Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on a Large-Scale Journal Database

Xinhai Liu
Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B3001, Leuven, Belgium, and Wuhan University of Science and Technology (WUST), College of Information Science and Engineering, Heping Road No. 947, 430081 Wuhan, Hubei, China. E-mail: Xinhai.liu@esat.kuleuven.be

Shi Yu and Frizo Janssens
Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B3001, Leuven, Belgium. E-mail: {Shi.Yu, Frizo.Janssens}@esat.kuleuven.be

Wolfgang Glänzel
Katholieke Universiteit Leuven, Centre for R&D Monitoring, Department of Managerial Economics, Strategy and Innovation, Dekenstraat 2, B3000, Leuven, Belgium and Hungarian Academy of Sciences, IRPS, Budapest, Hungary. E-mail: Wolfgang.Glanzel@econ.kuleuven.be

Yves Moreau and Bart De Moor
Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B3001, Leuven, Belgium. E-mail: {Yves.Moreau, Bart.DeMoor}@esat.kuleuven.be

We propose a new hybrid clustering framework to incorporate text mining with bibliometrics in journal set analysis. The framework integrates two different approaches: clustering ensemble and kernel-fusion clustering. To improve the flexibility and the efficiency of processing large-scale data, we propose an information-based weighting scheme to leverage the effect of multiple data sources in hybrid clustering. Three different algorithms are extended by the proposed weighting scheme and they are employed on a large journal set retrieved from the Web of Science (WoS) database. The clustering performance of the proposed algorithms is systematically evaluated using multiple evaluation methods, and they were cross-compared with alternative methods. Experimental results demonstrate that the proposed weighted hybrid clustering strategy is superior to other methods in clustering performance and efficiency. The proposed approach also provides a more refined structural mapping of journal sets, which is useful for monitoring and detecting new trends in different scientific fields.

Received July 7, 2009; revised October 31, 2009; accepted December 30, 2009

 • Published online 11 March 2010 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21312

Introduction

In scientometrics, information from journals can be categorized lexically or with citations. An important area of scientometric research is the clustering or mapping of scientific publications. The widely used method of cocitation clustering was introduced independently by Small (1973, 1978) and Marshakova (1973). Cross-citation-based cluster analysis for science mapping is different; while the former is usually based on links connecting individual documents, the latter requires aggregation of documents to units like journals or subject fields among which cross-citation links are established. Some advantages of this method (for instance, the possibility to analyze directed information flows) are undermined by possible biases. For example, bias could be caused by the use of predefined units (journals, subject categories, etc.), implying already certain structural classification. Journal cross-citation clustering has been used by Leydesdorff (2006), Leydesdorff and Rafols (2009), and Boyack, Börner, and Klavans (2009), while Moya-Anegón et al. (2007) applied subject cocitation analysis to visualize the structure of science and its dynamics.

The integration of lexical similarities and citation links has also attracted interest in other fields such as search engine design (i.e., Google combines text and links; Brin & Page,

1998). The combination of link-based clustering with a textual approach was suggested as early as 1990 to improve the efficiency and usability of cocitation and coword analysis. One of the aims was to improve the apparently low recall of cocitation analysis concerning current work (Braam, Moed, & Van Raan, 1991a, 1991b; Zitt & Bassecoulard, 1994). The combination of link-based and textual methods also makes it possible to cluster objects whenever links are weak or missing (e.g., in the case of poorly cited or uncited papers). The present article is based on a new combined citation/lexical-based clustering approach (Janssens, Glänzel, & De Moor, 2008), which forms a hybrid solution in two respects. First, it combines citations and text, and second, it uses individual papers to cluster the journals in which they appear. Furthermore, the lexical component is used to label the journal clusters obtained for interpretation.

Hybrid clustering has also been applied in various document analysis applications (Modha & Spangler, 2000; He, Zha, Ding, & Simon, 2002; Wang & Kitsuregawa, 2002; Bickel & Scheffer, 2004) as well as science mapping research (Glenisson, Glänzel, Janssens, & De Moor, 2005; Janssens, 2007; Liu et al., 2009). Although all the approaches combined lexical and citation information, the actual algorithms that were applied are quite diverse. For Web document 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 & 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 agglomerative) 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.

The present article advances the hybrid clustering approach in terms of using larger scale experimental data and combining more refined data models. Large-scale journal data presents a challenge to hybrid clustering, because the journal sets are usually expressed in a high-dimension vector space and a massive amount of journals usually represents a large number of scientific fields. Moreover, the present study combines the lexical and citation data into 10 heterogeneous representations for hybrid clustering. Therefore, when the dimensionality, the number of samples, and the number of categorizations are large, many existing algorithms become inefficient. To tackle this problem, we present a new hybrid clustering approach for large-scale journal data in terms of scalability and efficiency. The data used in this article was collected from the Web of Science (WoS) journal database from the period 2002–2006, which comprises over 6,000,000 publications. In our approach, the above-mentioned 10 data sources are combined in a weighted manner, where the weights are determined by the average normalized mutual information (ANMI) between the single source partitions and the hybrid clustering partitions based on combined data. To evaluate the reliability of the clustering obtained on journal sets, we compared the clustering results with the standard categorizations, Essential Science Indicators (ESI; http://www.esi-topics.com/fields/index.html), provided by Thomson Scientific (Philadelphia, PA). We systematically compare the automatic clustering results obtained by all methods with the standard ESI categorizations. We also apply some statistical evaluation methods to produce label-independent evaluations. In total, 12 different hybrid-clustering algorithms are investigated and benchmarked using two external and two internal validation measures. The experimental results show that the proposed algorithms have both improved clustering result and high efficiency.

This article is organized as follows. The adopted data set and the standard ESI categorizations are described next. We then present the proposed hybrid clustering methodologies and the ANMI weighting scheme. Next, the experimental results are analyzed, followed by illustrating and investigating the mapping of journal sets obtained from hybrid clustering. Finally, we draw the conclusions.

Journal Database Analysis

In this section, we briefly describe the WoS journal database, the related text mining analysis and citation analysis.

Data Sources and Data Processing

The original journal data contains more than six million published papers from 2002 to 2006 (i.e., articles, letters, notes, reviews, etc.) indexed in the WoS database provided by Thomson Scientific. Citations received by these papers have been determined for a variable citation window beginning with the publication year, up to 2006. An item-by-item procedure was used with special identification keys made up of bibliographic data elements, which were extracted from the first author names, journal title, publication year, volume, and the first page. To resolve ambiguities, journals were checked for the name changes and the papers were checked for name changes and merged accordingly. Journals not covered in the entire period (from 2002 to 2006) have been omitted. Two criteria were applied to select journals for clustering: at first, only the journals with at least 50 publications from 2002 to 2006 were investigated, and others were removed from the data set; then only those journals with more than 30 citations

DOI: 10.1002/asi

TABLE 1. The 22-field Essential Science Indicators (ESI) labels of the Web of Science journal database.

Field #	ESI field	Number of journals	Field #	ESI field	Number of journals
1	Agricultural Sciences	183	12	Mathematics	312
2	Biology & Biochemistry	342	13	Microbiology	87
3	Chemistry	441	14	Molecular Biology& Genetics	195
4	Clinical Medicine	1410	15	Multidisciplinary	25
5	Computer Science	242	16	NeroScience & Behavior	194
6	Economics & Business	299	17	Pharmacology & Toxicology	135
7	Engineering	704	18	Physics	264
8	Environment/Ecology	217	19	Plant & Animal Science	608
9	Geoscience	277	20	Psychology/Psychiatry	448
10	Immunology	73	21	Social Science	968
11	Materials Sciences	258	22	Space Science	47

from 2002 to 2006 were kept. With this kind of selection criteria, we obtained 8,305 journals as the data set adopted in this article.

Text Mining Analysis

The titles, abstracts, and keywords of the journal publications were indexed with a Jakarta Lucene-based (Gospodnetic & Hatcher, 2005) text mining program using no controlled vocabulary. The index contains 9,473,061 terms but we cut the Zipf curve of the indexed terms at the head and the tail to remove rare terms, stopwords, and common words (Janssens, Zhang, De Moor, & Glänzel, 2009). These words are known to be usually irrelevant and noisy for clustering purposes. After the Zipf cut, 669,860 meaningful terms were used to represent the journals in a vector space model where the terms are attributes and the weights are calculated using four weighting schemes: TF-IDF, IDF, TF, and binary. The paper-by-term vectors are then aggregated to journal-by-term vectors as the representations of the lexical data. Therefore, we have obtained four submodels as the textual data sources varied with the term-weighting scheme. We applied Latent Semantic Indexing (LSI) on the TF-IDF data to reduce the dimensionality to 200 LSI factors. LSI is implemented on the basis of the singular value decomposition (SVD) algorithm. The number of LSI factors was selected empirically in a way similar to the preliminary work of Janssens (2007). For the 8,305 journals, on a dual Opteron 250 with 16 GB RAM, time taken for LSI computation was around 105 minutes.

