IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON AI-DRIVEN BIG DATA PROCESSING: THEORY, METHODOLOGY, AND APPLICATIONS

Received May 7, 2019, accepted June 8, 2019, date of publication June 24, 2019, date of current version July 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923583

An Artificial Intelligence Driven Multi-Feature Extraction Scheme for Big Data Detection

JIAN WAN[1,2,3], PIAOPIAO ZHENG[1], HUAYOU SI[2,3], NEAL N. XIONG[4], WEI ZHANG[2,3], AND ATHANASIOS V. VASILAKOS[5]

[1]School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Zhejiang 310023, China
[2]School of Computer Science and Technology, Hangzhou Dianzi University, Zhejiang 310018, China
[3]Key Laboratory of Complex Systems Modeling and Simulation, Ministry of Education, Hangzhou 310018, China
[4]College of Intelligence and Computing, Tianjin University, Tianjin 300072, China
[5]Department of Computer and Telecommunications Engineering, University of Western Macedonia, 50100 Kozáni, Greece

Corresponding authors: Huayou Si (sihy@hdu.edu.cn) and Neal N. Xiong (xiongnaixue@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61472112 and Grant 61502129, and in part by the Key Research and Development Program of Zhejiang Science and Technology Department under Grant 2017C03047.

ABSTRACT The Internet improves the speed of information dissemination, and the scale of unstructured text data is expanding and increasingly being used for mass communication. Although these large amounts of data meet the infinite demand, it is difficult to find public focus in a timely manner. Therefore, information extraction from big data has become an important research issue, and there are many published studies on big data processing at home and abroad. In this paper, we propose a multi-feature keyword extraction method, and based on this, an artificial intelligence driven big data MFE scheme is designed, then an application example of the general scheme is expanded and detailed. Taking news as the carrier, this scheme is applied to the algorithm design of hot event detection. As a result, a multi-feature fusion clustering algorithm is proposed based on user attention with two main stages. In the first stage, a multi-feature fusion model is developed to evaluate keywords, and this model combines the term frequency and part of speech features. We use it to extract keywords for representing news and events. In the second stage, we perform clustering and detect hot events in accordance with the procedure, and during the composition of news clusters, we analyze several variadic parameters in order to explore the optimal effectiveness. Then, experiments on the news corpus are conducted, and the results show that the approach presented herein performs well.

INDEX TERMS Artificial intelligence, extraction scheme, big data, online news, news clustering.

I. INTRODUCTION

In recent years, the rapid development of the Internet has ushered in an era of explosive growth of information. With the comprehensive transfer of human life to the Internet, the era of big data has now become inevitable. As a frontier concept of the global Internet, big data mainly includes two characteristics: on the one hand, the amount of information in the whole society increases sharply; on the other hand, the amount of information available to individuals is growing exponentially. From the perspective of scientific and technological development, big data is an inevitable product under the trend of digitalization. As this trend continues, we will enter an era where everything is recorded and everything is digitized in the near future. Everyone's life is full of structured and unstructured data from the Internet, social media [1], [2] and mobile media [3], [4]. In the past few years, the big data industry has paid more attention to how to deal with massive, multi-source and heterogeneous data and obtain value from it, but most of them are structured data. It is undeniable that, although structured data volume is large enough, these are only the tip of the iceberg. According to IDC, data is growing at a rapid rate of about 50% per year. By 2020, the global data scale will reach 40ZB [5], among which text and other unstructured data account for up to 75-85%. Therefore, it is particularly urgent and important to mine and analyze unstructured text data.

In business practice, information extraction based on big data is widely used in all walks of life, using cognitive technology to gain new business insights and solve key intellectual problems, which is called cognitive business by IBM. For example, companies can obtain text data from

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

customer relationship data, social networks, news sites and shopping site reviews, then perform natural language processing through computers, ultimately get business insights at a whole new level across all businesses. Briefly, information extraction based on big data can reveal trends and associations hidden in textual information, providing strong support for business decisions, industry trend research, and hot content tracking. Actually, the development of information extraction in China is relatively early. Especially with the advent of the era of artificial intelligence, the importance of information extraction has gradually been valued by the market.

Information extraction is to mine and establish a monitoring model. Once the data capacity explodes or the popular vocabulary is updated, information extraction can update the learning model in real time and redefine the monitoring model. In this paper, we proposed a multi-feature keyword extraction (MFE) for calculating the relevance of words and phrases to the content of the article subject and returning the first N words or phrases that can best represent the article subject according to the order of relevance. The evaluation method of multi-feature fusion is used in the process of keyword extraction, which can extract high-quality keywords even if the article is short. Then combining with the features of user attention, such as article reading quantity, comment volume and comment growth rate, we adopted a new algorithm based on MFE to cluster text of various media, so as to facilitate subsequent analysis of social hot events.

Next, we will take online news as an example. Online news reports have increased considerably among enormous quantities of data on blogs, online newspapers and news websites. The massive amount of news can overwhelm people when they access reports of interest, although there are currently general classifications of news, such as business, technology, and sports. For an individual who frequently reads news, it would be desirable if they could access or abandon all similar news reports conveniently targeting events in which they are personally interested or uninterested. To accomplish this, the clustering model built in this paper combines characteristics of news for the static pool of news. Usually, a news report includes not only the contents of the report itself but also some incidental information, such as the number of comments, which is often overlooked. In addition to concentrating on seeking better utilization of news content [6], we expect incidental information to play its due role. Depending on this case, the following two aspects will be taken into consideration. One is an improved representation of news reports, which should effectively model the contents. The other is that hot news can better represent events, so we propose a method based on user attention for event extraction.

According to this case, our main contributions include the following: 1) Based on tens of thousands of online news from NetEase News, we analyzed the characteristics of static pool with news to extract events that most people focus on. 2) We proposed a new method to calculate the similarity between news and events, relying on an aligned set of basis vectors obtained using keyword correlation analysis. 3) We established a multi-feature fusion model for evaluating keywords that combines the term frequency and part of speech features together. 4) We designed an approach to detect hot events in accordance with the procedure, and during the composition of news clusters, we analyzed several variadic parameters in order to explore the optimal effectiveness.

