provide a cognitive analysis of the clustering results as well as the visualization as a mapping of the journal set.

Keywords Hybrid clustering · Multi-view data · Text mining · Bibliometric analysis

X. Liu (✉)
College of Information Science and Engineering & ERCMAMT, Wuhan University of Science and Technology, Heping Road No. 947, Wuhan 30081, Hubei, China
e-mail: xinhai.liu@esat.kuleuven.be

X. Liu · B. D. Moor
Department of Electronic Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

W. Glänzel
Center for R&D Monitoring (ECOOM), Department of MSI, Katholieke Universiteit Leuven, Waaistraat 6, 3000 Leuven, Belgium

W. Glänzel
IRPS, Hungarian Academy of Sciences, Budapest, Hungry

Introduction

Clustering is a fundamental task for scientific publication analysis because it can understand the impact of individual authors and who is collaborating with whom, estimate the type of information being published and by which venues, extract "hot topics" and predict trends. Hybrid clustering refers to the clustering of the same class of entities with multi-view representations, either from various information sources or from different feature generators (we call this kind of data multi-view data in this research). As modern computer technology develops, multi-view data abounds in a wide variety of real applications. For example, two types of data are often employed in journal database analysis: textual content and citation link, both of which describe the same journal entities but contain heterogeneous information. On the one hand, textual content provides rich information but contains much noise, for example, many overlapping terms exist among unrelated documents, which will cause "over-clustering", that is, some un-related documents might be clustered into the same group. On the other hand, citation link is informative as well but might be incomplete, for instance, some citations do not exist among the related documents due to the limitation of reference numbers and the citation time, which will lead to the "weak-clustering", that is, some related documents can not be grouped into the same cluster. Meanwhile, these two types of data are not entirely independent and actually they are closely correlated and supplement each other. Hybrid clustering of both data might provide a nice mapping of journal sets (Glenisson et al. 2005; Boyack and Klavans 2010).

Historically, bibliometric researchers have focused solely on citation analysis or text analysis (Small 1973; Callon et al. 1983; Braam et al. 1991a), but not both simultaneously. Meanwhile, clustering algorithms are originally designed for single-view data, such as hierarchical clustering (Jain and Dubes 1988), *k*-means clustering and spectral clustering (Luxburg 2007).

The integration of lexical similarities and citation links has also attracted interest in the scientific documents analysis (Braam et al. 1991a, b). Recently, several hybrid clustering algorithms by integrating multiple information sources have been proposed for Web and Bibliometric analysis. Modha and Spangler (2000) integrated similarity matrices from terms, out-links and in-links by a weighted linear combination, and the data partition was obtained from the combined similarity matrix using the toric *k*-means algorithm. He et al. (2002) incorporated three types of information (hyperlink, textual, and cocitation information) to cluster Web documents using a graph-cut algorithm. Bickel and Scheffer (2004) investigated Web documents and combined intrinsic views (page content) with extrinsic views (anchor texts of inbound hyperlinks). Three clustering algorithms (generic expectation-maximization [EM], *k*-means, and agglomerate) were applied to combine the different views as hybrid clustering. With the exception of Web page analysis, Glenisson et al. (2005) combined textual analysis and bibliometrics to improve the performance of journal publication clustering. Janssens (2007) proposed an unbiased combination of textual content and citation links on the basis of Fisher's inverse chi-square for agglomerative clustering. Liu et al. (2009) reviewed some popular hybrid clustering techniques within a unified computational framework and proposed an adaptive kernel *k*-means clustering (AKKC) algorithm to learn the optimal combination of kernels constructed from heterogeneous data sources. Our method shares the same flavor with AKKC regarding hybrid clustering, but AKKC only works for the medium or small-scale database. Liu et al. (2010) advanced the hybrid clustering method by putting forward a weighted hybrid clustering scheme which leverages the effect of multiple data sources in hybrid clustering based on information measurement. Tang et al. (2009) introduced a linked matrices

factorization method to combine citation and text information and applied it to the SIAM publication analysis. These hybrid clustering solutions might sound natural and can even achieve better performance, but the underlying principle is not clear. For instance, most algorithm are based on the combination of different similarity matrices (kernels) by linear averaging without further analysis, which would ignore the discriminating capability of each perspective. So even though the research of hybrid clustering has recently received considerable attention, it still seems to be at an early stage

Increasingly, tensors are becoming common in modern applications dealing with multi-way data. For instance, tensor is a natural model to integrate multi-view data (Kolda and Bader 2006); a tensor method named multi-linear singular value decomposition (MLSVD) can provide several meaningful matrix factors and some tensor methods can even offer an upper error boundary for the approximation analysis (Ding et al. 2008; De Lathauwer et al. 2000a). Tensors have been successfully applied to several domains, such as chemometrics (Smilde et al. 2004), signal processing (De Lathauwer and Vandewalle 2004; Comon 1994), Web search (Dunlavy et al. 2006; Kolda and Bader 2006) and data mining (Savas and Eldén 2007). Several tensor based strategies have been proposed to integrate multiple information sources for the analysis of scientific publications. Sun et al. (2006) introduced a dynamic tensor analysis (DTA) method and applied multi-way latent semantic indexing to the analysis of DBLP data. Dunlavy et al. (2006) applied CANDECOMP/PARAFAC (CP) decomposition for analyzing scientific publication data with multiple linkages. Selee et al. (2007) created a new tensor decomposition method called Implicit Slice Canonical Decomposition (IMSCAND) to group scientific publications of SIAM with multiple similarities. The last two methods are based on the tensor method of CP decomposition, but compared with MLSVD, the decomposed factors are un-orthogonal, which might be unsatisfactory for the partitioning requirement.

Spectral clustering has become a strong competitor for other clustering methods. Spectral methods are appealing to researchers because they are easy to implement and reasonably fast. Also they do not intrinsically suffer from the problem of local optima. Spectral clustering has been widely employed in many real applications, from image segmentation to community detection (Luxburg 2007). Although spectral clustering works well in single-view data, it is not well suited to the presentation of multi-view data. Therefore, by extending spectral clustering to handle multi-view data, we propose a hybrid clustering method based on a tensor method of Multi-linear Singular Value Decomposition (HC-MLSVD), which leverages the inherent consistency of multi-view data and integrates their information seamlessly. The conceptual overview of our hybrid clustering methods to integrate citation-links and lexical content is presented in Fig. 1. We also illustrate its comparison with spectral clustering of single-view data (both text and citation). Obviously, with tensor analysis, our HC-MLSVD can be regarded as a multi-view extension of spectral clustering.

We apply our method to the science publication analysis of a large-scale Web of Science (WoS) journal database and cross compare our method with alternative clustering methods. We provide a scientific mapping of the WoS journal database by integrating text and citation information. The visualization and cognitive analysis of the hybrid clustering result are presented as well.

The rest of the paper is organized as follows. "Spectral clustering" section introduces the concepts of spectral clustering. "Hybrid clustering by MLSVD" section presents our hybrid clustering based on MLSVD (HC-MLSVD). "Application" section illustrates the application on the WoS journal database and analyzes the clustering results from the bibliometric view. Finally, we conclude in "Conclusion and outlook" section.

