Does deep learning help topic extraction? A kernel k-means clustering method with word embedding

Yi Zhang[1], **Jie Lu**[1], **Feng Liu**[1], **Qian Liu**[1, 2], **Alan Porter**[3, 4], **Hongshu Chen**[1, *], **Guangquan Zhang**[1]

[1]Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

[2]School of Computer Science, Beijing Institute of Technology, China

[3]Technology Policy and Assessment Center, Georgia Institute of Technology, USA

[4]Search Technology, Inc., USA

Email Address: yi.zhang@uts.edu.au; jie.lu@uts.edu.au; feng.liu-2@student.uts.edu.au; qian.liu-9@student.uts.edu.au; alan.porter@isye.gatech.edu; hongsue1114@hotmail.com (*); guangquan.zhang@uts.edu.au.

Abstract

Topic extraction presents challenges for the bibliometric community, and its performance still depends on human intervention and its practical areas. This paper proposes a novel kernel k-means clustering method incorporated with a word embedding model to create a solution that effectively extracts topics from bibliometric data. The experimental results of a comparison of this method with four clustering baselines (i.e., k-means, fuzzy c-means, principal component analysis, and topic models) on two bibliometric datasets demonstrate its effectiveness across either a relatively broad range of disciplines or a given domain. An empirical study on bibliometric topic extraction from articles published by three top-tier bibliometric journals between 2000 and 2017, supported by expert knowledge-based evaluations, provides supplemental evidence of the method's ability on topic extraction. Additionally, this empirical analysis reveals insights into both overlapping and diverse research interests among the three journals that would benefit journal publishers, editorial boards, and research communities.

Keywords bibliometrics; topic analysis; cluster analysis; text mining.

1. Introduction

Topic extraction is attracting enormous attention from the bibliometric community – particularly, the techniques that rely on bibliometric indicators, such as citation statistics and co-word counts. Comparisons examining the performance of existing topic extraction models and new models oriented toward bibliometric data and real-world issues are well-documented in the literature (Boyack et al. 2011; Ding & Chen 2014; Suominen & Toivanen 2016; Zhang et al. 2016b). Many new algorithms and models, such as topic models and graph methods, have harnessed the power of modern computing techniques to enhance topic extraction over the past decade (Chen et al. 2007; Dong et al. 2013; de Paulo Faleiros & de Andrade Lopes 2015). The accomplishments in topic extraction have inspired a rising passion for further advancement. However, one unsolved issue has remained a challenge in topic extraction for decades – the human costs associated with data pre-processing, e.g., filtering the links among citations and co-citations, cleaning words and terms, and synthesizing technical synonyms (Zhang et al. 2014). While some contributing factors to these problems have been revealed and explored, effective solutions remain elusive. For example, researchers citing references may hold either positive or negative opionions to the sources (Rip 1988), and co-word-based approaches have difficulties processing technological synonyms, especially in emerging sectors (Peters & van Raan 1993).

Deep learning techniques provide approaches to represent complicated unstructured data through computational models with multiple processing layers (LeCun et al. 2015). Word embedding, as one such application of deep learning in natural language processing (NLP), maps words from a vocabulary to vectors of real numbers and, in doing so, creates a way to discover the latent semantics in large-scale text (Mikolov et al. 2013). Given these circumstances, replacing traditional word representations (e.g., word frequency vectors) with word embeddings holds great potential for topic extraction. For instance, word embedding techniques are able to extract a given

number of latent features that represent a document through neural networks, which might be considered as a way of feature extraction to take the place of human intervention required in data pre-processing. To the best of our knowledge, there are no existing text-based bibliometric studies that attempt to apply word embedding techniques to topic extraction, with the use of bibliometric indicators. More specifically, how to exploit word embedding in a text-based bibliometric method to help effectively extract topics from bibliometric data is unclear.

Aiming to address the above concerns by creating a solution to effectively extract topics from unstructured text data, this paper proposes a kernel k-means clustering methodology that incorporates word embeddings. 1) The Word2Vec method (Mikolov et al. 2013; Le & Mikolov 2014), a well-recognized word embedding technique, is exploited to handle data pre-processing, which could skip over the human costs for traditional data cleaning and, instead, generate a vector space to represent documents using a relatively small number of latent features. Note that we use Word2Vec as a representative technique of deep learning, but we are also fully aware that Word2Vec is a form of shallow neural networks since it attempts to simplify the use of neural networks to increase its computational efficiency. 2) A novel kernel k-means model, that introduces a polynomial kernel function with a k-means approach, is proposed to conduct topic extraction. The model is inspired by the potential benefits of kernel methods in bibliometric data-based clustering (Nieminen et al. 2013) and the effectiveness of polynomial kernel functions in handling NLP problems (Chang et al. 2010).

Two text datasets were applied. The principal text dataset was generated using 4770 articles downloaded from the Web of Science (WoS) database, with the WoS subject category serving as ground truth for each article, followed by a series of experiments to compare our method with four clustering baselines (i.e., k-means, fuzzy c-means, principal component analysis, and topic models) and with a radial basis function-based kernel k-means approach. The second text dataset exploited a dataset containing 557 computer science-related academic proposals granted by the National Science Foundation (NSF) of the United States (US) in 2009. The results of this experiment demonstrate that the proposed k-means method using a polynomial kernel function provides some superior benefits over the competing approaches. To further verify the method's performance in bibliometric topic extraction, we conducted an empirical study using 6767 articles published between 2000 and 2017 by three top-tier bibliometric journals: the *Journal of the Association for Information Science and Technology*[1], the *Journal of Informetrics*, and *Scientometrics*. Coupled with expert evaluations, this empirical analysis provides supplemental evidence to illustrate the ability of our method on topic extraction. Additionally, some of the resulting insights into both overlapping and diverse research interest among the three journals would benefit journal publishers, editorial boards, and research communities.

The main contribution of this manuscript is the development of a polynomial function-based kernel k-means clustering method, with the incorporation of the Word2Vec model. It skips over human costs in traditional data pre-processing, and it is superior to certain existing text-based clustering baselines on topic extraction in diverse bibliometric datasets.

The rest of this paper is organized as follows. We review previous studies in Section 2. Section 3 details our method and its three models: a word embedding model, a kernel k-means clustering model, and a validation measurement model. Section 4 presents the comparisons with several baselines in two test datasets. In Section 5, we apply our method to bibliometric topic extraction on a selection of articles published between 2000 and 2017 and qualitatively examine the performance of the method with the aid of expert knowledge. Finally, the technical implications and possible applications of the method are discussed, along with our current and future research directions in Section 6.

2. Related Works: Topic Extraction

[1] This title includes the former *Journal of the American Society for Information Science and Technology* and *Journal of the American Society for Information Science*.

This literature review specifically examines two aspects of topic extraction: clustering techniques and clustering with bibliometric indicators.

2.1. Clustering

Clustering aims to group similar patterns, e.g., observations, features, or data items, into certain categories (Manning et al. 2008) and is specific to unsupervised classification. Many clustering algorithms have been designed to meet a range of needs, such as efficiency, accuracy, or adaptability, and each has both strengths and weaknesses. For example, the initial parameter configuration (e.g., the number of clusters) and local optimization issues in k-means algorithms have attracted criticism; however, they achieve acceptable performance on large-scale data analytics and adapt well to a variety of real-world data sources (Kanungo et al. 2002; Jain 2010). Hierarchical clustering algorithms, a more traditional approach, typically generate high-quality clustering outputs but are time-consuming (Zhao et al. 2005). Kernel functions were introduced with the aim of enhancing clustering algorithms for using real-world, non-linear datasets. These functions map low-dimensional data into a high-dimensional, or even infinite, feature space (Dhillon et al. 2004), but these efforts would increase computational complexity and kernel function selection issues (Xu & Wunsch 2005). While traditional clustering algorithms assume that each object can only be assigned to one cluster, soft clustering solutions, such as fuzzy c-means algorithms, assign objects to all clusters with some degree of "membership", called a membership grade. These approaches are able to identify overlapping clusters (Zhao & Karypis 2004).

Topic models are also credited with raising interest in topic extraction. Latent Dirichlet allocation (LDA) is a fairly representative topic modeling approach, which construct a 3-level Bayesian model to discover the hidden thematic structures in large-scale document collections (Blei et al. 2003). Expansions of LDA that incorporate nonparametric Bayesian models have significantly strengthened the adaptability of topic models to unsupervised environments (Xuan et al. 2017).

2.2. Clustering with Bibliometric Indicators

Most clustering approaches are designed to meet a specific need, and repurposing a generic clustering approach to different tasks is still a complicated and difficult task (Jain et al. 1999). Hence, bibliometric researchers have devoted a great deal of effort to developing clustering models specifically for bibliometric data, e.g., scientific articles, patents, and academic proposals. Bibliometric indicators are fundamental to these efforts, mainly word co-occurrence (Zhang et al. 2017), citation/co-citation statistics (Funk & Owen-Smith 2016), co-authorship (Li et al. 2014), and bibliographic coupling links (Zhao & Strotmann 2014). Moreover, many combinations of clustering algorithms and bibliometric indicators have been compared on a vast range of datasets and tasks. For example, Boyack et al. (2011) examined the accuracy of five clustering approaches using term frequency, inverse document frequency-based cosine similarity, latent semantic analysis, topic models, and two Possion-based language models on biomedical articles derived from Medline; Ding and Chen (2014) compared the effectiveness among topic models co-word analysis, and co-citation analysis for topic detection and tracking; Zhang et al. (2016b) explored the usefulness of k-means, hierarchical clustering, and topic models for analyzing academic proposals granted by the National Science Foundation of the United States; Klavans and Boyack (2017) specifically focused on citation analysis and tested the ability of directional citations, bibliographic couplings, and co-citations to accurately represent the taxonomy of scientific and technical knowledge. Not surprisingly, no one algorithm stands out for performance in all these comparisons; the advantages and drawbacks of specific bibliometric indicators largely depend on the situation. Some of the factors influencing the bibliometric indicators to use include the scientific/technical domain, the degree of data pre-processing, and the level of human intervention. An on-going study, titled the Topic Extraction Challenge[2] and led by a group a leading bibliometric researchers, has posed an open invitation to the bibliometric community around the world to participate in a systematic comparison of topic extraction approaches in order to develop state-of-the-art methods for a variety of current and future applications.

