Scientometrics (2017) 111:1017–1031
DOI 10.1007/s11192-017-2298-x

Clustering articles based on semantic similarity

Shenghui Wang[1] · **Rob Koopman**[1]

Received: 6 June 2016 / Published online: 27 February 2017

Abstract Document clustering is generally the first step for topic identification. Since many clustering methods operate on the similarities between documents, it is important to build representations of these documents which keep their semantics as much as possible and are also suitable for efficient similarity calculation. As we describe in Koopman et al. (Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015. Bogaziçi University Printhouse. http://www.issi2015.org/files/downloads/all-papers/1042.pdf, 2015), the metadata of articles in the *Astro* dataset contribute to a semantic matrix, which uses a vector space to capture the semantics of entities derived from these articles and consequently supports the contextual exploration of these entities in *LittleAriadne*. However, this semantic matrix does not allow to calculate similarities between articles directly. In this paper, we will describe in detail how we build a semantic representation for an article from the entities that are associated with it. Base on such semantic representations of articles, we apply two standard clustering methods, K-Means and the Louvain community detection algorithm, which leads to our two clustering solutions labelled as OCLC-31 (standing for K-Means) and OCLC-Louvain (standing for Louvain). In this paper, we will give the implementation details and a basic comparison with other clustering solutions that are reported in this special issue.

Keywords Semantic indexing · Clustering · Visualisation · K-Means · Louvain community detection

✉ Shenghui Wang
shenghui.wang@oclc.org

Rob Koopman
rob.koopman@oclc.org

[1] OCLC Research, Schipholweg 99, Leiden, The Netherlands

Introduction

Topics, sub-fields, specialities build the core in the self-organised process of scientific knowledge production (Bruckner et al. 1990). There is a lot of ambiguity how to define these units of cognitive and social organisation (Sugimoto and Weingart 2015), and an ongoing debate about how to extract them in an automatic, algorithmic way (Gläser et al. 2017). Still, one way to identify *topics* is to cluster documents. There are different ways to determine if two documents address a related subject matter. Some well-known signals for a topical relatedness include citations (if one document cites another) (Garfield 1983), co-citations (if two documents are cited by a third document) (Small 1973), bibliographic coupling (if two documents share a reference in their bibliography) (Glänzel and Czerwon 1996), and co-word linkages (if two documents share certain words) (Leydesdorff 1989). Each of these signals or traces can be used to construct a different matrix of relatedness or similarity between documents, based on which clusters of documents or topics can be identified.

In the bibliometric literature advantages and disadvantages of different methods have been discussed in abundance. In general, one differentiates between citation-based and text-based metrics (Boyack et al. 2013). Although words are expected to be less codified than cited references, we share the belief that words, especially those in titles and abstracts, do carry a certain amount of a knowledge claim made by a paper (Leydesdorff and Hellsten 2006). Hence, in accordance to the programme of cognitive scientometrics (Rip and Courtial 1984) and more recent full-text based bibliometric studies (Boyack et al. 2013), we state that if two documents share enough lexical information, they are considered to be related.

For the clustering approaches detailed in this paper, we rely on a new semantic representation of articles to determine their similarities. Both the underlying method and an interactive search interface based on it has been named *Ariadne* (Koopman et al. 2015, 2015). Our approach has great resemblance to methods used in information retrieval, in as such that it operates in a word space. But in difference to methods based on Salton's word space of documents, we use information from all elements of a document (in our case, an article), and create a word space for all those elements or entities. The motivation for this is based on the assumption that using information from many different elements of an article provides a more accurate semantic representation of this article. We consequently assume that this also improves the basis on which the similarity/relatedness between articles is determined. When we use entities such as authors, journals, subjects-headings, or references we simultaneously search for semantic similarity/relatedness along perspectives of a social (authors), communicative (journals as publication venue), or knowledge exchange (references) organisation of scientific knowledge production.

Our **research questions** are therefore (a) whether we could reconstruct a valid semantic representation for articles from all the entities they are associated with and (b) identify article clusters using standard methods based on such a semantic representation.

In this paper, we first describe how to represent the semantics of articles based on the entities that are involved with these articles. Then we briefly introduce two standard clustering methods, K-Means and Louvain community detection algorithm before reporting the implementation details. At the end, we compare our two solutions with the other clustering solutions reported in this special issue and conclude the paper.

From semantics of entities to semantics of articles

For our approach, we adopt the notion of *Statistical Semantics* (Furnas et al. 1983; Weaver 1955) based on the assumption of "a word is characterized by the company it keeps" (Firth 1957) or in Linguistics the *Distributional Hypothesis* (Harris 1954; Sahlgren 2008): words that occur in similar contexts tend to have similar meanings. In *Ariadne*, we extend words to entities (such as authors, journals, subjects, citations) so that each entity is indexed by a vector in a semantic space reflecting their lexical context, i.e., their co-occurrences with certain terms (including topical terms extracted from title and abstract plus user-defined subjects) (Koopman et al. 2017).