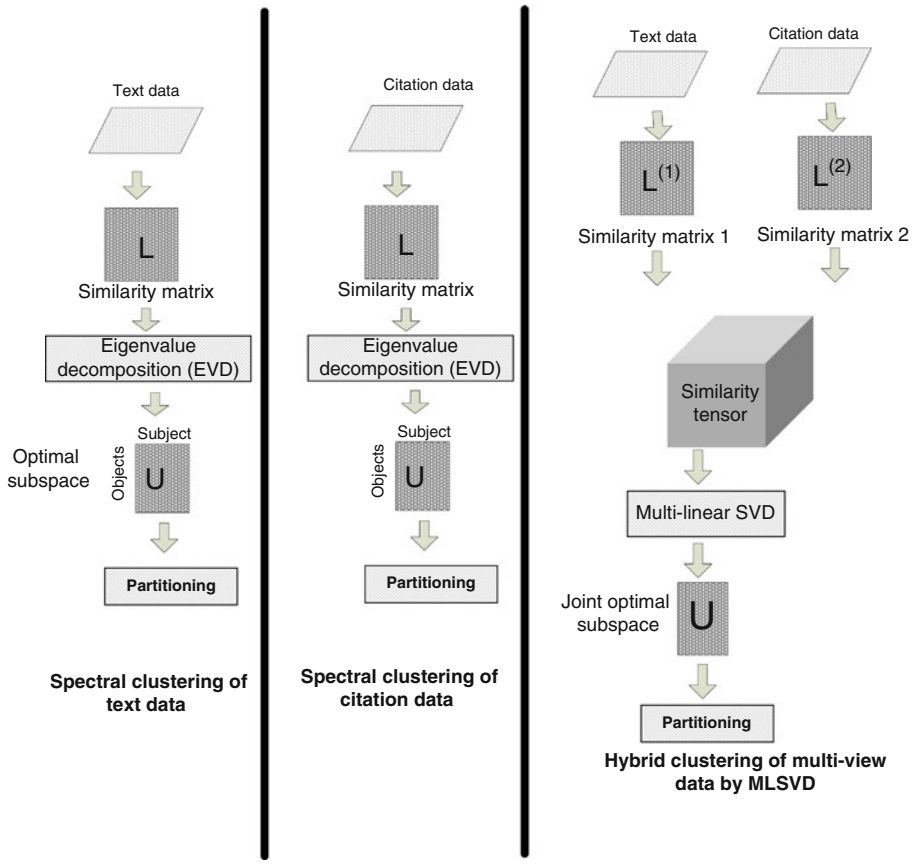

Fig. 1 Conceptual overview of spectral clustering and our hybrid clustering of multi-view data

Spectral clustering

Given a set of N data points $\{x_i\}_{i=1}^N$, $x_i \in R^d$ and some similarity measure $s_{ij} \geq 0$ between each pair of points x_i and x_j, an intuitive form of representing the data is using a graph $G = (V, E)$. The vertices V represent the data points and the edge $e_{ij} \in \varepsilon$ between two vertices v_i, v_j has a weight determined by s_{ij}. If the similarity measure is symmetric, then the graph is undirected. The affine matrix of the graph G is the matrix S with ij-entry $S_{ij} = s_{ij}$. The degree of vertex v_i is the sum of all the vertex weights adjacent to v_i and is defined as

$$d_i = \sum_{i=1}^{N} s_{ij}. \tag{1}$$

The degree matrix D is a diagonal matrix containing the vertex degrees $d_1, \ldots, d_N$ on the diagonal.

Akin to the formulation of spectral clustering (Ng and Jordan 2001), we define our similarity matrix as,

$$L = D^{-1/2} S D^{-1/2} \tag{2}$$

Suppose the relaxing indicator matrix $U \in \mathbb{R}^{N \times M}$ and M denotes the number of clusters, the optimization of our spectral clustering is

$$\begin{aligned} &\max_{U} \operatorname{tr}(U^T L U), \\ &\text{s.t. } U^T U = I, \end{aligned} \tag{3}$$

where tr refers to the trace function. The row of U can be taken as the new coordinates of each data point in the optimal subspace by spectral projection. Our spectral clustering formulation has close relationship with Normalized Cut (*NCut*) based spectral clustering (Shi and Malik 2000). The related NCut Laplacian matrix is defined as

$$L_{\text{NCut}} = D^{-1/2}(D - S)D^{-1/2}. \tag{4}$$

Then the *NCut* based spectral clustering can be formulated as

$$\begin{aligned} &\min_{U} \operatorname{tr}(U^T L_{NCut} U), \\ &\text{s.t. } U^T U = I. \end{aligned} \tag{5}$$

According to (4) and (5), we can get

$$L = I - L_{NCut}, \tag{6}$$

it is obvious that L shares the same set of eigenvectors as L_{NCut} and the eigenvalue relationship between L and L_{NCut} is $\lambda^{(L)} = 1 - \lambda^{(L_{NCut})}$.

For the convenience of tensor based optimization later, we prefer to adopt the Frobenius norm based optimization. Consider the following optimization,

$$\begin{aligned} &\max_{U} \|U^T L U\|_F^2, \\ &\text{s.t. } U^T U = I, \end{aligned} \tag{7}$$

if L is positive (semi)definite, the objective functions in (3) and (7) are different but happen to have their optima under the same matrix U, whose columns span the dominant eigenspace of U (Lay 2003; Yu 2009). Since the similarity matrix L is guaranteed to be positive (semi)definite (Luxburg 2007), the spectral clustering defined in (3) can be alternatively formulated by the optimization in (7).

Both (3) and (7) are only able to handle the spectral clustering of single-view data. Hence, we need to formulate the hybrid clustering of multi-view data.

Hybrid clustering by MLSVD

This section provides notation and minimal background on tensors and tensor methods adopted in this research. We refer readers to the relevant literature (Kolda and Bader 2009) for a more comprehensive review on tensors. The formulation of our algorithm is composed of three basic steps: (1) Building up the similarity tensor from multiple similarity matrices; (2) Tensor decomposition by MLSVD to get an optimal object subspace; (3) Final partition of the object subspace to obtain the cluster labels.

Basic conceptions of tensors

Tensor is a multidimensional array. The order of a tensor is the number of modes (or ways). A first-order tensor is a vector, a second order tensor is a matrix and a tensor of order three or higher is called a higher-order tensor. We only investigate 3-order tensor methods that are relevant to our problem.

The n-mode matrix unfolding Matrix unfolding is the process of re-ordering the elements of a 3-way array into a matrix. The n-mode ($n = 1, 2, 3$) matrix unfoldings of a tensor $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$ are denoted by $A_{(1)}, A_{(3)}$ and $A_{(3)}$ separately. For example, the matrix unfolding $A_{(1)}$ is a matrix with the number of rows I and the number of its columns is the product of dimensionalities of all other modes, that is, $J \times K$.

The n-mode product For instance, the 1-mode product of a tensor $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$ by a matrix $H \in \mathbb{R}^{I \times P}$, denoted by $\mathcal{A} \times_1 H$, is a $(P \times J \times K)$-tensor of which the entries are given by

$$(\mathcal{A} \times_1 H)_{pjk} = \sum_i a_{ijk} h_{ip}. \tag{8}$$

The analogous definitions are for 2-mode and 3-mode products.

MLSVD Multi-linear singular value decomposition (MLSVD) is a form of higher-order principle component analysis (PCA), which is also called Tucker decomposition or HOSVD (Tucker 1964, 1966; De Lathauwer et al. 2000a). It decomposes a tensor into a core tensor multiplied by a matrix along each mode. In the three-way case where $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$, we have

$$\mathcal{A} = \mathcal{S} \times_1 U \times_2 V \times_3 W, \tag{9}$$

where $U \in \mathbb{R}^{I \times I}, V \in \mathbb{R}^{J \times J}$ and $W \in \mathbb{R}^{K \times K}$ are called factor matrices or factors and can be thought of as the principle components of the original tensor along each mode. The factor matrices U, V and W are assumed to be column-wise orthonormal. The tensor $S \in \mathbb{R}^{I \times J \times K}$ is called the core tensor and its elements show the level of interaction between different components. According to De Lathauwer et al. (2000a), given a tensor $\mathcal{A}$, its matrix factors U, V and W as defined in (9) can be calculated as the left singular vectors of its matrix unfoldings $A_{(1)}, A_{(2)}$ and $A_{(3)}$ respectively. The tensor approximation by truncating the decomposition is named truncated MLSVD.

Integrating multi-view data by a tensor

As aforementioned in the introduction section, multi-view data can be modelled as a tensor naturally. The formulation of a tensor is a key step to devise our hybrid clustering algorithm. There are several options for this formulation, for example, Huang et al. (2008) formulated a tensor from different object-by-feature matrices. That formulation is only applicable to the scenario of homogeneous data sources, where the dimensionalities of different feature space are completely same. Taking into account the fact that text data and citation data are heterogeneous, we formulate the tensor from our multi-view data in the following way. Suppose a tensor $\mathcal{A}$ is built from several similarity matrices $\{L^{(1)}, L^{(2)}, \ldots, L^{(K)}\}$ as the frontal slices, instead of the object-feature matrices, the first and the second dimensions I and J of the tensor $\mathcal{A}$ are equal to the dimensions of the similarity matrices $L^{(i)}, (i = 1, \ldots, K)$, and its third dimension K is equal to the number of multiple

views (different similarity matrices). Such a formulation of a similarity tensor is illustrated in Fig. 2.