Citation Analysis

We investigated the citations among the selected publications in five aspects.

- Cross-citation (CRC): Cross-citation between two papers is defined as the frequency of citations between each other. We ignored the direction of citations by symmetrizing the cross-citation matrix.
- Binary cross-citation (BV-CRC): To neglect the side effect of the large amount of citations appearing in the journals, we used binary value 1 (0) to represent whether there is (no) citation between two journals, termed binary cross-citation.
- Cocitation (COC): Cocitation refers to the number of times two papers are cited together in subsequent literature. The cocitation frequency of two papers is equal to the number of papers that cite them simultaneously.
- Bibliographic coupling (BGC): Bibliographic coupling occurs when two papers reference a common third paper in their bibliographies. The coupling frequency is equal to the number of papers they simultaneously cite.
- Latent Semantic Indexing of cross-citation (LSI-CRC): We also applied LSI on the sparse matrix with cross-citations to reduce the dimensionality. The selection of the number of the LSI factors was also based on the previous work (Janssens, 2007) and was set to 150.

The citations among papers were all aggregated to the journal level. All the textual data sources and citation data sources were converted into kernels using a linear kernel function. In particular, for the textual data, the kernel matrices were normalized and their elements correspond to the cosine value of pairwise journal-by-term vectors.

Reference Labels of Journals

As is mentioned in last section, to evaluate the science mapping results, we refer to the 22 categorizations of ESI, which are curated by various professional experts. Our main objective is, thus, to compare the automatic mapping obtained by the proposed hybrid methods against the ESI categorizations. As shown in Table 1, the number of journals contained in the different ESI fields is quite imbalanced. For instance, the largest field (Clinical Medicine) contains 1410 journals, whereas the smallest (Multidisciplinary) only has 25 journals.

Weighted Hybrid Clustering for Large-Scale Data

The hybrid-clustering algorithms considered in our experiments can be divided into two approaches: clustering ensemble and kernel-fusion clustering. Clustering ensemble is also known as clustering aggregation or consensus clustering, which integrates different partitions into a consolidated partition with a consensus function. Kernel-fusion clustering maps the data sets into a high-dimensional feature space and combines them as kernel matrices. Then a kernel-based clustering algorithm is applied to the combined kernel matrix.

DOI: 10.1002/asi

The details about these two approaches are mentioned in our earlier work (Liu et al., 2009). The present article proposes a novel weighting scheme on the basis of ANMI to leverage the effect of multiple sources in hybrid clustering. For all submodels, the one with the largest ANMI value is expected to have the most relevant information, and, therefore, it should contribute dominantly to the hybrid clustering.

Definition of ANMI

ANMI has been employed in clustering ensemble algorithms (Strehl & Ghosh, 2002), where the optimal cluster ensemble is obtained by maximizing the ANMI value. Given a set of cluster labels $P = \{P_1, \ldots P_i, \ldots, P_N\}$, where P_i represents the labels obtained from a single submodel and N is the number of submodels. ANMI measures the average normalized mutual information between P_i and P, given by

$$ANMI(P_i, P) = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} NMI(P_i, P_j) \qquad (1)$$

where normalized mutual information (NMI) is the normalized mutual information indicating the common information shared by two partitions, given by

$$NMI(P_i, P_j) = \frac{\sum_{k=1}^{C} \sum_{m=1}^{C} c_{km} \log\left(\frac{nc_{km}}{a_k b_m}\right)}{\sqrt{\left(\sum_{k=1}^{C} e_k \log\left(\frac{e_k}{n}\right)\right)\left(\sum_{m=1}^{C} f_m \log\left(\frac{f_m}{n}\right)\right)}} \qquad (2)$$

In the formulation above, C is the cluster number; $e_k$is the number of data points contained in the k-th cluster in the partition P_i; f_m is the number of samples contained in the m-th cluster in the partition P_j; c_{km} is the number of intersection samples between the k-th cluster from P_i and the m-th cluster from P_j. In particular, if P_j is the standard reference labels, $NMI(P_i, P_j)$ evaluates the performance of P_i with the standard labels.

Comparison of ANMI With Other Evaluation Measures

In data fusion applications, the use of external validation indicators is an appropriate way to provide data-independent evaluations about the clustering quality; however, they rely on the known reference labels. In contrast, the statistical validation indicators (internal validation indicators) depend on the scales, the structures and the dimensionalities of data, and, thus, they are not suitable to be compared among multiple data sources. In this case, the reliability of the internal and the external validation indicators can be judged by cross-comparing with each other. The ANMI adopted in our approach belongs to the internal validation case because it does not require any reference labels. To prove its reliability, we compare the ANMI with external validation indicators (NMI and adjusted Rand index [ARI]), using the individual submodels of journal sets. Besides the ANMI, we also compare the other two internal validation indicators (mean silhouette value [MSV] and modularity). As illustrated in Figure 1, the ANMI shows almost the same trend as the NMI and the ARI when predicting the model performance. In contrast, the MSV and the modularity show some similar trends but are not very consistent with the curve of the NMI and the ARI. The merit of ANMI is that the performance is evaluated on the basis of information criterion, which avoids the data dependency on scales, structures, and dimensionalities. In our problem, the ANMI shows similar evaluation on submodels as the NMI and the ARI, which both need the extra reference labels for evaluation. Therefore, ANMI is reliable to apply in explorative data analysis. Furthermore, the validity of ANMI as an evaluation measure has also been introduced by Strehl and Ghosh (2002).

Weighting Scheme

As explained, our approach assumes that when different submodels are applied for the hybrid clustering, the more relevant submodels should contribute more to the hybrid clustering. A straightforward way to leverage the submodels is to weigh them according to the values of their indicators (i.e., the ANMI values, the MSV values, the modularity values, etc.). Based on this assumption, we propose an ANMI-based weighting scheme to combine the kernel matrices (similarity matrices) of multiple submodels as a weighted convex linear combination. The conceptual scheme of our proposed weighting strategies is depicted in Figure 2.

As illustrated in Figure 2, the weighted hybrid clustering comprises several steps that may be summarized as follows:

Step 1: The kernels of all submodels are constructed and clustered individually by ward's linkage based hierarchal clustering (Ward's linkage based hierarchical clustering ([WLHCl]; Jain, 1988). The obtained partition of each submodel is denoted as P_i. For all the submodels, the set of partitions is denoted as $P = \{P_1, P_2, \ldots, P_N\}$. As introduced, 10 submodels are involved so N is equal to 10.

Step 2: Based on P, the clustering result of each submodel is evaluated using the ANMI as defined in Equation 1. The ANMI index is denoted as a_i, given by

$$a_i = ANMI(P_i, P), i \in \{1, 2, \ldots, N\} \qquad (3)$$

Step 3: We compute the weights w_i of submodels as proportional to their ANMI values, given by

$$w_i = \frac{a_i}{a_1 + \ldots + a_i + \ldots + a_N}, \quad i \in \{1, 2, \ldots, N\} \qquad (4)$$

Step 4: Using the weights obtained in step 3, we combine the kernels in a weighted manner, and alternatively, we integrate the labels of submodels as weighted clustering ensemble. The algorithms are briefly described as follows:

- *Weighted kernel-fusion clustering method (WKFCM).* In kernel-fusion clustering, given a set of kernels $K_i, i = 1, \ldots, N$, constructed from N submodels, to