[2] See http://www.topic-challenge.info/ for more information.

Additionally, an issue raised in a recent paper stemming from this project is the lack of access to non-proprietary benchmark datasets to support such studies (Velden et al. 2017).

3. Methodology: Kernel K-means Clustering Method with Word Embedding

This paper presents a text-based bibliometric method for conducting topic extraction in a range of bibliometric datasets. Our method includes two innovations: 1) a word embedding technique (i.e., the Word2Vec method) is incorporated for effectively and efficeintly extracting a small set of key features (i.e., several hundred), which is able to skip over the human costs in traditional data clearning; 2) a polynomial kernel function is integrated into a cosine similarity-based k-means clustering algorithm to enhance the performance of topic extraction with bibliometric data sources. The framework of the kernel k-means clustering method with word embedding is shown in Figure 1.

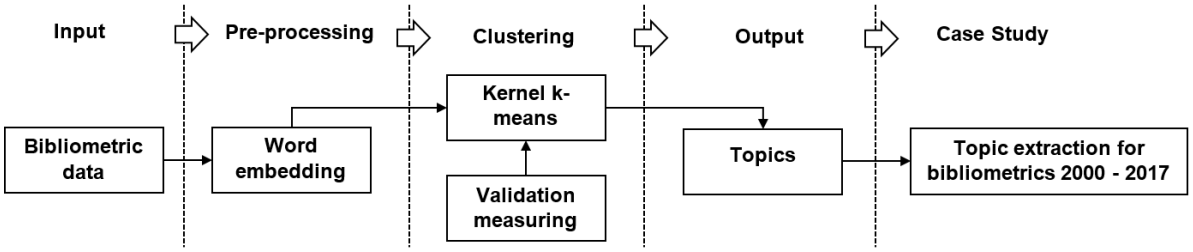

Figure. 1. Framework of the kernel k-means clustering method incorporated with word embedding.

3.1. Word Embedding

Traditional pre-processing steps usually focus on removing noise and selecting features. For example, a term clumping process (Zhang et al. 2014) has been developed to semi-automatically reduce the scale of the term-based feature space from approximately 100,000 (assuming a dataset with about 5000 articles) to several thousand. Human intervention directly influences performance with this approach, so we have further incorporated a word embedding technique to replace traditional pre-processing.

The basic assumption of word embedding is that words with a similar context tend to have similar meanings (Firth 1957). Among existing word embedding techniques, neural network algorithms are playing an increasingly crucial role in uncovering such word patterns with similar meanings (Levy & Goldberg 2014). In particular, the Word2Vec method has proven to be both efficient and effective in learning high-quality word embeddings from large-scale unstructured text data (Mikolov et al. 2013), and has been found to discover a latent factorization of a specific point-wise mutual information matrix (Levy & Goldberg 2014). These benefits could hold interest for the information retrieval community. The Word2Vec architecture contains two models: a continuous bag-of-words model (CBOW) and a Skip-Gram model. However, according to the independent benchmarking conducted by Levy et al. (2015), there is no fundamental performance difference between the two models. Therefore, we integrated both into our word embedding model to empirically examine their performance. Based on the Word2Vec method developed by Mikolov et al. (2013) and modifications given by Liu et al. (2018), the two models are described as follows.

Definition 1: Consider a corpus C with m records, where φ is the size of corpus C (i.e., the total number of words), and W represents a vocabulary that includes n words.

Definition 2: Given a word sequence $D = \{w_{i-k}, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_{i+k}\}$, where w_i is a target word and k is the context size of the target word, the window size (i.e., the length of a word sequence) is $S = 2k + 1$.

The CBOW model predicts a target word w_i using the context words in a sliding window, e.g., $\{w_{i-k}, \dots, w_{i-1}\}$ and $\{w_{i+1}, \dots, w_{i+k}\}$. The main objective is to maximize the average log probability $L(D)$, where the probability $Pr(w_i|w_{i-k}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+k})$ is formulated with a softmax function:

$$L(D) = \frac{1}{\varphi}\sum_{i=1}^{\varphi} \log Pr(w_i | w_{i-k}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+k})$$

$$Pr(w_i | w_{i-k}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+k}) = \frac{\exp(x_i \cdot x_{\text{avg}})}{\sum_{w \in W} \exp(x \cdot x_{\text{avg}})}$$

where x_i is the vector representation of the target word w_i, and x_{avg} is the average vector of all the context words.

The negative sampling technique is adopted for constructing the training objective function, which is defined as follows. The stochastic gradient descent (SGD) method is used for optimization and the gradients are calculated using the back propagation neural networks.

$$l = \frac{1}{\varphi}\sum_{i=1}^{\varphi} log\, \sigma(x_i \cdot x_{\text{c}}) + k \cdot \mathcal{N}(w' \sim w_i) \cdot log\sigma(x' \cdot x_{\text{c}})$$

where $\sigma(x) = 1/(1 + \exp(-x))$ and k is the number of negative samples, $\mathcal{N}(w' \sim w_i)$ denotes the sampled word collection of word w_i, w' represents one sampled word, and x' is its corresponding vector.

Conversely, the Skip-Gram model predicts the context words using the target word. Thus, the objective function $L(D)$ of the Skip-Gram model is:

$$L(D) = \frac{1}{\varphi}\sum_{i=1}^{\varphi} \sum_{-k \le c \le k, c \ne 0} \log Pr(w_{i+c} | w_i)$$

$$Pr(w_{i+c} | w_i) = \frac{\exp(x_{i+c} \cdot x_i)}{\sum_{w \in W} \exp(x \cdot x_i)}$$

where x_{i+c} is the vector representation of a context word for the target word w_i.

The Skip-Gram model also adopts the negative sampling and SGD approaches for optimization, and the objective function can be represented as:

$$l = \frac{1}{\varphi}\sum_{i=1}^{\varphi} \sum_{-k \le c \le k, c \ne 0} log\sigma(x_{i+c} \cdot x_{\text{i}}) + k \cdot \mathcal{N}(w' \sim w_{i+c}) \cdot log\sigma(x' \cdot x_{\text{i}})$$

where k is the number of negative samples, $\mathcal{N}(w' \sim w_{i+c})$ denotes the negative sample collection of context word w_{i+c} .

Considering corpus C (i.e., a list of m records where each record is a row) is the input to the pre-processing step, and the raw output is an $n \times \lambda$ matrix M_{raw}. λ is the parameter that determines the dimension of the word's vector representation, and $M_{\text{raw}} = \{x_1, x_2, \dots, x_i, \dots, x_{n-1}, x_n\}^T, 1 \le i \le n$. However, since the aim is to construct a clustering method that groups records, but the labels in real-world bibliometric data are also usually based on records, the embedding vector r_i of a record is constructed from a simple average of all the word embedding vectors that align with the record, i.e.,

$$r_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j$$

where n_i is the number of distinct words that a record contains.

The output of this step is a $m \times \lambda$ matrix $M = \{r_1, r_2, \dots, r_i, \dots, r_{m-1}, r_m\}^T, 1 \le i \le m$.

3.2. Kernel K-means Clustering

Traditional k-means clustering algorithms, and their extensions, have been widely exploited for bibliometric data and have received relatively positive feedback (Colavizza & Franceschet 2016; Zhang et al. 2016b). However, the need for direct or indirect human intervention has still been criticized – e.g., the need to use a category code assigned by experts, such as international patent classification codes. Therefore, to improve the performance of k-means algorithms without human intervention, we have introduced a kernel method to map relatively low-dimensional data into a high-dimensional feature space, which is able to identify non-linear relationships in bibliometric data.

Even though the radial basis function (RBF) kernel, also known as the Gaussian kernel, is commonly used in support vector machine classification and traditional kernel k-means approaches, experimental results conducted by Chang et al. (2010) indicate that a polynomial kernel function performs better in information retrieval, especially with natural language processing tasks. Therefore, we have exploited a polynomial kernel function to calculate the product of two vectors r_i and r_j:

$$K_{poly}(r_i, r_j) = \emptyset(r_i) \cdot \emptyset(r_j) = (\varepsilon r_i \cdot r_j + \tau)^d$$

where ε and τ are parameters, and d is the degree.

