IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 3, 2019, accepted October 22, 2019, date of publication October 29, 2019,
date of current version November 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950045

Text Mining Techniques to Capture Facts for Cloud Computing Adoption and Big Data Processing

MUHAMMAD INAAM UL HAQ[1,4], QIANMU LI[1,2,3], AND SHOAIB HASSAN[1,4]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]Intelligent Manufacturing Department, Wuyi University, Jiangmen 529020, China
[3]School of Information Engineering, Nanjing Xiaozhuang University, Nanjing 211171, China
[4]COMSATS University Islamabad, Sahiwal Campus, Sahiwal 57000, Pakistan

Corresponding author: Qianmu Li (qianmu@njust.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities (30918012204), Jiangsu province key research and development program (BE2017739), Jiangsu province key research and development program (BE2017100), 2018 Jiangsu Province Major Technical Research Project Information Security Simulation System.

ABSTRACT Digital libraries, journals and conference proceedings repositories are a great source of information. These sources are very useful for the purpose of research and development. This paper presents an overview of text mining and its application towards information extraction from literature. In this study, we used word cloud, term frequency analysis, similarity analysis, cluster analysis, and topic modeling to extract information from multi-domain research articles. Cloud computing and big data are new emerging trends. So it is important to extract useful patterns and knowledge from published articles in these domains and discover the relationship between them. Therefore, a total of two hundred research articles published from 2010 to 2018 in these domains, were selected. The source of these articles is high impact factor journals from reputed publishers namely IEEE, Springer, Wiley, Elsevier, and ACM. It is a cross-domain analysis in cloud computing and big data domains to find the latest trends, related topics, tools, terms, and author affiliation from extracted data. This study identifies the ten major areas of big data using cloud computing, fourteen factors towards cloud adoption, and hurdles in adoption. Moreover finding shows that IEEE has more sources for subject cloud computing application towards big data, then comes Springer, Wiley, and Elsevier. Furthermore, it has been observed in the analysis that the number of articles in these domains increased from 2013 onward.

INDEX TERMS Text mining, topic modeling, cloud computing, big data, information extraction, literature analysis, similarity.

I. INTRODUCTION

In the current era, most of the information is stored in the form of e-documents. This practice is adopted by several business organizations, institutes, media, and others. These documents are structured, un-structured and semi-structured formats [1]. The scientific literature is stored mostly in the form of e-documents in digital libraries, conference proceedings repositories, and journal databases. This is a great source of information and it provides the basis for future developments. We can use data mining techniques to unhide the important concepts, knowledge from scientific literature [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Afzal.

Text mining is a process in which we extract a variety of patterns and discovered useful facts from textual sources [3]. Data mining covers all types of techniques to mining data e.g. the techniques to process audio, video data. It is used to process big records stored in different databases to reveal important concepts and relationships among data. [4] There are several techniques to recover textual information, indexing is one of them. It is used to handle the documents which are in an un-structured format. In traditional research studies the user search for already defined concepts and relevant terms. The problem is, most of the time the retrieved results are not according to the requirements specified. To get rid of this situation we can used text mining methods to find the required information. The text mining process goes through

several steps. In the first step, the document from various sources is collected than in the next step the format of the document is verified and then it goes through the analysis stage. The analysis stage includes the semantic analysis and other techniques to get data processed according to the requirements. The result of this phase can be stored in the database management system for further processing. Text mining aims to get information that is not available before from various textual sources. Data mining can be used to handle structured data while text mining can be used to handle semi-structured and un-structured data. [5]. There are different types of text mining techniques they include clustering, classification, topic modeling, and summarization. The main purpose of these techniques is to extract knowledge from text [6]. The text mining techniques are applied in several areas, they include academia, web applications, internet, industry and other disciplines [7]. The text mining and data mining supposed to be similar methods but in fact, these are different because data mining needs structured data but in the case of text mining we are dealt with un-structured data that need pre-processing steps in addition and it involves the interaction with NLP. A lot of research is going on in NLP. It involves the interaction and relationship between the big amounts of the textual format of data. [8], [9]. A document in regular language consists of information that cannot be used for data mining. For information extraction to be successful, appropriate articles must be chosen with the condition that these documents should have some association [10], [11].

The induction of the concept of big data demands more requirements in terms of data processing in real-time, hence the traditional IT technology cannot full fill these requirements. There comes the concept of cloud computing technology which turns the task of huge data processing and data mining into reality. The cloud computing platform provides better services to the user as compared to traditional IT services. But it brings some security problems in addition [12]. In the past few years the enterprise's interest in the adoption of cloud computing is manifold. Cloud computing gives the opportunity for organizations to meet their computing requirements effectively [13]. In addition to providing shared services, it is an innovative solution that brings value to enterprises [14]. It helps them to shift their focus towards the real business which brings a positive impact on their productivity [15].

There are a variety of research domains having many research areas, techniques, and models. The latest trends topics and hot issues are discussed in research papers. The results of experiments in one domain may affect the associated domains having common topics. We have studied the interaction and cross-domain trends in this research. Cloud computing is an emerging computing paradigm having many advantages. This research focuses on IE of cloud computing application towards big data and cloud adoption trends using text mining techniques on the published research papers in these domains. This paper is divided into the following sections. In the I section we give the detail introduction of the text mining and the domain in which these techniques are applied. In the II section, we discuss the related work. In the third section, we explain in detail the research methodology, in the IV section we discuss the experiments and results, and in V section we have a conclusion.

II. RELATED WORK

The research work in the domain of information extraction being conducted by the utilization of many techniques. The main aim of the conduction of this research is to find the techniques which proved to be useful for extracting structured data from a text corpus. It starts with the introduction of the research topic, comparative analysis of previous research models, then major techniques of text mining and topic modeling are applied. The purpose is to determine the research topic and establish an association between them. The topics produced by information extraction are shown using visual tools and libraries. These methods are useful to reveal the hidden relationship between topics. In this way, the user can see the cross-domain topics and appreciate the research trends [16]. The text mining techniques for literature analysis were also applied by other researchers in the field of Bio-Medical domains. In this research, the literature of SCI journals from 2004 to 2013 was extracted and processed. The idea is to look into it from the institutional vision to extract year base changes, new research domains, scientific journals, important terms, and keywords. It has been observed that there is a rapid increase in published articles in this domain. The research outcome shows that most of the published literature focused on Name Entity identification, relationship extraction, text categorization, co-occurrence and cluster analysis [17].

A statistical method to extract specific information from a research paper is developed. It extracts algorithms, techniques, keywords from the research article. It also extracts "CueWords" to find the limitation of the article. It uses the Naive Bayes classifier to identify the area of the research papers [18]. The technique which uses conditional random fields to extract different common fields from the header and citation of research papers is proposed. This model shows F1 as 36 % and an error rate of 78% in comparison to the previous results of SVM [19]. A graph mining-based technique used to explore document content is proposed. Infect it is user define query which used to generates the graph. It is a good approach for extracting the same patterns from documents. There is no need for understanding of graphical models in a large collection of documents. This model shows 86.64% as precision and 90.8% as recall. But it has limitation e.g, it cannot provide correct patterns for a large collection of text fields [20]. A system is designed to understand the experimental results of a research paper. It takes a query as input and returns the best method against the parameters given in the query. This query, in fact, is the answer to the computational question the system is trained on 5 papers and tested on 61 papers [21]. Song and Kim applies text mining techniques to a larger set of fuller length papers to unhide the knowledge.

He did not depend upon citation data available in WOS and PUBMED. He performs a bibliometric examination of fuller text papers. In this approach, he prepared a custom citation database. The result shows that the papers that were published in the bioinformatics domain were not cited [22]. Top Cat (Topic Categories) technique is proposed to recognize the repeated papers into the collection set. The primary purpose of this method is to recognize the name entities in independent articles and to represent them as a collection set of papers. Top Cat recognizes the topics which are logically correct [23]. Another technique to identify named entities and extract useful information from the criminal record is proposed. It processes the 304 court judgments. In order to train this system, the judgments are manually tagged by nine entities. It used CRF and achieves significant precision 0.97, recall 0.87 and f-measure 0.89 [24].

