Journal of King Saud University – Computer and Information Sciences 34 (2022) 408–420

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

On evaluating the collaborative research areas: A case study

Mona Moradi, Mohammad Rahmanimanesh *, Ali Shahzadi

Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran

ARTICLE INFO

Article history:
Received 11 August 2019
Revised 23 September 2019
Accepted 10 November 2019
Available online 14 November 2019

Keywords:
Text mining
Text-similarity
Topic detection
Probability density function
Fuzzy clustering
Collaborative research

ABSTRACT

The growth of social networks is ever-increasing. Many available scientific publications evidence the interest of researchers in this area. Within a time span of eight years from 2011 to 2018, approximately 2600, 230, 150, and 110 scientific articles were published from the USA, Iran, Saudi Arabia, and Turkey, respectively around this area of research. To comprehensively survey all the sub-fields and interests within this research area, the present paper proposes a novel density-based method for finding topic descriptors from academic articles. By employing a robust to noise fuzzy clustering algorithm, the terms are clustered, and by utilizing a modified Parzen window, k topic descriptors from each cluster are extracted. Besides, an optimization problem has been designed to detect the similarity between word pairs. By conducting the experiments, the research priorities for four countries within this time span have been found. Moreover, the closeness of the research in developing countries to the developed country have been measured. The experimental results show that for four years, the research topics in Turkey were close to the research topics in the USA on average, and the research topics in Saudi Arabia were close to the USA topics during the past two years. Additionally, the experimental comparison of the proposed method with two clustering baselines indicates the superiority of the proposed method in terms of precision, recall, and accuracy.

1. Introduction

The interactions and communications between researchers are key issues worldwide. In spite of the geographical boundaries and cultural differences, researchers from different countries can collaborate to solve issues. It is worthwhile to mention that the researchers can address more issues if more cultural common points exist among them. Because of the indispensable role of researchers in scientific and technological advancements, it is necessary to use the experiences of others to avoid unnecessary duplication of effort and studies. Despite the fact that the increase in studies results in significant advances, some of them do not lead to worthwhile achievements because they might have been done in inappropriate directions. Moreover, good research ideas sometimes do not yield the expected consequences, although these issues are inevitable features of the scientific efforts. The research prioritization is a suitable solution to prevent wasteful studies. By prioritizing directions for future studies and present alternatives can be addressed. The importance of prioritization can be explained in two different aspects. First, it helps the researchers to identify key challenges. Second, it helps funders to decide how to investing in research projects.

Internationalization of research activities and also progression into the scientific, technical, and commercial links between developed and developing countries will bring academic and industrial growth. Detecting collaboration fields or even discovering the gap between national and international research can facilitate information exchange and provide opportunities for collaborations in universities, and research institutes in the future.

Today the world's dependence on the Internet, social networks, and network analysis significantly are increased. Hence, a historical review of related research and also finding the closest topics to the field of research are time-consuming enough. Text classification is a process that natural language texts are labeled with thematic categories from a pre-defined set (Manning et al., 1999). Supervised machine learning approaches like k-Nearest Neighbor (kNN), Support Vector Machines (SVM), Neural Networks (NN), Random Forest (RF), and Maximum Entropy (ME) (Elghannam, 2019; Francis and Sreenath, 2019) and also similarity/dissimilarity

* Corresponding author.
E-mail addresses: mmoradi@semnan.ac.ir (M. Moradi), rahmanimanesh@semnan.ac.ir (M. Rahmanimanesh), shahzadi@semnan.ac.ir (A. Shahzadi).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2019.11.005
1319-1578/

metrics (Al-Anzi and AbuZeina, 2017; Belazzoug et al., 2019) are widely-used for text classification and topic categorization.

Topic detection is defined as a task of discovering outstanding words from a collection of documents and any word that implies a matter dealt with a text is defined as the topic of the document. Up to now, many studies regarding natural language processing (NLP) have been proposed for topic detection (Choi and Park, 2019; Reihanian et al., 2016; Winarko and Pulungan, 2019). Majority of these studies have been resolved this issue from the prospect of the clustering techniques (Aiello et al., 2013; Dutta et al., 2019; Wartena and Brussee, 2008), particularly *k*-means (Li et al., 2016; Zhu et al., 2019). Fuzzy logic-based techniques can improve the performance of NLP systems by handling natural language characteristics like ambiguity and uncertainty (Lau et al., 2008; Rashid et al., 2019; Sheeba and Vivekanandan, 2014). Probabilistic approaches such as Latent Dirichlet Allocation (Blei et al., 2003; Ihou and Bouguila, 2019), Latent Semantic Analysis (LSA) (Chun-hong et al., 2011; Deerwester et al., 1990), and Probabilistic Latent Semantic Analysis (PLSA) (Brants et al., 2002; Hofmann, 1999) have been studied for a long time. Neural networks (Poria et al., 2016; Rajaraman and Tan, 2001), and Independent Component Analysis (ICA) (Grant et al., 2008) are another attractive techniques to reach some appropriate results. Despite the variety of topic detection methods, this area still faces some shortcomings:

- The traditional methods mostly use clustering techniques, but these techniques perform inefficiently when exposing to the noise and outliers.
- The probabilistic methods often tend to ignore the words with low frequencies, and this approach may lead to poor results, especially in the cases that prominent words occur with low frequency.

The proposed method aims to (1) extract the trend topics from the collection of scientific articles, automatically, (2) discover the research priorities in different countries, and (3) measure the closeness of the research areas in developing countries to a developed country. Accordingly, the articles from four countries (three neighboring countries, Iran, Saudi Arabia, Turkey as developing countries, and the USA as a developed country) were selected for evaluation.

For some reasons, Iran, Saudi Arabia, Turkey, and the USA were chosen as the countries of interest for this study. First, the nationality of the authors of the present paper is Iranian, and naturally, Iran's activities in science and technology are the authors' concerns. According to the statistics obtained from (World Population Review, 2019), the average population rank of this country is 18.5. Second, there are wide relevant factors across Iran, and its two neighboring countries, Saudi Arabia, and Turkey (e.g., economic and cultural structures). Saudi Arabia's average population rank is 41, and Turkey's rank is 18. Third, the USA is a country that stands at the frontier of knowledge. Its average population rank is 3. Comparing the scientific situations of Iran, Saudi Arabia, and Turkey with the USA helps these three countries facing limited resources to clarify directions for present and future studies.

In the present paper, a robust to noise fuzzy clustering method called Fuzzy *C*-Ordered-Means clustering (FCOM) (Leski, 2016) is used to summarize the main topics. The reason for selecting the fuzzy technique is that the topics of a particular knowledge area are associated with each other, mainly. Moreover, the topics may also be shared with other knowledge areas; thus, some words are related to different domains in varying degrees. Utilizing the word embedding technique associated with the probabilistic density function, a method is designed therein top-*k* words (topic descriptors) are extracted to inference the topic. Moreover, exploiting an optimization problem helps both to identify the research priorities in different countries and to determine the closeness of the research areas in developing to the developed country. The results of the experiments demonstrate that the proposed method provides superior benefits over the competing methods.

The contribution of this paper is threefold:

1) Annotating scientific publications and finding research priorities with keywords and abstracts have essential roles for searching, indexing, and categorizing such documents. Hence, a novel method is proposed to recognize and extract the topics from the collection of scientific articles; thereby, it reduces the time and cost needed for the manual annotation.
2) It identifies the research priorities in different countries. So, it helps researchers to discover appropriate directions for future studies. Also, it helps funders to decide how to invest in the research projects.
3) The proposed method allows quantifying how much the research areas in developing counties are close to the research areas in developed countries. The ability to detect trending topics within the countries standing at the frontier of knowledge has direct implications on the possibility of discovering active challenges and real-world open problems. Besides, this issue helps developing countries that face limited financial resources to clarify directions for present and future studies. Moreover, it reduces waste resulting from unnecessary duplication of studies.

The remainder of this paper is organized as follows: Sections 2 and 3 introduce the related work and foundations of the Fuzzy *c*-ordered means clustering, respectively. Section 4 gives a detailed description of the methodology. Section 5 illustrates the experimental results. Finally, Section 6 contains conclusions and future work.

2. Related work

In the literature devoted to the topic detection, methods can be divided into three main categories: the probabilistic topic models, document-pivot methods, and feature-pivot methods (Milioris, 2018). The probabilistic topic models are based on the probability distribution of some features like *n*-grams in the documents (Brants et al., 2002; Chun-hong et al., 2011; Hoffman et al., 2010; Wang et al., 2012). LSA, PLSA, and LDA are three widely used methods. LSA (Deerwester et al., 1990) is a foundational technique in topic modeling. Its core idea is to take a document-term matrix and decompose it into two separate matrices: document-topic matrix and topic-term matrix. PLSA (Hofmann, 1999) is the other popular topic modeling technique. Its core idea is to find a probabilistic model with latent topics that can generate the observable data in the document-term matrix. LDA model (Blei et al., 2003), as a well-established method for topic detection, is a Bayesian hierarchical probabilistic generative model in which each document is modeled as a discrete distribution over topics, and each topic is considered as a discrete distribution over terms. The methods mentioned above have difficulties in setting parameter values and knowing prior distribution.

Topic detection based on document-pivot is performed based on the semantic distance between documents (Hasan et al., 2016, 2019; Ozdikis et al., 2017; Panagiotou et al., 2016). To a better understanding of topics, some analytic techniques, including both topics evolving analysis and sentimental analysis, have been proposed. Various similarity detection techniques for document content have been proposed in the literature. Broadly speaking, there are three main approaches to compute relatedness (Niraula et al., 2015): (1) knowledge-based techniques use the ontology

concepts, e.g., WordNet, for finding the similarities between documents (Jayawardhana and Gorsevski, 2019; Juckett et al., 2019; Shenoy et al., 2012), (2) web-based techniques use search engines to collect cooccurrence statistics, e. g., Point-wise Mutual Information (PMI), for similarity detection (Recchia and Jones, 2009), and (3) corpus-based techniques use word vector representations obtained from a corpus to compute the distance between these features. Zhang et al. (2018b) proposed a polynomial function-based kernel k-means clustering method that incorporated the Word2Vec model for topic extraction. This method proved that word embedding techniques can skip over human costs in traditional data pre-processing. One of the drawbacks of document-pivot based methods is sensitivity to the noise and outliers. Also, these methods suffer from problems related to clustering.