Additionally, bibliometric researchers have discovered that a cosine measurement (Salton & McGill 1986) performs better than other approaches including Euclidean distance and Jaccard index (Leydesdorff 2008; Zhang et al. 2016a). Thus, the polynomial kernel function in our method is coupled with a cosine-based similarity measurement, rather than the Euclidean distance that is usually used in traditional kernel k-means models. Hence, the similarity measurement function $\mathrm{Sim}(r_i, r_j)$ in our method is calculated as

$$\mathrm{Sim}(r_i, r_j) = \cos(r_i, r_j) = \frac{r_i \cdot r_j}{\sqrt{r_i \cdot r_i} \cdot \sqrt{r_j \cdot r_j}} \rightarrow \frac{\mathrm{K_{poly}}(r_i, r_j)}{\sqrt{\mathrm{K_{poly}}(r_i, r_i)} \cdot \sqrt{\mathrm{K_{poly}}(r_j, r_j)}} = \frac{(\varepsilon r_i \cdot r_j + \tau)^d}{\sqrt{(\varepsilon r_i \cdot r_i + \tau)^d} \cdot \sqrt{(\varepsilon r_j \cdot r_j + \tau)^d}}$$

Further, some basic structures were adopted from an existing k-means clustering approach, outlined in the work of Zhang et al. (2016b). The resulting kernel k-means clustering algorithm is described below.

Step 1 Initialization: K records are randomly selected in the dataset as initial centroids $c_p, 1 \le p \le K$;

Step 2 Clustering: the similarity $\mathrm{Sim}(r_i, c_p)$ between both each record and each centroid is calculated and the record r_i is assigned to the most similar centroid c_p;

Step 3 Centroid re-calculation: the new centroid c'_p is re-calculated for each cluster by averaging the embedding vectors of all records r_i assigned to the cluster, i.e.,

$$c'_p = \frac{1}{n_{c_p}} \sum_{i=1}^{n_{c_p}} r_i$$

where n_{c_p} is the number of records in the cluster represented by centroid c_p.

Step 4 Terminative condition: the matrix for all old centroids is denoted as $\Theta = \{c_1, \dots, c_p, \dots, c_k\}^T, 1 \le p \le k$ and the matrix for all new centroids is denoted as $\Theta' = \{c'_1, \dots, c'_p, \dots, c'_k\}^T$. Then, the Euclidean distance $d(\Theta, \Theta')$ between Θ and $\Theta'$is measured. If $d(\Theta, \Theta') > \xi$, the algorithm returns to Step 2; otherwise, it terminates. Here, ξ is a parameter.

$$d(\Theta, \Theta') = \sqrt[2]{\sum_{p=1}^{k} (c_p - c'_p)^2} = \sqrt[2]{\sum_{p=1}^{k} \sum_{i=1}^{\lambda} (\vartheta_{p,i} - \vartheta'_{p,i})^2}$$

where $\vartheta_{p,i}$ and $\vartheta'_{p,i}$ is the ith element of centroid c_p and c'_p, respectively.

The output of the clustering step is a list of clusters, with each cluster containing a set of records.

4. Experiments

The experiments were designed to compare quantitatively the proposed word embedding-incorporated kernel k-means clustering algorithm to several clustering baselines, including a k-means algorithm, a principal component analysis (PCA) algorithm, a topic modeling algorithm, and a fuzzy c-means algorithm. Two bibliometric datasets (i.e., scientific articles from the Web of Science and academic proposals granted by the National Science Foundation of the United States) and two parallel validation measurements (i.e., counting pair-based clustering evaluation metrics and Herfindahl index) were conducted.

4.1. Experiment Design and Validation Measurements

Aiming to validate the results, three series of experiments were conducted with the use of two sets of evaluation metrics, adapting to the interest of both research communities, i.e., information retrieval and bibliometrics.

4.1.1. Experimental design

Three series of experiments were conducted with the following design and settings.

1) Experiment Series 1 compared the performance of the proposed method with four clustering baselines. All methods require the number of clusters to be set manually, so this was set to an interval of [5, 20]. The parameter ξ for the terminative condition in our kernel k-means method was set to 0.0001. The three selected baselines were:

- The traditional k-means (KM) clustering algorithm integrated within MATLAB's statistics and machine learning toolbox;[3]
- The traditional PCA clustering algorithm integrated within MATLAB's statistics and machine learning toolbox;
- The LDA algorithm of the topic model (TM) written by Steyvers and Griffiths (2007), which is considered to be the "official" topic modeling toolbox in MATLAB[4], with modified code to retrieve the topic distribution of articles as the output for evaluation; and
- The fuzzy c-means (FCM) algorithms integrated within MATLAB's fuzzy logic toolbox[5].

2) Experiment Series 2 compared the performance of four different word embedding models on clustering tasks. The four models were: both models in the Word2Vec method [i.e., CBOW and Skip-Gram (SG)], the Paragraph Vector (PV) model (Le & Mikolov 2014), and a pre-trained (PT) global vector model (Pennington et al. 2014). In addition, a word vector (WV) was generated for each article based on the term clumping process (Zhang et al. 2014), which was also used as a baseline input for the clustering approaches, with the parameter λ set to the default of 100 typical of most word embedding models.

- The PV model is an important baseline in word embedding. Its main idea is to learn feature representations from variable-length pieces of text, such as sentences, paragraphs, and documents, rather than the entire corpus. We set each article as a piece of text.
- The global vector model is pre-trained by Pennington et al. (2014), based on a corpus of 6 billion tokens collected from Wikipedia 2014 and Gigaword 5, which represents approximately 400 thousand words through 100-dimensional vectors.

---

[3] See http://au.mathworks.com/help/stats/kmeans.html for a description and more details
[4] See http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm for more information
[5] See https://au.mathworks.com/help/fuzzy/fcm.html for a description and more details

3) Experiment Series 3 compared the performance of two different kernel functions: a radial basis function (RBF) and a polynomial function (PF). The parameters in both functions were set according to the experiments conducted by Chang et al. (2010) as follows:

- RBF – $1/2\sigma^2 = 0.18$.
- PF – $\varepsilon = 0.18$ and $\tau = 0.3$, but considering negative numbers exist in the output of word embedding, we decided to set $d = 3$ rather than $d = 2$ as usual.

4.1.2. Validation Measurements

1) Counting pair-based clustering evaluation metrics

As a common validation measurement for information retrieval, we used counting pair-based clustering evaluation metrics (Xuan et al. 2018), including the Jaccard Coefficient (JC), the Folkes & Mallows (FM), and the F1 measure (F1). Given a clustering result,

- *a* is the count of pairs of articles that are grouped in the same cluster and also fall within the same category in golden standards.
- *b* is the count of pairs of articles that are not grouped in the same cluster but fall within the same category in golden standards.
- *c* is the count of pairs of articles that are grouped in the same cluster but do not fall within the same category in golden standards.

The three metrics were measured in the following way:

$$JC = \frac{a}{a+b+c}$$

$$FM = \left(\frac{a}{a+b} \cdot \frac{a}{a+c}\right)^{1/2}$$

$$F1 = \frac{2a^2}{2a^2+ac+ab}$$

Compared to the traditional set of matching-based evaluations (e.g., precision, recall, and F-measure), these three counting pair-based evaluation metrics better satisfy certain clustering constraints, such as clustering homogeneity and clustering completeness (Amigó et al. 2009).

2) Herfindahl index

Despite a close relationship with information retrieval, the bibliometric community has also proposed a number of its own validation measuring approaches. In this experiment, we used the Herfindahl index (H index) to measure the concentration of the results, following two representative studies conducted by Boyack et al. (2011) and Klavans and Boyack (2017). Both studies were based on global document models (or large-scale datasets) and certain specific bibliometric elements, e.g., the grant-to-article linkages and references, but the relatively small dataset used in this experiment would not be suitable for using such elements, e.g., most references are not within the dataset and the acknowledged grants of involved articles are very scattered. Therefore, we still used the WoS categories as the golden standards, but calculated the H index in a modified way, given as follows:

$$H_i = \sum \left({}^{n_{i,j}}/_{n_i}\right)^2$$

where $n_{i,j}$ is the number of articles with a WoS category i in cluster j, and n_i is the total number of articles in the category i.

An overall value for each cluster solution is then calculated as the weighted average over all categories:

$$H = \sum \frac{n_i}{N} H_i$$

where N is the total number of articles in the dataset.

4.2. Topic Extraction for Scientific Articles from the Web of Science

4.2.1. Data

The Web of Science (WoS) database[6] provides a disciplinary classification system in the form of WoS Categories. With the aid of WoS's own subject experts and journal editors (or publishers), every journal covered by the WoS Core Collection is assigned to at least one WoS category, and any articles published in a given journal are linked to the categories of the publishing journal. We strategically selected 10 categories and randomly retrieved approximately 500 articles from each category on 23 June 2017. The description of the dataset is provided in Table 1.

Table 1. Data description.

No	*WoS Category*	*Abbreviation*	*Num. of Articles*	*Multi-category Articles*
1	Biochemistry & Molecular Biology	BM	517	87 in CM
2	Mathematics	MA	516	13 in CA, and 3 in HP
3	Computer Science, Artificial Intelligence	CA	500	-
4	Nanoscience & Nanotechnology	NN	500	-
5	Chemistry, Medical	CM	497	-
6	Education & Educational Research	EE	489	13 in HP
7	History & Philosophy of Science	HP	489	-
8	Nursing	NU	464	-
9	Business	BU	459	-
10	Engineering, Aerospace	EA	455	-

Note that multi-category articles were assigned to the category with a largest number of articles. For example, we assigned the 87 articles that align with both BM and CM to BM, and the 13 articles that align with both EE and HP to EE.

We assembled 4770 articles in total, 116 of which were aligned to two disciplines.

The 10 categories were selected based on the following criteria:

- The dataset should contain a relatively broad range of disciplines, rather than a specific domain;
- The dataset should represent both fundamental and emerging disciplines;
- Some categories should contain terms that are unique to that discipline, while other categories should share common terms, so that the algorithms' ability to handle complicated real-world requirements can be assessed.