Additionally, because the similarity strength of each view is measured in various feature spaces, the normalization of each similarity matrix is required. Actually, our definition of similarity matrix in (2) could be regarded as a normalization step.

Solving hybrid clustering by MLSVD

The formulation of hybrid clustering

Given multiple similarity matrices from various data sources, we can integrate them for joint analysis. The simple way is to link their spectral optimization together

$$\begin{aligned} &\max_{U} \sum_{i=1}^{K} \|U^T L^{(i)} U\|_F^2, \\ &\text{s.t. } U^T U = I, \end{aligned} \tag{10}$$

which belongs to a global optimization but also takes the optimization of each single-view data into account. Therefore, although the global optimal U may not the best for each single-view data, it is good average while close to optimal of each single view. This simultaneous matrices analysis can be solved by tensor decomposition.

Suppose a similarity tensor $\mathcal{A}$ is built from similarity matrices $L^{(i)} \in \mathbb{R}^{N \times N}$ $(i = 1, \ldots, K)$ as illustrated in Fig. 2. According to the Frobenius matrix norm of tensors (De Lathauwer et al. 2000a),

$$\sum_{i=1}^{K} \|U^T L^{(i)} U\|_F^2 = \|\mathcal{A} \times_1 U^T \times_2 U^T\|_F^2, \tag{11}$$

Consequently, the hybrid clustering of multi-view data can be formulated as tensor based optimization,

$$\begin{aligned} &\max_{U} \|\mathcal{A} \times_1 U^T \times_2 U^T\|_F^2, \\ &\text{s.t. } U^T U = I. \end{aligned} \tag{12}$$

The above optimization can be understood as a truncated tensor decomposition of Tucker-2 model, which merges one factor in (9) with core tensor (Tucker 1966; Phan and

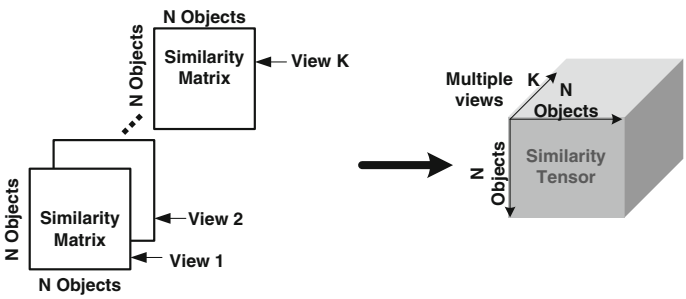

Fig. 2 The formulation of a similarity tensor

Cichocki 2010). The relevant tensor decomposition is shown in Fig. 2, where the columns of U can be regarded as the joint optimal subspace shared by multi-view data.

The solution of MLSVD

In fact, although the optimization in (12) is based on Tucker-2 model, it still can be approximated by MLSVD. Hence we call our strategy hybrid clustering by MLSVD (HC-MLSVD). Because projection by MLSVD on the dominant higher-order singular vectors usually gives a good approximation of the given tensor (De Lathauwer et al. 2000a), taking the columns of U equal to the dominant 1-mode singular vectors is expected to yield a large value of the objective function in (12). The dominant 1-mode singular vectors of U are equal to the dominant left singular vectors of $A_{(1)}$. The truncated MLSVD obtained this way, can be regarded as an approximation solution of (12). However, MLSVD usually performs well, and the algorithm is simple to implement and quite efficient. Moreover, there exists an upper bound on the approximation error (De Lathauwer et al. 2000a). The optimization of (12) can also be solved by other tensor methods, such as higher-order orthogonal iteration (HOOI) which offers an optimal solution (De Lathauwer et al. 2000b), but it will bring computation and memory burden.

Final partition

Regarding spectral clustering, in general, k-means clustering is employed to partition the obtained optimal subspace. The drawbacks of k-means is obvious: normally finding a local optimal and highly sensitive to the initial algorithm. To circumvent the shortcoming of k-means, we adopt other two partition methods that are believed to obtain better performance: k-means++ (Arthur and Vassilvitskii 2006) and hierarchical clustering based on Ward's linkage (HC-WL) (Jain and Dubes 1988).

k-means++: By augmenting k-means with a simple, randomized seeding technique, k-means++ is an algorithm that is $O(log\ k)$-competitive with the optimal clustering.
HC-WL: Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Here we adopt the agglomerative strategy that is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Ward's linkage uses the incremental sum of squares; that is, the increase in the total within-cluster sum of squares as a result of joining two clusters.

With this two partition methods, the final clustering result is unique, which is different from k-means.

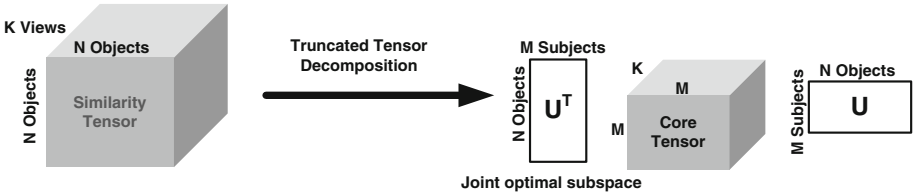

Fig. 3 Hybrid clustering by tensor decomposition of Tucker-2 model

Application

In this section, we apply our algorithms in the real application of the WoS journal set analysis. Our objective is to map these journals into different subjects by clustering algorithms. Recently, many researchers have applied text mining and citation analysis to the journal set analysis. The integration of lexical and citation information is a promising strategy towards better mappings (Janssens et al. 2009; Liu et al. 2009).

Journal data set

The journal data set is retrieved from a database containing more than 8,305 journals covered by the Web of Science (WoS) from the year 2002 till 2006. These journals are assigned to 22 fields in Thomson Reuters, referred to as Essential Science Indicator (ESI) fields.[1] We will employ ESI as a gold standard reference for clustering evaluation. We retrieve lexical and citation information from the those journals.

Text Mining Data The titles, abstracts and keywords of journal publications are indexed by a Jakarta Lucene based text mining program without any controlled vocabulary. The index result contains 9,473,061 terms and we cut the Zipf curve of terms at the head and the tail to remove the rare terms, stopwords and common words. After Zipf cut, 669,860 meaningful terms are kept as the attribute representations in vector space model and the weights of terms are calculated by the weighting scheme of TF-IDF. The publication-by-term vector is then aggregated to journal-by-term vector to construct the lexical data.

Bibliometric Data We investigate the cross-citations links among the selected journals. Citations among individual papers are aggregated to the journal level. We ignore the direction of citations links by symmetrizing the cross-citation matrix. Each row of this cross-citation matrix can be taken as a (citation-)link vector of the corresponding journal.

We apply our algorithms to combine two-view data (text and citation) and cluster the journals into 22 partitions. From text mining data, we take TF-IDF feature while from bibliometric data, we adopt cross-citation (CRC) feature.

Graph construction for spectral partition

For both views, we adopt Cosine similarity by normalizing each feature vector and computing the inner product between corresponding feature vectors. Given two vectors of attributes, x_i and x_j, the Cosine similarity, is represented using a dot product and magnitude as

$$s(x_i, x_j) = \cos(\theta) = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|}. \tag{13}$$

Regarding spectral clustering, we need to transform the given data points with pairwise similarity into a graph and model the local neighborhood relationships between data points (Luxburg 2007). In other words, the similarity should be local enough (more block-like) so as to carry out spectral analysis. The advantage of local similarity is that it decays rapidly, thus the eigenspectrum is more obvious. As known, Cosine similarity is not a typical local similarity and we need to make it local to fit in with the requirement of graph based spectral clustering. Therefore, the following four local similarities will be investigated.

[1] http://www.esi-topics.com/fields/index.html.

Gaussian similarity with normalized Euclidean distance: Gaussian similarity is a local similarity. Given a set of data points with pairwise similarities s_{ij}, the Gaussian similarity is defined as $s(x_i, x_j) = exp(-\|x_i - x_j\|^2/(2\delta^2))$, where the parameter δ controls the width of the neighborhoods. Since the text feature of TFIDF is aggregated from paper level to journal level , which needs to be normalized, we can not directly use the Euclidean distance of data points (journals) to compute the Gaussian similarity (The same to the cross-citation feature). Therefore, we normalized the row of each data vector and compute the Gaussian similarity with the normalized Euclidean distance. This local similarity can be formulated as

$$s(x_i, x_j) = exp\left(-\| \frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|} \|^2/(2\delta^2)\right), \tag{14}$$

where we choose $\delta = 1$.