The remainder of this paper is organized as follows. Section 2 gives a brief review of related work. Section 3 provides the necessary text representation model and presents our computation method of similarity. In section 4, the proposed approach is presented. Then, we report on the experimental methodologies and results in section 5, followed by conclusions and future work in section 6.

II. RELATED WORK

This research mainly relates to previous work on information extraction, news event detection and keyword extraction. Thus, we mainly review them in this section.

During the last few years, information extraction has attracted wide attention due to adding knowledge to the search results of major commercial search engines. Working with collections of articles in a particular domain to extract relevant information in a structured manner are often customized. This means some specific templates are used for filling in with information collected from texts. The main obstacle is poor portability because templates are designed for a specific purpose and they are difficult to reuse. Mooney and Bunescu [7] describe an information extraction algorithm for identifying entities and relationships in texts. Infoboxes and Wikipedia are used to solve the problem of named entity recognition [8]. To solve the problem of mixing structured texts, a method [9] for identifying terms in articles is proposed, using the terms to identify document-related fragments. The method [10] describes how to summarize a web entity based on the entity's occurring in web articles. We can refer to [11] for the knowledge generation. It describes a project that attempts to automatically extract information about artists from the Web, use it to generate personalized narrative biographies, and populate a knowledge base. YAGO [12] is well-known project related to extracting knowledge from WordNet and Wikipedia. An automatic query-based retrieval method [13] has been built to extract user-defined relationships from large text databases such as media databases. Gkatziaki *et al.* [14] contribute to the computation of objective evaluations with a tool that is targeted at extracting data about companies from Web-accessible, semi-structured resources in a way that fits the justly tight requirements of the platform. Based information extraction, machine learning has gained much more interest due to effectiveness and efficiency, particularly their success in many shared tasks [15]. Some machine learning methods were directly used for natural language processing task such as tumor information extraction [16]. Information extraction is a hot topic in the field of natural language processing in recent years, in which event extraction is one of the

three main tasks. The ACE (Automatic Content Extraction) evaluation conference, which promotes the development of event extraction, defines events as special things that involve participants [17]. In the MUC (Message Understanding Conference) evaluation meeting before the ACE meeting, there is a scenario template task, which focuses on the extraction of events. As early as 1958, a classical approach [18] based on the frequency of occurrence of words was proposed. The research on event extraction has strong domain relevance, emphasizing the extraction of relevant information from the text according to the specified event type and its template. Ji *et al.* [19] put forward a method for extracting events by taking frequently occurring named entities as core entities, which could perform cross-document event recognition tasks and then arrange the identified events on the timeline. Garrido *et al.* [20] proposed a method called TM-Gen, extracting information from any number of articles and representing them in a topic map format. Shinyama *et al.* [21] describes the process of using named entity recognition to obtain important definitions (citing different narrative styles of the same event) from news articles. Different natural language processing systems have been developed and utilized to extract events and clinical concepts form text, and several success stories in applying these tools have been reported widely [22], [23]. Yazdani *et al.* [24] proposed an approach relies on both atrial and ventricular activity analysis, based on a novel non-linear filtering technique recently proposed to extract short-term events from biomedical signals.

Although the definition of an event can vary significantly, the notion of an event is widely used in many related research fields in natural language processing [25]. According to WordNet [26], a general definition of a news event is ''a specific thing happens at a specific place and time'', which may be consecutively reported by various media within a period. Most prevailing approaches to news event detection were proposed based on Allan *et al.* [27]. They were mainly variants and improvements of the single pass method and agglomerative clustering algorithms [28]–[34]. One characteristic of news, time information, was not helpful for improving the results of news event detection [28], while another study [35] utilized aging theory and obtained good performance. In this paper, the boosting comment characteristic will be the focus. In addition, the classical news clustering algorithm could be classified into two groups: k-means variations and vector space models. The k-means algorithm and its extensions [36]–[39] are most popular for partitioned and hierarchical clustering with a number of defects: effectiveness depends on random initialization and decreases with big data [40]. The vector space model [41] is an approach that reveals better results for homogeneous events and requires ensuring the number of clusters in advance [42]. Some techniques related to ontologies, machine learning, and natural language processing were applied to help the documentalists of categorization and tagging of news [43]–[45]. It is not an easy task to extract specific information from a large number of articles with a journalistic writing style in natural language. Many studies [46]–[48] have faced such problems, which have not yet been solved. Many researchers applied events detection to a specific field [49], [50], working on detection of economic events that may influence the market, such as mergers. Yang *et al.* [51] presented an event detection framework to discover real-world events from multiple data domains, including online news media and social media.

Automatic keyword extraction is the process of identifying key paragraphs, key phrases, key terms, or keywords in an article that properly represent the document topic [52]. Due to keyword extraction provides a compact representation of article, some applications, such as automatic classification, automatic clustering, automatic indexing, automatic filtering, and automatic summarization, can benefit from the keyword extraction process [53]. Keyword is the smallest meaningful element of language that can move independently in Chinese, and the relationship between news and corresponding keywords has been studied extensively, falling into two categories: supervised and unsupervised approaches. In the former, the keyword extraction algorithm consists of two steps. The first step is training, in which models are trained to find keywords from labeled news reports. The second step is extraction, where the keywords are chosen from news reports that are not trained. The latter is composed of simple statistical approaches, linguistic approaches and machine learning approaches. Zhao and Zeng [54] extracted keywords from micro blogs using a three-feature graph model, semantic space and word location. Hromic *et al.* [55] focused on a structure approach based on exploded events and graph generation. Marujo *et al.* [56] proposed a method to find solutions for problems such as high variance and lexical variants. Kim *et al.* [57] detected exploded and popular keywords including abbreviations. Zimniewicz *et al.* [58] applied a scheduling model for a class of keyword extraction approaches and proposed methods for the overall performance evaluation of different algorithms, which are based on processing time and correctness (quality) of answers. Duari and Bhatnagar [59] proposed a parameterless keyword extraction method (sCAKE) based on semantic connectivity of words, combining with graph construction and scoring methods. In this paper, a multi-feature fusion will be used for keyword extraction.