DOI: 10.1002/asi

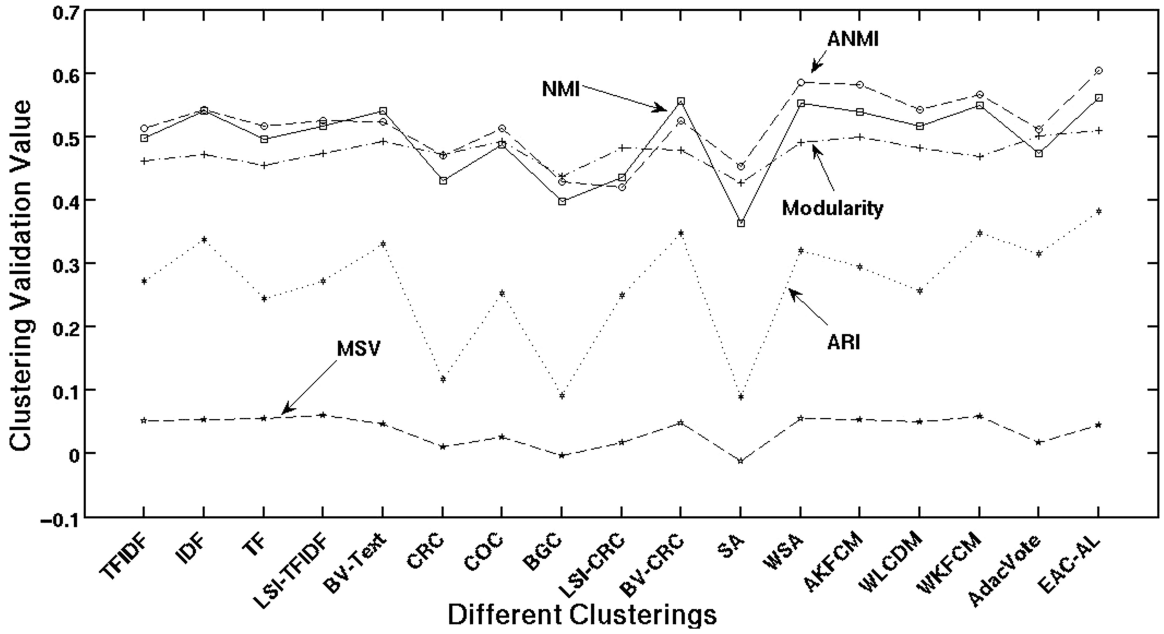

FIG. 1. Comparison of average normalized mutual information (ANMI) with the external-validation indicator (normalized mutual information [NMI] and adjusted Rand index [ARI]) and the internal-validation indicators (mean silhouette value [MSV] and modularity). The partitions of submodels are obtained by Ward's linkage-based hierarchical clustering (Jain, 1988).

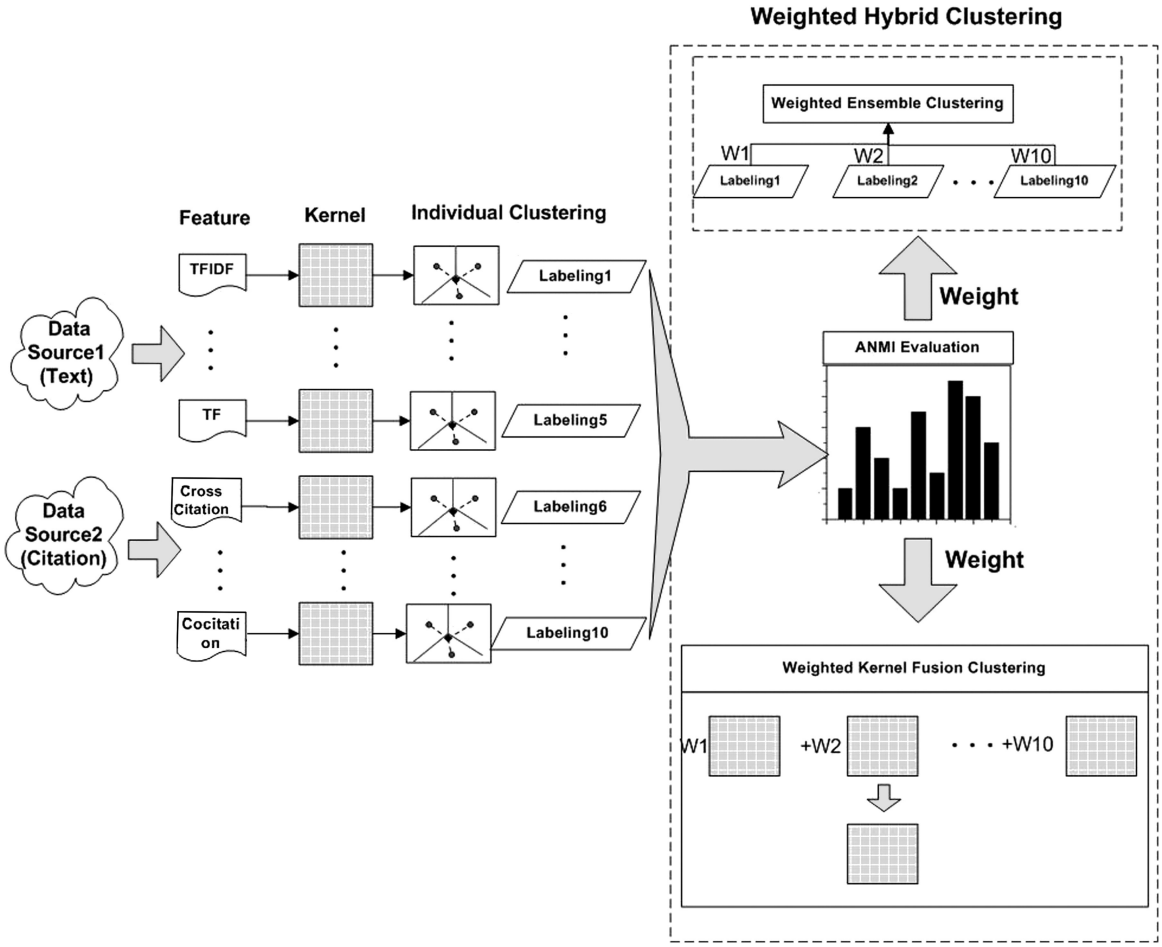

FIG. 2. Conceptual framework of the average normalized mutual information (ANMI)-based weighted hybrid clustering.

DOI: 10.1002/asi

leverage their effects in hybrid clustering, we integrate their kernels as a weighted combination, given by

$$K = \sum_{i=1}^{N} w_i K_i \tag{5}$$

The combined kernel K is further applied by single kernel-based clustering algorithms (i.e., kernel K-means, hierarchical clustering based on kernel space, spectral clustering, etc.).

- *Weighted clustering ensemble method of Strehl's algorithm (WSA) and weighted evidence accumulation clustering with average linkage (WEAC-AL).* In clustering ensemble, the partitions of all submodels $\{P_1, \ldots, P_N\}$ are usually considered as equally important. To incorporate the weights, we extend the algorithm of the clustering ensemble method (SA) proposed by Strehl and Ghosh (2002) as the WSA. Moreover, we also analogously extend the evidence accumulation clustering with average linkage (EAC-AL) algorithm proposed by Fred and Jain (2005) as the weighted EACA-AL algorithm (WEAC-AL). Both extensions are straightforward: in the original versions, the partitions of multiple submodels are considered as the input; in the weighted versions, the input is formulated as $\{w_1 P_1, \ldots, w_N P_N\}$.

Collectively, we have proposed three weighted hybrid-clustering methods on the basis of ANMI. The pseudo codes of these algorithms are combined together and illustrated as follows:

Weighted hybrid-clustering method based on ANMI.

Construct the kernels (similarity matrices) K_i for different submodels, $i \in 1, \ldots, N$.

Obtain the partition of each submodel using the base clustering algorithm (WLHC):

$$P_i(P_i \in P) \leftarrow K_i, i \in 1, \ldots, N$$

Compute the weights using ANMI:

$$a_i = ANMI(P_i, P), \quad i = 1, 2, \ldots N,$$

$$w_i = \frac{a_i}{a_1 + \ldots + a_i + \ldots + a_N}, i = 1, 2, \ldots N,$$

Obtain the overall partition using weighted hybrid clustering:

Method 1: weighted clustering ensemble, use $\{w_1 P_1, \ldots, w_N P_N\}$ as the input.

Method 2: weighted kernel-fusion clustering, use $K = \sum_{i=1}^{N} w_i K_i$ as the input.

Return the labels as the overall clustering partition.

Clustering Evaluation

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

$$S(i) = \frac{\min(B(i, C_j) - W(i))}{\max[\min(B(i, C_j), W(i))]} \tag{6}$$

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. The MSV for all objects is an intrinsic measure on the overall quality of a clustering solution. MSV may vary with the number of clusters, which is also useful to find the appropriate cluster number statistically. In the journal database, the dimensionality of lexical data is extremely high so the distance-based calculation of MSV is computationally expensive. As an alternative solution, we precompute the paired distances of all samples and store it as a kernel; in this way, the average distance required in the MSV value is directly computable in the kernel of paired distances.