The resulting entity-term co-occurrence matrix could become extremely big and sparse which makes any computation on top of it very expensive and impractical. Thanks to Random Projection (Achlioptas 2003; Johnson and Lindenstrauss 1984), we can dramatically reduce the dimensionality of this semantic space, obtaining a much smaller and manageable sized *semantic matrix* yet keeping the semantics of the entities as much as possible. With all entities represented as vectors in the same semantic space, it is possible to compute the distance or relatedness between any pairs of entities, no matter which types they are. Such freedom is a unique feature of *Ariadne*. It provides a contextual view about an entity or a query as a start of an exploratory journey. For a more detailed description please refer to other papers (Koopman et al. 2015, 2017).

In the semantic matrix each article contributes to the semantics of individual entities. When executed over a big corpus the statistics are reliable to calculate the similarity between entities, However, from this semantic matrix, we cannot directly calculate similarities between articles.

To be able to cluster articles, and thus be comparable to the other methods, we first construct an integrated representation of an article from the entities associated with it. To do so, for each article, we look up all entities associated with this article in the Semantic Matrix. Consequently we obtain a set of vectors $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for each article, where n is the number of entities associated with it and $\mathbf{v}_i$ is the vector for entity e_i. These entities can be the authors, subjects, journal, citations, topical terms (extracted from its title and abstract), etc. Each article is represented by a unique set of vectors. The size of the set n can vary, but each of the vectors inside of a set has the same length, in our case 600 (see Koopman et al. (2017) for more details).

For each article we now build a new vector $\mathbf{v}'$, the weighted centroid of its constituent vectors:

$$\mathbf{v}' = \frac{\sum_{i=1}^{n} w_i \cdot \mathbf{v}_i}{\sum_{i=1}^{n} w_i}, \tag{1}$$

where $w_i = log(N/f_i)^3$, N is the total number of articles and f_i is the number of articles which contain the entity e_i. With this specific weighting frequent entities are heavily penalized to have little contribution to the resulting representation of the article. In the end, each article is represented by a vector of 600 dimensions.

Feature selection We extend our results published in Koopman et al. (2015) by putting the citations as additional entities in the Semantic Matrix (see Koopman et al. (2017) for more details). In order to see which role the citation information plays in terms of clustering, we will experiment by including or excluding citation vectors when computing the semantic vectors for articles (Eq. 1). So, for each article, we generate 3 vectors, one is a weighted average of everything but citations (i.e., topical terms, subjects, authors, and

journals, the same in Koopman et al. (2015)), one is a weighted average of only citation entities, and one is a weighted average of all types of entities. In Sect. 3.1, we will report the comparison results.

Standard cluster algorithms

Once the article vectors are generated, the next step is to identify clusters of articles. Various clustering methods can be applied. We mainly experiment with K-Means because it is a simple and highly scalable clustering method which directly operate on the vectorial representations of the articles. Our goal is to check whether such semantic representations yields sensible clusters.

Network-based clustering methods are well used in the scientometrics community. Therefore, we also try to solve the clustering problem from a network point of view. As a further process of such semantic representations, we transform the similarities calculated based on such vectorial representations to a similarity network of articles from which communities (clusters) could be detected. We choose to apply the Louvain community detection method (Blondel et al. 2008) as it is widely used in the scientometrics community but mostly applied to citation-based data models. We are interested to check whether the Louvain method could also find communities based on semantic similarities of the articles, instead of citations between them.

We now briefly describe these two standard algorithms and the implementation details on our dataset.

Clustering using K-means

The K-Means algorithm is one of the simplest unsupervised learning algorithms that solves the well defined clustering problem (MacKay 2003; Witten et al. 2011). It scales well to large number of samples and has been used across a large range of application areas in many different fields including scientometrics (Boyack et al. 2005).

Given a set of data points or observations $(x_1, x_2, \ldots, x_n)$, where each data point is characterized by a d-dimensional real vector, k-means clustering aims to partition the n data points into $k(\leq n)$ sets or clusters $S = \{S_1, S_2, \ldots, S_k\}$ so that the Within-Cluster Sum of Squares (WCSS) is minimized. In other words, the objective of the K-Means algorithm is to find

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{2}$$

where μ_i is the centroid (mean) of points in S_i.

This algorithm requires the number of clusters to be specified a priori. It starts with an initial set of k centroids $m_1^{(1)}, \ldots, m_k^{(1)}$ and proceeds by alternating between two steps (MacKay 2003):

Assignment step: Assign each data point to the cluster whose mean yields the least WCCS.[1]

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\}, \tag{3}$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the data points in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{4}$$

The algorithm converges when the assignments no longer change, which leads to a (local) optimum while the global optimum is not guaranteed.