III. SYSTEM MODEL AND DEFINITIONS

To calculate the similarity between news and events, we should transform news reports into a form that a computer can understand, that is, to build a model to represent news reports. The commonly used text representation model is the vector space model, which we also use in this paper. Every dimension of vector is a feature item extracted from news. Using vectors to represent text, the relevant knowledge in mathematics to calculate the similarity between vectors can be used, and the similarity between vectors can be referred to as the similarity between news, thus reducing the difficulty of calculating similarity between news reports.

A. NEWS AND EVENT MODEL

For each news report N, we can extract keywords from news headlines and contents as feature items and construct news vector models based on these characteristics. Using the vector $N(n_1, m_1, n_2, m_2, \ldots, n_k, m_k)$ to represent a news report, n_k is the feature of the vector, and the feature term is the keyword extracted from the news. The variable m_k is the weight of the feature item, which will be introduced in more detail in section 4.1.

For each hot event, a vector $E(e_1, s_1, e_2, s_2, \ldots, e_i, s_i)$ can be constructed to represent the event, of which e_i is a keyword extracted from the event and s_i is the weight of the keyword related to the event. We specify a ten-dimensional vector for an event; then, the initial elements of the vector are set to the same as the first news in this event. When subsequent news is classified to the event, keywords change, and the corresponding weights are recalculated (see Section 4.2).

B. SIMILARITY CALCULATION

To support the functionality, which is finding groups of reports describing the same event, we take the similarity comparison method into consideration. This section explains how to calculate an approximate similarity between news reports and events. Then, the computation is based on an aligned set of basis vectors obtained using keyword correlation analysis.

In this paper, we propose a new method to calculate the similarity between news and events. The following is the relevant definition:

Definition N: N represents a piece of news.

Definition E: E represents an event.

Definition t_0: t_0 represents the current time.

Definition P_1:P_1 is a set, $P_1 = \{k_1, k_2, \ldots, k_m\}$, one of which k_i represents a keyword that appears at the same time in news N and event E.

Definition P_2: P_2 is a set, $P_2 = \{w_1, w_2, \ldots, w_m\}$, which contains the weight for each of the keywords in P_1.

Definition P_3: P_3 is a set, $P_3 = \{t_1, t_2, \ldots, t_m\}$, which contains the last time that each keyword in P_1 was recently updated.

Definition P_4: P_4 is a set, $P_4 = \{e_1, e_2, \ldots, e_n\}$, which contains all keywords for an event.

Definition P_5: P_5 is a set, $P_5 = \{s_1, s_2, \ldots, s_n\}$, which contains the weight of each keyword in P_4.

Definition P_6: P_6 is a set, $P_6 = \{q_1, q_2, \ldots, q_n\}$, which contains the last time that each keyword in P_4 was most recently updated.

From there, the similarity calculation formula is as follows.

$$Sim\,(N, E) = \frac{\sum_{i=1}^{m} \frac{1}{t_0 - t_i} \times w_i}{\sum_{j=1}^{n} \frac{1}{t_0 - q_j} \times s_j} \tag{1}$$

The similarity value we obtain with the above formula will be a fractional number between 0 and 1. The closer the value is to 1, the greater the similarity between news N and event E is and the greater the possibility that news N is classified into event E. Conversely, the closer the value is to 0, the smaller the possibility that news N is classified into event E. Finally, the calculated similarity is compared with a given threshold. If it is larger than the given threshold, the news is classified into the event. Otherwise, the news is stored as a new event in the event library.

IV. THE PROPOSED MFE SCHEME

A. MULTI-FEATURE KEYWORD EXTRACTION

TF is an acronym for term frequency, i.e., the number of times a term occurs in an article. It is generally believed that the higher the TF of a word, the more important the word is in the article. For search engine optimization, a large number of duplicate words are packed in one article. To avoid this, C_{total} is set as the total number of words in a news set, and $L_i = TF/C_{total}$. When $L_i > L$, we regard this word as useless information with low importance, in which L is set to 0.75.

$$TF_{new} = \begin{cases} TF \dfrac{TF}{C_{total}} \leq L \\ 0\ TF/C_{total} > L \end{cases} \tag{2}$$

According to the characteristics of the Chinese language, keywords are generally Nouns (n) and Verbs (v), and a few adjectives and adverbs are included. Prepositions and auxiliary words generally cannot express specific meaning. The Names (nr), Place Names (nt) and Institution Names (ns) are more likely to become keywords. Therefore, this paper uses part of speech (POS) to adjust the weights of keywords. Set A is a vector $\{nr, nt, ns, v\}$, t is the keyword, P_{weight} is the part of speech weight of words, p is the part of speech of candidate keywords, T is the keyword set, and P comprises the weight of the part of speech corresponding to the keywords. Adjustable variables a, b and c have general values 3, 2 and 1, respectively. When $\forall p \in A\ and\ p = nr$, $P_{weight} = TF_{new} * a$; when $p = (ns|nt)$, $P_{weight} = TF_{new} * b$; or when $p \in n \cap p \notin Aorp = v$; $P_{weight} = TF_{new} * c$, we add t to set T, and P_{weight} to set P. The final set T is the filtered keyword set, and P is the part of speech weight of corresponding words.