Modularity. Newman (2006) introduced modularity as a graph-based evaluation of the clustering quality. Up to a multiplicative constant, modularity calculates the number of intra-cluster links minus the expected number in an equivalent network with the same clusters, but with links given at random. It means good clustering may have more links within (and fewer links between) the clusters than could be expected from the random links. Modularity is defined as follows: a $k \times k$ symmetric matrix e is defined as the element, e_{ij} is the fraction of all the edges in the network that link vertices in community or cluster i to vertices in cluster j. The trace of this matrix $trace(e) = \sum_i e_{ii}$ represents the fraction of edges in the network that connect vertices in the same cluster. The sum of rows (or columns) $a_i = \sum_j e_{ij}$ represents the fraction of edges that connect to vertices in cluster i. The modularity Q is then defined as:

$$Q = \sum\nolimits_i (e_{ii} - a_i^2) = trace(e) - \left\| e^2 \right\| \tag{7}$$

where $\|x\|$ is the sum of the elements in matrix x and $\left\| e^2 \right\|$ refers to the expected fraction of edges that connect vertices in the same cluster with edges given at random in the network.

ARI. ARI is the corrected-for-chance version of the Rand index (Hubert & Arabie, 1985). The ARI measures the similarity between two partitions. Let us assume that two partitions X and Y are obtained from a given set of n elements $S = \{O_1, \ldots, O_n\}$, given by $X = \{x_1, \ldots, x_r\}$ and $Y = \{y_1, \ldots, y_s\}$, we define the following:

a, as the number of pairs of elements in S that are in the same set in X and in the same set in Y
b, as the number of pairs of elements in S that are in different sets in X and in different sets in Y
c, as the number of pairs of elements in S that are in the same set in X and in different sets in Y
d, as the number of pairs of elements in S that are in different sets in X and in the same set in Y

DOI: 10.1002/asi

TABLE 2. Comparison of different clustering methods by normalized mutual information and adjusted rand index.

	NMI		ARI			NMI		ARI	
Clustering	Mean	STD	Mean	STD	Clustering	Mean	STD	Mean	STD
TFIDF	0.5080	0.0084	0.2676	0.0173	WLCDM	0.5161	0.0079	0.2885	0.0118
IDF	0.5478	0.0088	0.3071	0.0186	AKFCM	0.5175	0.0057	0.2841	0.0118
TF	0.5124	0.0086	0.2816	0.0218	WKFCM	0.5495	0.0062	0.3246	0.0237
LSI-TFIDF	0.5242	0.0062	0.2925	0.0199	QMI	0.5477	0.0119	0.3069	0.0246
BV-Text	0.5399	0.0092	0.3213	0.0231	AdacVote	0.4851	0.0265	0.2824	0.056
CRC	0.4532	0.016	0.1604	0.0324	SA	0.4722	0.0245	0.1696	0.0656
COC	0.4672	0.0158	0.1786	0.0315	WSA	0.5532	0.0161	0.3057	0.0263
BGC	0.4191	0.0121	0.1256	0.0252	EAC-AL	0.5562	0.0062	*0.3387*	0.0187
LSI-CRC	0.4378	0.0099	0.2221	0.0184	WEAC-AL	0.5757	0.0084	0.3710	0.0137
BV-CRC	0.5544	0.0078	0.3350	0.0199					

Note. NMI = normalized mutual information; ARI = adjusted Rand index; STD = standard deviations; TFIDF = term frequency-inverse document frequency; IDF = inverse document frequency; TF = term frequency; LSI-TFIDF = latent semantic indexing; BV-Text = binary score of TFIDF; CRC = cross-citation; COC = cocitation; BGC = bibliographic coupling; LSI-CRC = latent semantic indexing of cross-citation; BV-CRC = binary cross-citation; WLCDM = weighted linear combination of distance matrices method (Janssens el al., 2008); AKFCM = average kernel-fusion clustering method; WKFCM = weighted kernel-fusion clustering method; QMI = the clustering ensemble method by Topchy, Jain, & Punch (2005); AdacVote = the cumulative vote weighting method by Ayad & Kamel (2008); SA = the clustering ensemble method by Strehl & Ghosh (2002); WSA = the weighted clustering ensemble method of Strehl's Algorithm; EAC-AL = evidence accumulation clustering with avergae linkage; WEAC-AL = weighted evidence accumulation clustering with average linkage.

The ARI R is defined as

$$R = \frac{2(ab - cd)}{((a+d)(b+d) + (a+c)(c+b))} \tag{8}$$

NMI. NMI is another external clustering validation measure which relies on the reference labels. NMI is defined in Equation 2.

All these four clustering validation measures will be employed together to evaluate the concerned clustering algorithms.

Other Hybrid-Clustering Algorithms

In addition to the three proposed hybrid-clustering algorithms, we also apply the following six hybrid-clustering algorithms for comparison.

SA: Strehl and Ghosh (2002) formulate the optimal consensus as the partition that shares the most information with the partitions to combine. The information is measured by ANMI. Three heuristic consensus algorithms (cluster-based similarity partition, hypergraph partition, metaclustering) based on graph partitioning are employed to obtain the combined partition.

EAC-AL: Fred and Jain (2005) introduce evidence accumulation clustering (EAC) that maps the individual data partitions as an clustering ensemble by constructing a coassociation matrix. The final data partition is obtained by applying average linkage-based (AL) hierarchical clustering algorithm on the co-association matrix.

Ayad and Kamel (2008) propose a cumulative vote weighting method (AdacVote) to compute an empirical probability distribution summarizing the ensemble.

Topchy, Jain, and Punch (2005) propose an clustering ensemble method based on quadratic mutual information (QMI). They phrase the combination of partitions as a categorical clustering problem. Their method adopts a category utility function, proposed by Mirkin (2001), that evaluates the quality of a "median partition" as a summary of the ensemble.

The above four algorithms belong to the category of clustering ensemble, whereas the next two algorithms are kernel-fusion clustering methods.

Average kernel-fusion clustering method (AKFCM): The averagely combined kernel is treated as a new individual data source and the partitions are obtained by standard clustering algorithms in the kernel space.

The weighted linear combination of distance matrices method (WLCDM) proposed by Janssens et al. (2008) is actually a simplified version of AKFCM: it is achieved by equally-weighted linear combination of a text based kernel and a citation based kernel.

Experiment Result

In this part, at first, we analyze our clustering result on WoS journal database. Then, we discuss the clustering under various number of clusters and the computational complexity of different clustering schemes.

Evaluation of Clustering Results

We applied all algorithms to combine the 10 submodels to cluster the journal data into 22 partitions. The 10 submodels were also clustered individually as single sources and the performance was compared with the hybrid clustering. To determine statistical significance, we used the bootstrap t-test (Efron & Tibshirani, 1993). The bootstrap sampling was repeated 30 times and for each repetition, approximately 80% of the journals were sampled. After bootstrapping, the duplicated samples were normalized as one sample for clustering. To evaluate the performance, we applied both ARI

DOI: 10.1002/asi

TABLE 3. Comparison of different clustering performance by *t*-test.

Compared clustering methods	P-value
WSA *vs.* SA	2.2205E-12
WKFCM *vs.* AKFCM	1.8458E-8
WEAC-AL *vs.* EAC-AL	5.8E- 03
WEAC-AL *vs.* BV-CRC	3.5E-03

Note. WSA = the weighted clustering ensemble method of Strehl's Algorithm; SA = the clustering ensemble method by Strehl & Ghosh (2002); WKFCM = weighted kernel-fusion clustering method; AKFCM = average kernel-fusion clustering method; WEAC-AL = weighted evidence accumulation clustering with average linkage; EAC-AL =evidence accumulation clustering with average linkage; BV-CRC = binary cross-citation.

TABLE 4. Comparison of different weighting scheme.

Weighted hybrid clustering method	NMI	ARI
MSV-based SA	0.5309	0.2866
ANMI-based SA (WSA)	0.5532	0.3057
MSV-based KFCM	0.5447	0.3067
ANMI-based KFCM (WKFCM)	0.5495	0.3246
MSV-based EAC-AL	0.5491	0.3414
ANMI-based EAC-AL (WEAC-AL)	0.5757	0.3710

Note. NMI = normalized mutual information; ARI = adjusted Rand index; MSV = mean silhouette value; SA = the clustering ensemble method by Strehl & Ghosh (2002); WSA = the weighted clustering ensemble method of Strehl's Algorithm; ANMI = average normalized mutual information; KFCM = kernel-fusion clustering method; WKFCM = weighted kernel-fusion clustering method; EAC-AL = evidence accumulation clustering with average linkage.

and NMI using the standard ESI categorizations. The mean evaluation values and the standard deviations (STD) of the 30 bootstrapped samples are shown in Table 2.