This study [26] analysis the 741 research articles of ''Food Policy'' the best agriculture policy journal in the field of agriculture economics. The objective is to find the author and co-author relationship and citation network on the basis of data extracted from these collected set of articles. The result shows that most of the paper written by a couple of authors. It means a vital role played by one typical author or group of authors. It also concluded that the group of authors site themselves often so that the research profile of the group and members can be enhanced. The study presented by [27] analyzed 300 research articles collected in the field of mobile learning using text mining techniques. This study used the term frequency analysis, similarity analysis, clustering techniques, and association algorithm to extract knowledge from articles. The result shows that 285 articles from selected papers focusing on the application of mobile learning in the context of higher education in the medical domain. The trend of publication in this area was on peak in 2015 to 2016. The topic modeling and text mining-based study are presented by [28] to find the big data trends in the marketing. In the study 1560 articles published from 2010 to 2015, belongs to the Business, management, accounting, computer, economics, econometrics and finance, domain sciences indexed in science direct were selected for analysis. The total eighteen topics containing 54 terms (each topic has three terms) were found. The frequency count of these terms in different selected publications also shown. The result shows that research is duplex between technology and selected research domains. It does not clearly find out novel approaches that are useful for marketing. Moreover, the author and co-author network and affiliation also shown from the selected set of publications. This study concluded that the research in the application of big data towards marketing needs further development.

The topic modeling and cluster analysis based study is proposed by [29] to find the research trends in the field of design research. The data set contain published papers from 2004 through 2015 which were processed. This study finds the important branches in the field of research design and 30 topics, showing the research trends in this field. The literature survey shows that still further research is required for information extraction from research articles. The approach of text mining and topic modeling can be applied in multi-dimension and in many kinds of domains. Cloud computing and big data are new emerging fields [12]–[15]. Through literature analysis, it has been found that IE and data mining techniques are never applied to literature published in these domains. So, we have selected 200 research articles from top rank journals in these domains and applied IE and data mining techniques to find out research trends, tools, technologies, concepts, related terms, key issues in cloud computing adoption, and its application trends towards big data processing.

III. MATERIALS AND METHODS

A. LITERATURE SELECTION

We collected two hundred articles from IEEE, Springer, Elsevier, Wiley, and ACM. The search query to get the first category articles is the ''application of cloud computing for big data'' and for a second category, the search query is ''cloud computing adoption trends''. In this way, we collected one hundred and nineteen papers in the first category and eighty-one papers in the second category, almost two hundred articles were collected in a data set. We placed these articles category wise in two folders.

B. PROCEDURE FOR TEXT MINING AND TOPIC MODELING

We designed a custom procedure for text mining of research articles in the light of steps followed by [27] and [28], as shown in Fig.2. The following are important phases that are essential for text mining, they include text pre-processing, text mining and text post-processing. The text pre-processing phase further divided into data selection, data cleaning, feature extraction, pattern matching and converting un-structured data into structured form. So that the data format will become compatible with data mining tools. In 2nd phase we use text mining techniques like finding word frequency, data clustering, measuring data similarity, visualization, and topic modeling, etc., to retrieve useful information. In the last step, some changes are may be incorporated in data through text mining function for analysis and visualization of knowledge. The text of collected articles first converted into tokens, this process is called tokenization [30] then all text is converted into lower case, then in the next step, stop words, and punctuations are removed from the text. In the last step, all tokens of length less than three characters and greater than twenty-five characters are removed from a text corpus.

The following techniques are applied on textual data for analysis, the brief description of these techniques is presented as follows:

C. WORD CLOUD

As per the study [25], the word cloud is the optic representation of word frequency in the text. This is a widely used

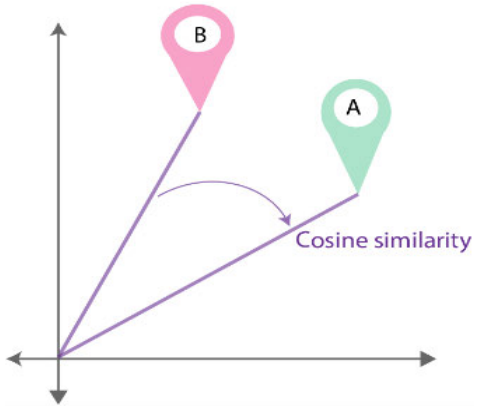

FIGURE 1. Angles between articles.

graphical representation of textual data to perform analysis [31]. Word cloud analysis is the starter stage for the detail textual analysis [32], [33]. This method can be used to find the relevancy in the text on the basic information provided. This technique is used in this study, its application detail is given in section (IV A&B).

D. SIMILARITY ANALYSIS

The measurement of similarity is an important concept, to understand and present the articles which are related to each other. A lot of methods and techniques are available to measure similarity. These include Euclidean distance, cosine similarity, and relative entropy. In mathematics, suppose that A= {a1,..., an} be a set of articles and T={t1,...tn} be the set of repeated terms in set A. In a simple way we can consider terms as words. The vector representation of an article having n dimensions is given by

$$ta = (tf(a, t1), ..tf(a, tn))$$

where tf(a,t) is the frequency of term $t \in T$ in articles $a \in A$.

If we consider articles as vectors the similarity is the measurement of cosine function of the angle between the corresponding vectors as shown in Fig.1. This technique is applied in the current study, application detail is given in section (IV, E) [34].

E. CLUSTER ANALYSIS

The theory behind the text cluster analysis suggests that the related documents are more similar as compared to non-related documents [34]. Clustering technique can be used for the analysis of huge data, through the state of the art literature it is confirmed that the clustering is an efficient tool for analyzing the subject of the text [34]. Clustering is very useful for managing named entities in similar groups on the basis of their co-existence [35]. The set of these named entities, having a relationship with any topic in a collection set, is denoted by a cluster. At the beginning of the clustering process, we do not know, cluster count, features and association among the similar groups. In this way, the different categories of documents can be identified [36]. In K-means algorithm data has been divided into the number of clusters denoted by k. The terms in the centroid of the cluster represent the subject of the cluster. An iterative process based on two steps is applied, in which all points assigned to the closet centroid, then the centroid is evaluated for the latest groups. This process continues until the constant value of the centroid is achieved [36]. In this study the K-means algorithm is used as shown in Fig2, application detail is given in section (IV, F).

TABLE 1. Word cloud terms distribution across IEEE.

Database	Terms	Frequency
IEEE	BigData	3198
	Cloud	1892
	Computing	1067
	Service	681
	time	435
	Application	414
	Security	410

F. LDA

As per the study [38], LDA is a useful technique to investigate the link between data and text documents. There are a variety of models used for topic modeling, but LDA is famous in this domain. We used this method for topic modeling as mention in Fig2. The application detail is given in section (IV, G).

IV. RESULT & DISCUSSION

The cloud computing application towards big data and its adoption trends are never investigated using text mining and topic modeling techniques. The induction of this kind of approach for information extraction, in this area, will be, value able for the research community.

A. CATEGORY 1 ARTICLES (WORD CLOUD & TERM FREQUENCY ANALYSIS)

The word cloud is an efficient visualization technique to represent the frequency of the words in the text corpus. This diagram can be plotted to view the basic idea about the concepts, related terms, and most frequent words [25]. The Fig. 7 shows that the vividness of "big data" word is clear as the most frequent word in the collection of category 1. The other categorical words in the list are cloud, computing, model, application, and service respectively. The cogent occurrence of the words "big data" and "cloud computing" in all collected articles under this category leads us to the conclusion that the cloud computing model is discussed in the context of its application for big data processing. The word cloud terms distribution is given database wise in tables (Table 1, 2, 3, 4, 5).