The Mini Batch K-Means (Sculley 2010) is a variant of the K-Means algorithm which uses mini-batches to reduce the computation time, while still attempting to optimize the same objective function. The algorithm takes small batches (randomly chosen) of the dataset for each iteration. It then assigns a cluster to each data point in the batch, depending on the previous locations of the cluster centroids. It updates the locations of cluster centroids based on the new points from the batch. The update is a gradient descent update, which is significantly faster than a normal Batch K-Means update.

Using mini-batches drastically reduce the amount of computation required to converge to a local solution, but the quality of the results is reduced. In practice this difference in quality can be quite small (Béjar 2013). Therefore, we choose to use a Mini-Batch K-Means implementation provided by an open source machine learning library to cluster the articles in the *Astro* Dataset, where each article is a data point in the 600 dimensional semantic space, as described in Sect. 2.

Clustering using the Louvain method for community detection

We consider each article as a node in a network, and there is a link between two articles when they are highly similar. Practically in our case, we connect each article to its top 40 the most similar/related articles based on the cosine similarities calculated from their vectorial representation. This results in an article similarity network where clusters or communities could be detected. The task is to partition the network into communities of densely connected nodes, with no or sparse connections between the nodes belonging to different communities.

The Louvain method (Blondel et al. 2008) is a simple, efficient and well-accepted method for identifying communities in large networks. It is widely used in many applications in different domains including scientometrics (Zhang et al. 2010; Glänzel and Thijs 2017; Zhang et al. 2010). We apply it to see how well it performs on a similarity network rather than a citation-based network as what the ECOOM team reported in this special issue.

The method itself is a greedy optimization method that attempts to optimize the "modularity" of a partition of the network. Modularity is a scale value between -1 and 1 that measures the density of edges inside communities compared to edges outside communities. It is defined as Newman (2006):

[1] Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.

$$Q = \frac{1}{2|E|}\Sigma_{ij}\left[A_{ij} - \frac{k_i k_j}{2|E|}\right]\delta(c_i, c_j) \quad (5)$$

where $|E|$ is the total number of edges in the network, k_i is the degree of node i, A_{ij} is an element of the adjacency matrix (e.g., the weight of the edge between i and j), and c_i is the community to which node i is assigned, and the δ function is 1 if $c_i = c_j$ and 0 otherwise.

The optimization is performed in two steps iteratively. In the first phase, the method looks for "small" communities by optimizing modularity locally. Each node is initially assigned to a different community, i.e., there are as many communities as there are nodes. Then, for each node i the gain of modularity is calculated by moving i from its own community into the community of each neighbour j of i. After this value is calculated for all communities i is connected to, i is placed into the community that resulted in the greatest modularity increase. If no positive gain is possible, i remains in its original community. This process is applied repeatedly and sequentially to all nodes until no modularity improvement can occur and then the first phase is complete.

In the second phase, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities from the previous phase. Then the first phase can be re-applied to this new network. This way, it iteratively optimizes local communities until a maximum of global modularity is reached.

Compared to K-Means, the advantage of using the Louvain method is that the number of partitions or clusters is decided by the data itself. Similar to K-Means, the Louvain method is also an approximate method which does not really guarantee a global maximum of modularity. But it is highly scalable and often produces good approximation of the optimal communities.

Experiments

We applied the above-mentioned two clustering methods to the *Astro* dataset, which contains 111,616 articles in astronomy and astrophysics from 2003 to 2010 (please see - Gläser et al. (2017) for a full description of the dataset).

Experiments with K-means

Determining K based on a Pseudo-ground-truth

Evaluating clustering results or detected communities is a complex problem. The results could be presented to experts who decide whether each cluster or community is valid or not. Alternatively, a ground truth, i.e., a reference cluster or community allocation, could be used to measure how well the clustering solution fits the ground truth. Unfortunately, either way is extremely labour intensive if not impossible in our case.

This causes a practical problem while applying K-Means. A ground truth, or prior knowledge of the data, would help to determine one of the most important parameters for K-Means, the choice of k. The lack of ground truth forces us to determine k pragmatically.

The average silhouette of the data (Rousseeuw 1987) is a measure which could be used for determining k. The silhouette measures how closely a data point is matched to other data points within its cluster and how loosely it is matched to data points of the neighbouring cluster, i.e. the cluster whose average distance from the data point is lowest. The silhouette ranges from -1 indicating a wrong assignment to 1, an appropriate one while

scores around zero indicating overlapping clusters. We calculated the average silhouette of a sample of 20,000 data points with k from 10 to 100. As shown in Fig. 1, although slowly climbing the average silhouette scores are still around zero. This means that any numbers of clusters from this dataset are highly overlapping and a clear boundary between clusters seems not possible. This may reflect the intrinsically intertwined scientific communications between different topics. Another possible reason is that these articles may focus on different topics of the astrophysical domain, but they might still use the overlapping vocabulary which makes a clear distinction based on lexical information difficult to detect.