We fully agree that WoS categories are related to disciplines rather than topics that usually focus on much smaller units (Klavans & Boyack 2017), but the ten selected categories share relatively low disciplinary similarities with each other. For example, 1) the resulting dataset spanned science, engineering, information technology, social science, and certain cross-disciplines; 2) It represents fundamental disciplines (e.g., mathematics) and emerging disciplines (e.g., nanoscience & nanotechnology); 3) some of these categories, like business, use relatively unique terms, while others, like computer science, artificial intelligence, and mathematics, share some common terms.

[6] See https://webofknowledge.com/ for more information

We intentionally designed such composition to examine the ability of our method to distinguish coupling content. Given the circumstances, it would be reasonable to assume that scientific articles within the same WoS category in our dataset could belong to a relatively general topic, and the WoS categories would be an acceptable choice for this experiment.

Regarding data pre-processing, we used the titles and abstracts of the 4770 articles, and constructed a 4770-row list for further analysis, in which each row represents one article through its combined title and abstract. No additional pre-processing activities were conducted.

4.2.2. Results

1) Counting pair-based clustering evaluation metrics

Twenty four groups were compared in the experiments, as shown in Table 2, along with the average values of the three validation indicators of counting pair-based clustering evaluation metrics (i.e., JC, FM, and F1), demonstrating their overall performance. The detailed evaluation results are shown in Figures 2 to 4, and we record the number of topics at which the top five groups with the related index reach their best performance. Such information would provide a reference for further studies on deciding how many clusters actually exist.

Table 2. Comparative groups and their overall performance with JC, FM, and F1 (the WoS data)

Group	*Description*	*JC*	*FM*	*F1*
#23	PFKM&CBOW	**0.3489**	**0.5278**	**0.5159**
#19	RBFKM&CBOW	0.3436	0.5199	0.5106
#2	KM&SG	0.3416	0.5201	0.5080
#22	PFKM&SG	0.3415	0.5195	0.5085
#3	KM&CBOW	0.3320	0.5101	0.4970
#18	RBFKM&SG	0.3316	0.5100	0.4972
#24	PFKM&PT	0.2570	0.4167	0.4078
#5	KM&PT	0.2325	0.3841	0.3766
#20	RBFKM&PT	0.2318	0.3828	0.3755
#16	TM&WV	0.2245	0.3732	0.3656
#13	FCM&CBOW	0.1867	0.4134	0.3146
#1	KM&PV	0.1777	0.3065	0.3008
#21	PFKM&PV	0.1744	0.3025	0.2962
#17	RBFKM&PV	0.1723	0.2993	0.2933
#15	FCM&PT	0.1618	0.3660	0.2786
#11	FCM&PV	0.1558	0.3325	0.2693
#8	PCA&CBOW	0.1332	0.2461	0.2350
#7	PCA&SG	0.1323	0.2874	0.2334
#14	FCM&WV	0.1149	0.2161	0.2060
#10	PCA&PT	0.1131	0.2380	0.2032
#4	KM&WV	0.1072	0.2551	0.1935
#12	FCM&SG	0.1061	0.3258	0.1919
#9	PCA&WV	0.0907	0.1842	0.1662
#6	PCA&PV	0.0797	0.1503	0.1473

Note that 1) the three abbreviations represent the three validation indicators respectively, i.e., JC: Jaccard Coefficient; FM: Folkes & Mallows; and F1 measure; 2) We sort the results based on JC to help highlight factors that boost the results; 3) Regarding the abbreviations of related methods: KM – k-means; PCA – principal component analysis; FCM – fuzzy c-means; TM – topic model; PFKM – polynomial function-based kernel k-means; RBFKM – radical basis function-based kernel k-

means; CBOW – the continuous bag-of-word model; SG – the skip-gram model; PT – the pre-trained model; PV – the paragraph vector model; and WV – word vector.

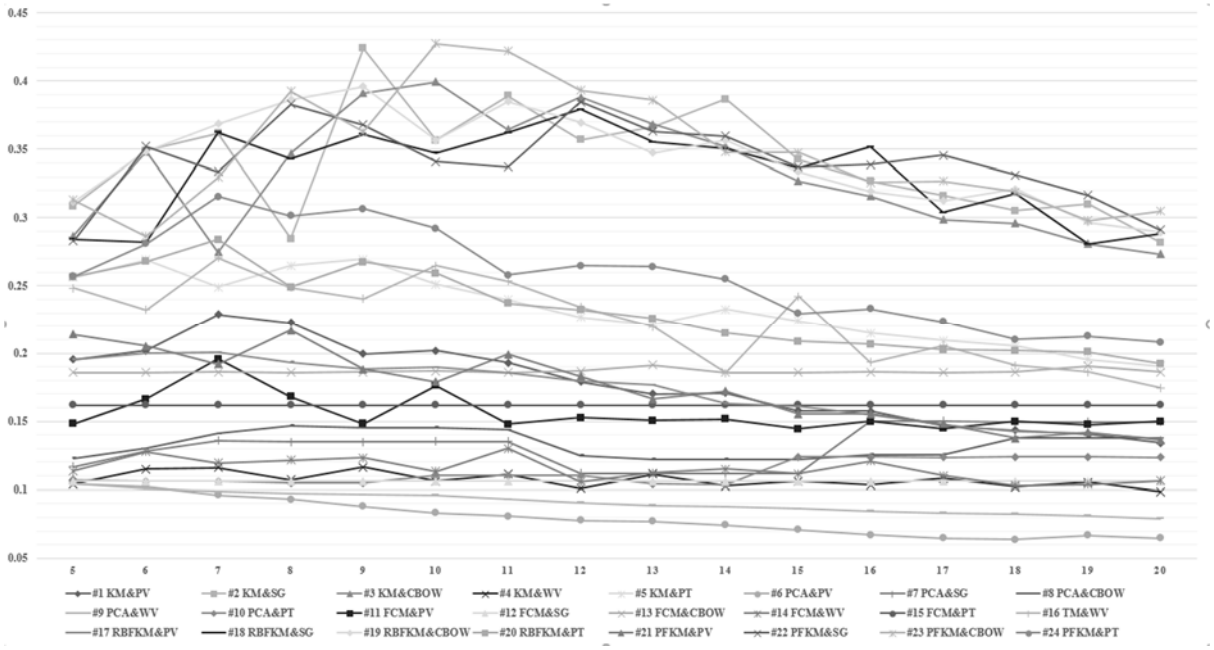

Figure. 2. JC validation results.

Note that #23 reaches its peak at 10 topics, #19 at 9 topics, #2 at 9 topics, #22 at 12 topics, and #3 at 10 topics.

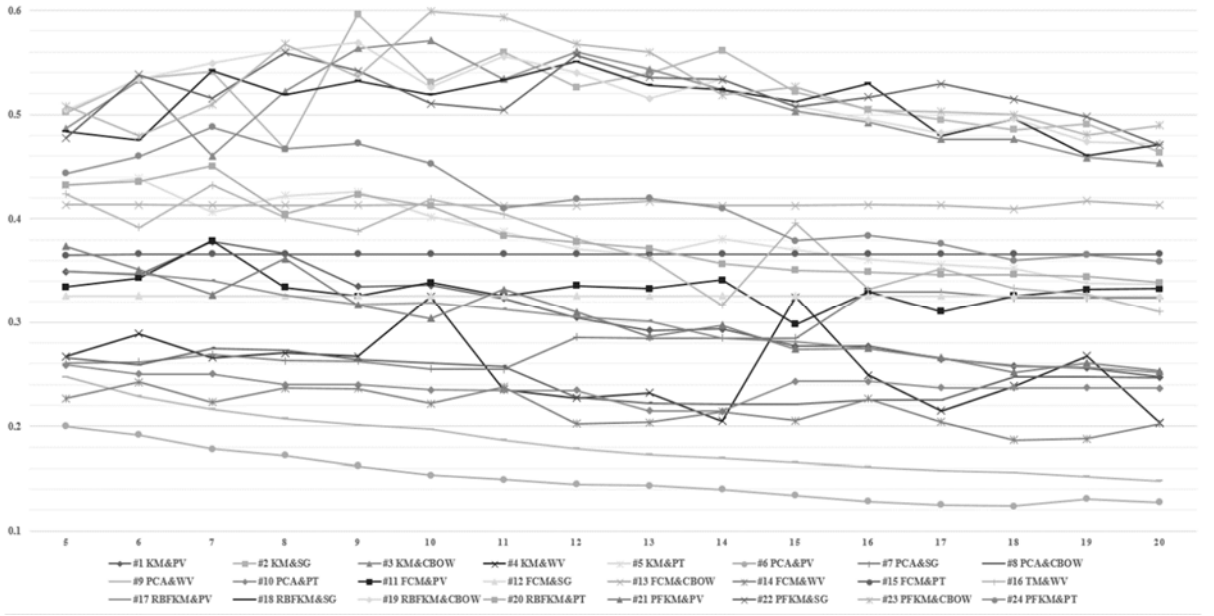

Figure. 3. FM validation results.

Note that #23 reaches its peak at 10 topics, #2 at 9 topics, #19 at 9 topics, #22 at 8 topics, and #3 at 10 topics.