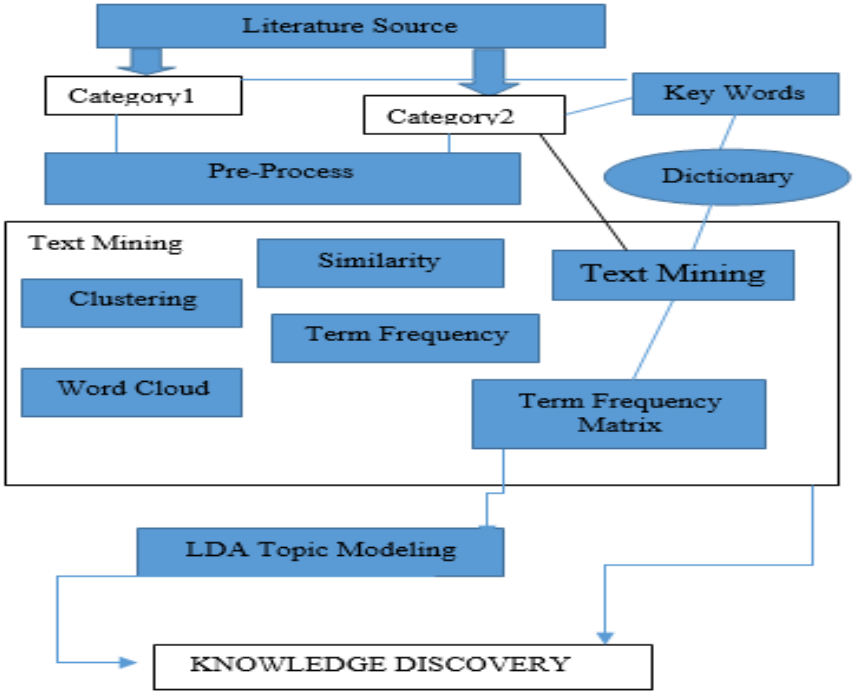

FIGURE 2. Text mining & topic modeling framework.

TABLE 2. Word cloud terms distribution across Springer.

Database	Terms	Frequency
Springer	BigData	2817
	Cloud	1606
	Computing	991
	Service	697
	system	648
	Application	351
	Analytics	307
	Information	298

TABLE 3. Word cloud terms distribution across Wiley.

Database	Terms	Frequency
Wiley	BigData	2720
	Cloud	1107
	System	1004
	Computing	979
	Service	400
	Application	329
	MapReduce	447
	Model	390

TABLE 4. Word cloud terms distribution across Elsevier.

Database	Terms	Frequency
Elsevier	BigData	2711
	Cloud	1083
	Computing	615
	Security	318
	Application	306
	System	350
	Information	271
	Model	201

TABLE 5. Word cloud terms distribution across ACM.

Database	Terms	Frequency
ACM	BigData	691
	Cloud	629
	computing	498
	Platform	261
	Management	204
	Storage	112
	Network	109
	Cluster	97

The word frequency technique is applied to the text of research papers, which are compiled under category1. The pictorial representation of (Fig. 6.) articulates that the following are the most frequent associated words surrounded by the collection of papers, which are ‘‘Big Data’’, ’’Cloud Computing’’,’’ Service’’,’’ Application’’,’’ Time’’,’’ Information’’,’’ Model’’,’’ Analysis’’,’’ Security’’ and ’’Technology’’ respectively. This result shows that the associated words focus on the application of the cloud computing model for big data analysis. The words mention here matched as depicted from the word cloud. The graph (Fig.6) shows that IEEE has more

TABLE 6. Word cloud terms distribution across IEEE.

Database	Terms	Frequency
IEEE	Cloud	2495
	Computing	1363
	Adoption	1068
	Security	469
	Information	424
	Technology	386

TABLE 7. Word cloud terms distribution across Springer.

Database	Terms	Frequency
Springer	Cloud	2435
	Computing	1782
	Adoption	828
	Service	743
	Technology	338

TABLE 8. Word cloud terms distribution across Wiley.

Database	Terms	Frequency
Wiley	Cloud	1796
	Data	1210
	Computing	954
	adoption	479
	Service	463
	Mobile	373

TABLE 9. Word cloud terms distribution across Elsevier.

Database	Terms	Frequency
Elsevier	Cloud	2315
	Computing	1460
	Adoption	848
	Information	630
	Service	509
	Data	482

TABLE 10. Word cloud terms distribution across ACM.

Database	Terms	Frequency
ACM	Cloud	595
	Computing	305
	Technology	211
	Adoption	158
	Data	77
	Security	71

TABLE 11. Author affiliation.

Country	No. of Authors
China	117
India	76
UK	54
USA	42
Australia	41
Malaysia	28
Saudi Arabia	21
Spain	20
South Africa	19
France	18

sources that have these words then comes Springer, Wiley, Elsevier, and ACM databases.

B. CATEGORY2 ARTICLES (WORD CLOUD & TERM FREQUENCY ANALYSIS)

As we applied the word cloud visualization technique to represent the frequency of the words in text corpus for the document of category 1. The same is applied for articles of category 2. Figure 10 shows that the vividness of the ''Cloud Computing'' word is clear as the most frequent word in the collection of category 2. The other categorical words in the list are adoption, data, service, application, and organization respectively. The cogent occurrence of the words ''Cloud Computing'' and ''cloud adoption'' in all collected articles under this category shows that the organization's focus is moving towards the adoption of cloud computing technology services. The word cloud terms distribution is given database wise in tables (Table 6, 7, 8, 9, and 10).

The word frequency technique is applied to the text of research papers, which are compiled under category2.

The pictorial representation of (Fig. 9) articulates that the following: are the most frequent associated words surrounded by the collection of papers, which are ''Cloud Computing'','' Cloud Adoption'','' Service'','' information'','' Technology'','' factors'','' Data'','' Services'','' Management'','' Organization'' and ''Security'' respectively. This result shows that the associated words focus on the trend of cloud computing adoption. The words mention here matched as depicted from the word cloud. The graph (Fig.9) shows that all databases have mixed trends of sources containing these words.

C. PUBLISHED PAPERS

Fig. 3 represents the distribution of the number of papers with respect to their publication year. It shows that the articles were published between 2010 through 2018. The prominent increase of publications is observed in this domain from 2013- onward.

D. AUTHOR AFFILIATION

The text in the papers of both categories consists of information about authors and their affiliations. As per the study [26], this information is extracted and processed. The resultant data contains 648 authors from 62 countries. The top ten records from this data are shown in Table. 11. The author's affiliation data analysis shows that most of the publications originated from Asia, Europe, North America, Australia and a few from Africa.

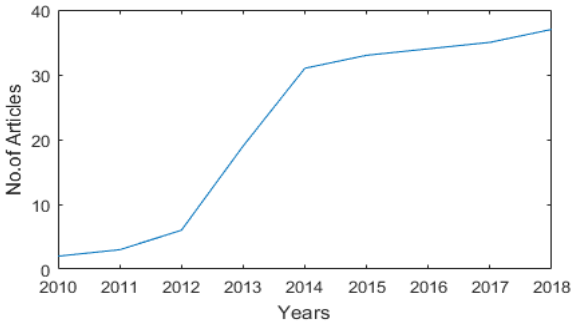

FIGURE 3. Distribution of research papers in terms of publication year.

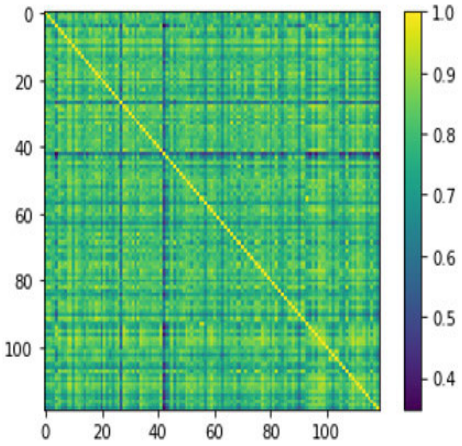

FIGURE 4. Heat map diagram (articles category 1).