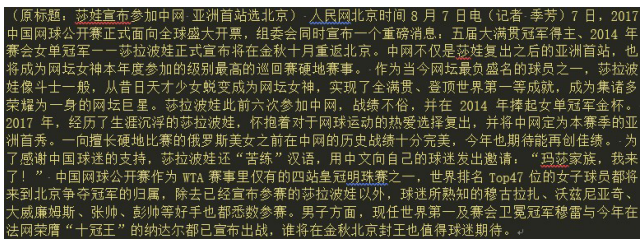
（原标题：莎娃宣布参加中网 亚洲首站选北京） 人民网北京时间 8 月 7 日电（记者 季芳）7 日，2017 中国网球公开赛正式面向全球盛大开票，组委会同时宣布一个重磅消息，五届大满贯冠军得主、2014 年赛会女单冠军——莎拉波娃正式宣布将在金秋十月重返北京。中网不仅是莎娃复出之后的亚洲首站，也将成为网坛女神本年度参加的级别最高的巡回赛硬地赛事。 作为当今网坛最负盛名的球员之一，莎拉波娃像斗士一般，从昔日天才少女蜕变成为网坛女神，实现了全满贯、登顶世界第一等成就，成为集诸多荣耀为一身的网坛巨星。莎拉波娃此前六次参加中网，战绩不俗，并在 2014 年捧起女单冠军金杯。2017 年，经历了生涯沉浮的莎拉波娃，怀抱着对于网球运动的热爱选择复出，并将中网定为本赛季的亚洲首秀。一向擅长硬地比赛的俄罗斯美女之前在中网的历史战绩十分完美，今年也期待能再创佳绩。“为了感谢中国球迷的支持，莎拉波娃还“苦练”汉语，用中文向自己的球迷发出邀请：“玛莎家族，我来了！”中国网球公开赛作为 WTA 赛事里仅有的四站皇冠明珠赛之一，世界排名 Top47 位的女子球员都将来到北京争夺冠军的归属，除去已经宣布参赛的莎拉波娃以外，球迷所熟知的穆古拉扎、沃兹尼亚奇、大威廉姆斯、张帅、彭帅等好手也都悉数参赛。男子方面，现任世界第一及赛会卫冕冠军穆雷与今年在法网荣膺“十冠王”的纳达尔都已宣布出战，谁将在金秋北京封王也值得球迷期待。

FIGURE 1. The example of a news reports.

Taking the news report shown in Figure 1 as an example, the keyword extraction is performed, and the result listed in table 1 shows that the combined feature of TF_{new} and POS is significantly better than the single TF_{new} feature. Under the influence of multi-feature keyword extraction, many irrelevant words were successfully removed.

TF_{new} and part of speech are two important features in evaluating the importance of keywords. According to the

TABLE 1. Evaluation of feature combination.

No.	TF_{new}		TF_{new}*POS(a,b,c)	
	keywords	weight	keywords	weight
1	'的'	18	'拉波娃'	18
2	'莎'	7	'北京'	10
3	'拉波娃'	6	'莎'	7
4	'宣布'	5	'亚洲'	6
5	'北京'	5	'中国'	6
6	'将'	5	'公开赛'	6
7	'在'	5	'金秋'	6
8	'冠军'	4	'宣布'	5
9	'也'	4	'冠军'	4
10	'网坛'	4	'网坛'	4

different degrees of importance of the two characteristics in the evaluation of keywords, we give the following formula:

$$w_i = k_1 TF_{new} + k_2 P_{weight} \tag{3}$$

In the above formula, k_i is an adjustable parameter, generally 0, 1, 2 or 3. In the process of keyword extraction, TF features and part of speech features were integrated into the unified multi-feature fusion model to determine the weights of the keywords. Then the multiplying parameters a/b/c are used to emphasize the possibility of different words being the keyword according to part of speech, and the integration of P_{weight} and the TF_{new} is used to balance the relative importance between the two features of TF_{new} and POS, then keyword extraction for different types of text is achieved by adjustment of a/b/c and k_i.

B. AN AI-DRIVEN BIG DATA MFE SCHEME

Applying the multi-feature keyword extraction in the previous section, an artificial intelligence driven big data MFE scheme is constructed. And the unstructured text big data mentioned in this section can come from various channels, such as the Internet, social media and mobile terminals. The main algorithm flow descriptions are shown as Algorithm 1.

The end condition of the algorithm is that no new features appear. In the specific application of MFE scheme, it can be summarized as the following general process:

Step 1: Data collection. The content of data collection covers all aspects of human activities such as scientific research, life, work and entertainment, and its forms include news, blog, BBS and microblog.

Step 2: Data cleaning. For the HTML raw text crawled by the crawler, data cleaning is required to filter out the label text. There are a lot of unnecessary information in the web page, such as advertisements, navigation bars, html code, javascript code, comments, etc. Information that we are not interested in can be deleted. If the main body extraction is required, the text can be extracted by using tag usage, tag density determination, data mining idea, visual web page block analysis technology and other strategies. Text data (usually spoken text records) may contain human emoji, such as [laughing], [Crying], [Audience paused]. These expressions are usually unrelated to what is being said and therefore need to be removed. This can be done with simple regular expressions. In addition, text data that people generate on social forums is essentially informal. Most tweets come with lots of sticky words like "RainyDay", "PlaingInTheCold" and so on. These also can be split into normal forms with simple rules and regular expressions. And all punctuation should be treated as a priority. For example, periods, commas, question marks are important punctuation that should be retained, while others need to be removed.

Algorithm 1 An AI-Driven Big Data MFE Scheme

Input: a set $D\{d_1, d_2, \ldots, d_n\}$ of texts
Output: a set $F\{f_1, f_2, \ldots, f_m\}$ of features; a set $G\{g_1, g_2, \ldots, g_m\}$ of evaluation measures
1: **For i = 1 : n do**
2: Extract keywords form d_i using MFE
3: **End for**
4: **For j = 1 : m do**
5: Compute the feature data f_j
6: Cluster the texts D with single pass algorithm in a particular order associated with feature f_j
7: Compute evaluation measures $\mathbf{g_j}\{v_1, v_2, v_3, v_4\}$
8: **End for**
9: **return** $\mathbf{G}\{g_1, g_2, \ldots, g_m\}$

Step 3: Data preparation. Data preparation can be divided into three parts: word segmentation, part-of-speech tagging, and text vector calculation. The data interaction between each step process is conducted through some data record files, so as to avoid the memory overflow error caused by adding all process data into memory at one time. For Chinese text data, such as a Chinese sentence, the words are continuous, and the minimum unit granularity of data analysis we want is words, so we need to work on word segmentation, so that we can prepare for the next step. The purpose of part-of-speech tagging is to allow the sentence to incorporate more useful language information into subsequent processing. Text vector calculation is prepared for the subsequent calculation of word weight and the screening of keywords.