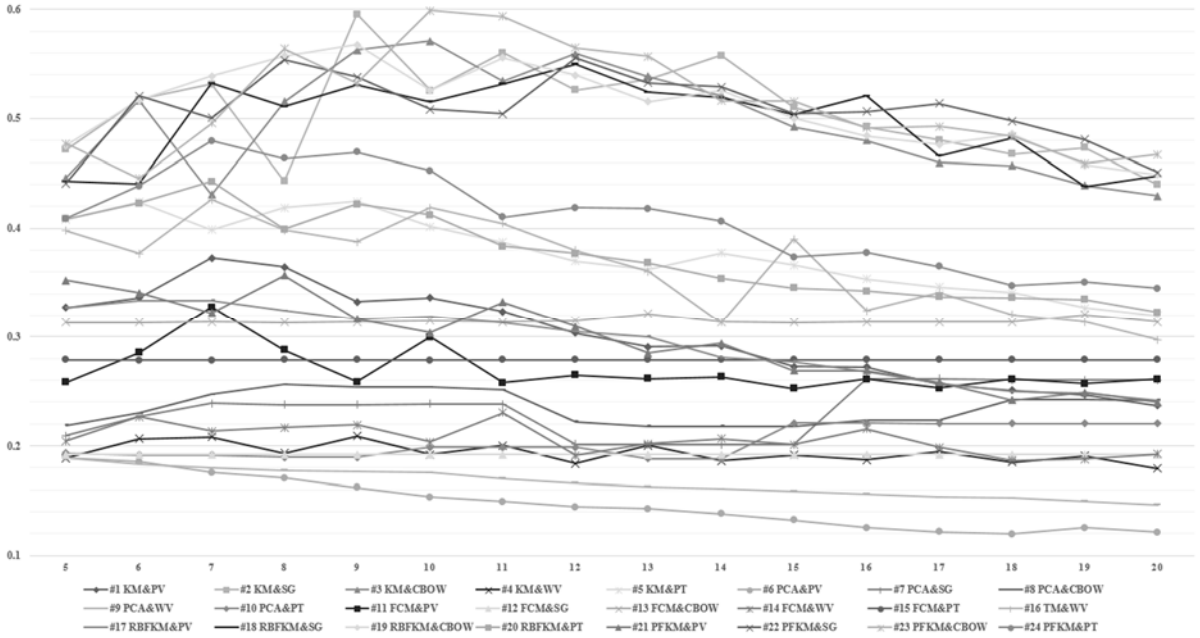

Figure. 4. F1 validation results.

Note that #23 reaches its peak at 10 topics, #19 at 9 topics, #22 at 12 topics, #2 at 9 topics, and #18 at 12 topics.

Some details on the experimental design are given below: 1) the classic topic model in Experiment Series 1 requires discrete values as input, but the output of word embedding is continuous. Thus, it is not applicable to directly integrate both. However, since the benefits of using topic models to conduct clustering tasks with word vectors have been widely approved (Ding & Chen 2014; Suominen & Toivanen 2016), it is reasonable to set Group 16 as a benchmark for traditional clustering approaches without word embedding. 2) We tested Groups 1 to 16 first and, as shown in Figures 2 to 4, word vectors did not work well with either the k-means, PCA, or the fuzzy c-means approaches (i.e., Groups 4, 9, and 14), and we conducted no further experiments assembled with the kernel k-means method and word vectors.

Based on the observations collected from Table 2 and Figures 2 to 4, certain insights follow:

- Word embedding techniques significantly increase the performance of clustering approaches. Evidence can be traced from the comparative pairs. Comparably, k-means approaches adapt to word embedding techniques better than PCA and fuzzy c-means approaches.
- The Word2Vec method provides more benefit to clustering approaches than the PV model and the pre-trained model. However, despite the slightly better performance observed for the CBOW model, the SG model is still competitive.
- The proposed kernel k-means method generated the highest values of all the groups. Compared to the RBF-based k-means approaches, the PF-based approaches showed the highest average values in all the three indicators (see Table 2). However, when analyzing the detailed results (see Figures 2 to 4), their performance was relatively less stable than the RBF-based approaches with less than 8 topics.
- Based on word vectors, topic models perform much better than the traditional k-means, PCA, and fuzzy c-means approaches. However, considering the gap between topic models and the use of word embedding techniques, it is not easy to judge whether such integrations would provide benefits. This is a tempting task for further study.

2) Herfindahl index

Table 3 provides the comparison among the twenty four groups with the average values of the H index. However, the results, compared with those of the counting pair-based clustering evaluation metrics, are relatively intriguing. As indicated in Table 3, #4, #7, #11-13, and #15 (underlined in Table 3) – the six groups are ranked in the relatively low level of Table 2 – achieve incredibly high H index values. Thus, we delved into the results and uncovered these reasons:

- The output of the fuzzy c-means approaches is a record-topic matrix, and each cell represents the membership grade of a record to a topic. Since we consider a record belongs to a topic with the highest membership grade, it results in the situation that records highly concentrate on several topics. Similarly, the imbalance of the results in #4 and #7 also results in concentrated 'big' topics. Definitely, a high concentration follows the basic concept of the H index, and then leads to a high value of H index.
- However, we manually checked the results of the six groups and compared them with the golden standards, and we do not consider they can exactly reflect the real situation of the experimental dataset, e.g., as a general plot, #12 (at all solutions with different number of topics) assigned all 4770 records into one topic.
- One additional understanding is that an alternative metric granularity can be used when one has cluster solutions with different distributional solutions (Waltman et al. 2017).

We, therefore, decided to discuss the results mostly based on the other groups and some insights are concluded as follows:

- Based on the H index, the SG model performed slightly better than the CBOW model, and both models are more beneficial to the topic extraction task than the PV model. Despite intriguing results for the traditional k-mean approaches (i.e., #4 is better than other three comparative groups with the traditional k-means approaches), the results of the PCA and fuzzy c-means approaches indicate the benefit of the Word2Vec methods compared to these clustering approaches.
- The PF-based approaches illustrate the best performance in the H index, when compared with the RBF-based approaches in all given scenarios and the traditional k-means in the groups incorporated with word embedding techniques. This endorses the findings observed from the validation based on the counting pair-based metrics.

Table 3. Comparative groups and their overall performance with H index (the WoS data)

		H Index
#12	FCM&SG	*0.9999*
#13	FCM&CBOW	*0.8972*
#15	FCM&PT	*0.7923*
#4	KM&WV	*0.6653*
#11	FCM&PV	*0.6567*
#7	PCA&SG	*0.5682*
#22	PFKM&SG	**0.5411**
#2	KM&SG	0.5393
#23	PFKM&CBOW	0.5390
#18	RBFKM&SG	0.5374
#19	RBFKM&CBOW	0.5315
#3	KM&CBOW	0.5248
#24	PFKM&PT	0.4352
#10	PCA&PT	0.4272
#5	KM&PT	0.3914

#20	RBFKM&PT	0.3901
#16	TM&WV	0.3805
#8	PCA&CBOW	0.3408
#1	KM&PV	0.3198
#21	PFKM&PV	0.3138
#17	RBFKM&PV	0.3129
#14	FCM&WV	0.2972
#9	PCA&WV	0.2923
#6	PCA&PV	0.1522

Note that regarding the abbreviations of related methods: KM – k-means; PCA – principal component analysis; FCM – fuzzy c-means; TM – topic model; PFKM – polynomial function-based kernel k-means; RBFKM – radical basis function-based kernel k-means; CBOW – the continuous bag-of-word model; SG – the skip-gram model; PT – the pre-trained model; PV – the paragraph vector model; and WV – word vector.

In conclusion, the experiment of topic extraction for scientific articles from the Web of Science quantitatively examines the performance of the proposed method through the comparisons with certain baselines, and we identify certain key findings as follows:

- Both validation measurements endorse that the incorporation of the Word2Vec methods would leverage the ability of clustering approaches for topic extraction, but as indicated in our case the performances of both models (i.e., the SG model and the CBOW models) are very close. However, the PV model would not be suitable for bibliometric data, since its main idea (i.e., learning feature representations from variable-length pieces of text) might not make good sense for bibliometrics.
- The experiment examines whether the use of the polynomial function in a kernel k-means clustering approach for topic extraction will achieve better performance than that of the radial basis function.
- In our designed dataset, the proposed kernel k-means method incorporated with the Word2Vec method achieves the best performance, compared with traditional k-means approaches, fuzzy c-means approaches, PCA, and topic models. However, topic models could be still competitive in word vector-based topic extraction.

Regarding the design of the experiments and related validation measurements, we raise some points on text-based topic extraction:

- The proposed method and the experiments align with text-based topic extraction, and the results might be different if citation data are applied.
- The relatively low values of the nineteen groups in the counting pair-based clustering evaluation metrics might result from the use of the specific dataset rather than the invalidity of the proposed method. As indicated in our experiments, the results support that our method will perform better for topic extraction than those existing baselines in the given dataset.
- While the use of the H index in measuring the concentration of topics has been extensively discussed (Boyack et al. 2011; Klavans & Boyack 2017), our results suggest that the H index could fully reflect its ability as a validation indicator in global data models. But, when a relatively small dataset contains topics with a similar number of records, a high value of the H index might result from the imbalance generated by some imperfect clustering approaches. Additionally, using the grant-to-article linkages or references as the golden standards might constrain the H index.