Since there are already a couple of clustering solutions on the same dataset from different research teams, we could build a *pseudo-ground-truth* based on the consensus of the available clustering solutions. We collected four clustering solutions, namely CWTS-C5, UMSI0, ECOOM-BC13 and STS-RG. Across all these four solutions, there are 93,986,261 pairs of articles, involving 96,072 articles (86% of the whole dataset), are always in the same clusters. We use these shared pairs as the pseudo-ground-truth. It is by no means the real ground truth, but a consensus we could use to tune our k to make a best guess.

As Table 1 shown, CWTS-C5 clusters provide the least number of article pairs while has the biggest proportion which is shared with the other three solutions. While the STS-RG clusters are quite the opposite: producing more than 940 K article pairs but only 10% of which are shared with others. It is mainly due to its largest 3 clusters which already contain 61% of the whole data set. They produce a large amount of within-cluster article pairs. But because these articles are in the same clusters, the total amount of shared pairs is not reduced much by including the STS-RG clusters. Note that the STS-RG clusters are generated using a rather different method from those used by the other three (Velden et al. 2017). Without the STS-RG clusters, there are 140M shared pairs and 100 K articles involved. However, as we find that including them does not have much effect on the choice of k, we decided to include the STS-RG clusters to build our pseudo-grounth-truth.

This simple comparison presented in Table 1 also suggests there might be a core set of articles whose cluster assignments are rather stable no matter which clustering method is used. Therefore, we argue that this set of 93 million *shared pairs* involving 96 K articles could be used to evaluate new clustering solutions, such as our own Louvain results.

With this pseudo-ground-truth, we are looking for an optimal k. On one hand these k clusters agree the most with the other four solutions, i.e., reproducing the most shared pairs. On the other hand large clusters are penalised if they put irrelevant articles into the same clusters. Formally we measure the precision (p) and recall (r) as follows:

$$p = \frac{\#\text{article pairs in common}}{\#\text{total produced pairs}}, r = \frac{\#\text{article pairs in common}}{\#\text{total shared pairs}} \tag{6}$$

where #total shared pairs is the total number of the article pairs in the pseudo-ground-truth, i.e. 93 million, #total produced pairs is the total number of within-cluster article pairs produced by the k clusters, and #article pairs in common is the number of article pairs which are produced by the k clusters and also shared by the other four solutions. A high p means a large proportion of produced article pairs are agreed by the other four solutions, while a high r indicates that a large proportion of the shared pairs in the pseudo-ground-truth are produced by the k clusters. A recall of 100% can be reached by putting all articles in one cluster, but that would give a very low precision, as majority of the within-cluster article pairs are not agreed by the other four solutions. Many small clusters could improve the precision as they only contain the articles which are considered to be in the

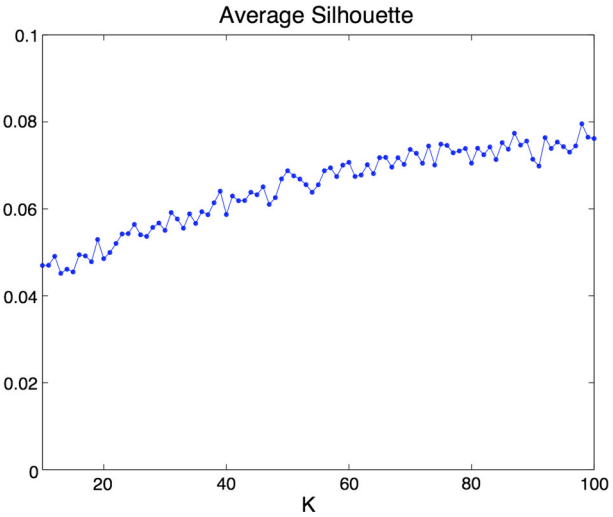

Fig. 1 Average Silhouette over 20,000 random chosen samples, with k from 10 to 100

Table 1 Statistics of the four clustering solutions for the pseudo-ground-truth

	#Cluster	#Total pairs	Of which are shared with the others (%)
CWTS-C5	22	337,151,232	28
UMSI0	22	453,492,311	21
ECOOM-BC13	13	498,846,580	18
STS-RG	556	940,553,592	10

same cluster by the other four solutions, however, many potentially related articles are distributed in different clusters which damages the recall.