The feature-pivot methods exploit the distributions of words and discover word-groups appearing together in the document (Guille and Favre, 2015; Padmaja et al., 2018). Ai et al. (2018) presented a two-phase topic detection method from a large number of textual digital materials for microblogs on spark. One drawback of these methods is that they depend on misleading term correlation.

Some of the feature-pivot methods utilize frequent pattern mining to find co-occurrence patterns containing more than two terms. Petkos et al. (2014) designed a soft frequent pattern mining algorithm for the topic detection problem. Choi and Park (2019) utilized HUPM (Liu and Qu, 2012) algorithm to detect candidate topics from Twitter streams. Their method took frequency and utility into account, simultaneously. In the end, to extract actual topic patterns from candidates, TP-tree was designed. Zhang et al. (2018a) proposed a pattern-based method to detect topics from a Twitter-like platform in China, which employs an FP-growth-like algorithm (Cao et al., 2014) to extract patterns and further summarizes them into topics by hierarchical clustering.

Graph-based methods are another type of feature-pivot methods. Chen et al. (2017) designed a graph-based method for topic detection using Markov decision processes. Clearly, all of these techniques have advantages and disadvantages, and a single method is not fully capable of quantifying the similarity/relatedness between text segments.

3. Foundations of the fuzzy c-ordered means clustering

Fuzzy C-means (FCM) is a clustering method in which each data object belongs to multiple clusters with varying degrees of membership. FCM is based on the minimization of the following objective function

$$J_{FCM} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} D^2(x_k, v_i) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \| x_k - v_i \|^2 \tag{1}$$

where n is the number of data points, c is the number of clusters, x_k is the k^{th} data object, v_i is the center of the i^{th} cluster, u_{ik} is the degree of membership of x_k to the i^{th} cluster, and m is the fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with $m > 1$. Fuzzy overlap refers to how fuzzy the boundaries between clusters are, that is the number of data objects that have significant membership in more than one cluster.

Although FCM is a well-established clustering algorithm, it has some deficiencies such as sensitivity to the fuzzification parameter, sensitivity to the initial guess for selecting clustering centers, and sensitivity to noise and outliers in data. Since all data objects have the same effect on the selection of the cluster center, the existence of even a few noises and outliers may lead to adverse impact on selecting the cluster centers. Various methods have been proposed in the literature to overcome these deficiencies. Recently, a new method named Fuzzy c-ordered-means clustering algorithm (FCOM) (Leski, 2016) was proposed to address the sensitivity to noises and outliers. FCOM employs two approaches. The first approach is using additional weighting β, and the second approach is utilizing the Huber's M-estimator $\mathscr{L}_{\text{HUB}}(\mathrm{e})$ as a robust loss function:

$$\mathscr{L}_{\text{HUB}}(\mathrm{e}) \begin{cases} \mathrm{e}^2/\delta^2, & |\mathrm{e}| \leqslant \delta \\ |\mathrm{e}|/\delta, & |\mathrm{e}| > \delta \end{cases} \tag{2}$$

where $\delta > 0$ is a tunable parameter. Considering these two approaches, the objective function of FCM is changed to:

$$J_{fcom}(\boldsymbol{V}, \boldsymbol{U}) = \sum_{i=1}^{c}\sum_{k=1}^{n} \beta_{ik} (u_{ik})^m \mathscr{D}(x_k,\ v_i) \tag{3}$$

where

$$\mathscr{D}(x_k, v_i) = \sum_{l=1}^{d} \mathscr{D}(x_{kl}, v_{il}),\ \mathscr{D}(x_{kl}, v_{il}) = \mathscr{L}(x_{kl} - v_{il}) \tag{4}$$

where $\mathscr{D}(\mathrm{x_k}, \mathrm{v_i})$ represents as a dissimilarity measure between k^{th} datum and i^{th} cluster center, d shows the dimension of data, $\mathscr{L}$ denotes Huber loss function, v_{il} is the l^{th} component of the i^{th} cluster center, m is fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with $m > 1$. Also, $\beta_{ik} \in [0, 1]$ denotes the typicality of the k^{th} datum regarding the i^{th} cluster; the lower $\beta_{ik} \in [0, 1]$ results in noise and outliers. Now, the set of all possible fuzzy partitions of n (d-dimensional) vectors into c clusters is defined by:

$$\mathscr{J}_{\text{fcom}} = \left\{ \mathrm{U} \in \mathbb{R}_{\text{cn}} |\ \forall_{\substack{1 \leqslant i \leqslant c \\ 1 \leqslant k \leqslant n}}\ \mathrm{u_{ik}} \in [0, 1];\ \sum_{i=1}^{c} \beta_{ik} \mathrm{u_{ik}} = \mathrm{f_k};\ 0 < \sum_{k=1}^{n} \mathrm{u_{ik}} < \mathrm{n} \right\} \tag{5}$$

where $\mathbb{R}_{cn}$ denotes a space of real $(c \times n)-$ dimensional matrices, and f_k parameter is an overall typicality of the k^{th} datum, which depends on typicality of the k^{th} datum concerning all the clusters. It is defined by Eq. (6).

$$\forall_{1 \leqslant k \leqslant n} f_k = \beta_{1k} \vee \beta_{2k} \vee \ldots \vee \beta_{ck} \tag{6}$$

where β_{ik} is determined for each cluster by ordering fixed values, called α_k, and $\vee$ denotes max operation. The form of parameter α_k can be sigmoidal as follows:

$$\alpha_k = 1 \Big/ \left\{ 1 + exp\left[\frac{2.944}{p_a n} (k - p_c n) \right] \right\},\ \ k \in \{1, 2, \ldots, n\} \tag{7}$$

where p_a and p_c are equal to 0.2 and 0.5, respectively. The obtained values for α_k vectors are in descending order $\alpha_1 > \alpha_2 > \ldots > \alpha_n$ where n is the number of data. To obtain the weights β of the i^{th} cluster by using the weights α, all data should be arranged according to their distance to the cluster center. In other words, the permutation function $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ is obtained in such a way that satisfies the following conditions:

$$\left| \mathrm{e}^{[\mathrm{r}-1]}_{\mathrm{i}\pi(1)\mathrm{l}} \right| \leqslant \left| \mathrm{e}^{[\mathrm{r}-1]}_{\mathrm{i}\pi(2)\mathrm{l}} \right| \leqslant \left| \mathrm{e}^{[\mathrm{r}-1]}_{\mathrm{i}\pi(3)\mathrm{l}} \right| \ldots \leqslant \left| \mathrm{e}^{[\mathrm{r}-1]}_{\mathrm{i}\pi(\mathrm{n})\mathrm{l}} \right| \tag{8}$$

where $\mathrm{e}^{[\mathrm{r}-1]}_{\mathrm{i}\pi(\mathrm{j})\mathrm{l}}$ is the distance of the $\pi(j)^{\text{th}}$ datum from i^{th} cluster center by considering the l^{th} component in the r^{th} iteration. After calculating the permutation function π, β parameters of i^{th} cluster are obtained as follows:

$$\beta_{ik} = \alpha_{\pi^{-1}(k)} \tag{9}$$

Generally, the high weights are assigned to the data that are close to the cluster centers. The degree of membership of the k^{th} datum in the i^{th} cluster is calculated from Eq. (10). Appendix A refers to the essential details (Leski, 2016).

$$\forall_{\substack{1 \leqslant i \leqslant c \\ 1 \leqslant k \leqslant n}} u_{ik} = f_k \mathscr{D}(\mathrm{x_k}, \mathrm{v_j})^{\frac{1}{1-\mathrm{m}}} \Big/ \left[\sum_{\mathrm{j}=1}^{\mathrm{c}} \mathrm{b_{jk}} \mathscr{D}(\mathrm{x_k}, \mathrm{v_j})^{\frac{1}{1-\mathrm{m}}} \right] \tag{10}$$

The l^{th} component of the cluster center i^{th} is calculated in the r^{th} iteration as follows:

$$\underset{\substack{1\leqslant i\leqslant c\\ 1\leqslant l\leqslant d}}{\forall} v_{il}^{[r]} = \left[\sum_{k=1}^{n}\beta_{ik}(u_{ik})^{m}h_{ikl}^{[r]}x_{kl}\right] \Big/ \left[\sum_{k=1}^{n}\beta_{ik}(u_{ik})^{m}h_{ikl}^{[r]}\right] \tag{11}$$

where

$$h_{ikl}^{[r]} = \begin{cases} 0, & x_{kl} - v_{il}^{[r-1]} = 0 \\ \dfrac{\mathscr{L}\left(x_{kl} - v_{il}^{[r-1]}\right)}{\left(x_{kl} - v_{il}^{[r-1]}\right)^{2}}, & x_{kl} - v_{il}^{[r-1]} \neq 0. \end{cases} \tag{12}$$

More details for obtaining Eq. (11) are provided in Appendix B (Leski, 2016). Algorithm 1 shows the FCOM clustering.

Algorithm 1 shows the FCOM clustering.

Algorithm 1. FCOM clustering (Leski, 2016).

1. Initialize $v_{il}^{[0]} = 0$. Set the iteration index $r \leftarrow 1$,
2. Calculate residuals $e_{ikl}^{[r-1]} = x_{kl} - v_{il}^{[r-1]}$ and then $h_{ikl}^{[r]}$ using Eq. (12),
3. Obtaining the permutation function $\pi(k)$ by considering Eq. (8),
4. Calculate α_k using Eq. (7),
5. Update the centers for the r^{th} iteration using Eq. (11),
6. If $\left|\left|v_{il}^{[r]} - v_{il}^{[r-1]}\right|\right|_2^2 > \varepsilon$ then $r \leftarrow r+1$ and go step 2, else stop.

4. Methodology

This section contains two subsections that focus on the proposed method for recognizing and extracting outstanding topics from the collection of scientific articles, and the method for detecting similarities. Now, the details of the proposed method are explained as the following.

4.1. Topic detection from the text

The block diagram of the proposed topic detection method is shown in Fig. 1, including document preprocessing, feature selection, vectorization, and inferring topic descriptors. The aforesaid stages will be explained in more detail below.