4.3. Topic Extraction for Academic Proposals Granted by the United States National Science Foundation

Aiming to further examine the effectiveness of the proposed kernel k-means method incorporated with a wording embedding model in a dataset within a relatively narrow area, we selected a training set we designed couple years

ago (Zhang et al. 2016b), which contains 557 academic proposals granted by the National Science Foundation (NSF) of the United States (US) in 2009, under the Division of Computer and Communication Foundation. The 557 proposals were archived by officers of the US NSF, and the original labels include 10 categories, such as RI (robust intelligence), SHF (software and hardware foundations), III (information integration and informatics), and NeTS (networking technology and systems). Similarly to the WoS case, we only used titles and abstracts of these records and no further pre-processing activities were conducted.

Considering our above understanding on the use of the H index, we only applied the counting pair-based clustering evaluation metrics to the nineteen groups. As indicated in Table 4, the results mostly coincide with the insights that we explore from the WoS case, e.g., 1) despite the SG model being prior to the CBOW model in this case, it supports our conclusion that both models are competitive for topic extraction; and 2) the incorporation of kernel functions and word embedding techniques enhance the performance of clustering approaches. Additionally, the results are also consistent with the experiments conducted by Zhang et al. (2016b) in some sense, e.g., k-means approaches perform better than topic models in this dataset. However, except the groups with PCA approaches, the experiments indicate that incorporation with the PV model negatively influences performance.

Table 4. Comparative groups and their overall performance with JC, FM, and F1 (NSF data)

Group	*Description*	*JC*	*FM*	*F1*
#22	PFKM&SG	**0.2281**	**0.3767**	**0.3712**
#18	RBFKM&SG	0.2230	0.3707	0.3645
#19	RBFKM&CBOW	0.2161	0.3607	0.3549
#3	KM&CBOW	0.2152	0.3595	0.3534
#2	KM&SG	0.2146	0.3591	0.3526
#23	PFKM&CBOW	0.2133	0.3576	0.3515
#13	FCM&CBOW	0.1525	0.2981	0.2644
#12	FCM&SG	0.1399	0.2794	0.2454
#24	PFKM&PT	0.1124	0.2055	0.2020
#14	FCM&WV	0.1097	0.2488	0.1977
#5	KM&PT	0.1078	0.1976	0.1943
#6	PCA&PV	0.1048	0.3238	0.1898
#20	RBFKM&PT	0.1043	0.1921	0.1889
#15	FCM&PT	0.1024	0.2057	0.1858
#7	PCA&SG	0.1017	0.2551	0.1846
#8	PCA&CBOW	0.0962	0.2140	0.1753
#10	PCA&PT	0.0944	0.1938	0.1725
#9	PCA&WV	0.0918	0.1923	0.1681
#4	KM&WV	0.0771	0.1467	0.1430
#21	PFKM&PV	0.0700	0.1335	0.1306
#16	TM&WV	0.0685	0.1306	0.1276
#17	RBFKM&PV	0.0679	0.1302	0.1269
#11	FCM&PV	0.0677	0.1300	0.1265
#1	KM&PV	0.0657	0.1261	0.1231

Note that we sort the results based on JC to help highlight factors that boost the results; Regarding the abbreviations of related methods: KM – k-means; PCA – principal component analysis; FCM – fuzzy c-means; TM – topic model; PFKM – polynomial function-based kernel k-means; RBFKM – radical basis function-based kernel k-means; CBOW – the continuous bag-of-word model; SG – the skip-gram model; PT – the pre-trained model; PV – the paragraph vector model; and WV – word vector.

5. Empirical Study: Topic Extraction for Bibliometrics from 2000 to 2017

The aim of the empirical study was to apply the proposed kernel k-means method incorporated with a word embedding model to a set of articles published by three top-tier bibliometric journals – the *Journal of the Association for Information Science and Technology* (JASIST), the *Journal of Informetrics* (JOI), and *Scientometrics* (SCIM). Compared with the quantitative validation measurements in Section 4, this emprical study is to qualitatively evaluate our method's performance, with the aid of leading bibliometric experts. Additionally, this study will also explore empirical insights of relevance to stakeholders, such as journal publishers, editorial boards, and the research community.

5.1. Data and Topic Extraction

We retrieved 6811 articles[7] from the WoS database on 24 August 2017. The collection comprised 3359 SCIM articles, 2784 JASIST articles, and 668 JOI articles, which was further narrowed to 6767 articles that contained both a title and an abstract. The word embedding process was completed using the CBOW model (with $\lambda = 100$) to generate a 6767×100 matrix.

Following the parameters for Group 19 in Table 2, the k-means method with a polynomial kernel function was used to extract topics from the dataset. We initially ran several clustering tests with different numbers of topics and finally set it to 8, considering it would result in minimum duplicate topics. However, since word embedding techniques represent words and articles via abstract feature vectors, we decided to use high-frequency terms to describe topics to help with the subsequent expert validation and visualization.

The natural language processing function within VantagePoint[8] was applied to the entire dataset, and 112,204 terms were extracted. The term clumping process (Zhang et al. 2014) was then used to remove noise and consolidate technological synonyms. The processing criteria included: 1) removing terms starting with non-alphabetic characters (e.g., "1.5%"); 2) removing meaningless terms (e.g., pronouns, prepositions, and conjunctions); 3) removing common terms in scientific articles (e.g., "method"); 4) consolidating terms with the same stem (e.g., singular/plural words); 5) removing terms appearing in only one article; and 6) removing single words (e.g., "dataset"). Ultimately, 12,776 terms were collected.

The strategy we followed for selecting "unique" terms to describe a topic follows. 1) The 12,776 terms were linked to the ten topics according to a "terms – articles – topics" structure; and 2) the top 10 highest-frequency terms in a topic were usually selected as descriptive terms, but the proportion of a term in a topic was also taken into consideration. Additionally, the following criteria were considered when labeling the topics: 1) whether the largest proportion of the term was contained in a topic, and 2) whether this was the highest-frequency term in a topic. The details of the eight topics are provided in Table 5.

Table 5. Details of the ten bibliometrics-related topics from 2000 to 2017

Topic	***#Art***	***Descriptive Terms***
1	896	information science; information seeking; information systems; information behavior*; digital library
2	1030	bibliometric analysis*; social network analysis; co-word analysis; co-citation analysis; case study
3	879	citation indicators; impact factor; journal citation report; citation impact; citation analysis*
4	789	information retrieval*; text mining; classification; semantic analysis; meta data
5	453	search engine*; web search; search process; search behavior; user satisfaction
6	888	h index*; g index; rankings; power law; citation distribution
7	957	social science; peer review; bibliometric indicators; research performance*; scientific community
8	875	international collaboration; co-authorship analysis; scientific production; R&D; scientific collaboration*

[7] Note the "articles" document type was selected to avoid review or other papers. Hence, all articles are research papers.

[8] VantagePoint is well-recognized in bibliometrics, especially for word or term analysis (see https://www.thevantagepoint.com/)

Note: #Art = the number of articles associated with a topic; underlined terms are consistent results compared with the case study conducted by Hou et al. (2018), see Section 5.3; terms with * were manually selected to represent their related topics.

5.2. Expert Knowledge-based Validation

Having already quantitatively validated the effectiveness of the proposed method in a dataset with relatively broad disciplines according to the two sets of validation indicators, it is interesting to qualitatively evaluate the method's performance in a specific domain, with the aid of leading domain experts.

A two-round expert evaluation was designed and seven leading bibliometric experts were engaged in the evaluation. Round 1 followed the way of traditional questionnaires to invite experts to mark the grouped topics in Table 5, and three criteria were raised. Each expert was asked to score the three criteria. Generally, 1 meant excellent agreement, 0 meant strong disagreement, while an intermediate judgment (e.g., 0.7) was fine as well.

- *Coherence*: How well do the terms of the Topic go together?
- *Distinctiveness*: Is the Topic separate from the others?
- *Significance*: Is the Topic important within bibliometrics? But note that this is an extra task for topic extraction, since we focus on clustering rather than identifying emerging topics.

Two leading bibliometric experts participated in this evaluation, and the average scores of their evaluation results and their correlation coefficient are presented in Table 6. It is interesting that both experts hold very consistent thoughts, e.g., Topic 5 (search engine) is coherent and distinct but not an exact bibliometric topic, and Topics 3 (citation analysis) and 6 (h index) are coherent and very important bibliometric areas, but their composing terms might be not distinct enough. However, differences also exist between the two experts, e.g., one expert marks Topic 7 (research performance) with "1" on distinctiveness and "0.6s" on coherence and significance, while the other expert only marks the topic with "0.2", "0.3" and "0.5" respectively. In general, we receive acceptable average scores on coherence and distinctiveness, and a passable score for significance.

Table 6. Results of the first round expert evaluation

#Topic	***Coher.***	***Distinct.***	***Signif.***	***Coefficient***
1 information behavior	0.75	0.6	0.15	0.1429
2 bibliometric analysis	0.55	0.45	0.8	*-0.0822*
3 citation analysis	0.85	0.45	0.95	**0.9449**
4 information retrieval	0.75	0.65	0.4	0.189
5 search engines	0.8	0.85	0.15	**0.9878**
6 h index	0.75	0.55	0.85	**0.866**
7 research performance	0.45	0.6	0.55	*-0.7559*
8 scientific collaboration	0.65	0.75	0.75	0.5
Average	0.6938	0.6125	0.5750	0.3491

Aiming to further evaluate the ability of our proposed method on topic extraction, we developed a novel way to conduct the Round 2 evaluation, in which five bibliometric experts (different from the ones in Round 1) were involved. Besides *Coherence*, we raise a criterion of *Relevance*, i.e., is the Topic relevant with your own research (i.e., bibliometrics)?