Gaussian similarity with Cosine distance: Cosine distance of pairwise data points equals one minus their Cosine similarity. We can employ Cosine distance to formulate Gaussian similarity,

$$s(x_i, x_j) = exp(-(1 - \cos(\theta))/(2\delta^2)). \tag{15}$$

Usually this local similarity is not sensitive to δ, so we put $\delta = 0.5$.

Cosine similarity with k -nearest neighbors (KNN): KNN is an efficient way to make the neighborhood of data points local, so we combine it with Cosine similarity. We only connect x_i with x_j when x_i is the k-nearest neighbor of x_j or x_j is the k-nearest neighbor of x_i. Here we select the k as 40 by empirical test.

Cosine similarity with cross-citation link: All the existing local strategies which can make the fully connected similarity local rely on a certain parameter which needs tuning to reach the optimal, so we seek a non-parameter local strategy in our application: only if there is cross-citation link between pairwise journals, we connect the two journals with edge strength same to the related Cosine similarity. Compared with other three local similarities, this local similarity is specific to our application.

As a result, the new similarity (affinity) matrix on each view is built up and a similarity tensor can be formulated accordingly.

Baseline hybrid clustering methods

In this and next Section, we will try to cross compare our tensor based clustering method and the other three baseline hybrid clustering methods which represent the hybrid strategies from different perspectives.

- Multiple kernel fusion (MKF): Joachims et al. (2001) integrated different kernels by linear combination for hybrid clustering. The similarity matrix defined in (2) can be regarded as a linear kernel as well.
- Strehl's clustering ensemble algorithm (SA): Strehl and Ghosh (2002) formulated the optimal consensus as the partition that shares the most information with the partitions to combine. Three heuristic consensus algorithms (cluster-based similarity partition, hyper-graph partition and meta-clustering) based on graph partitioning are employed to obtain the combined partition.

– CP-ALS: CANDECOMP/PARAFAC (CP) decomposition (Kolda and Bader 2009) is a tensor decomposition strategy that decomposes the original tensor into some rank-1 tensor and can be employed to obtain the optimal subspace as well. we adopt alternating least squares (ALS) algorithm to computer CP.

Clustering evaluation measure

We adopt the following three clustering validation measures to evaluate our clustering results.

– Normalized mutual information (NMI) (Strehl and Ghosh 2002): Mutual information is a symmetric measure to quantify the statistic information shared between two distributions. Let $\{c_i\}_{i=1}^n$ and $\{l_i\}_{i=1}^n$ be the set of indicators and the ground truth labels, respectively. The normalized mutual information is defined as:

$$NMI = \frac{2 \times H(\{c_i\}, \{l_i\})}{H(\{c_i\})H(\{l_i\})}, \tag{16}$$

where $H(\{c_i\}, \{l_i\})$ is the mutual information between $\{c_i\}_{i=1}^n$ and $\{l_i\}_{i=1}^n, H(\{c_i\})$ and $H(\{l_i\})$ are the entropy of indicators and labels. NMI is a measure between 0 and 1. NMI = 1 when two clusters are exactly the same.

– Mean silhouette value (MSV): The silhouette value of a clustered object measures its similarities with the objects within the cluster versus the objects outside of the cluster (Rousseeuw 1987), given by:

$$S(i) = \frac{min(B(i, C_j)) - W(i)}{max\{min(B(i, C_j), W(i))\}}, \tag{17}$$

where $W(i)$ is the average distance from object i to all other objects within its cluster, and $B(i, C_j)$ is the average distance from object i to all objects in another cluster C_j. MSV for all objects is an intrinsic measure on the overall quality of a clustering solution. Since Silhouette values are based on distances, depending on the chosen distance measure and reference data different, Silhouette values can be calculated. For instance, we use the complement of Cosine similarity applied to text data in this application.

– Modularity: For some sparse adjacency data, such as graph or network, the quality of a clustering can also be evaluated by calculating the modularity (Newman 2006). Up to a multiplicative constant, modularity measures the number of intra-cluster edges minus the expected number on an equivalent network with the same clusters but with edges given at random. Intuitively, in a good clustering there are more edges within (and fewer citations between) clusters than be expected from random edge. The modularity Q_q of a clustering into q partitions is defined as

$$Q_q = \sum_{s=1}^{q} (e_{ss} - a_s^2), \tag{18}$$

where $a_s = \sum_{r=1}^{q} e_{rs}$. Here, e_{rs} is the fraction of edges between nodes in partition r and s. Obviously, the larger the distortion between internal edges e_{ss} and expected internal edges a_s^2, the more modular the graph or network is (higher Q). By design, $Q < 1$.

Experiment results

First, we implement spectral clustering on text feature of TFIDF (SC-Text) as well as on citation feature of cross-citation (SC-Citation) with different local similarities. Secondly, we develop our HC-MLSVD algorithm and the related hybrid clustering algorithms. We make a cross-comparison by applying all these algorithms to WoS journal database. Afterwards, we investigate two different issues associated with HC-MLSVD: the clustering performance of our HC-MLSVD under varied cluster number and how to select the suitable number of decomposition factors for our HC-MLSVD. During the clustering, we employ the two final partition methods (k-means++ and HC-WL) to obtain the clustering labels respectively.

Comparison between spectral clustering and HC-MLSVD with different similarities

The NMI evaluations with ESI fields of the related clustering methods are presented in Table 1.

Firstly, the NMI values of HC-MLSVD are definitely higher than those of the spectral clustering of any single-view data. For instance, in the case of Cosine similarity with HC-WL partition, HC-MLSVD improves the NMI value of spectral clustering on text feature by 5.68%. This experiment shows that HC-MLSVD can utilize more information to facilitate the clustering performance and thus it is superior to spectral clustering of single-view data.

Secondly, regarding the four local similarities, based on hybrid clustering as well as spectral clustering, the NMI values of Cosine similarity with cross-citation link are beyond these of other local similarities because it makes use of the extra useful link information.

Thirdly, concerning the two partition methods, for all the clustering strategies listed in Table 1, the NMI value of HC-WL is better than that of k-means++. The reason why HC-WL is beyond k-means is that not only HC-WL is non-sensitive to the initialization but also it takes the balance of cluster size into account (Jain and Dubes 1988).

Table 1 Comparison between spectral clustering of single-view data and hybrid clustering of MLSVD with four local similarities

Local similarity	Partition	SC-text	SC-citation	HC-MLSVD
Cosine similarity + KNN	k-means++	0.1669	0.1705	0.4438
	HC-WL	0.4103	0.4172	0.5578
Gaussian similarity + Normalized E-distance	k-means++	0.4524	0.3867	0.4526
	HC-WL	0.5104	0.4926	0.5204
Gaussian similarity + Cosine distance	k-means++	0.4805	0.3414	0.4870
	HC-WL	0.5357	0.4358	0.5561
Cosine similarity + cross-citation link	k-means++	0.5220	0.4981	0.5468
	HC-WL	0.5402	0.5232	0.5709

KNN k -nearest neighbors, *HC-WL* hierarchical clustering based on Ward's linkage, *SC-text* spectral clustering based text data, *SC-citation* spectral clustering based citation data, *HC-MLSVD* hybrid clustering based on MLSVD

Cross-comparison of four hybrid clustering methods

The NMI evaluations of four hybrid clustering methods with two partition methods are illustrated in Fig. 4.

Firstly, it is clear that HC-MLSVD outperforms other three baseline hybrid clustering methods. For example, with HC-WL, HC-MLSVD improves the NMI value of MKF, which is the best among other baseline methods, by 2.68%.

The reason why the clustering ensemble method of SA, do not work well in this experiment, is probably because clustering ensemble methods rely more on the "agreement" among various partitions to find the optimal consensus partition. In our related work (Liu et al. 2009), the notion of "sufficient number" is also shown to be important for clustering ensemble whereas in current experiment, only two submodels (views) are involved. But HC-MLSVD can work with any number of multiple views.

Regarding MKF, by linearly averaging multiple kernels without further analysis, it only uses the global information from multi-view data while neglecting the effect of each single-view data. Whereas HC-MLSVD deals with the effect of each view simultaneously and can find a joint optimal subspace for them. Concerning CP-ALS, compared with HC-MLSVD, the subspace obtained by CP-ALS is not orthogonal which might lead to an unsatisfied partition result.