Weighted hybrid clustering performs better than its nonweighted counterpart. As shown in Table 2, all the weighted methods outperformed their nonweighted counterparts. For the EAC-AL algorithm, the weighted version improved the ARI value by 9.54% and the NMI value by 3.51%. For the kernel-fusion clustering, the weighted algorithm increased the ARI index by 14.23% and the NMI index by 5.99%. The weighted combination in WSA also improved the ARI value of the SA method by more than 50% and the NMI index by 18.32%. The improvement of the weighted methods was shown to be statistically significant and the *p*-values obtained from the bootstrapped *t*-test are presented in Table 3.

Weighted hybrid clustering performs better than the best individual submodel. We also compared the performance of individual submodels with the hybrid results. As shown in Table 2, WEAC-AL gained improvement by heterogeneous data fusion and led to better performance than the best individual submodel (BV-CRC). Compared with other hybrid-clustering algorithms listed in previouse section, WEAC-AL outperformed them as well.

Comparison of the lexical data and the citation data. When using the base algorithm on a single submodel, the lexical data generally performed better than the citation data. This was probably because the sparse structures in the citation data could be more thoroughly analyzed using the graph cut algorithms than using the kernel clustering methods. However, the main objective of this paper is to show the validity of the weighted hybrid-clustering scheme. To keep the problem simple and concise, we do not distinguish the heterogeneity of data structure. Combining different structures with different clustering algorithms is an interesting and novel problem, and it will be presented in our forthcoming publication.

The investigation of individual submodels also substantiated the validity of our proposed weighting scheme: the submodels with higher clustering performance were assigned larger weights. For example, the submodel IDF with the largest weight performed the second best individually, and the submodel (BV-CRC) with the second largest weight performed the best individually.

Comparison of kernel-fusion clustering with clustering ensemble. Our experiment compared six clustering ensemble and four kernel-fusion clustering methods on the same large-scale journal database. As shown in Table 2, the clustering ensemble methods generally showed better clustering performance. This was probably because the clustering ensemble relies more on the "agreement" among various partitions to find the optimal consensus partition. In our experiment, 10 submodels were combined and most of them were highly relevant, and so the combination of sufficient and correlated partitions was helpful in finding the optimal consensus partition. In our related work (Liu et al., 2009), the notion of "sufficient number" was also shown to be important for clustering ensemble. In contrast, kernel-fusion clustering algorithms were less affected by the number of submodels.

Comparison of ANMI-based and MSV-based weighting schemes. Alternatively, we could also base our weighting scheme on the MSV criterion to leverage different submodels in hybrid clustering. To compare the effects of MSV and ANMI in weight calculation, we applied the MSV-based weighting scheme to create three analogous hybrid-clustering methods. The comparison of the two weighting schemes is shown in Table 4. As illustrated, the weighting scheme by ANMI works better than that based on MSV.

Clustering by Various Number of Clusters

So far, the presented results were obtained for the number of clusters equal to the number of standard ESI categorizations. How to determine the appropriate cluster number from multiple data sources still remains an open issue. As known, in single data clustering, the optimal cluster number can be explored by comparing indices for various cluster numbers. In our approach, we compared the MSV and modularity indices from 2 clusters to 30 clusters. As depicted in Figure 3, almost all of the indices of the proposed algorithm are higher than

DOI: 10.1002/asi

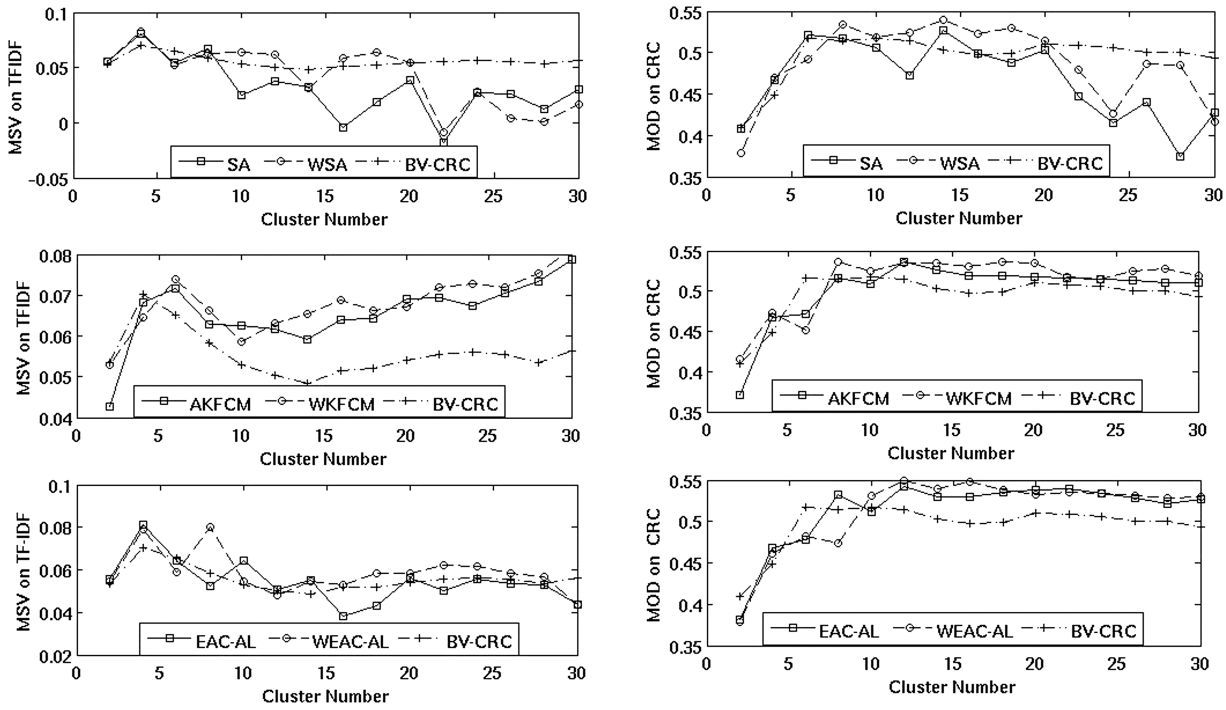

FIG. 3. Internal validations of weighted hybrid clustering methods on different cluster numbers. SA = the clustering ensemble method of by Strehl & Ghosh (2002); WSA = the weighted clustering ensemble method of Strehl's Algorithm; BV-CRC = binary cross-citation; AKFCM = average kernel-fusion clustering method; WKFCM = weighted kernel-fusion clustering method; EAC-AL = evidence accumulation clustering with average linkage; WEAC-AL = weighted evidence accumulation clustering with average linkage.

those of the nonweighted methods. Moreover, they are also generally better than the best individual data (BV-CRC).

The two figures on the top compare the weighted clustering ensemble methods. The figures in the middle evaluate the weighted kernel fusion clustering method of WSA. The figures on the bottom investigate the WEAC-AL clustering method. The figures on the left represent the MSV indices. The figures on the right side represent the modularity (MOD) indices. The MSV is calculated on the TF-IDF submodel and the MOD is verified on the CRC submodel.

Computational Complexity on Different Weighting Schemes

We also compared the computational time of the ANMI-based hybrid-clustering algorithms with the unweighted and the MSV-based weighted algorithms. The experiment was carried out on a CentOS 5.2 Linux system with a 2.4 GHz CPU and 16 GB memory. As illustrated in Figure 4, the ANMI-based weighting scheme is more efficient than the MSV-based weighting scheme. Moreover, the ANMI-based weighting method performs as efficiently as the unweighted version.

Mapping of the Journal Sets

To visualize the clustering result of journal sets, the structural mapping of the 22 categorizations obtained using the WEAC-AL method is presented in Figure 5.

For each cluster, the three most important terms are shown. The network is visualized by Pajek (Batagelj & 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 in each cluster.