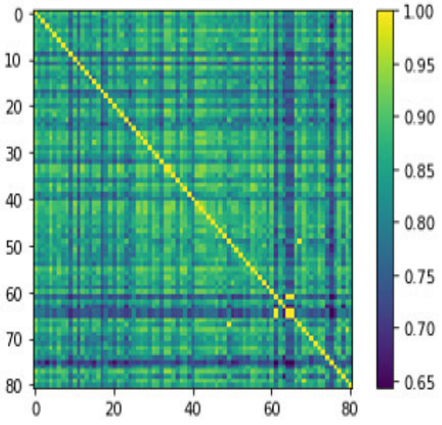

FIGURE 5. Heat map diagram (articles category 2).

E. SIMILARITY ANALYSIS

The similarity analysis technique presented in [23] was applied to the articles of both categories. The heat map shows resemblance among the articles. It is depicted from the (Fig.4 & Fig.5) that the articles in both categories belong to their particular domains as specified in the previous section. The similarity algorithm cannot find the exact correspondence between the articles, but all articles are interrelated to one another and having a likeness in meaning. Similarity analysis in terms of category shows that these articles are not exactly similar but interlinked in meaning in the context of their domains. The articles in the 1st category are more similar as compared to articles of the 2nd category.

TABLE 12. Cluster centroid terms for (articles category 1).

Clusters	Cluster Centroid Terms
0	Software, reliability, cloud, Big data OpenStack, computing, faults, rate, federation
1	Big data, Cloud computing, security, storage, analytics, service, information, applications
2	Big Data, cloud, computing, mobile, service, user, applications, platform, information, devices
3	Big data ,cloud computing, healthcare ,Hadoop ,cluster ,telehealth ,mapreduce
4	Histogram, event ,Big data, time, processing ,energy ,server ,allocation, intermediate ,performance

F. CLUSTER ANALYSIS

The techniques presented in [37] and [36] are applied to conduct cluster analysis in 1st category articles. The K-Means clustering algorithm is used to design clusters. We tried different values of K, 1 through 7 but finally selected k=5 because it gives the best results. Fig. 10 shows that there are five clusters, cluster 2 contains 108 articles, cluster 0 and 3 have one article each. The cluster 4 contains 5 and 1 contains 3 articles. The maximum number of (K=108) articles assembled in cluster 2. It means that these article discussing the main topic (cloud computing application towards big data). We tried and find out that the articles, assembled in other clusters (K=11) have a discussion on big data but slightly, from a different perspective as shown in Fig.8.

The terms that occurred in the centroid of the clusters are shown in Table12. As per study [36], the terms in the cluster centroid represents the subject being discussed in the articles of a cluster. As mention in (Table 12), the distinct centroid terms of cluster 0 are ''software'', ''reliability'' and ''OpenStack'' etc. that means the articles accumulated in this cluster are focusing on some sort of tools in the context of big data and cloud computing context. In a similar way, the distinct centroid terms of cluster 1 are ''security'', ''Storage'' and ''analytics'' etc. which means the articles accumulated in this cluster are focusing on big data storage, analytics, and applications of cloud computing services.

The distinct centroid terms of cluster 2 are ''mobile'','' devices'','' user'' etc. that means the articles accumulated in this cluster are focusing on big data, generated from user devices and mobile cloud platform services applications. The maximum number of articles accumulated in this cluster so, it means the big number of articles among the collected set are focusing in this subcategory. Cluster 3 distinct terms are health care, telehealth and ''map-reduce'', it means articles gathered in this cluster are focusing on the application of cloud computing in health care big data processing.

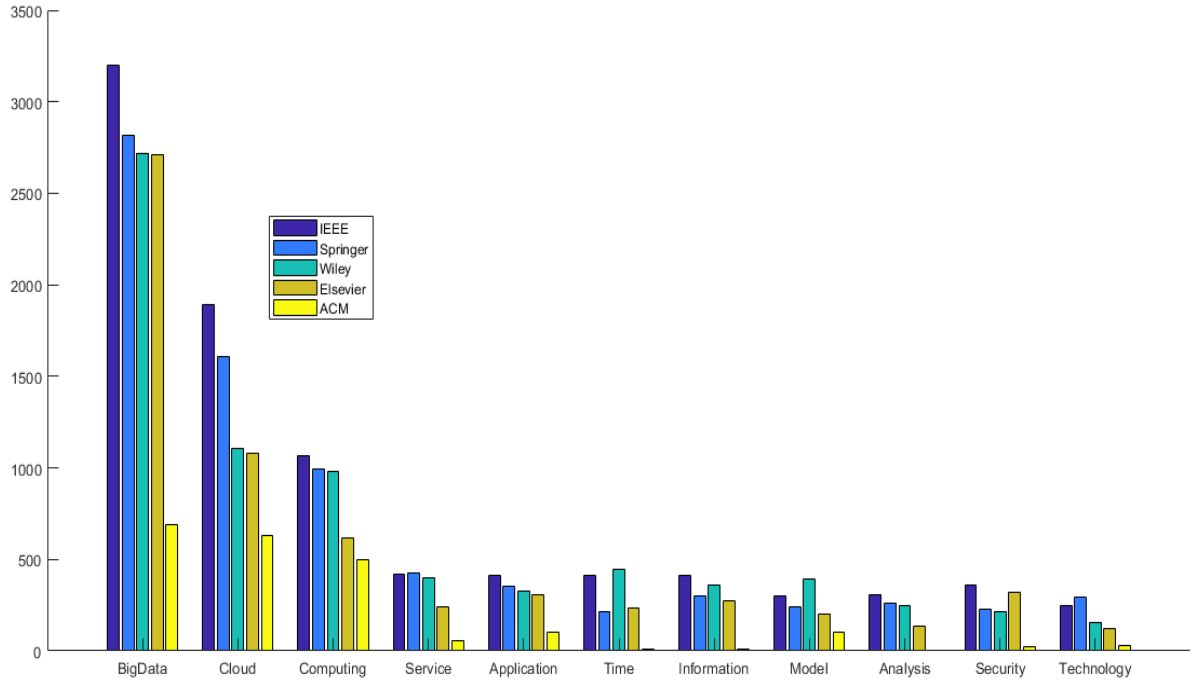

FIGURE 6. Word frequency distribution across all databases (category 1).

The cluster 4 distinct centroid terms are ''Histogram'', ''server'' and ''energy'', it means this cluster articles focused on Histogram techniques and its application. The centroid terms analysis shows that in every cluster the word Big Data and cloud computing is present so we can conclude that all these articles are interlinked and have a common context of the discussion.

The same process repeated to conduct cluster analysis for 2nd category articles. For this case k=5 i.e. there are five clusters, cluster 3 contains 70 articles, cluster 0 and 1 have one article each. Cluster 2 contains 5 and 4 contains 4 articles. The total number of articles (K=70) assembled in cluster 3. It means that these articles discussing the main topic (cloud computing adoption trends). We tried and find out that articles assembled in other clusters (K=11) have a discussion on cloud adoption trends but from a different perspective as shown in Fig. 11.

The terms that occurred in the centroid of the clusters are shown in Table13. As per study [36], the terms in the cluster centroid represents the subject being discussed in the articles of a cluster. As mentioned in (Table 13) the distinct centroid terms in cluster 0 are ''organization'','' capabilities'' and ''assessment'', it means the articles accumulated in this cluster are focusing on the capability of the organization for cloud computing adoption. The distinct centroid terms in cluster 1 are ''security'', ''compliance'','' checklist'' and ''gaps'' etc.it means the articles of this cluster focusing on security compliance and gaps. The cluster 2 distinct centroid terms are ''SMES'', ''factors'' and ''information'', it means the articles in this cluster are focusing on cloud adoption factors in SMEs. The distinct centroid terms of cluster 3 are

FIGURE 7. Word cloud across all databases (category 1).