To balance between p and r, we calculate the $F1$ measure[2] as widely used in the Information Retrieval community:

$$F1 = 2 \times \frac{p \times r}{p + r} \tag{7}$$

Furthermore, under the similar situation with respect to $F1$, we are aiming at a reasonably higher level of abstraction, i.e., the larger clusters the better, provided that a reasonable number of irrelevant articles are included. Therefore we reward bigger cluster by adding a parameter of the average size of the clusters into the calculation. Therefore our final score for a set of clusters is calculated as:

$$adjustedF1 = F \times \log(avgSize) \tag{8}$$

We therefore choose the best k which gives the highest *adjustedF*1 score. As mentioned before, we will later use the *adjustedF*1 score to evaluate the clustering results from the Louvain method as well.

[2] https://en.wikipedia.org/wiki/F1_score.

K-means clustering results

As mentioned in Sect. 2, we build for each article three vectorial representations: one averaging the semantic vectors of all entities, one with all entities except citations and one with only citation entities. We now search for the best k for these three representations of articles.

The K-Means algorithm is sensitive to the initialization step, i.e. where the k centroids are initially positioned. Therefore, for k from 10 to 60, we ran 10 times the Mini Batch K-Means algorithm provided in the scikit-learn python library[3] and chose the best solution which has the minimum WCSS. Then we used the *adjustedF*1 measure to evaluate our solutions against the pseudo-ground-truth. The *adjustedF*1 scores are plotted against k in Fig. 2.

If using all entities, the score climbs up until k is around 30 then decreases, with $k = 31$ giving the highest score. Therefore, we chose $k = 31$ as the best k if all entities are used for article semantic representation. Similarly, we found the best $k = 28$ if only citations are used and $k = 24$ if no citations are used. However, Fig. 2b presents, if using no citations, there are much bigger fluctuations when a similar up and down curve could be observed. While if using only citations, such curve is hardly seen.

Table 2 gives the detailed quality scores of these three clustering solutions based on the pseudo-ground-truth. The last column gives the average Adjusted Mutual Information scores (AMI) Vinh et al. (2010) between this solution and the other four solutions, namely CWTS-C5, UMSI0, ECOOM-BC13 and STS-RG. We see that if using only citations, the resulting clusters agree with the other clustering solutions more than those if no citations are used whose *adjustedF*1 score is also the lowest. It is not surprising as the other clustering solutions rely heavily on the citation information. So, even if the ways of using citations are different, the citation information still brings enough agreement between them. Using all entities to represent articles has the highest *adjustedF*1 score and agrees with the others the most.

Table 3 gives the AMI scores between these three solutions and the clusters based on the Louvain method. Again clusters based on only citations agree with the Louvain results almost to the same degree as those using all entities do. According to these measures, we decided to use all entities as the final selection of features, and keep these 31 clusters as our final K-Means results, labelled as OCLC-31. The size distribution of these 31 clusters is shown in Fig. 3a.

Community detection using the Louvain method

Different from the standard application of the Louvain method, whose input is a citation-based network, we apply the Louvain method on a semantic similarity network where each node is an article and there is an edge between two articles if they are highly similar/related. Based on the experiments with K-Means, we again use all entities to compute the semantic representation of articles. For each article, we calculated the top 40 most similar articles whose similarity values are higher than a certain threshold (in this case 0.6) and consider this article and its top 40 closest peers are connected. Once every article is connected to its peers, a similarity network is formed and then it becomes rather straightforward to apply the Louvain method to detect communities or clusters in this network.

[3] http://scikit-learn.org/.