4.1.1. Document preprocessing

At first, the stopwords and the punctuation marks of each document are removed. Also, for lexicon development, a set of n-grams (any sequence of n words) are extracted from the original document collection. We empirically found out that two or three consecutive words (bi-grams and tri-grams) have more information gains than uni-grams in collected documents. So, all uni-grams are disregarded. The generated set is comprised of bi-grams and tri-grams.

4.1.2. Feature selection and vectorization

Feature selection is performed by using the *tf-idf* weighting scheme (Salton and McGill, 1986). For constituting the *tf-idf* matrix, bi-grams and tri-grams of the collected documents are considered. In this case, a document collection is represented as a $\boldsymbol{U} \times N$ document-term matrix where each row corresponds to one term; each column corresponds to one document $\{Doc_1, \cdots, Doc_N\}$, and each matrix entry denotes *tf-idf* value for the corresponding term and document as:

$$w_{j,m} = (Term\_Freq)_{j,m} \times \log_2 \frac{N}{n} \tag{13}$$

where $w_{j,m}$ is the weight of term $\boldsymbol{u}_m$ in document Doc_j, $(Term\_Freq)_{j,m}$ is the frequency of term $\boldsymbol{u}_m$ in document Doc_j, N is the size of document collection (i.e., the number of articles collected from country c), and n is the number of documents where $\boldsymbol{u}_m$ occurs at least once.

In the next step, the *tf-idf* matrix is refined. Because all of the terms (n-grams) are not useful for information retrieval, redundant and irrelevant data should be ignored. To this end, all terms are ranked according to their *tf-idf* values. Then, for each document, M terms with highest *tf-idf* values are kept. In this method, M was set to 20.

After that, each selected term $\boldsymbol{u}$ is transformed to a d-dimensional vector space by utilizing Doc2Vec embedding approach (Le and Mikolov, 2014). In this method, d was set to 50. Lastly, all extracted vectors from all documents are collected together to form our term repository.

4.1.3. Density-based topic detection

We now present our topic detection method proposed based on fuzzy clustering and density estimation. At first, the terms are clustered by the FCOM (Leski, 2016) on the data vectors. As stated earlier, the FCOM has drawn attention towards handling noisy data and outliers. Due to the applied fuzzy clustering, the extracted terms overlap, i.e., multiple terms are shared among clusters, means that these terms are corresponding to the same topic and causing the clustering to be less sensitive to the number of clusters.

Relevant words constitute dense clusters which cause easy identification of outstanding topics. To recognize relevant words, we propose a modified Parzen window with a Gaussian kernel that places the window at each word. Parzen window (Parzen, 1962) is a popular non-parametric approach to estimate the Probability Density Function (PDF) of a random variable. The modified Parzen window equation is defined as follows:

$$p(\boldsymbol{u}_m^{cl_i}) \approx \frac{\sum_{\substack{n=1\\ n\neq m}}^{|cl_i|}\left(\mu_m^{cl_i}(.) \odot_H \mu_n^{cl_i}(.)\right)}{(2\pi)^{d/2}\left|\Sigma^{cl_i}\right|^{1/2}} \exp\left(-\frac{\left(\boldsymbol{u}_m^{cl_i} - m^{cl_i}\right)^{T}\Sigma^{cl_i^{-1}}\left(\boldsymbol{u}_m^{cl_i} - m^{cl_i}\right)}{2}\right) \tag{14}$$

where $|cl_i|$ denotes the size of cluster cl_i (i.e., the number of terms belongs to the cluster), $\mu_m^{cl_i}$ is the membership value of d-dimensional term $\boldsymbol{u}_m$ to cluster $cl_i \in \Omega$, $\Omega = \{cl_1, cl_2, \ldots, cl_T\}$, m^{cl_i} is the mean vector of cluster cl_i, Σ^{cl_i} is the covariance matrix of cluster cl_i, and $\odot$ denotes the product of two membership values. To assign a weight to each term in a particular cluster, we utilize the Hamacher product (Silambarasan and Sriram, 2017) of the membership values as a weighting factor to show the significant degree of a term in that cluster. The Hamacher product of two membership values is utilized to reflect how relevant two words are in a given cluster. It is defined by

$$\mu_m^{cl_i}(.) \odot \mu_n^{cl_i}(.) = \frac{\mu_m^{cl_i}(.) \times \mu_n^{cl_i}(.)}{\mu_m^{cl_i}(.) + \mu_n^{cl_i}(.) - \left[\mu_m^{cl_i}(.) \times \mu_n^{cl_i}(.)\right]} \tag{15}$$

where $\mu(.)$ is the distributed membership value of each term among a particular cluster cl_i. Note the higher Hamacher value, the more the term belongs to the cluster. Using this value helps to identify whether this term will constitute a dense cluster. Considering Eq. (14), the normalized probability density value is calculated as Eq. (16).

$$\zeta_m = \frac{freq(\boldsymbol{u}_m)}{\underset{m\in cl_i}{Max}\, freq(\boldsymbol{u}_m)} \times \frac{p\left(\boldsymbol{u}_m^{cl_i}\right)}{\underset{m\in cl_i}{Max}\; p\left(\boldsymbol{u}_m^{cl_i}\right)}, \tag{16}$$

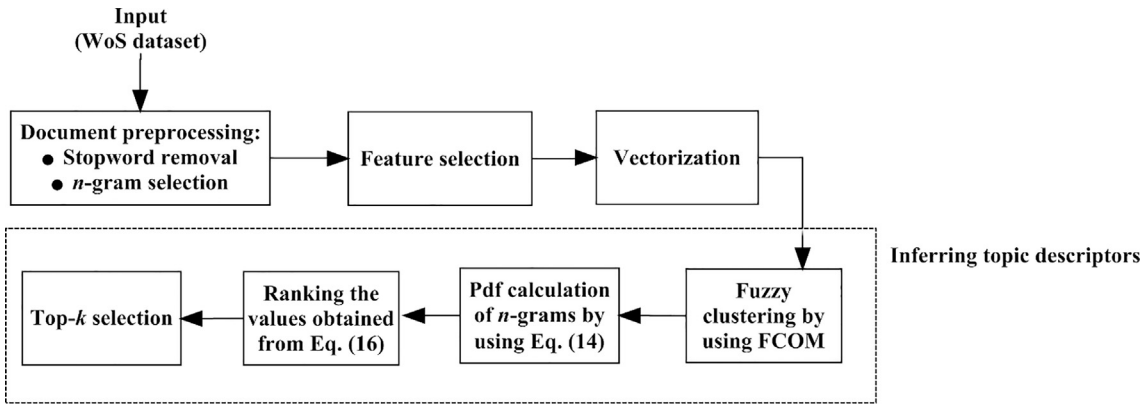

Fig. 1. The diagram of topic detection from scientific articles.

where $freq(\boldsymbol{u}_m)$ is a frequency of term $\boldsymbol{u}_m$ occurring in a document collection. Lastly, to recognize top-k topics for each cluster, all the obtained ζ_m are sorted in descending order, and k terms from the top of the list are selected. Then, each class $cl_i \in \Omega, \Omega = \{cl_1, cl_2, \ldots, cl_T\}$, is represented by a prototype $\pi = \{topic_1, ..., topic_k\}, i = 1, ..., T$. In the proposed method, clusters with fewer than four members are ignored.

4.2. Similarity detection

We now propose a distance-based optimizer to calculate the similarity between the terms extracted from each document and topic descriptors of each cluster. There are varied ways to compute features that perceive the semantics of documents, but one method that is surprisingly effective is to use the *tf-idf* values. The idea of the proposed optimizer is that the terms are represented as vectors of features, and they are compared by measuring the distance between these features.

Assume a repository contains N documents $\{Doc_1, ..., Doc_N\}$, and Doc_j includes M terms $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_M\}$. Let $\Omega^c = \{cl_1^c, cl_2^c, ..., cl_T^c\}$ be a set of class labels of country c, where each class is represented by a prototype $\pi_i = \{topic_1, ..., topic_k\}$, $i = 1, ..., T$ in which each term $topic_l$, $l = 1, ..., k$ is a d-dimensional observation, where d is the size of the vector obtained by the vectorization stage. The expected cost for classifying a document is computed by integrating *tf-idf* values with the Manhattan distance as follows.

$$\mathscr{J}_{\mathrm{doc_j}} = \frac{1}{|\pi_i| \times \mathrm{M}} \sum_{\mathrm{m=1}}^{\mathrm{M}} \sum_{\mathrm{l=1}}^{\pi_i} \sum_{\mathrm{k=1}}^{\mathrm{d}} \mathrm{e}^{-\mathrm{W_{rot_i}} \times \mathrm{W_{j,m}}} |\mathrm{u_{m.k}} - \mathrm{u_{l.k}}| \tag{17}$$

where $|\pi_i|$ denotes the number of topics in the i^{th} class, $w_{j,m}$ indicates the weight of term $\boldsymbol{u}$, $m = 1, ..., m$ in Doc_j, and w_{tot_i} is the mean of *tf_idf* values of the topics of the prototype of each class. w_{tot_i} is defined as:

$$w_{tot_i} = mean\left(w_{topic_{l,i}}\right), l = 1, \cdots, k; i = 1, \cdots, T \tag{18}$$

where $w_{topic_{l,i}}$ is the weight of $topic_l$ in cluster cl_i^c. Finally, the class of Doc_j is derived by:

$$cl_{j0}^c = \arg\min_j J_{doc_j} \tag{19}$$

5. Experimental results

In this section, the experimental results of the proposed method are reported. The used datasets are introduced in Subsection 5.1, concisely. Experimental setup and evaluation metrics are expressed in Subsections 5.2, and 5.3, respectively. Experiments on topic detection are presented in Subsection 5.4. Subsection 5.5 is comprised of several experiments for (1) identifying the research priorities in four countries, (2) comparing the performance of the Manhattan metric (employed in the proposed method) with three different metrics, (3) comparing the performance of the FCOM (employed in the proposed method) with three fuzzy clustering methods, and (4) comparing the performance of the proposed method with two baseline methods. Finally, the closeness of the research areas in developing countries to a developed country is evaluated in Subsection 5.6.