''data'','' services'','' security'' and management, etc. this cluster contains maximum articles, it means the major topic of discussion in this cluster is cloud computing technology services adoption and management of data security. The cluster 4 distinct terms are ''migration'', ''tenant'', ''analytics'' etc.it means articles in this cluster focusing on data migration towards Cloud environment and tenant thought towards the cloud as a concern.

G. TOPIC MODELING

As per the study of [28] and [38], we used the LDA model to find out topics for the articles of both categories. As per the study [28], we analyzed all the research articles and developed dictionaries of keywords as mention in Fig2. We also calculate the frequency of keywords as some of the word frequencies mention in section (IV A&B). Then using dictionaries and word frequency values designed a term frequency matrix, this matrix is applied as input to LDA to generate results (i.e. Topics) as shown in appendix A and

```
In [139]: model.labels_

Out[139]: array([2, 2, 2, 2, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
                 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2,
                 2, 2, 2, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 4,
                 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
                 2, 2, 2, 2, 2, 2, 2, 2, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 4, 2, 2, 2,
                 2, 2, 2, 2, 1, 4, 2, 1, 1])

In [140]: from collections import Counter, defaultdict
          print(Counter(model.labels_))

          Counter({2: 108, 4: 5, 1: 4, 0: 1, 3: 1})
```

FIGURE 8. Cluster analysis (articles category 1).

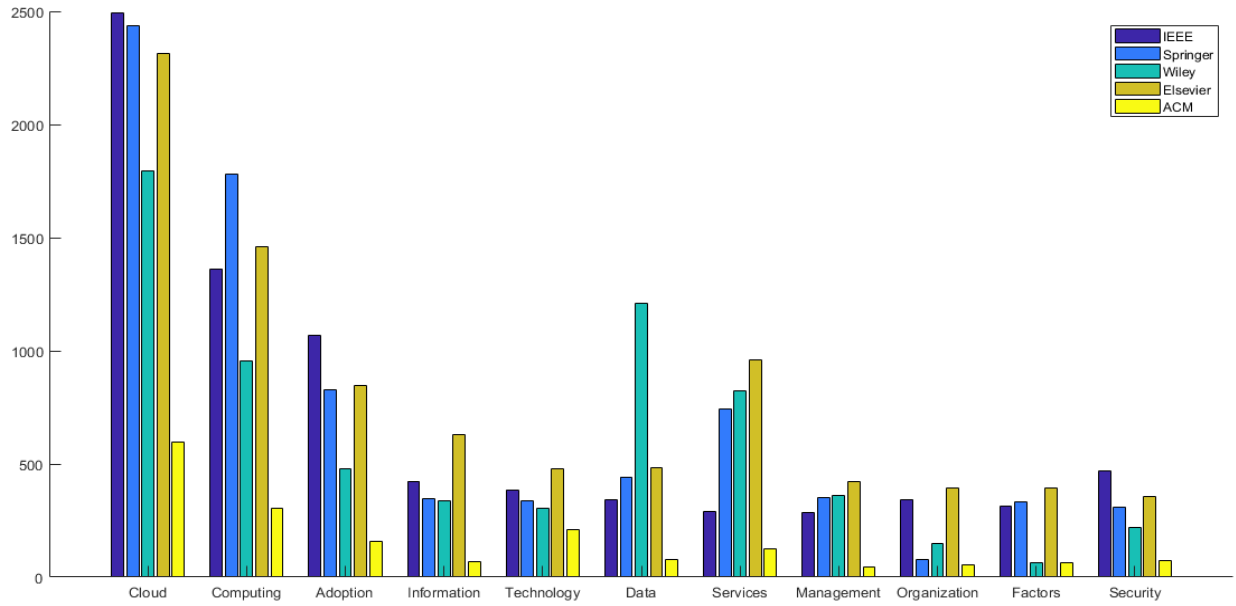

FIGURE 9. Word frequency distribution across all databases (articles category2).

FIGURE 10. Word cloud across all databases (articles category 2).

appendix B. The terms multiplied by values (b) is shown in both tables. The value b of the term close to zero means that term contributes more towards the topic. In articles of the 1st category, as per the study [29], we analyzed keywords that appeared in topics and cluster centroid and identified the big data areas discussed in these articles, in which cloud

```
In [207]: model.labels_

Out[207]: array([3, 3, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3,
                 3, 1, 2, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2,
                 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
                 3, 3, 2, 3, 3, 4, 3, 3, 4, 4, 4, 3, 3, 3, 3])

In [208]: from collections import Counter, defaultdict
          print(Counter(model.labels_))

          Counter({3: 70, 2: 5, 4: 4, 0: 1, 1: 1})
```

FIGURE 11. Cluster analysis (articles category 2).

computing technology is used for data processing. These are Enterprise big data, health care,business intelligence, bioinformatics, medical images, remote sensing image processing, telehealth, CCTV application data, and education learning system, mobile cloud processing, IOT big data, traffic hotline GPS data, big data image processing and internet of vehicles (IOV) big data processing. (Appendix Table 14 (Topic 1-10 and 18) and Table 12).Similarly (Topic 11-17 and Topic19 -24) (Appendix A (Table 14) and Table 11) we identified tools, techniques, terms, and algorithms discussed in the context of the application of the cloud computing model towards big data processing. These are NSGAII-algorithm, NoSQL, Open MP, OpenStack, FPGA, Hadoop, Spark, K means -Algorithm, SDN, Map Reduce and anonymization technique for data security. In articles of 2nd category, as per study [29], we analyzed keywords appeared in topics (Appendix B (Table 15)) and cluster centroid (Table 13) and found the following key points as given below:

Topic 1: discusses the understanding of the organizations and their management towards cloud computing technology that, cloud mechanism is outsourced, so what are data and processes backup recovery mechanism in case of disaster.

Topic 2 discusses the small enterprises' study and research for the adoption of cloud computing services and analysis of its factors that need to be taken care of.

Topic 3 & 9 discuss forms of cloud computing services SAAS, PAAS and IAAS, alternative systems and approaches, it concludes that decision can be made according to the requirements.

Topic 4 &14 identify the discussion from these articles that what is the cost and benefits of adopting this technology to business and security it provides, it also discussed about the infrastructure and software provided by a cloud provider.

Topic 5 & 8 discuss the organizations adopting cloud services for enterprise resource planning (ERP), medical and healthcare data services.

Topic 6 talks about the success of the cloud computing model in adoption because it provides all needful services to the business for the analysis, management of data and process the data.

Topic 7 talks about cloud computing success towards processing and analysis of big data systems.

TABLE 13. Cluster centroid terms for (articles category 2).

Clusters	Cluster Centroid Terms
0	adoption, cloud computing,clouds,organization,capabilities,assessment,enterprise
1	Cloud ,security,phase,assessment,checklist,gaps,methodology,computing,compliance
2	Cloud ,SMES, computing ,adoption ,factors ,information, data, technology
3	Cloud, computing, adoption, technology, data, services, security, management
4	Data,cloud,migration,tenant,application,analytics,computing,components,node,service

Topic 10 talks about the survey conducted in Saudi Arabia for cloud computing adoption. It concludes that cloud computing is adopted in health organizations and industry for its financial benefits.

Topic 11 talks about the maturity of the system. It discussed that the system is outsourced there may be some risk related to the capability and maturity of the framework.

Topic 12 & 13 depict one of the main factors of cloud computing adoption that, it is an innovative model which provides business services at a lower cost, this feature convinces the top management of organizations to adopt this system.