Meanwhile, our MLSVD based hybrid clustering can be thought of a "multi-view PCA" analysis, which integrates multi-view information seamlessly and form a joint optimal subspace, therefore it can extract the latent pattern shared by all views and filter out irrelevant information or noise.

In addition, concerning the two partition schemes, as can be observed in Fig. 4, the NMI value of HC-WL is also better than that of k-means++, which echoes the performance in last experiment.

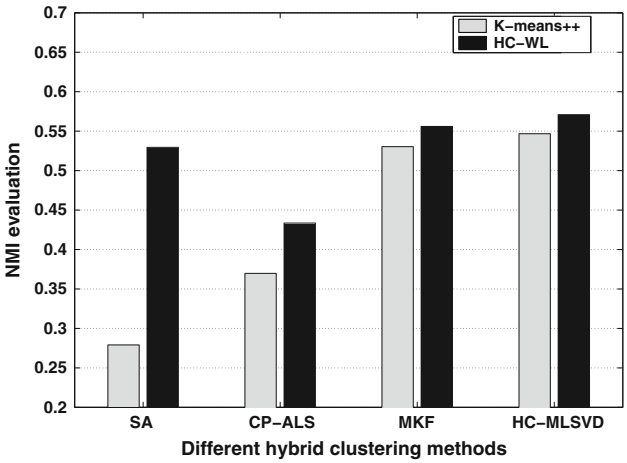

Fig. 4 Comparison among various hybrid clustering methods with two different partition schemes on WoS journal database. *NMI* normalized mutual information, *HC-WL* hierarchical clustering based on Ward's linkage, *SA* Strehl's ensemble clustering algorithm, *CP-ALS* CANDECOMP/PARAFAC (CP) decomposition based on alternating least squares (ALS) algorithm, *MKF* multiple kernel fusion, *HC-MLSVD* hybrid clustering based on MLSVD

Clustering by various number of clusters

So far, the presented results were obtained under the condition that the number of clusters equals the number of standard ESI fields. Similar to spectral clustering, how to determine the appropriate cluster number from multiple data sources is an open issue as well. In this experiment, we also compare the clustering performance of our method with that of the single-view clustering methods under varied number of clusters. We adopt two internal validation methods: mean silhouette value (MSV) and modularity which do not rely on the ground truth labels. we compare both indices from 2 clusters to 30 clusters. The MSV is calculated on the TF-IDF feature and the modularity is verified on the cross-citation feature. As depicted in Fig. 5, with different cluster number, most of the clustering indices of our proposed HC-MLSVD method is higher than the two spectral clustering strategies based on single-view data, which is consistent with the former clustering evaluation by NMI.

Clustering by varied number of components from MLSVD

In the case of spectral clustering, the number of components in the subspace for partition is believed to equal the number of clusters expected (Luxburg 2007). Although our HC-MLSVD can be considered an extension of spectral clustering, it is different from usual spectral clustering because it involves tensor decomposition. Therefore, regarding our hybrid clustering, we need to investigate how to select the suitable number of components in the joint subspace for partition. The NMI clustering evaluation of various number of components with two final partition methods is illustrated in Fig. 6. We assume the cluster number is 22 as the number of ESI fields and we change number of factors from 2 to 50.

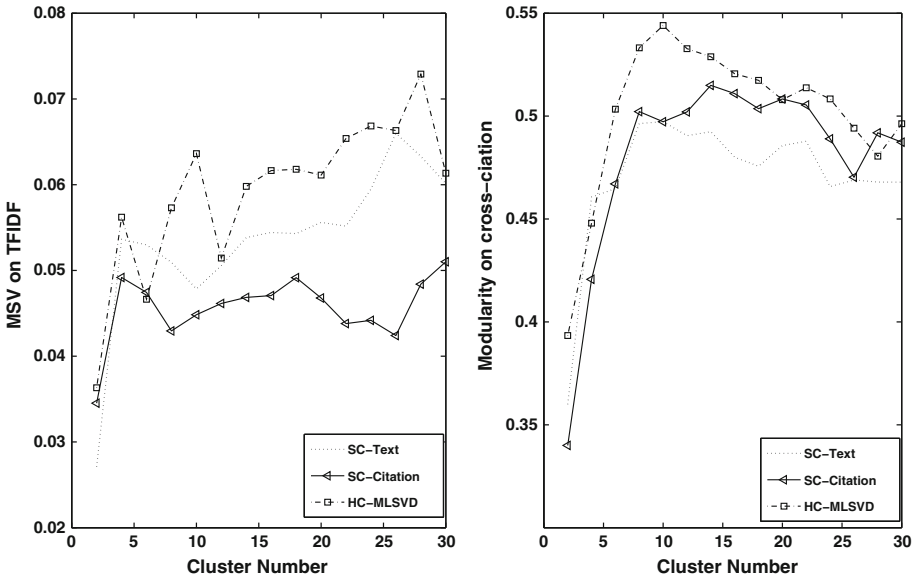

Fig. 5 Validation on clustering methods with various cluster number. *SC-text* spectral clustering based text data, *SC-citation* spectral clustering based citation data, *HC-MLSVD* hybrid clustering based on MLSVD, *MSV* mean silhouette value)

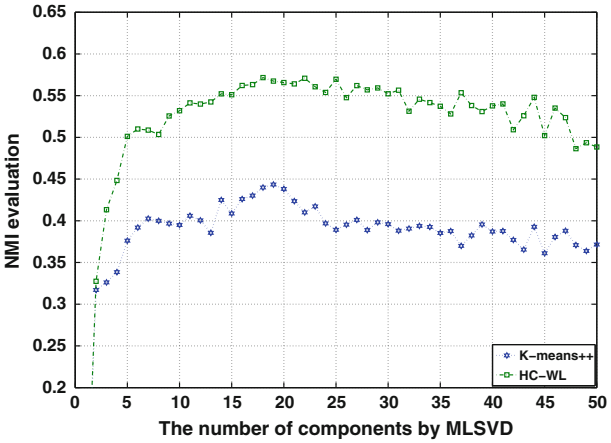

Fig. 6 NMI validation on HC-MLSVD with varied components. *NMI* normalized mutual information, *HC-WL* hierarchical clustering based on Ward's linkage

We can see that, with HC-WL partition, the clustering performance becomes stable when the number of components is around 22 which has the highest NMI value, and at the beginning (from 2 to 15) or in the end (from 30 to 50), the NMI values are lower. The NMI evaluation with k-means++ also shows the similar trend. This experiment demonstrates that in our HC-MLSVD, the suitable number of components can be assumed to be the number of clusters as well.

Cognitive characteristics of clusters

Our above clustering evaluation mainly depends on the validation index of NMI. The NMI validation measure is not always reliable because the reference category is not unique. Hence, we also evaluate the results of the journal database analysis in another way: by means of cognitive analysis from a bibliometric perspective.

To better understand the structure of clustering, we applied a modified Google PageRank algorithm to analyze the journals within each cluster (Janssens et al. 2009). The algorithm is also applied to rank a journal within each cluster according to the number of papers it published and the number of cross-citations it received. The algorithm is defined as follows:

$$PR_i = \frac{1-\alpha}{n} + \alpha \sum_j PR_j \frac{a_{ji}/P_i}{\sum_k \frac{a_{jk}}{P_k}} \tag{19}$$