5.1. Dataset

The area of the proposed method concerns academic publications; on May 16th, 2019, by providing a data included papers from Iran, Saudi Arabia, Turkey, and the USA, the datasets were collected from Web of Science (WoS, 2019) considering 'social network' as the searched words, 'English' as language, 'Article' as document type, and '2011–2018′ as time span. The datasets characteristics are depicted in Fig. 2. Text mining analysis was carried out based on the titles, abstracts, and keywords of the publications. Moreover, the papers with no available title, abstract, or keyword(s) were excluded.

5.2. Experimental setup

For fair comparisons, in all experiments, the number of topics retrieved for each cluster was set to 4. The number of clusters should be large enough to discover more distinctive topics. Hence, the number of clusters was set to ten. For LDA, the total number of topics retrieved was set to 40. Also, all the cluster centers were initialized, randomly. For the FCOM, the Fuzzy C-Means (FCM), the Possibilistic Clustering (PC) (Krishnapuram and Keller, 1993), the Fuzzy Clustering with Polynomial Fuzzifier (FPCM) (Winkler et al., 2011) the weighting exponent m = 2 was used. For the weighting function presented in Eq. (7), $p_a = 0.2$ and $p_c = 0.5$ were used. For FPCM β = 0.5 was used. The iterations were stopped as soon as the Frobenius norm of the successive $\boldsymbol{V}$ matrices difference was less than 10^{-4} for FCOM, FCM, PC, FPCM, and k-means. For PC, η_i values were calculated by using Eq. (9) that was defined in (Krishnapuram and Keller, 1993). In that equation, $K = 1$ was used.

5.3. Evaluation metrics

Aiming to provide ground truth data for evaluations, the scientific articles were labeled manually by three, independent human annotators. We provided the lists of candidate class labels from

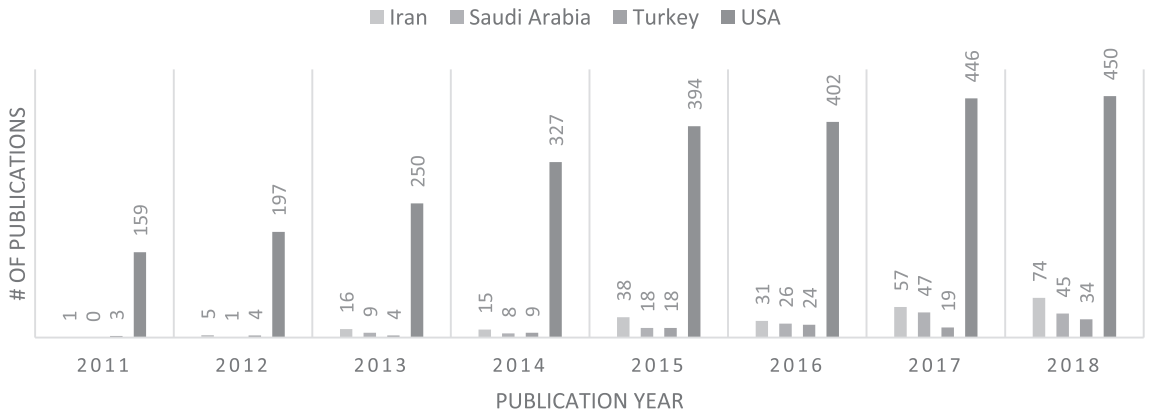

Fig. 2. Overview of the publications related to 'social network' by publication year.

Table 2, $\Omega^c = \{cl_1^c, cl_2^c, \cdots, cl_T^c\}$, and asked them to select the best label cl_i^c for each article. Lastly, we considered the union of the lists provided by the human annotators as the ground truth. Several examples of the ground truth are shown in Table 1.

To compare the efficiency of the proposed method, evaluation metrics such as precision, recall, and accuracy have been applied. These estimators are determined by Eqs. (20)–(22).

$$Precision = \frac{SS}{SS + SD} \tag{20}$$

$$Recall = \frac{SS}{SS + DS} \tag{21}$$

$$Accuracy = \frac{SS + DD}{SS + SD + DS + DD} \tag{22}$$

where *SS*, *SD*, *DS*, and *DD* are introduced in Table 2. For each pair of documents, one of the following four states may hold.

5.4. Detecting outstanding topics of research

The topics extracted from the collected articles are shown in Table 3; wherein blank cells denote an ignored cluster that its members are less than four. The terms collected (*n*-grams) are grouped into ten clusters, and top-4 terms of each cluster are extracted as cluster prototypes. The cluster prototypes of each country show which terms are closely related together.

The findings of the clustering analysis demonstrate that Iran, Saudi Arabia, and Turkey have nine distinct groups of topics in contrast to the USA that has ten distinct groups of topics. The most striking feature of Table 3 is that some clusters of a particular country share more terms (e.g., the USA), while some countries share fewer terms (e.g., Saudi Arabia). It is worth noting that the 'social network' domain is challenging because it extends across a wide range of disciplines sharing common topics (e.g., community detection, recommender systems, opinion mining, etc.). This domain includes not only disciplines containing unique terms but also disciplines sharing common terms. Accordingly, it would be reasonable to assume that the articles from the USA share relatively high disciplinary similarities with each other, whilst, the articles from Saudi Arabia share relatively low disciplinary similarities with each other.

Table 2
Confusion matrix.

		Labeled by the proposed method	
		Same cluster	**Different cluster**
Labeled by the human annotators	**Same cluster**	SS	DS
	Different cluster	SD	DD

5.5. Research priorities identification

The results of research priorities identification are reported in Tables 4–7. These tables contain the comparison between the Manhattan metric (applied in the proposed method) and three well-known metrics, named the Euclidean, the Cosine, and the Pearson correlation coefficient. The values represent the

Table 1
Examples of the ground truth.

Dataset	Title	Abstract	Keywords	Labeled by the human annotators
Iran	Energy and spectrum efficient mobility-aware resource management ...	The rapid growth of cellular traffic which is mostly due to increasing demands for multimedia and ...	Device to Device Communication; Multicasting; Energy Efficiency; ...	cl_3^{IRN}
Saudi Arabia	Social Commerce as a Driver to Enhance Trust and Intention ...	The deployment of cryptocurrencies in e-commerce has ...	Cryptocurrencies; trust; social commerce; social support; ...	cl_6^{SAU}
Turkey	Using swarm intelligence algorithms to detect influential ...	People use online social networks to exchange information, ...	Social influence analysis; Influential individuals; Influence maximization; ...	cl_4^{TUR}
The USA	Towards the implementation of the Social Internet of Vehicles	The Internet of Vehicles (IoV) represents the emerging paradigm ...	Vehicular Ad-hoc NETwork; Internet of Things; ...	cl_{10}^{USA}

Table 3
The cluster prototypes.

Country	Cluster									
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}
Iran	trust network	community detection	genetic algorithm	–	network influence	closeness centrality	opinion formation	future link	neural network	malware propagation scale
	detecting community	shortest path	influence maximization	–	centrality measure	finding influential node	multi objective	maximization problem	community detection dynamic	file sharing network
	meta heuristic algorithm	diffusion model	maximum clique stochastic	–	interest social network	predict future link	fuzzy set	similarity metric	particle swarm	user privacy concern
	fuzzy graph	link prediction	objective optimization	–	artificial immune network	automaton link prediction	recommender system	link prediction	semantic web	mobile social network
Saudi Arabia	citizen community structure	frequent subgraph mining	linguistic trust function	big data	recommender system	decision making	emotion recognition	twitter spam	feature space	–
	chaotic system	ad hoc	sentiment analysis	social network analysis	opinion dynamic	fuzzy set	privacy protection	hoc social network	privacy security	–
	credibility analysis system	content filtering	complex network analysis	internet thing iot	cold start	cyber harassment	fuzzy preference relation	community detection	socio cyber network	–
	confidence level	overlapping community	fuzzy inference system	matrix factorization	content based filtering	heterogeneous bibliographic information	multi criterion	tag recommendation	content analysis	–
Turkey	user behavior	supply chain network	socio economic variable	social network theory	–	content shared	based classifier	mobile local search	analytic hierarchy process	collaborative filtering
	use facebook	detection algorithm	prediction evolving	community detection algorithm	–	fuzzy set	heterogeneous social network	multiple classifier system	sustainable supply chain	psychiatric demographic socio
	temporal social network	criterion decision	business customer segmentation	seed selection heuristic	–	node similarity	depressed adolescent	fuzzy analytic hierarchy	type fuzzy	topic based microblogging
	graph processing	energy investment	scientific productivity cooperation	swarm intelligence	–	complex network	nomophobia level	sentiment analysis	anxiety social	link prediction evolving
USA	delay tolerant network	network analysis	information network	location recommendation	recommender system	community detection	twitter sentiment	information system	diffusion social	management system
	network formation	trust network	heterogeneous information	betweenness centrality	convolutional neural	neural network	trust propagation	information diffusion	influence social	mobile social network
	health social network	tolerant network	clustering algorithm	support vector machine	health informatics	recommendation system	evolution social network	ad hoc network	privacy setting	internet social
	web graph	online health community	mobile phone	analysis opinion mining	twitter social network	collaborative filtering	maximum likelihood estimation	communication technology	semantic web	hoc network

Table 4
Research priority in the USA (%) by using different metrics.