Step 4: Algorithm using. For a single feature, the algorithm arranges texts in descending order according to the feature values, and prioritizes text with significant features. Texts with a large amount of reading, more comments, or faster commentary speeds can better represent the hot spots of public concern, that is, where social hot spots are. After all the features are traversed, the clustering results are evaluated according to the common evaluation indicators of text analysis, such as recall and precision, and the optimal results and their corresponding features can be obtained.

Then the following section expands the example.

C. SPECIAL HOT EVENT DETECTION SCHEME

As we mentioned earlier, this paper uses keyword notation as the basis for the algorithm. By calculating the similarity between a news report and all existing events, we can obtain the maximum value. If the max is larger than a given Similarity Threshold (ST), we assume that this news could be classified into the corresponding event. However, if it is not, we would create a new event based on this news in the range of the Proportion of Prepositive Clustering (POPC) or abandon it out of range. Afterwards, the weights of the keywords on the event must be handled carefully. Then, in descending order, according to the number of comments attached to the news, the algorithm could handle massive online news effectively. Generally, the algorithm flow is specified in Algorithm 2.

Algorithm 2 Hot Event Detection

Input: a set $S\{s_1, s_2, \ldots, s_n\}$ of news reports
Output:a set $E\{e_1, e_2, \ldots, e_m\}$ of events
For each news report s_i in S by a particular order
  If $i == 1$ **then**
    Set $e_1 = s_1$
    Add e_1 to E
  Else if $i < POPC * size(S)$ **then**
    If $sim(s_i, e_k) > ST$ **then**
      Add s_i to e_k
      Recalculate the weight of e_k
    Else
      Create a new event e_c
      Add e_c to E
    End if
  Else if $sim(s_i, e_k) > ST$ **then**
      Add s_i to e_k
      Recalculate the weight of e_k
  Else
      Abandon the news report s_i
  End if
End for
Return E

The detailed explanation is as follows.

Step 1: Collect news reports by crawling the news portals with a Web crawler. In addition, extract keywords from every piece of news and simultaneously set the Number of Keywords (NOK). Then, news will be represented as a fixed-length list of keywords from the content of the news itself.

Step 2: Sort all news in descending order according to the number of comments attached to the news.

Step 3: Set the first piece of news with the most comments as the initial event and the news' weights as that of the event.

Step 4: Fetch a news report from the news corpus in order and calculate the maximum similarity between the news and all the events based on the list of keywords. In this paper, we compare the news obtained with the first event and obtain a similarity value. However, we are not sure whether it is the maximum or not, and we then calculate another similarity value between the news and the next event so that we can record the larger one and its corresponding event. With that methodology, the maximum similarity can be calculated.

Step 5: If the maximum similarity is larger than the given similarity threshold, classify this news to the corresponding event and add the weights of same keywords representing news to the corresponding weights of the event. Then, divide all the keyword weights of the event by the number of news items belonging to this event, and the newly calculated event keywords are obtained. After that, if the weight of the news keywords is larger than that of the event, the algorithm will replace the smaller weight and its keyword with the larger pair. Besides, we assume that news with a larger number of comments is most likely to be an event. And In the initial cluster, the news with the most comments firstly constitute the initial event set, so for subsequent news, which key words haven't appeared in the key words of events, the algorithm will directly remove it.

Step 6: If the maximum similarity is not larger than the given similarity threshold, a new event for the news is created, and the keywords and weights of the news are regarded as the event's, with the news report in the range of POPC. In addition, the news report will be abandoned if it is out of range.

Step 7: Repeat the processing steps from step 4 to step 6 until all news is handled.

D. THREE PARAMETERS

There are three impact factors in our algorithm that may influence the result of news clustering, including the number of keywords, the similarity threshold and the proportion of prepositive clustering.

One of the key issues in this algorithm is the similarity calculation, which is based on keywords representing news. According to the formula calculating similarity, we assume that the increment or decrement to the number of keywords directly changes the result of the similarity calculation, thereby possibly impacting the maximum similarity.

To handle this, a simple comparison of the max to the similarity threshold is performed to determine whether the news is classified to an existing event stored in the event library or not. It will be hard to classify the news to a certain event if the similarity threshold is too high. Thus, the similarity threshold will have a significant impact on the result of news clustering.

Through the initial local clustering, the algorithm first automatically generates a hot event library in which the global number of hot events and their contents are determined by the proportion of prepositive clustering. If POPC has a higher proportion, more news will be clustered with more events extracted. In this paper, the proportion of prepositive clustering is also explored.

V. EXPERIMENTS AND ANALYSIS

In this section, we evaluate our proposed approach to learn the event mining. We use Excel to plot figures and Python

TABLE 2. The standard events.

Events	The Number of News
Court to Freeze Assets of LeEco Founder Jia Yueting.	504
Baoding Rongda threatens to quit Chinese league after controversial draw.	110
Logistics War between Jingdong and Suning.	37
Chinese teacher from Fujian traveling in Japan has gone missing.	32
Zou Shiming loses the WBO flyweight title as he's stunned in the 11th round by Japan's Sho Kimura.	54
Earthquake strikes China's remote Sichuan province.	258
Death of university graduate sparks anger at Chinese pyramid scam gangs.	129
Actor Xu Zheng was exposed and wounded female.	10
The resignation of Anti-GMO activist Cui Yongyuan.	8
Wolf Warrior 2 is the highest grossing movie in the world, beating out Hollywood blockbusters and epic European sci-fi Valerian.	103

to implement all algorithms. The compared approaches are respectively classical clustering algorithms and the multi-feature model (ours).

A. DATA PREPARATION

Our corpus is built from part of NetEase News, which contains 19681 news articles from July 1 through August 9 of 2017. Each piece of news is marked as on-event or off-event for every one of the events extracted artificially. We defined a piece of news as on-event if it is related to one of the standard events and off-event if it is not. Taken absolutely, an on-event news article belongs only to a certain event, not to two or more. table 2 lists some statistics about all the standard events and shows the number of news belonging to each event.

B. EVALUATION MEASURES

Methods considering relevance were used. Then, the evaluation indices system for testing the efficiency of clustering news were established, and its four components were the generation rate, precision, recall and F1-score.