where PR_i is the PageRank of the journal i, α is a scalar between 0 and 1 (we set $\alpha = 0.9$) in our implementation), n is the number of journals in the cluster, a_{ij} is the number of citations from journal j to journal i, and P_i is the number of papers published by the journal i. the self-citations among all the journals were removed before the algorithm was applied. Using the algorithm, we investigated the five most highly ranked journals in each cluster and presented them in Fig. 7. Moreover, for the journals presented in Fig. 7, we reinvestigated the titles, abstracts, and keywords that have been indexed in the text mining process, the indexed terms were sorted by their frequencies, and for each cluster, the thirty most frequent terms were used to label the obtained clusters. The textual labels of each journal cluster are shown in Fig. 8.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
(1) QUARTERLY JOURNAL OF ECONOMICS (2) JOURNAL OF ECONOMIC LITERATURE (3) JOURNAL OF FINANCE (4) JOURNAL OF FINANCIAL ECONOMICS (5) JOURNAL OF POLITICAL ECONOMY	(1) PROGRESS IN MATERIALS SCIENCE (2) INTERNATIONAL MATERIALS REVIEWS (3) ACTA MATERIALIA (4) COMPOSITES SCIENCE AND TECHNOLOGY (5) CORROSION	(1) ANNUAL REVIEW OF PHYTOPATHOLOGY (2) ENVIRONMENTAL MICROBIOLOGY (3) PLANT BIOTECHNOLOGY JOURNAL (4) CRITICAL REVIEWS IN PLANT SCIENCES (5) BIOTECHNOLOGY ADVANCES	(1) REVIEWS IN MINERALOGY & GEOCHEMISTRY (2) EARTH-SCIENCE REVIEWS (3) ANNUAL REVIEW OF EARTH AND PLANETARY SCIENCES	(0.00042089) (4) PROGRESS IN OCEANOGRAPHY (5) QUATERNARY SCIENCE REVIEWS
Cluster 5	**Cluster 6**	**Cluster 7**	**Cluster 8**
(1) LANCET NEUROLOGY (2) NEW ENGLAND JOURNAL OF MEDICINE (3) JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION (4) JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY (5) LANCET	(1) PSYCHOLOGICAL REVIEW (2) BEHAVIORAL AND BRAIN SCIENCES (3) TRENDS IN COGNITIVE SCIENCES (4) JOURNAL OF EXPERIMENTAL PSYCHOLOGY-GENERAL (5) COGNITIVE PSYCHOLOGY	(1) ACM COMPUTING SURVEYS (2) INFORMATION SYSTEMS RESEARCH (3) STATISTICAL SCIENCE (4) JOURNAL OF THE ACM (5) JOURNAL OF MACHINE LEARNING RESEARCH	(1) NATURE REVIEWS IMMUNOLOGY (2) ANNUAL REVIEW OF IMMUNOLOGY (3) NATURE REVIEWS MOLECULAR CELL BIOLOGY (4) NATURE IMMUNOLOGY (5) NATURE REVIEWS GENETICS
Cluster 9	**Cluster 10**	**Cluster 11**	**Cluster 12**
(1) CHEMICAL REVIEWS (2) PROGRESS IN POLYMER SCIENCE (3) ACCOUNTS OF CHEMICAL RESEARCH (4) SINGLE MOLECULES (5) MASS SPECTROMETRY REVIEWS	(1) PHYSICS IN PERSPECTIV;PHYSICS IN PERSPECTIVE (2) CLASSICAL ANTIQUITY (3) CRITICAL INQUIRY (4) TRANSACTIONS OF THE AMERICAN PHILOLOGICAL ASSOCIATION (5) DEGRES-REVUE DE SYNTHESE A ORIENTATION SEMIOLOGIQUE	(1) ANNUAL REVIEW OF ECOLOGY EVOLUTION AND SYSTEMATICS (2) OCEANOGRAPHY AND MARINE BIOLOGY (3) SYSTEMATIC BIOLOGY (4) AMERICAN MUSEUM NOVITATES (5) ANNUAL REVIEW OF ENTOMOLOGY	(1) GLOBAL CHANGE BIOLOGY (2) JOURNAL OF HYDROMETEOROLOGY (3) REMOTE SENSING OF ENVIRONMENT (4) ADVANCES IN ENVIRONMENTAL RESEARCH (5) JOURNAL OF ENVIRONMENTAL QUALITY
Cluster 13	**Cluster 14**	**Cluster 15**	**Cluster 16**
(1) ANNUAL REVIEW OF FLUID MECHANICS (2) PROGRESS IN ENERGY AND COMBUSTION SCIENCE (3) JOURNAL OF THE MECHANICS AND PHYSICS OF SOLIDS (4) MARINE STRUCTURES (5) PROGRESS IN AEROSPACE SCIENCES	(1) ANNUAL REVIEW OF ASTRONOMY AND ASTROPHYSICS (2) ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES (3) ASTROPHYSICAL JOURNAL (4) ASTRONOMICAL JOURNAL (5) MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	(1) REVIEWS OF MODERN PHYSICS (2) PHYSICS REPORTS-REVIEW SECTION OF PHYSICS LETTERS (3) ADVANCES IN PHYSICS (4) ANNUAL REVIEW OF NUCLEAR AND PARTICLE SCIENCE (5) REPORTS ON PROGRESS IN PHYSICS	(1) AMERICAN POLITICAL SCIENCE REVIEW (2) ANNUAL REVIEW OF SOCIOLOGY (3) AMERICAN SOCIOLOGICAL REVIEW (4) AMERICAN JOURNAL OF SOCIOLOGY (5) WORLD POLITICS
Cluster 17	**Cluster 18**	**Cluster 19**	**Cluster 20**
(1) NATURE MATERIALS (2) MATERIALS SCIENCE & ENGINEERING R-REPORTS (3) NANO LETTERS (4) ANNUAL REVIEW OF MATERIALS RESEARCH;ANNUAL REVIEW OF MATERIALS SCIENCE (5) SURFACE SCIENCE REPORTS	(1) NATURE REVIEWS CANCER (2) CA-A CANCER JOURNAL FOR CLINICIANS (3) ANNUAL REVIEW OF MEDICINE (4) BIOSTATISTICS (5) LANCET ONCOLOGY	(1) ANNUAL REVIEW OF PSYCHOLOGY (2) PSYCHOLOGICAL METHODS (3) PSYCHOLOGICAL BULLETIN (4) REVIEW OF EDUCATIONAL RESEARCH (5) STRUCTURAL EQUATION MODELING	(1) JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY (2) FOUNDATIONS OF COMPUTATIONAL MATHEMATICS (3) JOURNAL OF THE AMERICAN MATHEMATICAL SOCIETY (4) ANNALS OF MATHEMATICS (5) ACTA MATHEMATICA
Cluster 21	**Cluster 22**		
(1) ARCHIVES OF GENERAL PSYCHIATRY (2) JOURNAL OF CONSULTING AND CLINICAL PSYCHOLOGY (3) JOURNAL OF HEALTH AND SOCIAL BEHAVIOR (4) MILBANK QUARTERLY (5) ANNUAL REVIEW OF PUBLIC HEALTH	(1) CLINICAL MICROBIOLOGY REVIEWS (2) EMERGING INFECTIOUS DISEASES (3) AMERICAN JOURNAL OF CLINICAL NUTRITION (4) JOURNAL OF NUTRITION (5) ENVIRONMENTAL HEALTH PERSPECTIVES		

Fig. 7 The five most important journals of each cluster ranked by the modified pagerank algorithm. Data source: Thomson Reuters, Web of Science

In order to visualize the clustering result of the journal set, the structural mapping of the 22 clusters obtained using HC-MLSVD algorithm is presented in Fig. 9. For each cluster, the three most important terms are shown. The network is visualized by Pajek (Batagelj and Mrvar 2003). The edges represent cross-citation links and darker color represents more links between the paired clusters. The circle size represents the number of journals within each cluster.

The results presented in Fig. 9 resembles at first sight the solution by (Janssens et al. 2009); however, an important deviation can be observed. The present cluster solution differentiates more between theoretical and practice-oriented research than the previous work. This can be seen above all in the natural and medical sciences. A reliable cognitive assignment of the 22 hybrid clusters can be based on the best TF-IDF terms describing the clusters and the most "relevant" journals according to the PageRank algorithm(We omit this analysis due to the page limitation). Six clusters are related to the social sciences and humanities, particularly, #1 represents economics and business, #16 political sciences, #19 social and educational sciences, #6 psychology and behavioral sciences, #21 health sciences and #10 humanities. In the sciences and life sciences, the picture is more