Year	Metric	cl_1^{USA}	cl_2^{USA}	cl_3^{USA}	cl_4^{USA}	cl_5^{USA}	cl_6^{USA}	cl_7^{USA}	cl_8^{USA}	cl_9^{USA}	cl_{10}^{USA}
2011	Manhattan	0.49	0.42	0.53	0.49	0.49	0.57	0.69	**1.03**	0.69	0.65
	Euclidean	0.37	0.57	0.69	0.42	0.53	0.50	0.61	**1.03**	0.84	0.50
	Cosine	0.57	0.53	0.42	0.69	0.65	0.49	0.53	0.69	**0.76**	0.72
	Pearson	0.46	0.69	0.34	0.72	0.65	0.53	0.53	0.69	0.69	**0.86**
2012	Manhattan	0.80	0.91	0.88	0.91	0.61	0.30	0.91	0.49	0.69	**1.01**
	Euclidean	0.76	0.88	0.84	0.84	0.65	0.42	0.80	0.61	0.72	**0.99**
	Cosine	**0.99**	0.76	**0.99**	0.69	0.69	0.46	0.65	0.76	0.57	0.95
	Pearson	0.91	0.69	0.95	0.69	0.84	0.49	0.65	0.76	0.53	**0.99**
2013	Manhattan	0.91	1.07	0.99	0.76	1.14	0.69	**1.22**	1.14	0.84	0.76
	Euclidean	0.88	1.07	0.95	0.84	0.88	0.88	0.91	1.10	**1.18**	0.84
	Cosine	0.84	1.18	0.91	1.14	0.80	0.72	1.07	**1.22**	0.95	0.69
	Pearson	0.80	1.18	0.91	1.14	0.84	0.76	1.02	**1.22**	0.95	0.69
2014	Manhattan	1.34	1.11	**1.45**	0.95	1.34	**1.45**	1.22	1.25	1.22	1.14
	Euclidean	1.26	1.03	**1.71**	0.95	1.37	1.26	1.30	1.03	1.18	1.37
	Cosine	1.07	1.07	1.22	1.30	0.76	1.44	1.37	**1.53**	**1.53**	1.18
	Pearson	1.02	1.02	1.30	1.44	0.69	1.22	1.34	1.48	**1.67**	1.25
2015	Manhattan	**1.71**	1.48	1.34	1.44	1.53	1.30	**1.71**	1.44	1.44	1.60
	Euclidean	1.52	1.48	1.33	1.60	1.71	1.10	1.68	1.52	1.26	**1.79**
	Cosine	1.34	1.44	1.53	1.60	1.25	1.37	1.34	1.44	1.64	**2.06**
	Pearson	1.48	1.48	1.60	1.64	1.30	1.41	1.22	1.44	1.60	**1.83**
2016	Manhattan	1.67	1.37	**1.95**	1.53	1.64	1.44	1.44	1.25	1.48	1.53
	Euclidean	**1.68**	1.52	1.52	1.60	1.56	1.45	1.52	1.49	**1.68**	1.30
	Cosine	1.56	**1.94**	1.71	1.34	1.87	1.48	0.84	1.56	1.79	1.22
	Pearson	1.44	1.79	1.79	1.37	**1.90**	1.41	1.11	1.60	1.64	1.25
2017	Manhattan	1.67	1.53	1.71	1.71	1.83	1.56	1.67	1.53	**2.03**	1.67
	Euclidean	1.45	1.52	1.83	1.90	1.90	1.41	1.79	1.30	**2.10**	1.78
	Cosine	1.41	1.41	1.83	1.48	1.79	1.41	**2.25**	1.37	**2.25**	1.79
	Pearson	1.30	1.25	1.79	1.56	1.79	1.44	**2.32**	1.56	2.17	1.79
2018	Manhattan	**2.18**	1.87	1.37	1.87	1.64	1.64	1.76	1.83	1.56	1.44
	Euclidean	**2.10**	1.94	1.30	1.87	1.64	1.83	1.71	1.60	1.71	1.45
	Cosine	2.02	1.71	1.30	1.94	1.44	**2.06**	2.02	1.53	1.48	1.64
	Pearson	1.79	1.64	1.14	**2.17**	1.48	2.06	1.94	1.60	1.48	1.83
Average	Manhattan	**1.35**	1.22	1.28	1.21	1.28	1.12	1.33	1.25	1.24	1.24
	Euclidean	1.25	1.25	1.27	1.25	1.28	1.11	1.29	1.21	**1.33**	1.25
	Cosine	1.23	1.26	1.24	1.27	1.16	1.18	1.26	1.26	**1.37**	1.28
	Pearson	1.15	1.22	1.23	**1.34**	1.19	1.17	1.27	1.29	**1.34**	1.31

percentage of publications that are assigned to each of ten clusters, and entries '-' denote the clusters with fewer than four members. Also, for each column, the average is calculated. The highest value for each row is shown in **bold** font.

The research priorities in the USA are listed in Table 4. By employing the Manhattan metric, the highest research priorities were assigned to topics presented in *cl*1. Moreover, by employing the Euclidean and Cosine, topics presented in *cl*9 achieved the highest priorities. By using the Pearson, topics presented in *cl*4 and *cl*9 obtained the highest priorities.

The research priorities in Iran are listed in Table 5. When Manhattan or the Euclidean were employed, the highest research priorities were assigned to topics presented in *cl*1. Besides, by employing the Cosine and Pearson, topics presented in *cl*7 achieved the highest priorities.

The research priorities of Saudi Arabia are listed in Table 6. It should be noted that, because the information on papers published in 2011 was incomplete, those records were ignored in the preprocessing stage. So, the first four rows of the table do not contain any data. As shown, when Manhattan was used, the highest research priorities were assigned to topics presented in *cl*6. Furthermore, by employing the Euclidean, and Cosine, topics presented in *cl*3 and *cl*6 achieved the highest priorities, respectively. By utilizing the Pearson, both *cl*6 and *cl*7 achieved the highest priorities.

The research priorities in Turkey are presented in Table 7. As shown, by employing the Manhattan, the highest research priorities were assigned to topics presented in *cl*7. Also, by employing the Euclidean, topics presented in *cl*1 achieved the highest priorities. By utilizing the Cosine and Pearson, topics presented in *cl*9 achieved the highest priorities.

Aiming to examine the effectiveness of the Manhattan dissimilarity metric applied in Eq. (17), this metric was compared with three metrics, named the Euclidean, the Cosine, and the Pearson correlation coefficient. The results obtained are listed in Table 8. Since the ranking provides a reasonable comparison of the methods, the performance of each method is ranked from the highest to the lowest, and they are reported in enclosed parenthesis. As can be seen, the Manhattan dissimilarity metric always achieves the best precision, recall, and accuracy. This is because small differences in that will not be ignored. The difference in performance with the Euclidean dissimilarity metric is smaller in terms of recall and accuracy. However, the Euclidean negatively influences the precision because this metric exaggerates large differences and ignores small differences. Also, the difference in performance with the Pearson similarity metric is smaller in terms of precision.

Besides, the effectiveness of the FCOM (the fuzzy clustering algorithm that has been employed in the proposed method) was compared with three existing fuzzy clustering methods, named, FCM, PFCM, and PC (see Table 9). The PFCM and PC are the fuzzy clustering methods that reduce the effect of outliers. In the PFCM method, for each data point, its distances to the prototypes are ordered. In the FCOM method, for each prototype, the data points will be ordered with respect to their distances from the prototype.

Table 5
Research priority in Iran (%) by using different metrics.

Year	Metric	cl_1^{IRN}	cl_2^{IRN}	cl_3^{IRN}	cl_4^{IRN}	cl_5^{IRN}	cl_6^{IRN}	cl_7^{IRN}	cl_8^{IRN}	cl_9^{IRN}	cl_{10}^{IRN}
2011	Manhattan	**0.43**	0	0	–	0	0	0	0	0	0
	Euclidean	**0.43**	0	0	–	0	0	0	0	0	0
	Cosine	**0.43**	0	0	–	0	0	0	0	0	0
	Pearson	**0.43**	0	0	–	0	0	0	0	0	0
2012	Manhattan	0	0	0.42	–	0.42	0.42	0	0	**0.85**	0
	Euclidean	0	0	0.42	–	0	0.42	0	0	**1.27**	0
	Cosine	0	0	0	–	0	0.42	0	0.42	0.42	**0.85**
	Pearson	0	0	0	–	**0.85**	0.42	0	0	**0.85**	0
2013	Manhattan	0.42	0.42	0.84	–	0.42	0	**2.96**	1.26	0.42	0
	Euclidean	0.42	0.42	1.27	–	0.84	0	**2.11**	0.84	0.84	0
	Cosine	0.42	0	0	–	0.42	0.42	**2.54**	1.69	0.42	0.84
	Pearson	0.42	0.42	0.42	–	0.42	0	**2.12**	1.26	0.42	1.26
2014	Manhattan	0.84	1.26	**1.70**	–	0.42	0.84	0	0	0.42	0.84
	Euclidean	0.84	0.84	0.84	–	0.84	0.84	**1.27**	0	0.42	0.42
	Cosine	0	0	**1.70**	–	0	1.26	1.26	1.26	0	0.84
	Pearson	0	0	**1.27**	–	0	0.84	**1.27**	**1.27**	0.42	**1.27**
2015	Manhattan	0.42	**2.96**	2.11	–	2.53	2.53	2.11	0.84	1.26	1.26
	Euclidean	0.42	2.53	**2.95**	–	2.11	2.11	1.69	1.27	1.27	1.69
	Cosine	0.84	2.95	0.84	–	2.95	0.84	1.69	1.26	**3.39**	1.26
	Pearson	1.26	**2.95**	1.26	–	**2.95**	1.26	1.26	0.84	2.11	2.11
2016	Manhattan	1.69	0	**2.13**	–	1.69	**2.13**	1.69	1.26	1.69	0.84
	Euclidean	1.69	0	2.11	–	**2.53**	1.27	2.11	0.84	1.69	0.84
	Cosine	1.26	1.69	2.11	–	0.84	0.42	0.42	2.11	1.69	**2.53**
	Pearson	1.26	1.69	2.11	–	0.84	0.42	0.42	2.11	1.69	**2.53**
2017	Manhattan	1.69	1.26	2.11	–	2.95	1.26	2.11	4.64	**5.07**	2.95
	Euclidean	1.27	0.84	2.95	–	2.11	1.69	1.69	4.64	**5.49**	3.38
	Cosine	1.69	2.53	2.53	–	3.80	0.84	**4.23**	2.95	3.38	2.11
	Pearson	1.69	2.53	2.11	–	3.80	1.69	**4.22**	2.11	3.80	2.11
2018	Manhattan	**5.90**	**5.90**	3.38	–	3.38	2.11	2.53	2.95	1.69	3.38
	Euclidean	**5.91**	**5.91**	4.22	–	2.11	2.11	1.69	3.38	2.95	2.95
	Cosine	4.64	**5.07**	4.64	–	3.38	1.69	3.80	3.38	1.69	2.95
	Pearson	4.64	4.64	3.80	–	3.38	1.69	**5.07**	3.80	1.69	2.53
Average	Manhattan	1.42	1.48	**1.59**	–	1.48	1.16	1.43	1.37	1.43	1.16
	Euclidean	1.37	1.32	**1.85**	–	1.32	1.06	1.32	1.37	1.74	1.16
	Cosine	1.16	1.53	1.48	–	1.42	0.74	**1.74**	1.63	1.37	1.42
	Pearson	1.21	1.53	1.37	–	1.53	0.79	**1.80**	1.42	1.37	1.48

In experiments, the FCOM was replaced by one of the algorithms mentioned above, and for fair comparisons, the rest of the steps remained unchanged. Also, the performance of each method was ranked, and the ranks were reported in enclosed parenthesis. As shown in Table 9, in the case that the FCOM is employed, the highest precision, recall, and accuracy are achieved.