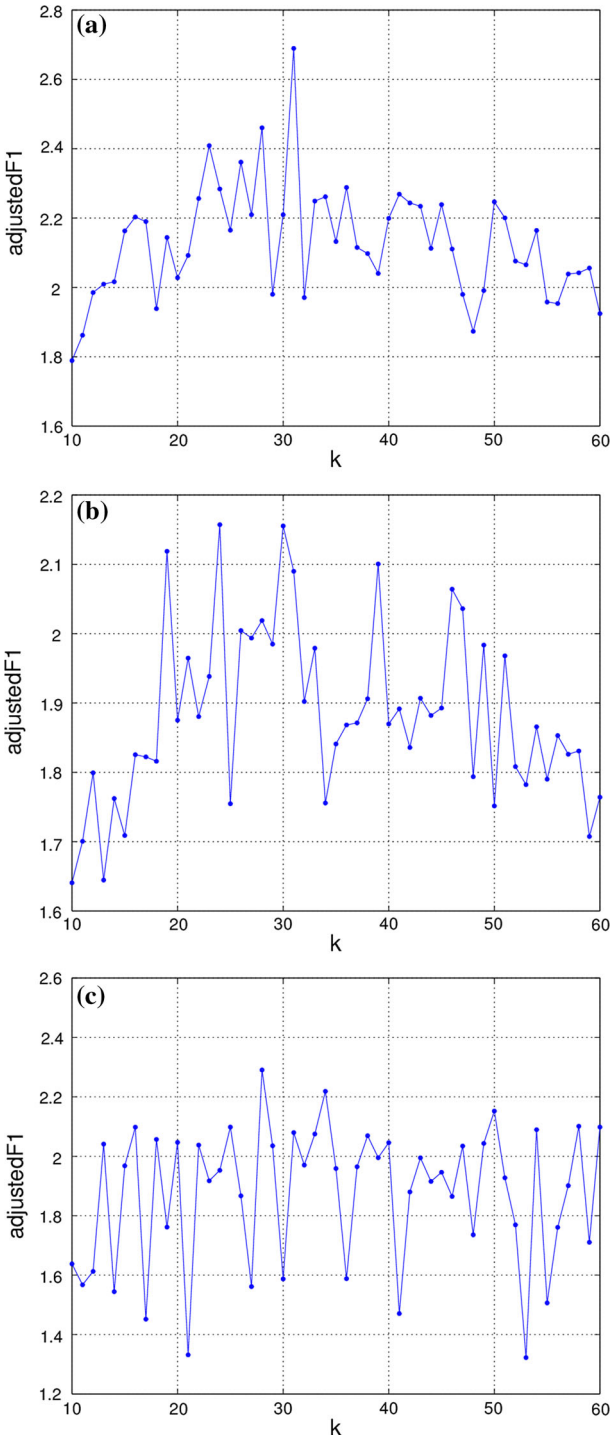

Fig. 2 Looking for the best k based on *adjustedF*1, using different sets of entities **a** all entities, **b** no citations, **c** only citations

Table 2 Quality comparison among different feature selections

	#Clusters	*r*	*p*	*f*1	*Adjusted*F1	Average AMI to others
No citations	24	0.53	0.17	0.26	2.16	0.44
Only citations	28	0.58	0.18	0.28	2.29	0.47
All entities	31	0.56	0.23	0.33	2.69	0.47
OCLC_Louvain	32	0.61	0.21	0.31	2.57	0.49

Table 3 Adjusted mutual information between solutions

	No citations	Only citations	All entities	Louvain
No citations	1.00	0.59	0.63	0.56
Only citations		1.00	0.69	0.65
All entities			1.00	0.67
OCLC_Louvain				1.00

We use the python library networkx[4] and its community detection module which implements community detection using the louvain method.[5] This results in 32 best partitions (clusters), labelled as OCLC-Louvain, with the largest partition containing 9646 articles, the smallest 86 articles and in average 3488 articles, see Fig. 3b. Its quality against the pseudo-ground-truth is given in Table 2.

The Louvain clusters perform similarly to the K-Means clusters, and actually agrees more with the other clustering solutions. However, the disadvantage of using the Louvain method is that it is not scalable for a bigger dataset as the similarity network is expensive to generate using a distance metric, even if the Louvain algorithm itself is relatively scalable.

Consensus checking

Now we can use standard consensus measures such as Adjusted Mutual Information (AMI) (Vinh et al. 2010) to check how much these two clustering solutions agree with each other. Table 4 gives the consensus score between these two solutions and the other five solutions reported in this special issue.[6] The last row gives the average AMI between one clustering solution and all the other solutions.

These numbers suggest that the data model has more impact on the solution than the algorithm chosen because OCLC-31 and OCLC-Louvain have the second highest value in terms of agreement with each other (the highest agreement is between CWTS-C5 and UMSI0, which also use the same data model). Comparing to STS-RG and ECOOM

[4] https://networkx.github.io/.

[5] http://perso.crans.org/aynaud/communities/.

[6] The CWTS-C5 and UMSI0 are the clustering solutions generated by two different methods, Infomap and the Smart Local Moving Algorithm (SLMA) respectively, applied on the direct citation network of articles. The two ECOOM clustering solutions are generated by applying the Louvain method to find communities among bibliographic coupled articles where ECOOM-NLP11 also incorporates the keywords information. The STS-RG clusters are generated by first projecting the small *Astro* dataset to the full Scopus database and