Cluster	50 best terms
1	firm price market tax wage busi polici economi capit trade invest ltd organiz monetari earn compani investor corpor financi asset employe forecast stock incom welfar countri portfolio incent sector retail brand bank custom employ equiti auction household labour inc labor econom wilei industri strateg unemploy profit inflat cointegr debt manageri
2	alloi steel microstructur corros sinter ceram coat temperatur weld ltd materialia grain powder crack film wear tensil sic particl al2o3 austenit metal thermal deform oxid fractur ni melt martensit friction glass ferrit fabric creep carbid nitrid anneal alumina fatigu resin surfac titanium shear aluminum crystal specimen diffract fe bond cu
3	cultivar plant milk acid protein wheat gene leaf cow soil seed ferment diet chees ltd fruit crop seedl rice shoot meat broiler starch soybean carcass enzym weed breed speci flower strain qtl germin maiz cell calv barlei dairi silag coli corn ph genotyp temperatur inocul potato fat pcr fed dry
4	ocean seismic rock basin fault sediment magma tecton mantl earthquak sea crustal volcan subduct isotop magmat metamorph crust basalt cloud lithospher wind sedimentari climat melt faci holocen deposit cretac aerosol atmospher granit ic water temperatur continent lake ltd geochem river assemblag zircon tropospher miner geolog orogen atlant stratigraph monsoon rift
5	patient surgeri clinic arteri postop pain diseas surgic therapi coronari hospit tumor cancer women lesion stent children diabet laparoscop ventricular resect symptom aneurysm infect renal pulmonari cardiac injuri blood preoper myocardi syndrom bone diagnosi graft acut aortic hypertens month mortal stroke knee placebo implant infarct risk lung dose underw fractur
6	semant fmri phonolog cortex patient lexic speech auditori stimulu task stimuli word sentenc verb eeg brain saccad perceptu cognit cue ltd memori erp cortic languag prefront children motor schizophrenia speaker noun frontal vowel distractor syntact inc visual pariet aphasia hemispher verbal linguist emot hear gyru syllabl neuron latenc listen deficit
7	algorithm fuzzi wireless queri semant graph robot qo ltd user packet xml web network traffic multicast server servic fault bit cach architectur cdma scheme watermark bandwidth machin wilei languag ontolog simul processor nois multimedia video schedul node queue neural antenna scalabl circuit heurist decod filter softwar ofdm paper logic comput
8	protein cell gene receptor mice rat kinas neuron bind transcript mutant mrna dna tumor phosphoryl mutat apoptosi il peptid ca2 inhibit inhibitor acid ltd mous patient rna enzym infect genom cancer membran beta vitro inc vivo antibodi viru tissu brain chromosom induc pathwai alpha subunit assai immun antigen diseas amino
9	catalyst polym acid crystal ligand wilei angstrom nmr ltd ion hydrogen bond solvent adsorpt poli atom copolym oxid temperatur polymer film molecul metal spectroscopi chiral nanoparticl aqueou compound carbon electrod cation anion electrochem catalyt synthesi reaction spectra methyl mol cu diffract h2o molecular silica ph surfact liquid tio2 alkyl ru
10	music literari essai poetri narr christian polit artist poem god text leprosi fiction poetic aesthet religi religion theologi philosoph theatr poet hi shakespear church discours roman art moral philosophi book centuri ethic archaeolog writer divin genr write war danc greek jesu paint metaphor rhetor theolog stori mediev gospel biblic ritual
11	speci habitat forest fish predat larva prei plant lake egg nov genu taxa biomass soil femal season bird forag male river larval seedl breed fisheri tree ecosystem phylogenet abund parasitoid nest mate ecolog veget beetl assemblag spawn seed phytoplankton sea leaf clade reproduct reef parasit diet sp juvenil pollin sediment
12	soil ltd water sludg crop sediment wastewat plant groundwat forest biomass runoff river manur pollut irrig compost tillag pah metal ha carbon ozon rainfal emiss pcb contamin adsorpt sorption effluent land nutrient season agricultur nitrogen aquif watersh fertil hydrolog ph zn concentr wast reactor aerosol moistur catchment lake wetland leach
13	ltd crack turbul finit heat flame vibrat shear concret reynold beam steel veloc elast acoust vortex equat temperatur convect combust flow nonlinear plate thermal jet load wave actuat buckl fatigu vortic cylind simul fuel turbin piezoelectr rotor wilei deform seismic cement friction fluid wind weld damp unsteadi boundari motion stiff
14	galaxi star stellar luminos redshift galact ngc solar telescop ionospher supernova dwarf accret quasar pulsar rai cospar emiss nebula cosmic disk magnetospher radio cloud agn halo wind orbit interstellar grb kpc flare planet cosmolog neutrino dust magnet kev flux chandra veloc hubbl photometr xmm observatori photometri jet maser auror photospher
15	quantum quark neutrino brane neutron qcd meson spin nucleon boson hadron beam gev fermion detector atom photon mev supersymmetr entangl cosmolog relativist magnet particl energi soliton lattic higg decai wave spacetim collis plasma gluon ion scatter pion baryon symmetri momentum einstein string lepton laser interperiodica scalar bose qubit equat gaug
16	polit polici social court war democraci parti democrat women reform discours crime polic elector sociolog urban religi vote justic ideolog geographi elect labour civil immigr argu citi citizen violenc voter welfar crimin gender legal feder govern transnat actor ethnic capit economi feminist market racial soviet law moral militari rural citizenship
17	film dope temperatur crystal optic magnet quantum si dielectr alloi gan laser anneal epitaxi atom diffract nm beam silicon gaa spin ion electron superconduct fabric semiconductor layer deposit ferromagnet nanotub voltag waveguid phonon substrat metal sputter zno etch spectroscopi ferroelectr thermal photoluminesc oxid thin lattic surfac exciton sio2 transistor spectra
18	patient tumor cancer cell carcinoma diseas clinic therapi transplant renal rat gene liver serum receptor protein infect dose blood mice breast diabet chemotherapi hepat mutat tumour insulin il arteri bone antibodi tissu lung prostat lesion hcv malign women lymphoma hiv mrna liss apoptosi biopsi hypertens plasma chronic endotheli syndrom gastric
19	student teacher children school educ adolesc social emot psycholog classroom teach cognit child learn ltd curriculum attitud skill anxieti peer women parent gender colleg autism esteem learner instruct wilei undergradu youth career score academ person particip girl questionnair item interview disabl languag literaci satisfact perceiv violenc boi behavior adhd mother
20	algebra theorem finit infin asymptot equat manifold polynomi graph inc let banach nonlinear ltd semigroup inequ singular cohomolog convex conjectur omega lambda ellipt eigenvalu infinit abelian integ hilbert automorph hyperbol algorithm phi invari sigma commut dirichlet epsilon bound holomorph isomorph math hamiltonian riemannian sobolev boundari topolog subspac converg matric stochast
21	patient nurs schizophrenia health children disord depress women adolesc symptom psychiatr clinic mental suicid hospit smoke abus anxieti interview care ptsd sleep ltd physician cognit social alcohol intervent hiv medic caregiv questionnair child drug sexual antipsychot infant score dementia servic educ violenc adhd men disabl bipolar psycholog risk therapi youth
22	dog infect hors diet cow rat vaccin pcr serum clinic patient dietari intak calv cattl viru cat strain blood protein acid diseas herd milk gene pig assai plasma vitamin anim ltd cell antibodi serotyp dose mic mice veterinari semen fat liver dairi mare wk coli sheep isol sperm equin oocyt

Fig. 8 The textual labels of the journal clusters. Data source: Thomson Reuters, Web of Science

differentiated. While clusters #20, #9 and #4 can be uniquely assigned to mathematics, chemistry and geosciences, physics is split up into cluster #15 with theoretical and high-energy physics, # 17 with crystallography and condensed-matter physics and the small cluster #14 with astronomy and astrophysics. The technical sciences are also represented by three clusters, #2 (chemical engineering), #13 (physical engineering) and #7 (electric & electronic engineering and computer science). The biological sciences form three clusters as well, #11 (biology), #3 (applied biology and agriculture) and #12 (environmental