We mixed up the 40 core bibliometric terms in Table 5 and asked experts to come up with N clusters (they can decide the N) based on their expertise. Each cluster should represent an area of research in bibliometrics (in some cases, information & library sciences), and each term can only be used once. After that, the experts would consider their own research and interest, and score these clusters, in which 1 meant excellent relevance; 0 meant strong

irrelevance, and an intermediate judgment is acceptable as well. This evaluation provides a relatively fair way to generate golden standards for the topic evaluation.

Briefly, 2 terms [i.e., 'text mining' and 'classification' -- 5%] were only selected by two experts, 10 terms [e.g., 'information science', 'power law', and 'information behavior' -- 25%] were selected by three experts, and the other 28 terms (70%) were selected by all of the five experts. We consider those unselected terms are either too general or within the area of information science rather than bibliometrics.

We then set the topics given by the five experts as the golden standards respectively and evaluated our generated eight topics as follows (similar with the JC index): 1) each topic contains five descriptive terms, i.e., 10 distinct term pairs; 2) we then looked for the pairs in the golden standards and confirmed whether the two terms of a pair are within one topic or not. If so, we consider this pair is grouped correctly; and 3) the percentage of correct pairs in the 10 distinct pairs of each topic is considered as the performance of our proposed method. The results of the Round 2 evaluation are given in Table 7.

Table 7. Results of the second round expert evaluation

Topic	***Expert 1***	***Expert 2***	***Expert 3***	***Expert 4***	***Expert 5***
1 information behavior	1	0	1	1	0.1
2 bibliometric analysis	0.2	0.3	0.3	0.6	0.6
3 citation analysis	0.3	0.2	1	1	0.4
4 information retrieval	0.6	0.1	0.3	1	0.3
5 search engines	0.3	0.4	0.6	1	0.4
6 h index	0.3	1	0.1	0.6	0.3
7 research performance	0.4	0.3	0.1	0.4	0.6
8 scientific collaboration	0.3	0.4	0.2	0.6	0.6

Note that as a reference, the five experts generated 5, 7, 8, 2, and 5 topics respectively, in which Expert 4 only splits the terms into two parts, i.e., bibliometrics, and general information science.

Despite the five experts being leading bibliometric researchers, it seems still not easy to make them coincide on all of the topics – some of them might focus on research evaluation, while some others concentrate on bibliometric methodologies. Given the circumstances, we explore certain meaningful observations from Table 7:

- Topics 1 (information behavior), 3 (citation analysis), and 5 (search engine) receive relatively good marks from the experts. One key reason for such performance would be their relatively independent descriptive terms, e.g., Topics 1 and 5 topics align with information science and Topic 3 involves most citation-related terms.
- Topics 4 (information retrieval) and 6 (h index) could be sound. The majority of both can match the golden standards, and the reason behind that could be that terms in these topics are still unique but some interdisciplinary or relatively general terms exist, e.g., citation distribution and classification.
- Topics 2 (bibliometric analysis) and 8 (scientific collaboration) are sort of confusing, and Topic 7 (research performance) receives the lowest overall score. We try to understand the situation and one explanation could be that experts assigned those bibliometric methodologies with certain specific applications, e.g., co-word analysis was linked with text mining, and social network analysis was considered as an approach for investigating international collaborations.

In summary, the two-round expert knowledge-based validation empirically examined the performance of the proposed method (i.e., the polynomial function-based kernel k-means clustering method incorporated with the Word2Vec method) on topic extraction.

5.3. A Comparison Study-based Validation

One empirical study investigating emerging trends and developments in information science disciplines[9] was published recently (Hou et al. 2018), in which 10 core journals were carefully selected and articles published between 1996 and 2006 were analyzed based on a co-citation approach. Despite the slight difference (i.e., time period and source journals), it is still interesting to compare our results with theirs (we abbreviate as Hou's work).

Our comparison was mostly based on their results presented in Table 4, and Figures 3, 6, and 8, and we specifically emphasized the inspection that *whether the eight topics and their descriptive terms appear in Hou's work as well.* Following the sequence of the eight topics given in Table 5, the results of the comparison are discussed in Table 8.

Table 8. Results of the comparison with the case study conducted by Hou et al. (2018)

Topic	***Comparison Comments***
1 information behavior	Information behavior and digital libraries were identified as two topics in Hou's work, and information seeking was assigned within the topic information behavior. Additionally, since Hou's work aimed to concentrate on information science disciplines, they excluded information systems journals, which might result in the missing of this term.
2 bibliometric analysis	Despite that there is no such a topic in Hou's work, involving different bibliometric methodologies; they identified bibliometric analysis as an inactive topic between 2003 and 2013, but specific methods could be traced in other topics, e.g., co-citation analysis was associated with citation analysis.
3 citation analysis	We both highlight the significance of citation analysis in bibliometrics, but Hou's work split it into topics, such as citation data, citation count, and citation performance, in which impact factor appeared as a key term.
4 information retrieval	Information retrieval system was identified as a topic in Hou's work, but intriguingly, classification was assigned to citation performance, which might be used as a tool for citation analysis.
5 search engines	Query log and academic web site were two topics in Hou's work. Despite not having the same terms, it is promising to consider that we coincide on this topic. In addition, Hou's work also considered this topic closely relates to information behavior, which in some sense might match our terms, such as search behavior and search process.
6 h index	Definitely, this topic is clearly identified in both our study and Hou's work. The only difference is Hou's work considered social science was a topic evolved from h index in 2016.
7 research performance	Despite most terms being found in Hou's work, their related topics are not consistent, e.g., Hou's work classified research performance into triple helix (a topic stands for collaboration) and bibliometric indicators belonged to the topic "citation performance."
8 scientific collaboration	A topic, triple helix, was clearly identified in Hou's work, which coincides with our results on collaboration.

Based on the discussion given in Table 8, we consider the majority of our results match Hou's work reasonably, and differences can be explained as follows: 1) there were 11 topics in Hou's work while we generated 8 topics, so it is acceptable that one our topic might relate to two or more of Hou's topics; 2) the inconsistent results are Topics 2 (bibliometric analysis) and 7 (research performance), which might result from our different clustering approaches, i.e., our method is based on semantic similarities while Hou's work exploits the document co-citation relationships. However, we also notice that the two topics also received very low scores in the expert knowledge-based validation, which would indicate potential limitations of using semantic similarities to extract topics in a relatively narrow domain; and 3) it seems that Hou's work might underestimate the importance of text-based approaches in bibliometrics, or all related terms were grouped in their topic information retrieval system.

5.4. Empirical Insights

[9] Bibliometrics were considered as an subarea of information science in the study of Hou et al. (2018), and they divided the timeline into two periods, i.e., 1996 – 2008 and 2008 – 2016. Additionally, the three bibliometric journals used in our studies are parts of their source journals.

With the reliability of the proposed method assessed quantitatively and qualitatively, our next task was to explore any empirical insights that may be relevant for stakeholders. The main focus of this analysis was to discover the main research interests of JASIST, JOI, and SCIM, and discuss their similarities and differences.

Undoubtedly, JASIST, JOI, and SCIM are the three top-tier journals for the bibliometric community, but each journal maintains unique coverage of the field. In an effort to distinguish these differences, we specifically recorded the composition of each of the eight topics by journal and the distributions in Table 9.

Table 9. Journal composition of ten bibliometric topics

Topic*	***SCIM***		***JASIST***		***JOI***	
	#A	***%***	***#A***	***%***	***#A***	***%***
1 information behavior	69	0.021	822	0.297	5	0.007
2 bibliometric analysis	731	0.219	189	0.068	110	0.165
3 citation analysis	640	0.192	162	0.059	77	0.115
4 information retrieval	127	0.038	632	0.228	30	0.045
5 search engines	8	0.002	442	0.160	3	0.004
6 h index	402	0.121	207	0.075	279	0.418
7 research performance	598	0.180	253	0.091	106	0.159
8 scientific collaboration	756	0.227	61	0.022	58	0.087

Note that: 1) here we manually selected one descriptive term to label its related topic; 2) #A = number of articles and % = proportion in the total number of the journal's articles.

It is interesting that a relatively scattered topic pattern and significant diversity among the three journals can be observed as follows.

- JASIST is the only journal with a clear interest in information behavior, information retrieval, and search engines.

In WoS, JASIST is aligned with two categories: Information Science & Library Science, and Computer Science & Information Systems. From this perspective, the composition of the JASIST community (i.e., the Association for Information Science and Technology) not only contains bibliometric researchers, but also a majority of members from broad disciplines in information technology. As a result, Topics 1 (information behavior) and 5 (search engine) make JASIST unique among the three journals. However, considering the descriptive terms of Topic 4 (information retrieval) in Table 5, text mining techniques are closely related, and the JOI and SCIM communities are interested in those topics as well, but not too much so far.

- JOI prefers theoretical studies, especially those associated with methodological innovations in bibliometrics, while SCIM holds interests in empirical studies as well, e.g., applications for bibliometrics.[10]

The research communities for JOI and SCIM overlap in many mainstream topics of bibliometrics, e.g., Topics 2 (bibliometric analysis), 3 (citation analysis), and 7 (research performance), but some differences between the two journals still can be identified. 1) SCIM has published a large number of empirical studies into bibliometrics, and a large proportion of those studies focus on Topic 8 (scientific collaboration). Whereas, JOI has published a very limited number of related articles in that area. 2) JOI dominates Topic 6 (h index), which also includes g indexes and some other extensions, and is also more likely to publish studies on research performance (i.e., Topic 7). 3) SCIM's strength in Topics 2 (bibliometric analysis) and 8 (scientific collaboration) indicates its interest in applying bibliometric methodologies (e.g., social network analysis) to investigate collaborations between specific entities, such as countries, research organizations, and individual researchers.