V. CONCLUSION

This research focuses on the application of text mining techniques for information extraction in the domain of cloud computing and its application towards big data processing. It finds key factors for successful adoption, hurdles in its adoption and other linked knowledge from different dimensions. We selected 200 research articles published from 2010 to 2018. The sources of these articles are IEEE, Springer, Elsevier, Wiley and ACM databases. The diverse text mining techniques are applied such as term frequency analysis, similarity analysis, cluster analysis and topic modeling (LDA). Using term frequency analysis we found the high-frequency words in literature and linked words in both categories of articles. Similarity analysis in terms of category shows that

TABLE 14. Top words in topics (articles category 1).

Topic	Top Words
1	(0.056*"business" + 0.053*"platform" + 0.049*"service" + 0.044*"enterprise" + 0.035*"data" + 0.031*"cloudbased" + 0.020*"management" + 0.019*"big" + 0.015*"industry"+0.011*"university"')
2	(0.063*"system" + 0.058*"healthcare" + 0.045*"information" + 0.042*"data" + 0.035*"service" + 0.028*"model" + 0.022*"computing" + 0.019*"medical" + 0.019*"care"+ 0.015*"analytics" + 0.013*"cloud" ')
3	(0.047*"cloud" + 0.045*"service" + 0.041*"data" + 0.038*"management" + 0.032*"bi" + 0.025*"analytics" + 0.025*"business"+ 0.023*"computing" + 0.017*"information"+ 0.016*"intelligence" +0.014*"big"')
4	('0.094*"data" + 0.070*"bioinformatics" + 0.037*"analysis" + 0.036*"biomedical" + 0.035*"computing" + 0.032*"large" + 0.024*"medical" + 0.022*"cloud" + 0.015*"image" + 0.015*"research")
5	('0.079*"data" + 0.067*"remote" + 0.054*"sensing" + 0.048*"cloud" + 0.047*"computing" + 0.046*"big" + 0.041*"image" + 0.022*"processing" + 0.019*"arcgis" + 0.015*"analytics")
6	('0.084*"data" + 0.040*"cloud" + 0.035*"computing" + 0.021*"big" + 0.020*"telehealth" + 0.016*"cctv" + 0.016*"application" + 0.016*"result" + 0.015*"sharing" + 0.014*"realtime"'),
7	('0.066*"learning" + 0.059*"system" + 0.056*"education" + 0.042*"computing" + 0.032*"video" + 0.026*"big" + 0.024*"data" + 0.019*"cloud" + 0.019*"retrieval" + 0.015*"online" + 0.015*"model")
8	('0.154*"mobile" + 0.065*"cloud" + 0.048*"computing" + 0.037*"network" + 0.036*"application" + 0.028*"system" + 0.027*"iot" + 0.027*"information" + 0.025*"smart" + 0.024*"architecture" + 0.023*"sensor" + 0.021*"big")
9	('0.227*"traffic" + 0.066*"hotline" + 0.062*"taxi" + 0.059*"gps" + 0.050*"data" + 0.041*"discovery" + 0.038*"method" + 0.036*"planning" + 0.029*"cloud"+ 0.020*"big" + 0.017*"transport" +0.015*"analysis")
10	('0.176*"method" + 0.139*"image" + 0.058*"cloud" + 0.050*"wavelet" + 0.050*"local" + 0.041*"algorithm" + 0.040*"based" + 0.026*"mean" + 0.023*"information" + 0.020*"dataset" + 0.019*"clustering" + 0.016*"analysis")
11	('0.002*"openstack" + 0.002*"nsgaii" + 0.002*"networked" + 0.002*"nosql" + 0.002*"opensource" + 0.002*"ortho" + 0.002*"paradigm"+ 0.002*"pcm" +0.002*"openmp" + 0.002*"obd"')
12	('0.150*"computing" + 0.043*"fpga" + 0.042*"cipher" + 0.042*"system" + 0.036*"hardware" + 0.029*"data" + 0.023*"platform" + 0.023*"control" + 0.023*"heterogeneous" + 0.022*"lightweight"')
13	('0.098*"scheduling" + 0.068*"vm" + 0.051*"time" + 0.050*"task" + 0.045*"execution" + 0.035*"data" + 0.022*"service" + 0.022*"big" + 0.022*"qos" + 0.021*"migration" + 0.020*"energyaware" + 0.018*"machine" + 0.017*"virtual"')
14	('0.116*"data" + 0.102*"cloud" + 0.080*"computing" + 0.047*"service" + 0.035*"big" + 0.027*"transmission" + 0.024*"based" + 0.022*"model" + 0.022*"method" + 0.022*"analysis" + 0.013*"hadoop" + 0.012*"security" + 0.012*"system"')

TABLE 14. *(Continued.)* Top words in topics (articles category 1).

15	('0.163*"data" + 0.046*"cloud" + 0.041*"algorithm" + 0.034*"based"+ 0.033*"graph" + 0.030*"technique" + 0.028*"series" + 0.028*"scheduling" + 0.022*"resource" + 0.021*"network" + 0.019*"computation" + 0.014*"model" + 0.014*"mapreduce" + 0.014*"twophase"')
16	('0.120*"mapreduce" + 0.062*"hadoop" + 0.053*"result" + 0.051*"data" + 0.048*"map" + 0.038*"framework" + 0.037*"cluster"+ 0.030*"intermediate" + 0.026*"performance" + 0.022*"parallel" + 0.020*"2" + 0.018*"execution"+ 0.016*"distributed" + 0.016*"function"),
17	(0.087*"data" + 0.070*"cluster"+ 0.033*"big" + 0.030*"computing" + 0.030*"cost" + 0.026*"time" + 0.024*"spark" + 0.021*"cloud" + 0.020*"mining"+ 0.015*"kmeans" + 0.014*"computation" + 0.013*"parallel"),
18	('0.228*"data" + 0.121*"big" + 0.034*"processing" + 0.032*"network" + 0.020*"technology" + 0.018*"application" + 0.018*"sdn" + 0.018*"cloud" + 0.013*"analysis" + 0.012*"stream" + 0.010*"computing"+ 0.010*"iov" + 0.009*"hadoop" + 0.009*"sensor")
19	('0.260*"data" + 0.103*"big" + 0.038*"service" + 0.037*"analytics" + 0.035*"storage" + 0.025*"platform" + 0.017*"management" + 0.012*"large" + 0.012*"business" + 0.012*"information" + 0.011*"access" + 0.010*"processing" + 0.010*"warehouse" + 0.009*"hadoop")
20	('0.272*"data" + 0.075*"cloud" + 0.055*"privacy" + 0.041*"big" + 0.025*"system" + 0.019*"model" + 0.019*"user" + 0.015*"access" + 0.013*"time" + 0.012*"computation" + 0.012*"framework" + 0.012*"anonymization" + 0.012*"service")
21	('0.128*"data" + 0.035*"processing" + 0.027*"model" + 0.025*"mapreduce" + 0.023*"application" + 0.021*"cloud" + 0.018*"stream" + 0.017*"computing" + 0.013*"large" + 0.012*"nosql" + 0.010*"graph"'),
22	('0.103*"cloud" + 0.075*"data" + 0.058*"application" + 0.058*"analytics" + 0.052*"big" + 0.037*"solution" + 0.030*"integration" + 0.024*"hybrid" + 0.021*"analysis" + 0.021*"business" + 0.020*"service" + 0.019*"framework")
23	(0.077*"cloud" + 0.077*"data"+ 0.046*"big" +0.034*"analysis" + 0.029*"model" + 0.027*"network" + 0.027*"computing" + 0.025*"time" + 0.024*"assessment" + 0.022*"method" + 0.021*"openstack")
24	('0.093*"data"+ 0.042*"big" + 0.036*"application" + 0.036*"management" + 0.035*"histogram" + 0.033*"openstack" + 0.033*"experimental" +0.030*"machine" + 0.028*"network" + 0.027*"migration" + 0.024*"technology" + 0.020*"computing")

these articles are not exactly similar but interlinked in meaning in the context of their domains. The articles in the 1st category are more similar as compared to articles of the 2nd category. Cluster analysis technique shows that highly related documents accumulated in one cluster, it means that these articles discussing the same topic. The topic modeling technique grouping papers into logically related topics.