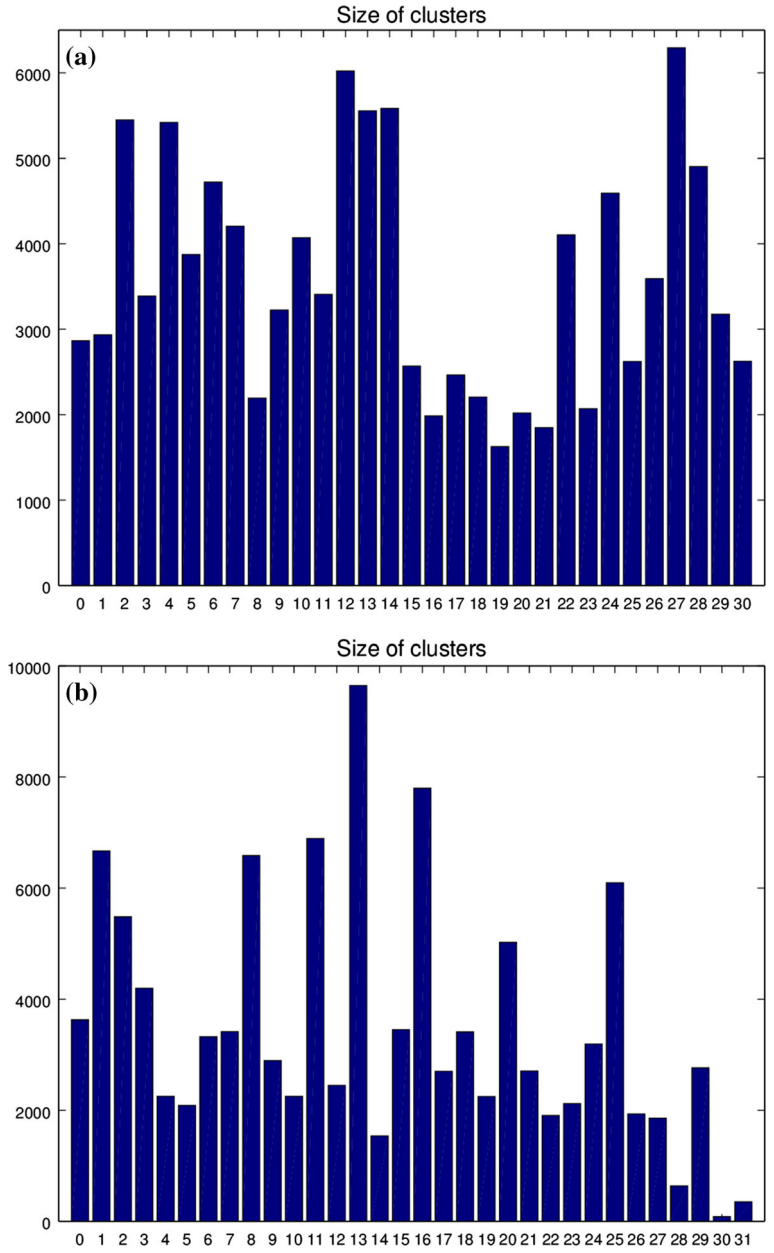

Fig. 3 The size distribution of our two clustering solutions **a** OCLC-31, **b** OCLC-Louvain

solutions, our two solutions agree more with CWTS-C5 and UMSI0, which indicates that

Footnote 6 continued
collecting their cluster assignments after the full Scopus articles are clustered using SLMA on the direct citation network. More detailed account can be found in Velden et al. (2017).

Table 4 Consensus checking using adjusted mutual information (AMI)

	sr	c	u	eb	en	ok	ol
STS-RG (sr)	1.0	0.44	0.46	0.43	0.34	0.41	0.42
CWTS-C5 (c)		1.0	0.77	0.47	0.39	0.56	0.61
UMSI0 (u)			1.0	0.47	0.38	0.51	0.55
ECOOM-BC13 (eb)				1.0	0.46	0.46	0.46
ECOOM-NLP11 (en)					1.0	0.41	0.39
OCLC-31 (ok)						1.0	0.67
OCLC-Louvain (ol)							1.0
Average AMI	0.42	0.54	0.52	0.46	0.40	0.50	0.52

even with a different data model, the results are still highly comparable. More detailed comparison can be found in Velden et al. (2017).

Conclusion

In this paper, we applied two clustering methods to identify clusters in the Astro dataset. Different from the other methods presented in this special issue, we built semantic representation for articles and tried to detect clusters of articles based on their semantic similarity. We gave technical details and the decision path towards our two clustering solutions, one based on K-Means and one based on Louvain community detection method.

The semantic representation of articles is built on a semantic matrix to which these articles contribute. Each entity (topical terms, subject, author, journal, citation) is represented by its lexical environment extracted and highly reduced from the corpus. We integrated the semantic vectors of all entities involved in one article as the representation of this article. Our experiments show that such integration of the semantics of the individual entities reflects the semantics of articles and the clustering results are competitive with other clustering solutions which are mainly based on citation information.

We would like to emphasise that the two clustering methods used in this paper are only two options we tried on such semantic representation. K-Means is highly scalable and produces results with high agreement with other solutions. One advantage is that it is applicable when citation data is missing. It could be a first step of clustering to separate articles based on their lexical information, before diving into relevant subsets with more delicate and complex clustering methods.

Acknowledgements Part of this work has been funded by the COST Action TD1210 Knowescape. We would like to thank Jochen Gläser and Andrea Scharnhorst for extended comments on earlier versions of the text. We would also like to thank the internal reviewer Michael Heinz as well as the anonymous external referees for their valuable comments and suggestions.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, *66*(4), 671–687. doi:10.1016/S0022-0000(03)00025-4.