As table 2 shows, the sum of events in the standard event set is 10, which is marked as a constant variable n. We let the set of n standard events E_n be $\{E_1, \ldots, E_n\}$ and the set of events detected automatically be $T_r = \{T_1, \ldots, T_r\}$, of which each T consists of keywords $\{t_1, \ldots, t_k\}$ as well as each $E = \{e_1, \ldots, e_k\}$ and k is below limit ten. If there is an event E_i that has a percentage of the same keywords between E_i and T_j greater than 30%, we define $E_i \in E_n \cap T_r$. Then, the generation rate could be expressed as follows:

$$generation - rate = \frac{|E_n \cap T_r|}{n} \tag{4}$$

where it reaches its best value at 1 and its worst value at 0.

In this paper, we define precision and recall based on the standard events and clustering events detected by the method we propose. That is,

$$precision = \frac{\sum_{i=1}^{n} \frac{|E_i \cap T_k|}{|E_i|}}{n} \tag{5}$$

and

$$recall = \frac{\sum_{i=1}^{n} \frac{|E_i \cap T_k|}{|T_k|}}{n} \tag{6}$$

where if an event T exists in the set of events T_r and it makes $E_i \cap (\forall T \in T_r)$ take the maximum value, we regard it as T_k. Intuitively, precision is the ratio of events that are clustered exactly to the standard events, and recall is the ratio of events clustered correctly to all the events detected.

To combine the two indicators mentioned above, we take the F_1-score into account. In the equation below, the precision and recall are weighted equally.

$$F_1 - score = \frac{2 * precision * recall}{prcision + recall} \tag{7}$$

Thus, the closer the F1-score is to 1, the higher the clustering quality.

C. EXPERIMENTAL DESIGN

Two sets of experiments were performed in our study. The first experiment investigated the measures for evaluating our approach with the aforementioned dataset. In this experiment, there were a total of 224 parallel tests, and among them, differentiators depended on three variable parameters, including the number of keywords, similarity threshold and proportion of prepositive clustering. We set the number of keywords representing a news article from 4 to 18 with the specific interval 2, the similarity threshold from 0.1 to 0.7 with the specific interval 0.1, and the proportions of prepositive clustering at 1%, 3%, 5% and 7%, yielding the 224(8*7*4) pairs of different combinations. For example, in one test, we clustered the news based on the number of keywords as 6, a similarity threshold of 0.3 and the proportion of prepositive clustering at 1%. In the second experiment, we examined the performance of the classical clustering algorithms on the same set of data.

To compare our approach, some other clustering algorithms were chosen as the baseline. In k-means algorithm, news articles are placed into k partitions, and the initialization methods Forgy and Random Partition [60] are used; the Forgy method randomly chooses k observations from the corpus and uses these as the initial means, while the Random Partition method first randomly assigns a cluster to each observation and proceeds to the update step and computing the initial mean to be the centroid of the cluster's randomly assigned points until it is improved. We chose initial centers in a way that gives a provable upper bound on the WCSS object, and the filtering algorithm used kd-trees to speed up each k-means step [31]. Besides, mini batch k-means, dbscan clustering, birch clustering, spectral clustering, agglomerative clustering, mean shift clustering and affinity propagation clustering are also on the list.

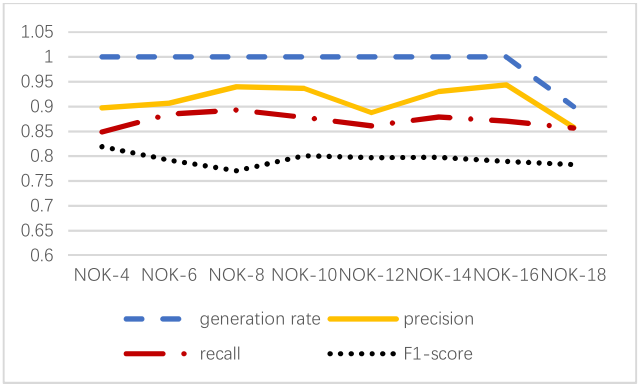

FIGURE 2. The tradeoff between different representations of news in terms of generation rates, precision, recall and F1-score.

D. EXPERIMENTAL RESULTS AND ANALYSIS

As can be seen from Figures 2, the linear correlations of the data are similar in shape. The data using ten keyword representations appear balanced and a little superior to the data using other keyword representations. This phenomenon reflects the small amount of available information, which means that using ten keyword representations could lead to better clustering results.

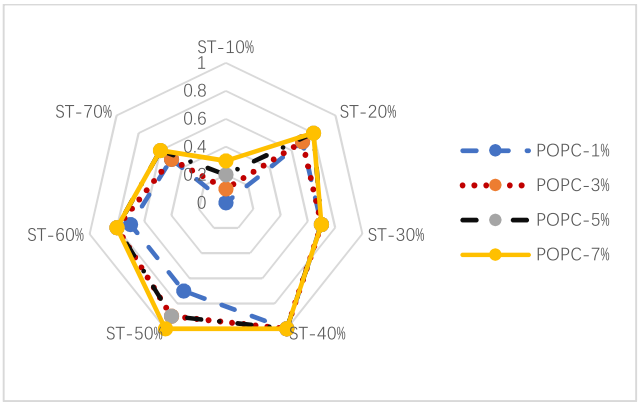

FIGURE 3. The tradeoff between POPC and ST in terms of generation rates, where NOK is ten.

Figures 3-6 show the effectiveness of the approach proposed in this paper with various parameters, comparing the generation rate, precision, recall and F1-score. In Figures 3-6, the more significantly the line extends outwards, the better the clustering results. The graphs in these figures lead to two preliminary conclusions.

1) For the proportion of prepositive clustering, we found that the highest percentage 7% outperforms any other one. Furthermore, the combinations of POPC 7% and ST 40% had greater influence on the experimental results (that is, evaluation indicators reached larger values).

In table 3, events keywords extracted by the algorithm 2 are listed in details and the listing order corresponds to the standard events.

2) In the direction of the similarity threshold, an inverse relationship between precision and recall is evidenced, where it is possible to increase one at the cost of reducing the other.