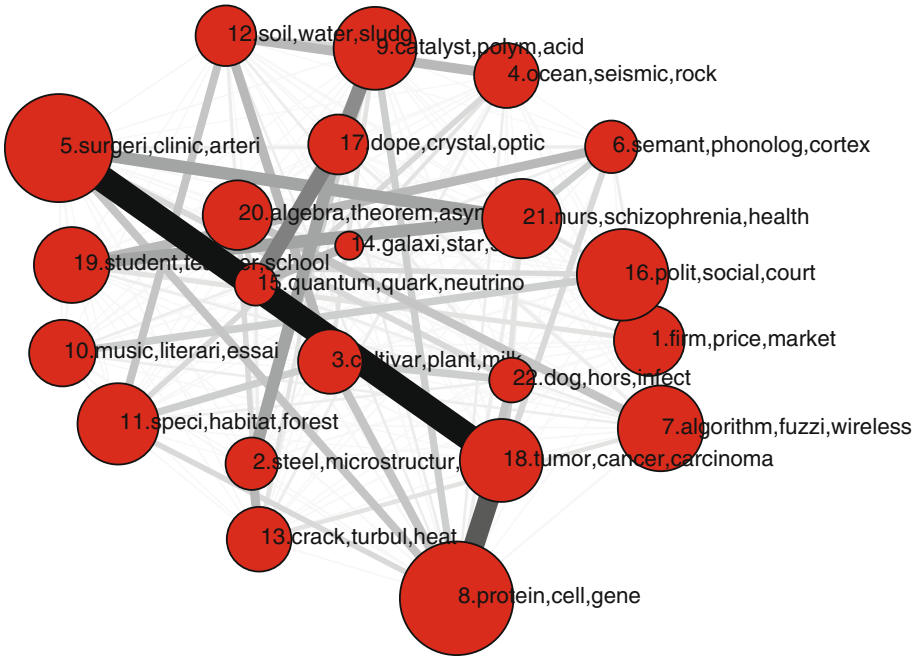

Fig. 9 Visualization of 22 clusters on the WoS journal database. Data source: Thomson Reuters, Web of Science

science & ecology). The medical sciences form another large group: cluster #8 could be identified as representing biomedical research, # 22 as veterinary sciences and clusters #5 and #18 stand for clinical and experimental medicine. While cluster #18 rather represents the research and experimental aspect of this field, cluster #5 emphasizes medical practice. The two aspects of clinical medicine are nicely visualized by Fig. 9. On the one hand, the strongest link among the 22 clusters is found between #18 and #5, and, on the other hand, the links between #18 and #8 (biomedical research) and between #5 and #21 (health science), respectively, reflect the orientation of the two clusters within clinical medicine.

Conclusion and outlook

With aid of tensor methods, we integrate multi-view data by a tensor. Then we extend spectral clustering to a hybrid clustering method of multi-view data named HC-MLSVD. The advantages of HC-MLSVD are obvious: integrating multi-view data seamlessly while taking the effect of each view into account.

We employed our algorithms to the bibliometric application of WoS journal database. We cross compared our methods with the single-view spectral clustering strategies as well as other three baseline hybrid clustering methods. The clustering performance demonstrated that our algorithm is superior to not only single-view spectral clustering methods, but also other baseline hybrid clustering methods. The cognitive analysis of the clustering results verified the effectiveness of our algorithm as well.

In later research, we will carry on the following directions: (1) we will explore more efficient tensor methods underpinning our hybrid clustering framework for scalable application; (2) we will expand our hybrid clustering algorithm to higher-order data (we only use three-order data in this research), such as, adding another time-order for dynamic hybrid clustering; (3) Besides our application, our proposed method is equally applicable to many other tasks, such as Web mining, social network analysis and multiple sensor fusion. Hence, we will generalize our methods to meet the requirements of a wide variety of applications.

Acknowledgements Thanks for the relevant work with Professor Lieven De Lathauwer at the K.U. Leuven. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by (1) Engineering Research Center of Metallurgical Automation and Measurement Technology (ERCMAMT), Ministry of Education, 430081, Hubei, China and China Scholarship Council (CSC, No. 2006153005); (2) Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; (3) FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC); (4) IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; (5) Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007–2011); (6) EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); (7)Contract Research: AMINAL; Other: Helmholtz: viCERP; ACCM; Bauknecht; Hoerbiger; (8) Flemish Government: Center for R&D Monitoring. (9) Thanks for discussion with Dr. Carlos Alzate in K.U. Leuven. The scientific responsibility is assumed by its authors.

References

Arthur, D., & Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. Technical Report 2006-13, Stanford InfoLab.

Batagelj, V., & Mrvar, A. (2003). Pajek—analysis and visualization of large networks. *Graph Drawing Software, 2265*, 77–103.

Bickel, S., & Scheffer, T. (2004). Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining* (pp. 19–26). IEEE Computer Society, Washington, DC, USA.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389–2404.

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis, part i: Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991b). Mapping of science by combined co-citation and word analysis, part ii: Dynamical aspects. *Journal of the American Society for Information Science, 42*(4), 252–266.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*(3), 287–314.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications, 21*(4), 1253–1278.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000b). On the best rank-1 and rank $(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications, 21*(4), 1324–1342.

De Lathauwer, L., & Vandewalle, J. (2004). Dimensionality reduction in higher-order signal processing and rank-$(r_1, r_2, \ldots, r_n)$ reduction in multilinear algebra. *Linear Algebra and its Applications, 391*, 31–55.

Ding, C., Huang, H., & Luo, D. (2008). Tensor reduction error analysis applications to video compression and classification. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). Washington, DC: IEEE Computer Society.

Dunlavy, D. M., Kolda, T. G., & Kegelmeyer, W. P. (2006). Multilinear algebra for analyzing data with multiple linkages. Tech. Rep. SAND2006-2079, Sandia National Laboratories.

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

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

Huang, H., Ding, C., Luo, D., & Li, T. (2008). Simultaneous tensor subspace selection and clustering: The equivalence of high order svd and k-means clustering. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 327–335). New York: ACM.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. New York: Prentice Hall.

Janssens, F. (2007). Clustering of scientific fields by integrating text mining and bibliometrics. PhD thesis, Faculty of Engineering, K.U. Leuven, Leuven, Belgium.

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and 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* (pp. 250–257). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Kolda, T. G., & Bader, W. B. (2006). The TOPHITS model for higher-order web link analysis. In *Proceedings of the SIAM Data Mining Conference Workshop on Link Analysis, Counterterrorism and Security*.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review, 51*(3), 455–500.

Lay, D. C. (2003). *Linear Algebra and Its Applications* (3rd ed.). Boston: Addition Wesley.

Liu, X., Yu, S., Moreau, Y., De Moor, B., Glänzel, W., & Janssens, F. (2009). Hybrid clustering of text mining and bibliometrics applied to journal sets. In *Proceedings of the SIAM International Conference on Data Mining*. Philadelphia, PA: SIAM.

Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology, 61*(6), 1105–1119.

Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395–416.

Modha, D. S., & Spangler, W. S. (2000). Clustering hypertext with applications to web searching. In *Proceedings of the 7th ACM on Hypertext and Hypermedia* (pp. 143–152). New York: ACM Press.

Newman, M. E. J. (2006). Modularity and community structure in networks. *PNAS, 103*(23), 8577–8582.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (pp. 849–856). Cambridge: MIT Press.

Phan, A., & Cichocki, A. (2010). Tensor decompositions for feature extraction and classification of high dimensional datasets. *Nonlinear Theory and Its Applications, IEICE* (in print).

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

Savas, B., & Eldén, L. (2007). Handwritten digit classification using higher order singular value decomposition. *Pattern Recognition, 40*(3), 993–1003.

Selee, T. M., Kolda, T. G., Kegelmeyer, W. P., & Griffin, J. D. (2007). Extracting clusters from large datasets with multiple similarity measures using IMSCAND. In M. L. Parks & S. S. Collis (Eds.), *CSRI Summer Proceedings 2007* (pp. 87–103). Technical Report SAND2007-7977. Albuquerque, NM and Livermore, CA: Sandia National Laboratories.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888–905.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269.

Smilde, A., Bro, R., & Geladi, P. (2004). *Multi-way analysis: Applications in the chemical sciences*. West Sussex, England: Wiley.

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

Sun, J., Tao, D., & Faloutsos, C. (2006). Beyond streams and graphs: dynamic tensor analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 374–383) New York: ACM.

Tang, W., Lu, Z., & Dhillon, I. S. (2009). Clustering with multiple graphs. In *ICDM '09: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining* (pp. 1016–1021). Washington, DC: IEEE Computer Society.

Tucker, L. (1964). The extension of factor analysis to three-dimensional matrices. In H. Gulliksen & N. Frederiksen (Eds.), *Contributions to mathematical psychology* (pp. 109–127). New York: Holt, Rinehart & Winston.

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*, 279–311.

Yu, S. (2009). Kernel-based data fusion for machine learning: Methods and applications in bioinformatics and text mining. PhD thesis, Faculty of Engineering, K.U. Leuven, Leuven, Belgium.