To better understand the structure of clustering, we applied a modified Google PageRank algorithm (Janssens, Zhang, De Moor, & Glänzel, 2009) to analyze the journals within each cluster. 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:

$$PR_i = \frac{1-\alpha}{n} + \alpha \sum\nolimits_j PR_j \frac{a_{ji}/P_i}{\sum_k \frac{a_{jk}}{P_k}} \qquad (9)$$

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_{ji} 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, as Equation 9, we investigated the five most highly ranked journals in each cluster and presented them in Table 5. Moreover, for the journals presented in Table 5, 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

DOI: 10.1002/asi

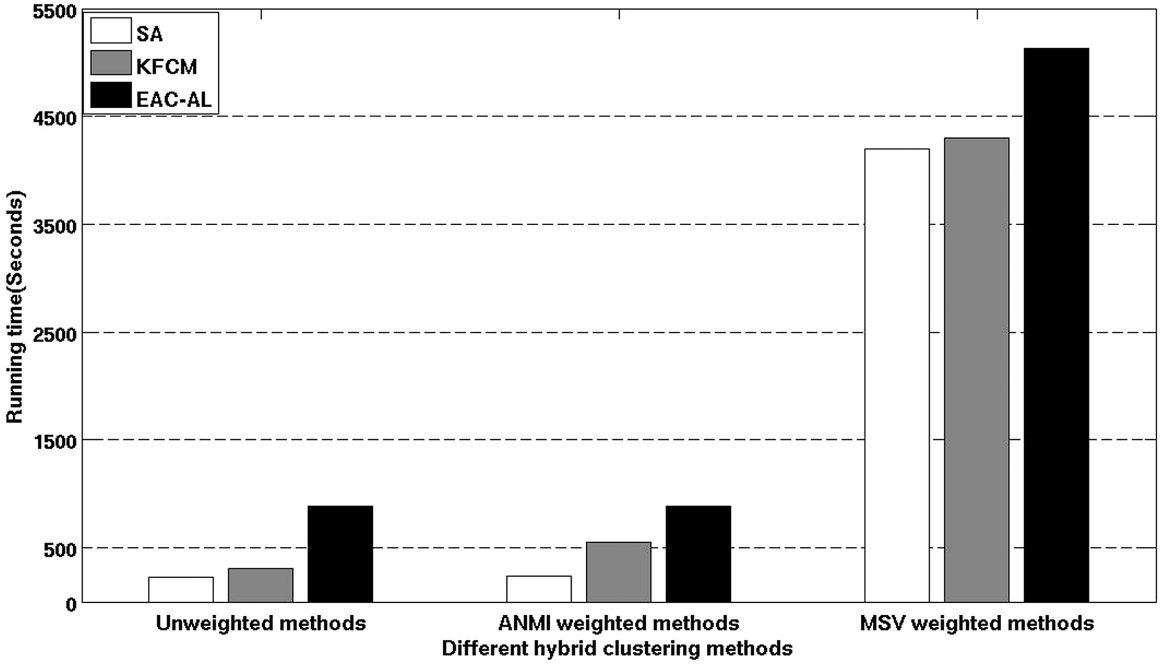

FIG. 4. Comparison of the running time of different hybrid clustering methods.
Note. The running time is measured when clustering all the journals to 22 partitions. SA = the clustering ensemble method of by Strehl & Ghosh (2002); kFCM = kernel-fusion clustering method; EAC-AL = evidence accumulation clustering with average linkage; ANMI = average normalized mutual information; MSV = mean silhouette value.

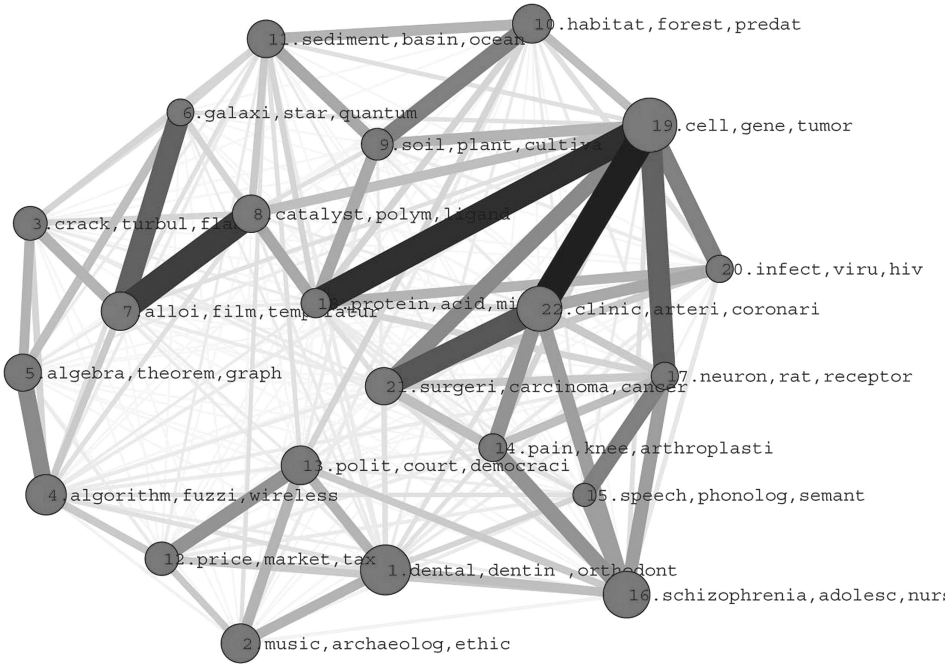

FIG. 5. Network structure of the 22 journal clusters.

clusters. The textual labels of each journal cluster are shown in Table 6.

According to Tables 5 and 6, we interpret the journal network structure (Figure 5) obtained by our clustering algorithm from a scientometric view. In the natural and applied sciences, we found nine clusters, particularly, clusters #3 through #11. On the basis of the most important journals and terms, we labeled them as follows: engineering (ENGN), computer science (COMP), mathematics (MATH), astronomy, astrophysics, physics of particles and fields (ASTR), physics (PHYS), chemistry (CHEM), agriculture, environmental science (AGRI), biology (BIOL), and geosciences

DOI: 10.1002/asi

TABLE 5. The five most important journals of each cluster ranked by the modified pagerank algorithm.

Cluster 1 (1) *Teaching in Higher Education* (2) *Strojarstvo* (3) *Veterinary Economics* (4) *Urban Education* (5) *Theoretical Linguistics*	Cluster 2 (1) *Public Historian* (2) *History of European Ideas* (3) *Public Culture* (4) *Rev Du Lou Rev Des Mus Franc* (5) *Antiquity*	Cluster 3 (1) *Acoustics Rese L. Online-Arlo* (2) *J. Appl Mech T. The Asme* (3) *Zamm-zei Ange Math und Mech* (4) *Applied Energy* (5) *AIAA J.*	Cluster 4 (1) *Australian Computer J.* (2) *J. Research & Prac in Infor Tech* (3) *Technometrics* (4) *IEEE Multimedia* (5) *J. Quality Technology*
Cluster 5 (1) *P. London Math Soc* (2) *Graphs & Combinatorics* (3) *P. Japan Aca S A-math Sci* (4) *ALGEB & GEOM Topology* (5) *Statis Meth in Medical Research*	Cluster 6 (1) *Physical Rev A* (2) *Astronomy & Astrophysics* (3) *A. Rev Nuclear & Parti Sci* (4) *Astrophysical J.* (5) *Jetp L.*	Cluster 7 (1) *Plating & Surface finishing* (2) *J. Applied physics* (3) *Plastics Rubber & Composites* (4) *Applied Physics L.* (5) *J. Phase Equilibria*	Cluster 8 (1) *Polymer International* (2) *Indian J. Chem Sec A-Inorganic Bio-inorganic Phy Theo & Analy Che* (3) *Polymer Engi & Sci* (4) *Afinidad* (5) *Studies in Surf SCI & Cata*
Cluster 9 (1) *J. Plant Growth Regul* (2) *A. J. Enology & Viticul* (3) *Agronomie;* (4) *J. Range Management* (5) *A Rev Phytopatho*	Cluster 10 (1) *Neotropical Entomology* (2) *Environ Entomology* (3) *Nautilus* (4) *Ameghiniana* (5) *Wilson J. Ornithology*	Cluster 11 (1) *Physics Earth & Plane Inter* (2) *IEEE T. Geosci & REMO Sensing* (3) *Phys & CHE of the Earth* (4) *Aquatic Geochemistry* (5) *Spe Drilling & Completion*	Cluster 12 (1) *J. Corporate Finance* (2) *Finance a Uver* (3) *A. J. Agricul & Reso econom* (4) *A. Occupational Hygiene* (5) *Management Learning*
Cluster 13 (1) *Popul & Environ* (2) *Geogra Zeitschrift* (3) *Politische Vierteljahresschrift* (4) *A. Rev of Public Administration* (5) *Washington Quarterly*	Cluster 14 (1) *J. A. Board of Family Medicine* (2) *Arthroscopy* (3) *Archives of Environ Health* (4) *Birth-issues in Perinatal Care* (5) *I. J. Geriatric psychiatry*	Cluster 15 (1) *Brain & Language* (2) *Behavior Research Methods* (3) *Clinical Linguistics & Phone* (4) *J. Neurolinguistics* (5) *Behavioral & Brain sci*	Cluster 16 (1) *Work & Stress* (2) *Telemedicine J. & E-health* (3) *Medecine et Hygiene* (4) *Fami Soc J. Contem Hum Serv* (5) *Zeits Entwicklungsp Padagogische Psychologie*
Cluster 17 (1) *Neuromolecular Med* (2) *Behavioural Brain Research* (3) *Archives Italiennes De Biologie* (4) *Brain* (5) *I. J. Neuroscience*	Cluster 18 (1) *J. Food Sci & Tech-Mysore* (2) *Archiv Fur Geflugelkunde* (3) *APPL & Environ Microbiology* (4) *Worlds Poultry Sci J.* (5) *Arch Latin Oameri de Nutricion*	Cluster 19 (1) *Math BIOSCI* (2) *Lab Animal* (3) *Methods in Enzymology* (4) *Methods-a Companion to Methods in Enzymology* (5) *Maydica*	Cluster 20 (1) *Rev in Med Microbio* (2) *Archives of Virology* (3) *A. Agricultural & Environmental Med* (4) *Avian Pathology*
Cluster 21 (1) *Pathology* (2) *Grae Arch Clin & Experi Ophthal* (3) *Pathologe* (4) *A. J . Neuroradiology* (5) *Skull Base Surgery*	Cluster 22 (1) *J. Aero Med-depo Clea & EFFE Lung* (2) *Obstetrics & Gynecology* (3) *Clin J. A. Soc. Nephrology* (4) *J. Des Maladies Vasculaires*		