To further verify, we compared the performance of the proposed method with two baseline methods (k-means and LDA) (see Table 10). By applying k-means, the centroids of clusters are treated as representative of latent topics. The experiments indicate that k-means obtains the lowest yield; meanwhile, the LDA surpasses the performance of k-means. As shown in Table 10, the proposed method provides some superior benefits over k-means and LDA.

5.6. Evaluating the closeness of the research areas in developing countries to a developed country

To evaluate the closeness of the research areas in developing countries to the developed country, the proposed method utilized a method similar to Subsection 4.2. Three countries, i.e., Iran, Saudi Arabia, and Turkey, are considered as developing countries, and the USA is considered a developed country. Let N be the size of document collection of developing country c (e.g., Iran, Saudi Arabia or Turkey) and M be the number of terms of document $Doc_j, j = 1, \cdots, N$ collected from that country. To measure the closeness an article from a developing country to the class cl_i^{USA}, $i = 1, \cdots, T$ of the USA, Eq. (17) is substituted into Eq. (19). Table 11 lists the results. Also, the highest average value in each year is shown in **bold**. As can be observed, Iranian research areas in 2013, and 2015 were close to the research areas in the USA. Turkey obtained the highest average values in 2011–2012, 2014, and 2016, that means the research trends in Turkey were more close to the research trends in the USA. Also, research areas in 2017 and 2018 in Saudi Arabia had the highest closeness to the research trends in the USA.

6. Conclusions and future work

In the present paper, we proposed a method that integrated a robust to noise fuzzy clustering algorithm with the probabilistic density function to attain three objectives: first, extracting the trend topics from the collection of scientific articles, automatically. Second, discovering the research priorities in different countries, and third, measuring the closeness of the research areas in developing countries to a developed country. The proposed method was carried out on the vast number of publications that emphasize 'social network'. By implementing the proposed method, the research priorities for four countries, (the USA, Iran, Saudi Arabia, and Turkey) were found, and the results revealed that for four years the research topics in Turkey were close to the research topics in the USA, and the research topics in Saudi Arabia were close to the USA topics during the past two years. The proposed method is applicable to any topic detection task, and it has

Table 6
Research priority in Saudi Arabia (%) by using different metrics.

Year	Metric	cl_1^{SAU}	cl_2^{SAU}	cl_3^{SAU}	cl_4^{SAU}	cl_5^{SAU}	cl_6^{SAU}	cl_7^{SAU}	cl_8^{SAU}	cl_9^{SAU}	cl_{10}^{SAU}
2011	Manhattan	0	0	0	0	0	0	0	0	0	–
	Euclidean	0	0	0	0	0	0	0	0	0	–
	Cosine	0	0	0	0	0	0	0	0	0	–
	Pearson	0	0	0	0	0	0	0	0	0	–
2012	Manhattan	0	**0.65**	0	0	0	0	0	0	0	–
	Euclidean	0	**0.64**	0	0	0	0	0	0	0	–
	Cosine	0	**0.65**	0	0	0	0	0	0	0	–
	Pearson	0	**0.66**	0	0	0	0	0	0	0	–
2013	Manhattan	0	**1.95**	0.65	0	1.30	0	1.30	0.65	0	–
	Euclidean	0	**1.95**	1.30	0	0.64	0	0.64	0.64	0.64	–
	Cosine	**1.30**	0.65	**1.30**	0.65	0	0	**1.30**	0	0.65	–
	Pearson	0	0.65	**1.30**	0.65	0.65	0	**1.30**	0.65	0.65	–
2014	Manhattan	0	0.65	0	0	**1.95**	0.65	1.30	0	0.65	–
	Euclidean	0	0	0.65	0.65	**1.30**	0.65	**1.30**	0	0.65	–
	Cosine	**1.30**	0	0.65	0	0.65	**1.30**	**1.30**	0	0	–
	Pearson	0.65	0.65	0.65	0	0.65	**1.30**	**1.30**	0	0	–
2015	Manhattan	0.65	0.65	1.30	0	1.95	1.95	**2.59**	1.30	1.30	–
	Euclidean	0	1.30	0.65	0.65	1.95	1.30	**3.90**	1.95	0	–
	Cosine	0.65	0	1.30	0.65	1.30	1.95	1.95	0.65	**3.25**	–
	Pearson	0.65	0	1.30	0.65	1.30	**2.59**	1.95	0.65	**2.59**	–
2016	Manhattan	1.30	1.30	1.95	1.95	1.95	**4.55**	2.59	0.65	0.65	–
	Euclidean	0.65	1.30	2.60	1.95	1.95	**3.90**	1.95	1.30	1.30	–
	Cosine	1.95	1.30	1.95	1.95	1.95	**3.25**	1.95	0.65	1.95	–
	Pearson	1.95	1.30	1.95	**2.59**	1.95	1.95	1.30	1.95	1.95	–
2017	Manhattan	3.89	**5.84**	2.59	0.65	3.25	**5.84**	3.25	1.95	3.25	–
	Euclidean	2.60	**5.84**	3.90	0.64	3.25	**5.84**	3.25	1.95	3.25	–
	Cosine	4.55	3.24	2.59	1.30	2.59	**5.84**	3.89	1.95	4.55	–
	Pearson	**5.19**	3.89	2.59	0.65	3.25	**5.19**	4.55	1.95	3.25	–
2018	Manhattan	3.25	**5.84**	3.25	0	3.89	5.19	3.25	1.95	2.59	–
	Euclidean	4.55	3.25	3.25	0	**5.19**	3.90	**5.19**	1.30	2.60	–
	Cosine	0.64	4.55	1.95	1.95	3.89	**5.19**	4.55	2.59	3.89	–
	Pearson	0.65	**5.19**	2.59	1.95	3.89	3.89	4.55	3.25	3.25	–
Average	Manhattan	1.14	2.11	1.22	0.33	1.79	**2.27**	1.79	0.81	1.06	–
	Euclidean	1.54	0.97	**1.87**	1.62	1.14	1.46	1.06	1.54	1.30	–
	Cosine	1.30	1.30	1.22	0.81	1.30	**2.19**	1.87	0.73	1.79	–
	Pearson	1.14	1.54	1.30	0.81	1.46	**1.87**	**1.87**	1.06	1.46	–

implications for knowledge management. The experimental results showed the proposed method outperformed several competing methods in terms of precision, recall, and accuracy. However, its drawback is that computing the density for each *n*-gram is computationally intensive; this means that the complexity grows with the size of *n*-gram collection.

For future work, the authors have decided to optimize the proposed method to be applied to real-time tasks. Also, they have decided to test the designed method on taken datasets from various knowledge area instead of examining a particular knowledge area ('social network').

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

If $\mathbb{R}_{cp}$ is fixed, the columns of $\boldsymbol{U}$ are independent, and the minimization of Eq. (3) can be done as follows:

$$J(\boldsymbol{V}.\boldsymbol{U}) = \sum_{k=1}^{n} g_k(\boldsymbol{U}), \tag{A.1}$$

where

$$\forall_{1\leqslant k\leqslant n} g_k(\boldsymbol{U}) = \sum_{i=1}^{c} \beta_{ik}(u_{ik})^m \mathscr{D}(x_k, v_i) \tag{A.2}$$

The Lagrangian of (A.2) with constraints from Eq. (5) is:

$$\forall_{1\leqslant k\leqslant n} g_k(\boldsymbol{U}\cdot\lambda_k) = \sum_{i=1}^{c} \beta_{ik}(u_{ik})^m \mathscr{D}(\mathrm{x_k}, \mathrm{v_i}) - \lambda_k\left[\sum_{i=1}^{c} \beta_{ik}\mathrm{u_{ik}} - \mathrm{f_k}\right] \tag{A.3}$$

where λ_k is the Lagrangian multiplier. The partial derivations of u_{ik} and f_k have been set to zero as follows:

$$\forall_{\substack{1\leqslant i\leqslant c\\ 1\leqslant k\leqslant n}} \frac{\partial G_k(U,\lambda_k)}{\partial u_{ik}} = m\beta_{ik}(u_{ik})^{m-1}\mathscr{D}(\mathrm{x_k}, \mathrm{v_i}) - \lambda_k\beta_{ik} = 0 \quad \rightarrow \mathrm{u_{ik}} = \left(\frac{\lambda_k}{m}\right)^{m-1} \mathscr{D}(\mathrm{x_k}, \mathrm{v_i})^{\frac{1}{1-m}} \tag{A.4}$$

$$\forall_{\substack{1\leqslant i\leqslant c\\ 1\leqslant k\leqslant n}} \frac{\partial G_k(U,\lambda_k)}{\partial \lambda_k} = \sum_{i=1}^{c} \beta_{ik}u_{ik} - f_k = 0 \tag{A.5}$$

By substituting u_{ik} into Eq. (A.5), f_k is calculated as follows:

$$f_k = \left(\frac{\lambda_k}{m}\right)^{m-1} \sum_{i=1}^{c} \beta_{ik}\mathscr{D}(x_k, v_i)^{\frac{1}{1-m}} \tag{A.6}$$

Table 7
Research priority in Turkey (%) by using different metrics.