Béjar, J. (2013). K-means vs mini batch k-means: A comparison. Tech. rep., Universitat Politècnica de Catalunya. http://upcommons.upc.edu/bitstream/handle/2117/23414/R13-8.pdf.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*, P10008. (12pp).

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*(9), 1759–1767. doi:10.1002/asi.22896.

Bruckner, E., Ebeling, W., & Scharnhorst, A. (1990). The application of evolution models in scientometrics. *Scientometrics*, *18*(1–2), 21–41. doi:10.1007/BF02019160.

Firth, J.R. (1957). A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis* pp. 1–32.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1983). Statistical semantics: Analysis of the potential performance of keyword information systems. *Bell System Technical Journal*, *62*(6), 17531806. doi:10.1002/j.1538-7305.1983.tb03513.x.

Garfield, E. (1983). *Citation indexing—Its theory and application in science, technology and humanities*. Philadelphia: ISI Press.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, *37*, 195–221.

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics. the astronomy dataset. *Scientometrics*. doi:10.1007/s11192-017-2301-6.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data: different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*. doi:10.1007/s11192-017-2296-z.

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146162.

Johnson, W., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, *26*, 189–206.

Koopman, R., Wang, S., & Scharnhorst, A. (2015) .Contextualization of topics—browsing through terms, authors, journals and cluster allocations. In: Salah, A.A., Tonta, Y., Salah, A.A.A., Sugimoto, C.R., Al, U., (Eds.), *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey*, 29 June to 3 July, 2015. Bogaziçi University Printhouse. http://www.issi2015.org/files/downloads/all-papers/1042.pdf.

Koopman, R., Wang, S., & Scharnhorst, A. (2017). Contextualization of topics—browsing through the universe of bibliographic information. In J. Gläser, A. Scharnhorst, & W. Glänzel (Eds.), *Same data—different results? Towards a comparative approach to the identification of thematic structures in science*, Special Issue of Scientometrics.

Koopman, R., Wang, S., Scharnhorst, A., & Englebienne, G. (2015). Ariadne's thread: Interactive navigation in a world of networked information. In: Begole, B., Kim, J., Inkpen, K., Woo, W., (Eds.), *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, CHI 2015 Extended Abstracts, Republic of Korea*, April 18–23, 2015, pp. 1833–1838. ACM doi:10.1145/2702613.2732781.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, *18*(4), 209–223. doi:10.1016/0048-7333(89)90016-4.

Leydesdorff, L., & Hellsten, I. (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about'monarch butterflies",frankenfoods',and'stem cells'. *Scientometrics*, *67*(2), 231–258.

MacKay, D. (2003). Information Theory, Inference and Learning Algorithms, chap. Chapter 20. An Example Inference Task: Clustering, p. 284292. Cambridge University Press.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci USA*, *103*(23), 8577–8582. doi:10.1073/pnas.0601602103. http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=pubmed&list_uids=16723398&dopt=AbstractPlus.

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, *6*(6), 381–400.

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. doi:10.1016/0377-0427(87)90125-7.

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguistica*, *20*(1), 3353.

Sculley, D. (2016). Web scale k-means clustering. In: *Proceedings of the 19th International Conference on World Wide Web*, p. 11771178. Raleigh, NC, USA.

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*, 265–269.

Sugimoto, C. R., & Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation*, *71*(4), 775–794. doi:10.1108/JD-06-2014-0082. http://www.scopus.com/inward/record.url?eid=2-s2.0-84933503812&partnerID=tZOtx3y1.

Velden, T., Boyack, K., van Eck, N., Glänzel, W., Gläser, J., Havemann, F., Heinz, M., Koopman, R., Scharnhorst, A., Thijs, B., & Wang, S. (2017). Comparison of topic extraction approaches and their results. In J. Gläser, A. Scharnhorst, & W. Glänzel (Eds.), *Same data—different results? Towards a comparative approach to the identification of thematic structures in science*, Special Issue of Scientometrics.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 28372854.

Weaver, W. (1955). Translation. In W. Locke & D. Booth (Eds.), *Machine translation of languages* (pp. 15–23). Cambridge, Massachusetts: MIT Press.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques, third edition edn. The Morgan Kaufmann series in data management systems*. Burlington: Morgan Kaufmann.

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, *4*(2), 185–193. doi:10.1016/j.joi.2009.11.005. http://www.sciencedirect.com/science/article/pii/S1751157709000832.

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics4*(2), 185–193. doi:10.1016/j.joi.2009.11.005. http://www.sciencedirect.com/science/article/pii/S1751157709000832. The ASIS&ISSI "metrics" pre-conference seminar and the Global Alliance.