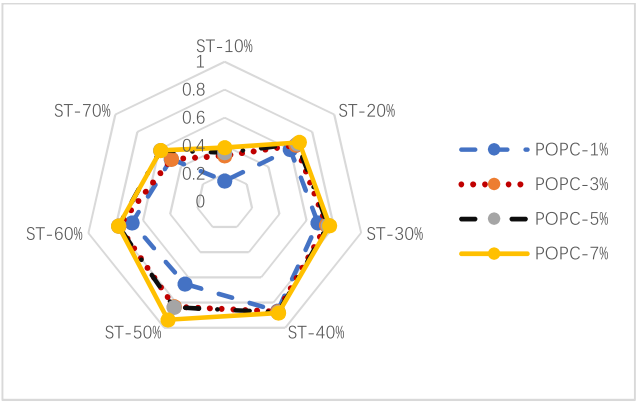

FIGURE 4. The tradeoff between POPC and ST in terms of precision, where NOK is ten.

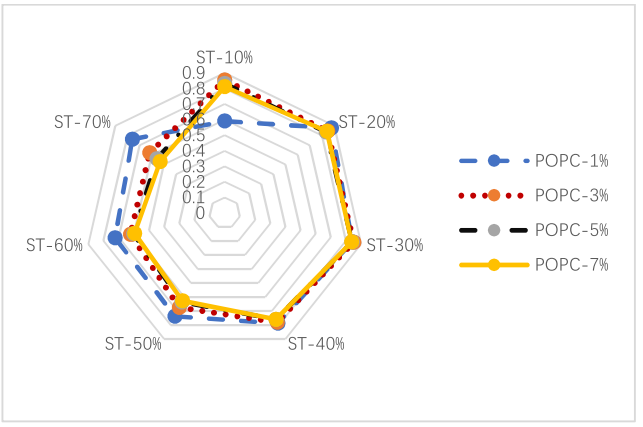

FIGURE 5. The tradeoff between POPC and ST in terms of recall, where NOK is ten.

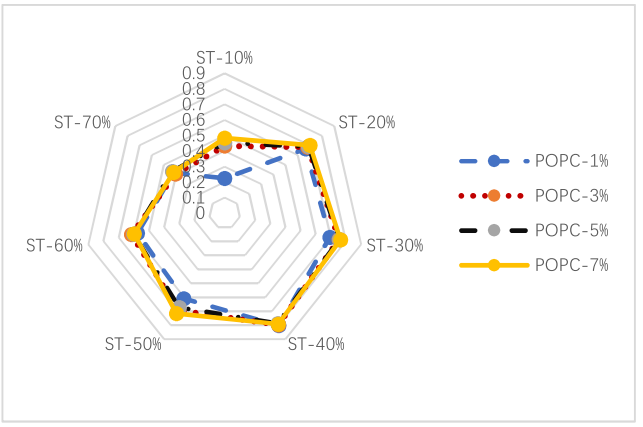

FIGURE 6. The tradeoff between POPC and ST in terms of F1-score, where NOK is ten.

the similarity threshold. Table 4 shows the results when POPC is 7% and NOK is 10. Usually, precision and recall are not dissected and discussed in isolation. Instead, the measure that is a combination of precision and recall is the F1-score (the weighted harmonic mean of precision and recall) [61], which is more useful for reference and often used in the field of information retrieval for query classification performance.

We will further explore the approach to verify whether the performance improved from this approach by setting up other

TABLE 3. Events keywords in experiment with NOK 10, POPC 7% and ST 40%.

Events No.	Event keywords
1	Letv, Jia Yueting, corporation, freeze, holding, justice, supplier, media, financial institution
2	club, football, soccer fans, competition, event, Baoding, quit, punishment, hope
3	expressage, China, physical distribution, Jingdong, Suning, e-commerce, Tonglu County, service, speed
4	Wei Qiujie, Japan, loss of communication, discover, journey, journalist, hotel, Fujian, teacher
5	Zou Shiming, Kimura, opponent, bout, boxing, offensive, world, occupation, stamina
6	earthquake, photograph, net friend, China, happen, aba prefecture, Jiuzhaigou county, Sichuan Province
7	pyramid scheme, activity, crime, Ministry of Public Security, public security organ, Shan Xin Hui, pursuant to the law, mastermind, organization, safeguard
8	female, reconciliation, Xu Zheng, deny, expose, exclusive, video, injury, actor
9	Cui Yongyuan, store, food, transgenosis, position, interest group, offend, statement, resign
10	passport, Wolf Warrior, movie, Wang Cailiang, box office, China, market, corporation, theme

TABLE 4. Experimental results for POPC 7% and NOK 10.

	precision	recall
ST-10%	0.385612102	0.812311218
ST-20%	0.680086392	0.840866789
ST-30%	0.768732519	0.837736526
ST-40%	0.883824456	0.759592619
ST-50%	0.936733762	0.626990142

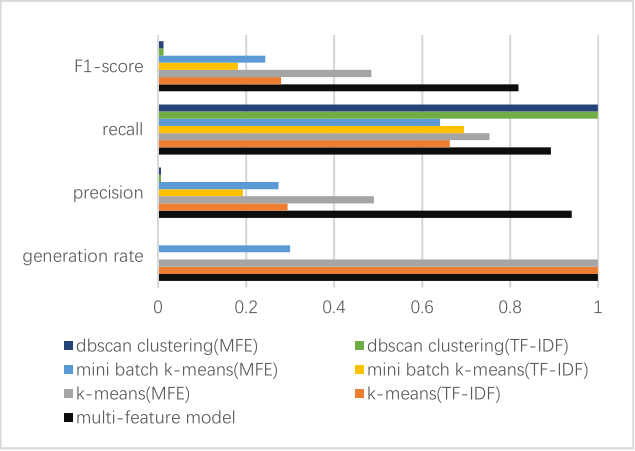

FIGURE 7. Experimental results of the three approaches.

experiments for comparison. In the experiment, we regard the keyword vector extracted by the general feature of TF-IDF and MFE schema as the inputs for clustering algorithms baselines. The comparison results between the clustering algorithms and the multi-feature model are shown in Figure 7. Unfortunately, there are five algorithms (including birch clustering, spectral clustering, agglomerative clustering, mean shift clustering and affinity propagation clustering) failing to cluster with same hardware, and the reason for the failure is memory not enough. It means these algorithms are memory intensive compared to our approach.