Year	Metric	cl_1^{TUR}	cl_2^{TUR}	cl_3^{TUR}	cl_4^{TUR}	cl_5^{TUR}	cl_6^{TUR}	cl_7^{TUR}	cl_8^{TUR}	cl_9^{TUR}	cl_{10}^{TUR}
2011	Manhattan	**0.87**	0	0	**0.87**	–	0	0	**0.87**	0	0
	Euclidean	**0.86**	0	**0.86**	**0.86**	–	0	0	0	0	0
	Cosine	**0.88**	0	**0.88**	0	–	0	0	**0.88**	0	0
	Pearson	**0.97**	0	**0.97**	0	–	0	0	**0.97**	0	0
2012	Manhattan	**1.83**	0	0	0.87	–	0	0.87	0	0	0
	Euclidean	**1.81**	0	0	0.86	–	0	0	0	0.86	0
	Cosine	**1.73**	0	0	0	–	0	0.87	0	0.87	0
	Pearson	**1.73**	0	0	0	–	0	0.87	0	0.87	0
2013	Manhattan	**0.87**	0	**0.87**	**0.87**	–	**0.87**	0	0	0	0
	Euclidean	**1.71**	0	0	0.75	–	0.75	0	0	0	0
	Cosine	**1.73**	0	0	0	–	0	0	0	0.87	0.87
	Pearson	**1.73**	0	0	0	–	0	0	0	0.86	0.86
2014	Manhattan	0.87	**1.73**	**1.73**	**1.73**	–	0.87	0	0	0.87	0
	Euclidean	0.86	**1.81**	**1.81**	0.86	–	0.86	0	0.86	0.86	0
	Cosine	0.87	**1.73**	**1.73**	0.87	–	0.87	0	0	**1.73**	0
	Pearson	0.86	1.73	0.86	0.86	–	0.87	0	0	**2.60**	0
2015	Manhattan	2.70	1.73	0.87	0.87	–	**4.35**	0.87	0	1.73	2.60
	Euclidean	2.56	1.81	0	0.86	–	**3.43**	1.81	0.85	2.56	1.81
	Cosine	0.87	**2.70**	0.87	1.73	–	2.60	1.73	1.73	1.73	1.73
	Pearson	0.86	**2.60**	0.86	1.73	–	**2.60**	1.73	1.73	1.73	1.73
2016	Manhattan	2.60	0	0.87	1.73	–	3.47	**5.21**	4.34	1.73	0.87
	Euclidean	1.81	0	1.81	2.56	–	3.42	**5.14**	3.42	1.81	0.86
	Cosine	0.87	1.73	1.73	**4.44**	–	4.34	3.48	1.73	0.87	1.73
	Pearson	1.73	1.73	0.86	3.48	–	**4.34**	3.48	2.60	0.87	1.73
2017	Manhattan	0	1.73	**3.48**	0.87	–	2.60	**3.48**	1.73	1.73	0.87
	Euclidean	0	2.56	**4.38**	1.81	–	1.81	3.42	0.86	0.86	0.86
	Cosine	0	0.87	**3.48**	1.73	–	1.73	1.73	1.73	1.73	**3.48**
	Pearson	0	0.87	**3.48**	1.73	–	0.87	2.60	1.73	1.73	**3.48**
2018	Manhattan	**6.08**	1.73	1.73	1.73	–	3.48	4.34	0.87	4.34	5.21
	Euclidean	4.28	2.56	2.56	2.56	–	3.42	3.42	1.81	**5.24**	3.42
	Cosine	4.34	4.34	0.87	2.60	–	3.48	4.34	0	**6.08**	3.48
	Pearson	4.34	4.34	0.87	2.60	–	0.86	3.48	1.73	**6.08**	5.21
Average	Manhattan	1.95	0.87	1.19	1.19	–	**1.96**	1.85	0.98	1.30	1.19
	Euclidean	**1.75**	1.09	1.43	1.40	–	1.73	1.72	0.98	1.52	0.87
	Cosine	1.41	1.42	1.20	1.42	–	1.63	1.52	0.76	**1.74**	1.41
	Pearson	1.53	1.41	0.99	1.30	–	1.19	1.52	1.10	**1.84**	1.63

Table 8
Comparison of the performance of different similarity/dissimilarity metrics (%).

Similarity/Dissimilarity metric	Evaluation metric	the USA	Iran	Saudi Arabia	Turkey	Average
Manhattan	Precision	**88.23**	**84.60**	**82.62**	**73.02**	**82.12(1)**
	Recall	**85.43**	**79.43**	**77.06**	**75.77**	**79.42(1)**
	Accuracy	**87.09**	**82.92**	**80.24**	**75.37**	**81.41(1)**
Euclidean	Precision	85.07	74.89	76.06	72.20	77.06(4)
	Recall	81.50	75.15	80.09	75.99	78.18(2)
	Accuracy	84.52	72.24	79.54	74.59	77.72(2)
Cosine	Precision	85.88	81.97	76.55	73.69	79.52(3)
	Recall	81.58	79.05	71.52	70.08	75.56(3)
	Accuracy	83.62	77.35	73.50	73.96	77.11(3)
Pearson	Precision	85.88	83.15	75.16	77.20	80.35(2)
	Recall	81.71	71.11	70.22	70.78	73.46(4)
	Accuracy	83.07	74.07	72.06	77.87	76.77(4)

Combining u_{ik} and f_k equations yields:

$$\forall_{\substack{1\leqslant i\leqslant c\\ 1\leqslant k\leqslant n}} u_{ik} = f_k \mathscr{D}(x_k, v_i)^{\frac{1}{1-m}} \Big/ \sum_{i=1}^{c} \beta_{ik} \mathscr{D}(x_k, v_i)^{\frac{1}{1-m}} \tag{A.7}$$

By defining Eq. (A.8) as follows:

$$\forall_{\substack{1\leqslant i\leqslant c\\ 1\leqslant k\leqslant n}} \begin{cases} \Gamma_k = \{i | 1 \leqslant i \leqslant c;\ \mathscr{D}(x_k, v_i) = 0 \\ \tilde{\Gamma}_k = \{1, 2, \cdots, c\} \backslash \Gamma_k, \end{cases} \tag{A.8}$$

If $\Gamma_k \neq \varnothing$, the minimization of the Eq. (3) is reachable by selecting $u_{ik} = 0$ for $i \notin \Gamma_k$ and $\sum_{i\in\Gamma_k} \beta_{ik} u_{ik}$ for $i \in \Gamma_k$, because elements of the partition matrix are zero for non-zero dissimilarities, and non-zero for zero dissimilarities.

Appendix B

By substituting $\mathscr{D}(x_{kl}, v_{il})$ from Eq. (4) into Eq. (3), $J_{fcom}(\boldsymbol{V}, \boldsymbol{U})$ is obtained as follows:

$$J_{fcom}(\boldsymbol{V}, \boldsymbol{U}) = \sum_{i=1}^{c} \sum_{k=1}^{n} \beta_{ik} (u_{ik})^m \sum_{l=1}^{d} \mathscr{D}(x_{kl}, v_{il}) = \sum_{i=1}^{c} \sum_{l=1}^{d} g_{il}(v_{il}), \tag{B.1}$$

Table 9
Comparison of the performance of fuzzy-based clustering methods (%).

Clustering method	Evaluation metric	the USA	Iran	Saudi Arabia	Turkey	Average
FCM	Precision	71.82	72.75	69.10	73.83	71.88(4)
	Recall	74.59	66.01	65.79	64.84	67.81(4)
	Accuracy	74.57	66.00	68.71	69.89	69.79(4)
PFCM	Precision	80.44	80.76	77.68	74.17	78.26(3)
	Recall	74.30	72.24	69.72	68.40	71.17(3)
	Accuracy	79.62	74.60	72.71	71.11	74.51(3)
PC	Precision	80.43	81.35	83.46	75.49	80.18(2)
	Recall	80.31	72.29	78.10	77.91	77.15(2)
	Accuracy	82.95	74.06	79.32	76.64	78.24(2)
FCOM	Precision	**88.23**	**84.60**	**82.62**	**73.02**	**82.12(1)**
	Recall	**85.43**	**79.43**	**77.06**	**75.77**	**79.42(1)**
	Accuracy	**87.09**	**82.92**	**80.24**	**75.37**	**81.41(1)**

Table 10
Comparison of the performance of different methods (%).

Method	Evaluation metric	the USA	Iran	Saudi Arabia	Turkey	Average
k-means	Precision	75.60	72.33	71.04	73.68	73.16(3)
	Recall	74.17	68.56	69.26	69.70	70.42(3)
	Accuracy	73.78	69.37	68.04	69.67	70.22(3)
LDA	Precision	79.49	80.81	78.31	75.43	78.51(2)
	Recall	79.61	71.28	78.10	77.59	76.65(2)
	Accuracy	82.64	73.22	79.20	75.69	77.69(2)
The proposed method	Precision	**88.23**	**84.60**	**82.62**	**73.02**	**82.12(1)**
	Recall	**85.43**	**79.43**	**77.06**	**75.77**	**79.42(1)**
	Accuracy	**87.09**	**82.92**	**80.24**	**75.37**	**81.41(1)**

Table 11
The closeness evaluation of the research areas in developing countries to the USA (%).

Year	Country	cl_1^{USA}	cl_2^{USA}	cl_3^{USA}	cl_4^{USA}	cl_5^{USA}	cl_6^{USA}	cl_7^{USA}	cl_8^{USA}	cl_9^{USA}	cl_{10}^{USA}	Average
2011	Iran	0	0	0	0	0	0	0	0	0	0.46	0.05
	Saudi Arabia	0	0	0	0	0	0	0	0	0	0	0
	Turkey	0	0.97	0	0	0	0.97	0.97	0	0	0	**0.29**
2012	Iran	0.46	0	0	0	0.92	0	0	0.46	0.46	0	0.23
	Saudi Arabia	0	0	0	0.76	0	0	0	0	0	0	0.08
	Turkey	0	0	0	0	0	0	0.97	0	0.97	1.93	**0.39**
2013	Iran	0.92	0.46	0	0	0.92	0.92	1.38	2.31	0.46	0	**0.74**
	Saudi Arabia	2.28	0	0.76	0.76	1.52	0	0.76	0.76	0	0	0.68
	Turkey	0	0	0	0.97	1.93	0	0	0.97	0	0	0.39
2014	Iran	0.92	0.92	0.46	0.92	0	1.38	0.46	0.92	0	0.92	0.69
	Saudi Arabia	1.52	0.76	0.76	0	0.76	0	0.76	0	0.76	0.76	0.61
	Turkey	0	0.97	1.93	0	0.97	0.97	0.97	0.97	1.93	0	**0.87**
2015	Iran	0.46	0.92	0.92	1.38	0.92	1.85	3.71	2.31	3.71	1.38	**1.76**
	Saudi Arabia	3.8	0.76	0.76	0.76	0.76	1.52	0.76	0	3.8	0.76	1.37
	Turkey	2.89	4.83	1.93	0.97	0.97	1.93	1.93	1.93	0	0	1.74
2016	Iran	2.31	1.38	0.92	0.92	2.31	0.46	2.78	0.92	0.46	1.85	1.43
	Saudi Arabia	3.03	1.52	0	3.80	3.03	0.76	1.52	2.28	2.28	1.52	1.97
	Turkey	4.83	2.89	0.97	2.89	0.97	2.89	0.97	2.89	0.97	2.89	**2.32**
2017	Iran	2.78	0.92	2.78	4.63	3.24	2.31	2.78	1.38	2.78	2.78	2.64
	Saudi Arabia	5.32	3.03	0.76	4.55	3.03	1.52	4.55	3.80	5.32	3.80	**3.57**
	Turkey	0	0.97	0	2.89	0.97	2.89	0	4.83	0.97	4.83	1.84
2018	Iran	4.84	2.78	2.31	2.78	4.17	1.85	4.63	2.78	2.78	5.09	3.40
	Saudi Arabia	4.55	2.28	1.52	5.32	2.28	3.03	2.28	3.03	6.07	3.80	**3.42**
	Turkey	0.97	5.80	0	3.87	4.83	3.87	5.80	1.93	2.89	2.89	3.29