TABLE 6. The textual labels of the journal clusters.

Cluster	30 best terms	Subject
1	teacher dental student dentin teeth school patient educ cari orthodont implant resin dentur enamel tooth mandibular classroom maxillari polit children social bond teach dentist discours cement librari incisor endodont learner	SCO1
2	music archaeolog polit ethic moral religi literari christian essai god philosoph religion church philosophi artist war centuri poetri historian hi roman text narr poem aesthet social theologi fiction argu kant spiritu	HUMA
3	crack turbul finit flame heat shear concret combust vibrat beam reynold temperatur veloc elast steel thermal vortex wilei fuel acoust convect coal load plate flow equat lamin fatigu jet buckl	ENGN
4	algorithm fuzzi wireless robot queri semant ltd qo packet traffic xml user graph network multicast fault wilei machin cdma web server bit servic cach bandwidth scheme architectur watermark sensor simul circuit	CSCI
5	algebra theorem finit graph asymptot polynomi infin equat inc manifold let banach nonlinear algorithm semigroup ltd singular cohomolog inequ conjectur convex omega lambda integ infinit ellipt eigenvalu abelian automorph hilbert bound hyperbol epsilon sigma	MATH
6	galaxi star quantum optic neutrino quark stellar brane luminos magnet laser redshift galact beam solar cosmolog photon superconduct qcd spin ngc atom meson neutron nucleon rai boson temperatur ion hadron	ASTR
7	alloi film temperatur dope crystal magnet si anneal dielectr diffract microstructur gan quantum silicon epitaxi steel metal ceram sinter atom nanotub fabric oxid nm layer spin thermal ion electron coat	PHYS
8	catalyst polym ligand acid crystal bond ion atom nmr hydrogen solvent adsorpt wilei angstrom copolym oxid ltd poli temperatur molecul polymer electrochem metal chiral film spectroscopi aqueou electrod anion compound	CHEM
9	soil plant cultivar leaf crop seedl seed arabidopsi shoot wheat gene speci flower rice weed biomass ha tillag germin fruit irrig maiz forest protein acid fertil manur water pollen root	AGRI
10	speci habitat forest predat fish larva prei nov egg lake genu femal taxa bird plant forag male larval biomass season river breed parasitoid nest phylogenet abund mate fisheri soil beetl	BIOC
11	sediment basin soil ocean ltd seismic rock fault water sea magma tecton earthquak mantl isotop river crustal aerosol volcan subduct groundwat lake magmat atmospher climat wind cloud crust metamorph temperatur ozon	GEOS
12	firm price market tax wage busi polici capit organiz economi trade worker employe invest monetari earn investor financi auction asset brand inc corpor compani stock welfar incom job employ retail bank	ECON
13	polit polici social ltd court parti democraci democrat urban reform forest elector women vote discours war sociolog land tourism geographi market welfar crime voter labour elect poverti econom economi govern citi	SOC2
14	patient pain knee arthroplasti hip injuri fractur tendon athlet clinic muscl ligament femor women ankl bone exercis cruciat arthroscop rehabilit surgeri flexion tibial hospit shoulder score dementia radiograph cancer nurs	CLI1
15	speech phonolog semant lexic word task children sentenc auditori memori cognit perceptu verb cue languag stimuli stimulu ltd speaker patient vowel neuropsycholog erp aphasia verbal noun hear distractor syllabl stutter listen	COGN
16	patient schizophrenia adolesc children nurs women health disord depress symptom psychiatr clinic anxieti mental student suicid social smoke abus ptsd emot hospit interview cognit psycholog child physician ltd questionnair sexual school	PSYC
17	neuron rat patient receptor brain cortex mice seizur epilepsi hippocamp synapt cell axon gaba hippocampu cortic protein ltd cerebr stroke dopamin nmda sleep astrocyt spinal inc motor nerv diseas gene glutam eeg	NEUR
18	protein acid milk diet gene ferment cell cow chees intak enzym meat starch fat dietari coli ltd strain broiler ph dna food carcass fed bacteria fatti rat antioxid dairi mutant yeast	BIOC
19	cell protein gene receptor mice rat tumor kinas patient bind transcript mrna cancer apoptosi dna mutat il phosphoryl mutant inhibitor inhibit ca2 peptid insulin acid enzym mous tissu beta vitro	BIOS
20	infect viru hiv vaccin patient dog protein cell antibodi viral gene pcr clinic hors mice strain antigen immun hcv parasit diseas rna malaria cd4 tuberculosi assai serotyp influenza virus pneumonia	MBIO
21	patient tumor surgeri carcinoma cancer postop lesion surgic clinic resect liver cell laparoscop diseas hepat endoscop arteri therapi ct gastric pancreat flap tissu preoper biopsi histolog mri malign tumour bone corneal	CLI2
22	patient cancer clinic arteri coronari renal diseas therapi transplant tumor diabet blood cell ventricular hypertens surgeri cardiac asthma hospit myocardi pulmonari lung children stent dose women prostat serum aortic graft	CLI3

DOI: 10.1002/asi

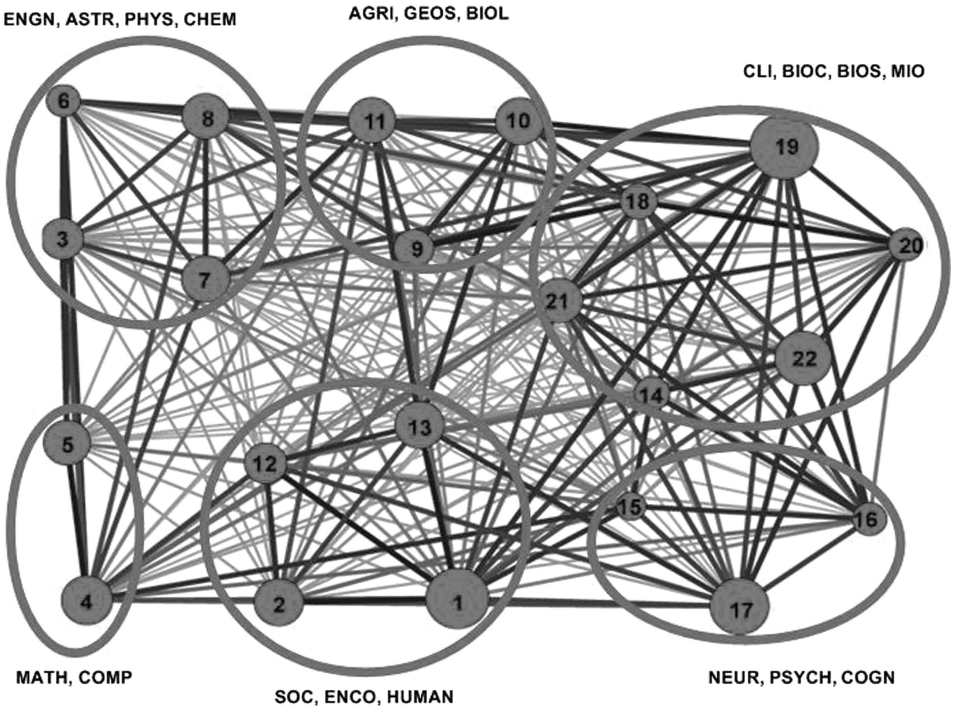

FIG. 6. Subgroups of the Web of Science journal network by weighted hybrid clustering.