TABLE 15. Top words in topics (articles category 2).

Topic	Top Words
1	('0.005*"programming" + 0.005*"research" + 0.005*"outsourcing" + 0.005*"resource" + 0.005*"return" + 0.005*"reality" + 0.005*"resilience" + 0.005*"provider" + 0.005*"paas" + 0.005*"professional" + 0.005*"process" + 0.005*"privacy" + 0.005*"potentially" + 0.005*"possible" + 0.005*"planning" + 0.005*"public" + 0.005*"readiness" + 0.005*"ranking"')
2	('0.191*"smes" + 0.066*"research" + 0.064*"adoption" + 0.061*"study" + 0.054*"issue" + 0.052*"information" + 0.050*"enterprise" + 0.046*"ict" + 0.036*"theory" + 0.035*"management" + 0.031*"small" + 0.030*"developing" + 0.025*"technology" + 0.021*"computing" + 0.018*"sme" + 0.017*"approach" + 0.015*"network" + 0.015*"data" + 0.015*"analysis" + 0.014*"factor"')
3	('0.090*"cloud" + 0.067*"software" + 0.058*"model" + 0.057*"decision" + 0.051*"public" + 0.047*"study" + 0.040*"analysis" + 0.035*"need" + 0.033*"medium" + 0.033*"computing" + 0.027*"provider" + 0.026*"service" + 0.025*"paas" + 0.024*"alternative" + 0.024*"system" + 0.024*"process" + 0.022*"saas" + 0.022*"iaas" + 0.022*"approach" + 0.021*"based"'),
4	('0.241*"cloud" + 0.139*"computing" + 0.049*"service" + 0.032*"adoption" + 0.029*"data" + 0.026*"cost" + 0.021*"enterprise" + 0.021*"technology" + 0.020*"business" + 0.018*"security" + 0.017*"information" + 0.017*"provider" + 0.017*"resource" + 0.015*"benefit" + 0.015*"infrastructure" + 0.014*"organization" + 0.013*"need"+ 0.012*"software"')
5	('0.179*"erp" + 0.135*"system" + 0.088*"organization" + 0.059*"service" + 0.047*"cloudbased" + 0.029*"saas" + 0.027*"cloud" + 0.025*"process" + 0.023*"business" + 0.019*"approach" + 0.019*"factor" + 0.019*"cost" + 0.018*"model" + 0.018*"provider" + 0.018*"iaas" + 0.018*"resource" + 0.016*"security" + 0.016*"csp" + 0.014*"infrastructure" + 0.014*"application"')
6	('0.097*"service" + 0.094*"system" + 0.085*"cloud" + 0.069*"model" + 0.050*"organisation" + 0.043*"information" + 0.041*"success" + 0.028*"provider" + 0.021*"computing" + 0.020*"business" + 0.019*"management" + 0.018*"need" + 0.018*"application" + 0.017*"resource" + 0.015*"technology" + 0.015*"science" + 0.014*"analysis" + 0.014*"data" + 0.013*"infrastructure" + 0.013*"technical"')
7	('0.233*"data" + 0.062*"big" + 0.055*"system" + 0.054*"cloud" + 0.048*"computing" + 0.035*"service" + 0.029*"resource"

TABLE 15. *(Cointinued.)* Top words in topics (articles category 2).

	+ 0.027*"model" + 0.026*"analytics" + 0.024*"application" + 0.019*"technology" + 0.019*"management" + 0.017*"infrastructure" + 0.015*"architecture" + 0.015*"network" + 0.014*"interface" + 0.014*"analysis" + 0.014*"programming" + 0.014*"information" + 0.013*"based"'),
8	('0.102*"information" + 0.093*"organization" + 0.077*"adoption" + 0.052*"research" + 0.037*"system" + 0.034*"industry" + 0.034*"factor" + 0.033*"health" + 0.033*"medical" + 0.029*"study" + 0.025*"management" + 0.025*"data" + 0.024*"service" + 0.023*"resource" + 0.023*"theory" + 0.017*"cloud" + 0.016*"technology" + 0.015*"time" + 0.014*"framework" + 0.014*"care"')
9	('0.116*"adoption" + 0.089*"saas" + 0.059*"study" + 0.052*"management" + 0.049*"institutional" + 0.041*"factor" + 0.039*"information" + 0.036*"research" + 0.033*"framework" + 0.032*"theory" + 0.029*"technology" + 0.029*"model" + 0.026*"organizational" + 0.025*"top" + 0.024*"environment" + 0.020*"system" + 0.016*"coercive" + 0.016*"mimetic"')
10	('0.170*"cloud" + 0.157*"computing" + 0.105*"sector" + 0.061*"saudi" + 0.058*"study" + 0.049*"public" + 0.043*"knowledge" + 0.041*"service" + 0.036*"industry" + 0.028*"model" + 0.026*"data" + 0.026*"science" + 0.024*"adoption" + 0.019*"technology" + 0.018*"financial" + 0.017*"mean" + 0.017*"organization" + 0.017*"health" + 0.015*"empirical" + 0.013*"deployment"')
11	('0.120*"process" + 0.069*"management" + 0.051*"model" + 0.048*"information" + 0.042*"cloud" + 0.041*"supply" + 0.037*"system" + 0.037*"architecture" + 0.036*"business" + 0.035*"chain" + 0.033*"service" + 0.029*"computing" + 0.025*"outsourcing" + 0.022*"enterprise" + 0.020*"assessment" + 0.018*"approach" + 0.017*"maturity" + 0.017*"risk" + 0.016*"capability" + 0.016*"framework"'),
12	('0.158*"cloud" + 0.125*"computing" + 0.105*"technology" + 0.074*"factor" + 0.061*"acceptance" + 0.043*"influence" + 0.042*"research" + 0.040*"model" + 0.031*"information" + 0.027*"study" + 0.026*"system" + 0.025*"science" + 0.019*"diffusion" + + 0.017*"empirical" + 0.016*"organization" + 0.015*"innovation" + 0.013*"theory" + 0.013*"higher" + 0.011*"data"'),

TABLE 15. *(Cointinued.)* **Top words in topics (articles category 2).**

13	('0.218*"adoption" + 0.101*"factor" + 0.077*"computing" + 0.065*"study" + 0.049*"cloud" + 0.044*"decision" + 0.037*"technology" + 0.035*"model" + 0.034*"innovation" + 0.033*"framework" + 0.031*"service" + 0.030*"organizational" + 0.028*"system" + 0.026*"top" + 0.024*"management" + 0.021*"information" + 0.021*"research" + 0.020*"based" + 0.016*"theory"'),
14	('0.207*"application" + 0.125*"cloud" + 0.111*"cost" + 0.094*"migration" + 0.093*"service" + 0.043*"time" + 0.043*"network" + 0.041*"enterprise" + 0.039*"communication" + 0.028*"resource" + 0.027*"architecture" + 0.023*"approach" + 0.020*"data" + 0.014*"benefit" + 0.012*"virtual" + 0.012*"model" + 0.010*"deployment" + 0.007*"information" + 0.007*"possible"')

By the analysis of keywords in topics and cluster centroid terms, this study depicted the ten big data areas in 1st category articles, in which cloud computing technology is being used. In articles of the 2nd category, this study discovered fourteen cloud adoption factors and hurdles in adoption. The author's affiliation data analysis shows that most of the publications originated from Asia, Europe, North America, Australia and a few from Africa. It has been observed in the analysis that the number of articles in these domains increased from 2013 onward which shows an increase research trend in these domains. In the future, we will extend this study with a larger corpus and a wider scope to discover knowledge from published sources of literature.

APPENDIX A

See table 14.

APPENDIX B

See table 15.