where

$$g_{il}(v_{il}) = \sum_{k=1}^{n} \beta_{ik}(u_{ik})^m \sum_{l=1}^{d} \mathscr{D}(c_{kl}, v_{il}) = \sum_{k=1}^{n} \beta_{ik}(u_{ik})^m \mathscr{L}(x_{kl} - v_{il}) \tag{B.2}$$

By decomposing the minimization problem in Eq. (3) into $c \times d$ minimization sub-problems, Eq. (B.2) can be rewritten as follows:

$$g_{il}(v_{il}) = \sum_{k=1}^{n} \beta_{ik}(u_{ik})^m h_{ikl}(x_{kl} - v_{il})^2 \tag{B.3}$$

where

$$h_{ikl} = \begin{cases} 0, & x_{kl} - v_{il} = 0, \\ \frac{\mathscr{L}(x_{kl} - v_{il})}{(x_{kl} - v_{il})}, & x_{kl} - v_{il} \neq 0. \end{cases} \tag{B.4}$$

References

Ai, W., Li, K., Li, K., 2018. An effective hot topic detection method for microblog on spark. Appl. Soft Comput. 70, 1010–1023.

Aiello, L.M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., Jaimes, A., 2013. Sensing trending topics in Twitter. IEEE Trans. Multimedia 15, 1268–1282.

Al-Anzi, F.S., AbuZeina, D., 2017. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. J. King Saud Univ. Comput. Inf. Sci. 29, 189–195.

Belazzoug, M., Touahria, M., Nouioua, F., Brahimi, M., 2019. ISCA: an improved sine cosine algorithm to select features for text categorization. J. King Saud Univ. Comput. Inf. Sci.. In Press.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Brants, T., Chen, F., Tsochantaridis, I., 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management. ACM, pp. 211–218.

Cao, J., Wu, Z., Wu, J., 2014. Scaling up cosine interesting pattern discovery: a depth-first method. Inf. Sci. 266, 31–46.

Chen, Q., Guo, X., Bai, H., 2017. Semantic-based topic detection using Markov decision processes. Neurocomputing 242, 40–50.

Choi, H.J., Park, C.H., 2019. Emerging topic detection in twitter stream based on high utility pattern mining. Expert Syst. Appl. 115, 27–36.

Chun-hong, W., Li-Li, N., Yao-Peng, R., 2011. Research on the text clustering algorithm based on latent semantic analysis and optimization. In: Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on. IEEE, pp. 470–473.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407.

Dutta, S., Ghatak, S., Das, A.K., Gupta, M., Dasgupta, S., 2019. Feature selection-based clustering on micro-blogging data. In: Computational Intelligence in Data Mining. Springer, pp. 885–895.

Elghannam, F., 2019. Text representation and classification based on bi-gram alphabet. J. King Saud Univ. Comput. Inf. Sci.. In Press.

Francis, L.M., Sreenath, N., 2019. Robust scene text recognition: using manifold regularized Twin-Support Vector Machine. J. King Saud Univ. Comput. Inf. Sci.. In Press..

Grant, S., Skillicorn, D., Cordy, J.R., 2008. Topic detection using independent component analysis. In: Proceedings of the 2008 Workshop on Link Analysis, Counterterrorism and Security (LACTS'08), pp. 23–28..

Guille, A., Favre, C., 2015. Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach. Social Network Anal. Mining 5, 18.

Hasan, M., Orgun, M.A., Schwitter, R., 2016. TwitterNews: real time event detection from the Twitter data stream. PeerJ PrePrints 4, e2297v2291.

Hasan, M., Orgun, M.A., Schwitter, R., 2019. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. Inf. Process. Manage. 56, 1146–1165.

Hoffman, M., Bach, F.R., Blei, D.M., 2010. Online learning for latent dirichlet allocation, advances in neural information processing systems, pp. 856-864..

Hofmann, T., 1999. Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp. 289–296.

Ihou, K.E., Bouguila, N., 2019. Variational-based latent generalized Dirichlet allocation model in the collapsed space and applications. Neurocomputing 332, 372–395.

Jayawardhana, U.K., Gorsevski, P.V., 2019. An ontology-based framework for extracting spatio-temporal influenza data using Twitter. Int. J. Digital Earth 12, 2–24.

Juckett, D.A., Kasten, E.P., Davis, F.N., Gostine, M., 2019. Concept detection using text exemplars aligned with a specialized ontology. Data Knowl. Eng. 119, 22–35.

Krishnapuram, R., Keller, J.M., 1993. A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. 1, 98–110.

Lau, R.Y., Song, D., Li, Y., Cheung, T.C., Hao, J.X., 2008. Toward a fuzzy domain ontology extraction method for adaptive e-learning. IEEE Trans. Knowl. Data Eng. 21, 800–813.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. Int. Conf. Mach. Learn., 1188–1196.

Leski, J.M., 2016. Fuzzy c-ordered-means clustering. Fuzzy Sets Syst. 286, 114–133.

Li, W., Feng, Y., Li, D., Yu, Z., 2016. Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. Autom. Control Comput. Sci. 50, 271–277.

Liu, M., Qu, J., 2012. Mining high utility item sets without candidate generation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, pp. 55–64.

Manning, C.D., Manning, C.D., Schütze, H., 1999. Foundations of statistical natural language processing. MIT press.

Milioris, D., 2018. Topic Detection and Classification in Social Networks..

Niraula, N.B., Gautam, D., Banjade, R., Maharjan, N., Rus, V., 2015. Combining word representations for measuring word relatedness and similarity, The Twenty-Eighth International Flairs Conference..

Ozdikis, O., Karagoz, P., Oğuztüzün, H., 2017. Incremental clustering with vector expansion for online event detection in microblogs. Social Network Anal. Mining 7, 56.

Padmaja, C., Narayana, S.L., Divakar, C., 2018. Probabilistic topic modeling and its variants: a survey. Int. J. Adv. Res. Comput. Sci. 9.

Panagiotou, N., Katakis, I., Gunopulos, D., 2016. Detecting events in online social networks: Definitions, trends and challenges, Solving Large Scale Learning Tasks. In: Challenges and Algorithms. Springer, pp. 42–84.

Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 33, 1065–1076.

Petkos, G., Papadopoulos, S., Aiello, L., Skraba, R., Kompatsiaris, Y., 2014. A soft frequent pattern mining approach for textual topic detection. In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14). ACM, p. 25.

Poria, S., Cambria, E., Gelbukh, A., 2016. Aspect extraction for opinion mining with a deep convolutional neural network. Knowl. Based Syst. 108, 42–49.

Rajaraman, K., Tan, A.H., 2001. Topic detection, tracking, and trend analysis using self-organizing neural networks. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 102–107.

Rashid, J., Shah, S.M.A., Irtaza, A., 2019. Fuzzy topic modeling approach for text mining over short text. Inf. Process. Manage. 56, 102060.

Recchia, G., Jones, M.N., 2009. More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. Behav. Res. Meth. 41, 647–656.

Reihanian, A., Minaei-Bidgoli, B., Alizadeh, H., 2016. Topic-oriented community detection of rating-based social networks. J. King Saud Univ. Comput. Inf. Sci. 28, 303–310.

Salton, G., McGill, M.J., 1986. Introduction to modern information retrieval..

Sheeba, J., Vivekanandan, K., 2014. A fuzzy logic based on sentiment classification. Int. J. Data Mining Knowl. Manage. Process 4, 27.

Shenoy, M.K., Shet, K., Acharya, U.D., 2012. Semantic plagiarism detection system using ontology mapping. Advanced Computing 3, 59.

Silambarasan, I., Sriram, S., 2017. Hamacher Sum and Hamacher Product of fuzzy matrices. Intern. J. Fuzzy Mathe. Archive 13, 191–198.

Wang, Y., Agichtein, E., Benzi, M., 2012. TM-LDA: efficient online modeling of latent topic transitions in social media. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 123–131.

Wartena, C., Brussee, R., 2008. Topic detection by clustering keywords, Database and Expert Systems Application, 2008. DEXA'08. In: 19th International Workshop on. IEEE, pp. 54–58.

Winarko, E., Pulungan, R., 2019. Trending topics detection of Indonesian tweets using BN-grams and Doc-p. J. King Saud Univ. Comput. Inf. Sci. 31, 266–274.

Winkler, R., Klawonn, F., Kruse, R., 2011. Fuzzy clustering with polynomial fuzzifier function in connection with M-estimators. Appl. Comput. Math 10, 2011.

World Population Review. http://worldpopulationreview.com/. (accessed 15 September 2019)..

Web of Science. https://clarivate.com/products/web-of-science/. (accessed 16 May 2019)..

Zhang, L., Wu, Z., Bu, Z., Jiang, Y., Cao, J., 2018a. A pattern-based topic detection and analysis system on Chinese tweets. J. Comput. Sci. 28, 369–381.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., Zhang, G., 2018b. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. J. Inf. 12, 1099–1117.

Zhu, Z., Liang, J., Li, D., Yu, H., Liu, G., 2019. Hot topic detection based on a refined TF-IDF algorithm. IEEE Access 7, 26996–27007.