(GEOS). The interpretation of the most characteristic terms of the nine life science and medical clusters is somewhat more complicated. In particular, we have a biomedical group, a clinical group, and a psychological group. The latter one has some overlap with another group, the social sciences and humanities clusters that we will discuss later. Although the overlap of the most important terms within the life science and medical clusters is considerable, the terms analysis in Table 6 provide an excellent description for at least some of the medical clusters. Thus, cluster #16 (PSYC) stands for psychology, #17 (NEUR) for neuroscience, and #15 (COGN) for cognitive science. Although NEUR represents the medical and clinical of neuro and behavioural sciences, COGN comprises cognitive psychology and neuroscience and PSYC contains psychology and psychiatry, which is traditionally considered part of the social sciences. Clusters #14, #21, and #22 represent different subfields of clinical and experimental medicine, and are therefore labeled (CLI1 through CLI3). CLI1 represents issues like health care, physiotherapy, sport science, and pain therapy, while CLI2 and CLI3 share many terms (cf. Table 6) but have a somewhat different focus as can be seen on the basis of the most important journals (cf. Table 5). Finally, clusters #18 (BIOC), #19 (BIOS), and #20 (MBIO) stand for biochemistry, biosciences, and microbiology, respectively (see Glänzel and Schubert, 2003). It should be noted that links and overlaps among the life science clusters are rather strong. The last group is formed by the social sciences and humanities (four clusters in total). Cluster #12 (ECON) is labeled as economics and business, cluster #2 (HUMA) represents the humanities, and clusters #1 (SOC1) and #13 (SOC2) two different subfields on the social sciences. Within the subject of social science, SOC1 stands for educational sciences, cultural sciences and linguistics while SOC2 represents sociology, geography, urban studies, political science and law.

The 22 clusters are more or less strongly interlinked (cf. Figure 5). The strong links between clusters #6 and #7, #7 and #8, or the "chain" leading from #18 to #21 via #19 and #22 might just serve as an example. Therefore, we have combined those clusters that are strongly interlinked to larger structures. These "mega-clusters" are presented in Figure 6. The first mega-cluster is formed by the social sciences clusters (SOC1, SOC2, ECON, and HUMA). The second one comprises MATH and COMP and the third one is formed by the natural and engineering sciences (without mathematics and computer science). Biology, agricultural, environmental, and geosciences (BIOL, AGRI, GEOS) form the fourth megastructure. The fifth and sixth one are formed by the biomedical clusters and the neuroscience clusters, respectively. The large neuroscience cluster (#15–#17) acts as a bridge connecting the life science mega-cluster with the social sciences and humanities, whereas the agricultural/environmental megacluster connects the life sciences with the natural and applied sciences (cf. Figure 6).

Conclusion

We proposed an ANMI-based weighting scheme for hybrid clustering and applied this scheme to a real application to obtain the structural mapping of a large-scale journal database. The main contributions are concluded as follows.

We presented an open framework of hybrid clustering to combine heterogeneous lexical and citation data for journal sets analysis from the scientometric point of view. We exploited two main approaches in this framework as

DOI: 10.1002/asi

clustering ensemble and kernel-fusion clustering. The performance of all approaches has been cross-compared and evaluated using multiple statistical and information based indices.

The analysis of lexical and citation information in this article was carried out at more refined granularities. The lexical information was represented in five independent data sources by the different weighting schemes of text mining. The citation information was also investigated with five different views, resulting in five independent citation data sources. These lexical and citation data sources were combined in hybrid clustering as refined representations of journals.

On the basis of the ANMI, we proposed an efficient weighting scheme for hybrid clustering. Three clustering algorithms were extended using the weighting scheme and they were systematically compared with the concerned algorithms using multiple evaluations.

To thoroughly investigate the journal clustering result, we visualized the structural network of journals on the basis of citation information. We also ranked the journals of each partition using a modified PageRank algorithm. Furthermore, we provided multiple textual labels for each cluster on the basis of text mining results. The obtained journal network integrates lexical and citation information and can be employed as a good reference for journal categorization. The proposed method is also efficient to be applied in large-scale data to detect new trends in different scientific fields. The proposed weighted hybrid-clustering framework can also be applied to retrieve multiaspect information, which is useful for a wide range of applications pertaining to heterogeneous data fusion (i.e., bioinformatics research and Web mining).

Acknowledgments

Xinhai Liu and Shi Yu made equal contributions to this article. The authors would like to thank Professor Blaise Cronin and the anonymous reviewers for fruitful comments. The authors also give thanks to Mr. Tunde (Adeshola) Adefioye and Mr. Ernesto Iacucci for their proofreading and acknowledge support from the China Scholarship Council (CSC No. 2006153005), Engineering Research Center of Metallurgical Automation and Measurement Technology, Ministry of Education, 430081, Hubei, China; Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc and Fellow Grants; Flemish Government: Steunpunt O&O Indicatoren; FWO: PhD/postdoc Grants, Projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBMEndometriosis, TBM-IOTA3, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformaticsand Modeling: from Genomes to Networks, 2007–2011); EU-RTD: ERNSI: European Research Network on System Identification;FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

References

Ayad, H.G., & Kamel, M.S. (2008). Cumulative voting consensus method for partitions with a variable number of clusters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(1),160–173.

Batagelj, V., & Mrvar, A. (2003). Pajek – analysis and visualization of large Networks. Graph Drawing Software, 2265, 77–103, Berlin, Germany: Springer.

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

Boyack, K.W., Börner, K., & Klavans, R. (2009). Mapping the structure and evolution of chemistry research. Scientometrics, 79(1), 45–60.

Braam, R.R., Moed, H.F., & Van Raan, A.F.J. (1991a). Mapping of science by combined cocitation and word analysis, Part 1: 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 cocitation and word analysis, Part II: Dynamical aspects. Journal of the American Society for Information Science, 42(4), 252–266.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–1), 107–117.

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall/CRC.

Fred, A.L.N., & Jain, A.K. (2005). Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(6), 835–850.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics, 56(3), 357–367.

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.

Gospodnetic, O., & Hatcher, E. (2005). Lucene in action. New York: Manning Publications.

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

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218.

Jain, A.K., & Dubes, R.C. (1988). Algorithms for clustering data. New Jersey: Prentice-Hall.

Janssens, F. (2007). Clustering of scientific fields by integrating text mining and bibliometrics. Doctoral dissertation. Faculty of Engineering, Katholieke Universiteit Leuven, Belgium.

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, special Issue on Informatics, 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, 683–702.

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal–journal citation relations using the journal citation reports? Journal of the American Society for Information Science and Technology, 57(5), 601–613.

Leydesdorff, L., & Rafols, I. (2009). A Global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology, 60(2), 348–362.

Liu, X.H., Yu, S., Moreau,Y., De Moor, B., Glänzel, W., & Janssens, F. (2009). Hybrid clustering of text mining and bibliometrics applied to journal sets. Proceedings of the Ninth SIAM International Conference

on Data Mining (pp. 49–60). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Marshakova, I.V. (1973). System of connections between documents based on references (as the science citation index). Nauchno-Tekhnicheskaya Informatsiya Seriya, 2(6), 3–8.

Mirkin, B. (2001). Reinterpreting the category utility function. Machine Learning, 45(2), 219–228.

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

Moya-Anegón, F. De., Vargas-Quesada, B., Chinchilla Rodríguez, Z., Corera-Álvarez, E., Munoz Fernández, F.J., & Herrero-Solana, V. (2007). Visualizing the marrow science. Journal of the American Society for Information Science and Technology, 58(14), 2167–2179.

Newman, M.E.J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23), 8577–8582.

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

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.

Small, H. (1978). Cited documents as concept symbols. Social Studies of Science, 8(3), 327–340.

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

Topchy, A., Jain, A.K., & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(12), 1866–1881.

Wang, Y., & Kitsuregawa, M. (2002). Evaluating contents-link coupled Web page clustering for Web search results. Proceedings of the 11th International Conference on Information and Knowledge Management (pp. 490–506). New York: ACM Press.

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

DOI: 10.1002/asi