[10] Note that even though we highlight SCIM's interest in empirical studies here, SCIM does also include theoretical studies. However, as indicated in Tables 5 and 9, the majority of JOI articles relates to theoretical studies.

- When concentrating on bibliometrics, the proportion of articles published in JOI and SCIM largely exceeds that of JASIST.

Considering the number of articles grouped in the "pure" bibliometrics-related topics [i.e., excluding Topics 1 (information behavior) and 5 (search engines)] and its proportion in a journal's total published articles, JOI and SCIM dominate the field. However, based on the best knowledge of the authors, JASIST prefers bibliometrics-related papers that involve modern information technologies, such as big data analytics, machine learning, and pattern recognition, and social media data, like Tweets.

6. Discussion and Conclusions

This paper proposes a polynomial function-based kernel k-means clustering method that incorporates the Word2Vec model with quantitative and qualitative demonstrations of its effectiveness in topic extraction. These demonstrations prove that 1) word embedding techniques can be exploited to skip over human costs in traditional data pre-processing and 2) the incorporation of word embedding techniques with the polynomial function-based kernel k-means clustering method is superior to certain existing text-based clustering baselines (i.e., k-means, PCA, fuzzy c-means, and topic models) on topic extraction with two labeled test datasets (i.e., scientific articles from the WoS and academic proposals from the US NSF).

A qualitative evaluation was made through an empirical study on bibliometric topic extraction supported by expert knowledge. Further, several insights for stakeholders were revealed during the qualitative investigation of the similarities and differences between the JASIST, JOI, and SCIM journals. Several key findings include: 1) JASIST is the only journal with a clear interest in topics such as information behavior, information retrieval, and search engines, since its range covers the entire area of library and information sciences; 2) JOI prefers theoretical studies, especially those associated with methodological innovations in bibliometrics, while SCIM holds interests in empirical studies as well, e.g., applications for bibliometrics; and 3) when concentrating on bibliometrics, the proportion of articles published in JOI and SCIM largely exceeds that of JASIST.

6.1. Technical Implications

With the rapid development of text analytics, word- or term-based bibliometrics are attracting increasing interest. However, one remaining challenge with co-word-based topic analysis is how to effectively and efficiently extract key features – i.e., how to remove noise, consolidate synonyms, and weight key features. Despite efforts to semi-automatically reduce the level of human intervention required with techniques like term clumping, a good deal of manual effort is still required. By contrast, word embedding, as an application of deep learning techniques on NLP, creates a solution to achieve this goal by extracting a relatively small number of latent features that represent word semantics rather than the simple co-occurrence relationships reflected by traditional word vectors. Such an accomplishment would not only reduce the workload associated with further clustering approaches but would also generate a representative feature space for clustering.

Despite tremendous efforts to introduce kernel functions in clustering approaches oriented toward bibliometric data, the selection of kernel functions has not been thoroughly discussed. Many existing models simply apply the popular radial basis function. However, based on some frontier research in the field of machine learning, our kernel k-means clustering method integrates a polynomial kernel function, and the experimental results demonstrate its advantages in handling text data – especially bibliometric data in this case. Additionally, from a theoretical perspective, word embedding techniques prefer to exploit large-scale datasets (e.g., a corpus of several million or billion tokens) and k-means approaches have been proven that they are superior on handling large-scale clustering tasks. Therefore, we consider our proposed method would be feasible for large-scale topic extraction.

6.2. Possible Applications

It is conceivable that a kernel k-means clustering approach, with the use of a word embedding model, could be applied to a wide range of topic extraction tasks.

- As a basic clustering tool, several aspects of the proposed method could be extended from bibliometric data to general text data, which would provide competitive advantages when compared to other clustering approaches in text analytics.
- Our method could be integrated with other analytical approaches, such as science maps and network analysis, to conduct specific bibliometric tasks, e.g., investigating multidisciplinary interaction and evaluating research performance.
- The engagement of data analytic models in handling issues in science, technology & innovation policy (STIP) is also an emergent trend in related fields. It appears that combining topic analysis with technology management tools (e.g., technology roadmapping) could create complementary benefits.

6.3. Limitation and Future Study

Several limitations of the current research and related future directions are summarized as follows. 1) When word embedding techniques create an abstract feature vector to represent words and documents, it leads to difficulties with describing topics in an explainable way, which would be considered as a limit in practices. Even though several high-frequency terms were selected to describe and label topics in our case study, there could be a more coherent way to achieve this goal. 2) Word embedding techniques with kernel functions and k-means approaches both require a number of parameters. However, methods of training these parameters for optimal benefit is a task that falls into the field of machine learning. 3) Although our test and empirical datasets comprised relatively broad disciplines and a very specific and narrow research area, it would be interesting to test our method with some public datasets and/or to compare the results with clustering approaches that rely on other bibliometric indicators, such as citations/co-citation statistics.

Acknowledgements

We acknowledge Arho Suominen and Ying Huang for their efforts in the pre-round expert knowledge-based evaluation, and our heartfelt appreciation goes to Lutz Bornmann, Kevin Boyack, Andrea D'Angelo, Ying Ding, Richard Klavans, and Ismael Rafols, Anthony F.J. van Raan for their expertise on evaluating the empirical results. We also thank the two anonymous reviewers for their professional comments and suggestions.

This work is partially supported by the Australian Research Council under Discovery Grant DP150101645 and the United States National Science Foundation Award #1759960.

References

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval, 12*(4), 461-486.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., . . . Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One, 6*(3), e18029.

Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M., & Lin, C.-J. (2010). Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research, 11*(Apr), 1471-1490.

Chen, K.-Y., Luesukprasert, L., & Seng-cho, T. C. (2007). Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Transactions on Knowledge and Data Engineering, 19*(8).

Colavizza, G., & Franceschet, M. (2016). Clustering citation histories in the Physical Review. *Journal of Informetrics, 10*(4), 1037-1051.

de Paulo Faleiros, T., & de Andrade Lopes, A. (2015). Bipartite Graph for Topic Extraction. *International Joint Conference on Artificial Intelligence*, 4363-4364.

Dhillon, I. S., Guan, Y., & Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 551-556.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology, 65*(10), 2084-2097.
Dong, R., Schaal, M., O'Mahony, M. P., & Smyth, B. (2013). Topic extraction from online reviews for classification and recommendation. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 1310-1316.
Firth, J. R. (1957). A synopsis of linguistic theory 1930-55, *Studies in Linguistic Analysis* The Philological Society.
Funk, R. J., & Owen-Smith, J. (2016). A dynamic network measure of technological change. *Management Science, 63*(3), 791-817. doi: 10.1287/mnsc.2015.2366
Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: a document co-citation analysis (2009–2016). *Scientometrics, 115*(2), 869-892.
Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR), 31*(3), 264-323.
Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.
Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 881-892.
Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984-998.
Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning* (ICML-14), 1188-1196.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.
Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 2177-2185.
Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3*, 211-225.
Leydesdorff, L. (2008). On the normalization and visualization of author co‐citation data: Salton's Cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology, 59*(1), 77-85.
Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., . . . Fleming, L. (2014). Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy, 43*(6), 941-955.
Liu, Q., Huang, H., Zhang, G., Guo, Y., Xuan, J., & Lu, J. (2018). Semantic structure-based word embedding by incorporating concept convergence and word divergence. *The 32th AAAI Conference on Artificial Intelligence*, New Orleans, the US.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). Flat clustering *Introduction to information retrieval* (pp. 350-374): Cambridge University Press.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.
Nieminen, P., Pölönen, I., & Sipola, T. (2013). Research literature clustering using diffusion maps. *Journal of Informetrics, 7*(4), 874-886.
Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation.* The 2014 Conference on Empirical Methods on Natural Language Processing.
Peters, H., & van Raan, A. F. (1993). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy, 22*(1), 23-45.
Rip, A. (1988). Mapping of science: Possibilities and limitations. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 253-273). North-Holland: Elsevier Science Publishers B.V.
Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. Auckland: McGraw-Hill.
Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.
Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human‐assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(19), 2464‐2476.
Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics, 111*(2), 1169-1221. doi: 10.1007/s11192-017-2306-1
Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645-678.
Xuan, J., Lu, J., Zhang, G., Xu, D., Yi, R., & Luo, X. (2017). Dependent Indian Buffet process-based sparse

nonparametric nonnegative matrix factorization. *IEEE Transactions on Knowledge and Data Engineering, 29*(7), 1357-1369.

Xuan, J., Lu, J., Zhang, G., Da Xu, R. Y., & Luo, X. (2018). Doubly nonparametric sparse nonnegative matrix factorization based on dependent indian buffet processes. *IEEE Transactions on Neural Networks and Learning Systems, 29*(5), 1835-1849.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.

Zhang, Y., Shang, L., Huang, L., Porter, A. L., Lu, J., & Zhu, D. (2016a). A hybrid similarity measure method for patent portfolio analysis *Journal of Informetrics, 10*(4), 1108-1130. doi: 10.1016/j.joi.2016.09.006

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016b). Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179-191.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Science evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925-1939.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology, 65*(5), 995-1006.

Zhao, Y., & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning, 55*(3), 311-331.

Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery, 10*(2), 141-168. doi: 10.1007/s10618-005-0361-3