REFERENCES

[1] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text mining methods and techniques," *Int. J. Comput. Appl.*, vol. 85, no. 17, pp. 42–45, 2014.
[2] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Burlington, MA, USA: Elsevier, 2011.
[3] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Commun. ACM*, vol. 49, no. 9, pp. 76–82, 2006.
[4] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, pp. 60–76, Aug. 2009.
[5] S. Gupta, G. E. Kaiser, P. Grimm, M. F. Chiang, and J. Starren, "Automating content extraction of html documents," *World Wide Web*, vol. 8, no. 2, pp. 179–224, 2005.
[6] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, 2010. [Online]. Available: https://www.springer.com/gp/book/9780387954332
[7] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications—A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.
[8] A. Akilan, "Text mining: Challenges and future directions," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 1679–1684.
[9] M. Sukanya and S. Biruntha, "Techniques on text mining," in *Proc. IEEE Int. Conf. Adv. Commun. Control Comput. Technol. (ICACCCT)*, Aug. 2012, pp. 269–271.
[10] S. A. Salloum, M. Al-Emran, and K. Shaalan, "A Survey of lexical functional grammar in the Arabic context," *Int. J. Comput. Netw. Technol.*, vol. 4, no. 3, pp. 141–147, 2016.
[11] M. T. Pazienza, Ed, *Information Extraction: Towards Scalable, Adaptable Systems*. Springer, 2003. [Online]. Available: https://www.springer.com/us/book/9783540666257
[12] R. Ban, L. Tu, and H. Liu, "Cloud computing platform for big data security strategy research," *J. Post Des. Technol.*, vol. 10, pp. 74–78, 2017.
[13] A. Elragal and M. Haddara, "The future of ERP systems: Look backward before moving forward," *Procedia Technol.*, vol. 5, pp. 21–30, Oct. 2012.
[14] N. Su, R. Akkiraju, N. Nayak, and R. Goodwin, "Shared services transformation: Conceptualization and valuation from the perspective of real options," *Decis. Sci.*, vol. 40, pp. 381–402, Aug. 2009.
[15] G. Garrison, S. Kim, and R. L. Wakefield, "Success factors for deploying cloud computing," *Commun. ACM*, vol. 55, no. 9, pp. 62–68, Sep. 2012.
[16] X. Jiang and J. Zhang, "A text visualization method for cross-domain research topic mining," *J. Vis.*, vol. 19, no. 3, pp. 561–576, 2016.
[17] X. Zhai, Z. Li, K. Gao, Y. Huang, L. Lin, and L. Wang, "Research status and trend analysis of global biomedical text mining studies in recent 10 years," *Scientometrics*, vol. 105, no. 1, pp. 509–523, 2015.
[18] S. D. Kavila and D. F. Rani, "Information extraction from research papers based on statistical methods," in *Proc. 2nd Int. Conf. Comput. Commun. Technol.*, in Advances in Intelligent Systems and Computing, vol. 381, S. Satapathy, K. Raju, J. Mandal, and V. Bhateja, Eds. Springer, 2016, pp. 573–580. [Online]. Available: https://link.springer.com/chapter/10.1007/978-81-322-2526-3_59
[19] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields," *Inf. Process. Manage.*, vol. 42, pp. 963–979, Jul. 2006.
[20] K. C. Santosh, "g-DICE: Graph mining-based document information content exploitation," *Int. J. Document Anal. Recognit.*, vol. 18, no. 4, pp. 337–355, 2015.
[21] E. Choi, M. Horvat, J. May, K. Knight, and D. Marcu, "Extracting structured scholarly information from the machine translation literature," in *Proc. 10th Int. Conf. Lang. Resour. Eval.*, 2016, pp. 421–425.
[22] M. Song and S. Y. Kim, "Detecting the knowledge structure of bioinformatics by mining full-text collections," *Scientometrics*, vol. 96, no. 1, pp. 183–201, 2013.
[23] C. Clifton, R. Cooley, and J. Rennie, "TopCat: Data mining for topic identification in a text corpus," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 8, pp. 949–964, Aug. 2004.
[24] A. Iftikhar, S. W. Ul Qounain Jaffry, and M. K. Malik, "Information mining from criminal judgments of lahore high court," *IEEE Access*, vol. 7, pp. 59539–59547, 2019.
[25] S. Jayashankar and R. Sridaran, "Superlative model using word cloud for short answers evaluation in eLearning," *Educ. Inf. Technol.*, vol. 22, no. 5, pp. 2383–2402, 2016.
[26] J. Popp, P. Balogh, J. S. Oláh Kot, M. H. Rakos, and P. Lengyel, "Social network analysis of scientific articles published by food policy," *Sustainability*, vol. 10, no. 3, p. 577, 2018.
[27] K. Shaalan *et al.*, Eds., *Intelligent Natural Language Processing: Trends and Applications* (Studies in Computational Intelligence), vol. 740. Springer, 2018.
[28] A. Amadoa, P. Cortezb, P. Rita, and S. Moro, "Research trends on big data in marketing: A text mining and topic modeling based literature analysis," *Eur. Res. Manage. Bus. Econ.*, vol. 24, no. 1, pp. 1–7, 2017.
[29] B. Nie and S. Sun, "Using text mining techniques to identify research trends: A case study of design research," *Appl. Sci.*, vol. 7, no. 4, p. 401, 2017.
[30] (Jul. 24, 2019). *Processing Raw Text*. [Online]. Available: https://www.nltk.org/book/ch03.html
[31] C. A. DePaolo and K. Wilkinson, "Get your head into the clouds: Using word clouds for analyzing qualitative assessment data," *TechTrends*, vol. 58, no. 3, pp. 38–44, 2014.
[32] J. Sinclair and M. Cardew-Hall, "The folksonomy tag cloud: When is it useful?" *J. Inf. Sci.*, vol. 34, no. 1, pp. 15–29, 2008.

[33] F. B. Viegas, M. Wattenberg, F. V. Ham, J. Kriss, and M. McKeon, "ManyEyes: A site for visualization at Internet scale," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1121–1128, Nov. 2007.
[34] A. Huang, "Similarity measures for text document clustering," in *Proc. 6th New Zealand Comput. Sci. Res. Student Conf. (NZCSRSC)*, Christchurch, New Zealand, 2008, pp. 49–56.
[35] E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher, "Clustering based on association rule hypergraphs," in *Proc. DMKD*, 1997, pp. 1–5.
[36] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, N. Tziritas, C.-Z. Xu, A. Y. Zomaya, A. S. Alzahrani, and H. Li, "A survey on text mining in social networks," *Knowl. Eng. Rev.*, vol. 30, no. 2, pp. 157–170, 2015.
[37] S. Zaza and M. Al-Emran, "Mining and exploration of credit cards data in UAE," in *Proc. 5th Int. Conf. e-Learn. (ECONF)*, 2015, pp. 275–279.
[38] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2018.

MUHAMMAD INAAM UL HAQ received the B.S. degree in computer science from the National University of Computer and Emerging Sciences (FAST-NUCES), in 2010, and the M.Phil. degree from Lahore Leads University, Pakistan, in 2014. He is currently pursuing the Ph.D. degree in computer science with the School of Computer science and Engineering, Nanjing University of Science and Technology (NJUST), China. His research interests include data mining, text mining, NLP, and machine learning.

QIANMU LI received the B.Sc. and Ph.D. degrees from the Nanjing University of Science and Technology, China, in 2001 and 2005, respectively. He is currently a Full Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. He is the author or the coauthor of more than 100 high indexed (SCIE/E-SCI/EI) journal/conference articles and eight books. His research interests include information security, computing system management, and data mining. He was a recipient of the China Network and Information Security Outstanding Talent Award, in 2016, and multiple Education Ministry Science and Technology Awards, in 2012.

SHOAIB HASSAN received the B.Sc. degree in computer system engineering from UCET, IUB, in 2012, and the M.S. degree in software engineering from NUST, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in computer science with the School of Computer science and Engineering, Nanjing University of Science and Technology, China. His research interests include text mining, data mining, software engineering, and bio-informatics.