There are significant differences among the above seven groups of experiments, and our approach obviously achieves the optimized result as well as the black part (see Figure 7). The results demonstrate that our approach offers good performance and MFE schema is superior to TF-IDF as input.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a multi-feature keyword extraction method, based on which we designed an artificial intelligence driven big data MFE scheme and expanded an application example of this universal scheme. We have discussed the task of clustering algorithms with application to news overloading on the Internet. Through the experiments we designed, we conducted a detailed empirical study on the feasibility and validity of our approach. According to the analysis of the experimental results, we can obtain a better event generation rate, precision, recall and F_1-score when using the specified POPC, ST and NOK, which provides an effective clustering method for detecting hot events from the massive amount of online news. Thus, for practical applications, the approach provides a reference value. However, this method also has the following shortcomings. The threshold set in this paper is a fixed value, while the news flow is dynamic data. If the threshold can be dynamically adjusted according to the event detection, the detection effect of the hot event can be improved. Moreover, the source of hot events is limited to news only. However, in daily life, many hot events may first appear in microblogging, forums and so on. If we want to provide users with more comprehensive social hot events, the source of information should not be confined to the news, but should be integrated with multiple data sources. We will solve these problems in future work and study the feasibility of applying this method to knowledge discovery.

REFERENCES

[1] J. He and N. Xiong, "An effective information detection method for social big data," in *Multimedia Tools Appl.*, vol. 77, no. 9, pp. 11277–11305, 2018.
[2] H. Si, Z. Chen, W. Zhang, J. Wan, J. Zhang, and N. N. Xiong, "A member recognition approach for specific organizations based on relationships among users in social networking Twitter," *Future Gener. Comput. Syst.*, vol. 92, pp. 1009–1020, Mar. 2019.
[3] P. Zhong, Y.-T. Li, W.-R. Liu, G.-H. Duan, Y.-W. Chen, and N. Xiong, "Joint mobile data collection and wireless energy transfer in wireless rechargeable sensor networks," *Sensors*, vol. 17, no. 8, p. 1881, 2017.
[4] Y. Xu, H. Yang, J. Li, J. Liu, and N. Xiong, "An effective dictionary learning algorithm based on FMRI data for mobile medical disease analysis," *IEEE Access*, vol. 7, pp. 3958–3966, 2019.
[5] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the Future, 2007, pp. 1–16. [Online]. Available: https://www.emc.com/leadership/digital-universe/2012iview/big-data-2020.htm
[6] Y. Yang, T. Pierce, and J. G. Carbonell, "A study on retrospective and on-line event detection," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1998, pp. 28–36.
[7] R. J. Mooney and R. Bunescu, "Mining knowledge from text using information extraction," *ACM SIGKDD Explor. Newslett.*, vol. 7, no. 1, pp. 3–10, 2005.

[8] G. Chrupała and D. Klakow, "A named entity labeler for German: Exploiting Wikipedia and distributional clusters," in *Proc. Conf. Int. Lang. Resour. Eval. (LREC)*, May 2010, pp. 552–556.

[9] V. Chakravarthy, H. Gupta, M. K. Mohania, and P. Roy, "Automatically linking documents with relevant structured information," U.S. Patent 7 899 822, Mar. 1, 2011.

[10] Z. Nie, J. R. Wen, and L. Yang, "Web-scale entity summarization," U.S. Patent 8 229 960, Jul. 24, 2012.

[11] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt, "Automatic ontology-based knowledge extraction from Web documents," *IEEE Intell. Syst.*, vol. 18, no. 1, pp. 14–21, Jan. 2003.

[12] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *Proc. 16th Int. Conf. World Wide Web*, May 2007, pp. 697–706.

[13] E. Agichtein and L. Gravano, "Querying text databases for efficient information extraction," in *Proc. 19th Int. Conf. Data Eng.*, Mar. 2003, pp. 113–124.

[14] V. Gkatziaki, S. Papadopoulos, R. Mills, S. Diplaris, and I. K. I. Tsampoulatidis, "easIE: Easy-to-use information extraction for constructing CSR databases from the Web," *ACM Trans. Internet Technol.*, vol. 18, no. 4, p. 45, 2018.

[15] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, "UIMA Ruta: Rapid development of rule-based information extraction applications," *Natural Lang. Eng.*, vol. 22, no. 1, pp. 1–40, 2016.

[16] W. W. Yim, T. Denman, S. W. Kwan, and M. Yetisgen, "Tumor information extraction in radiology reports for hepatocellular carcinoma patients," in *Proc. AMIA Summits Transl. Sci.*, 2016, p. 455.

[17] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, and R. M. Weischedel, "The automatic content extraction (ACE) program-tasks, data, and evaluation," in *Proc. LREC*, vol. 2, May 2004, p. 1.

[18] H. P. Luhn, "Auto-encoding of documents for information retrieval systems," IBM Res. Center, York, U.K., Tech. Rep. NBS # NBS00778, 1958.

[19] H. Ji, R. Grishman, Z. Chen, and P. Gupta, "Cross-document event extraction and tracking: Task, evaluation, techniques and challenges," in *Proc. Int. Conf. RANLP*, 2009, pp. 166–172.

[20] A. L. Garrido, M. G. Buey, S. Escudero, S. Ilarri, E. Mena, and S. B. Silveira, "TM-gen: A topic map generator from text documents," in *Proc. IEEE 25th Int. Conf. Tools Artif. Intell.*, Nov. 2013, pp. 735–740.

[21] Y. Shinyama, S. Sekine, and K. Sudo, "Automatic paraphrase acquisition from news articles," in *Proc. 2nd Int. Conf. Hum. Lang. Technol. Res.* Morgan Kaufmann, Mar. 2002, pp. 313–318.

[22] C. I. Wi, S. Sohn, M. C. Rolfes, A. Seabright, E. Ryu, G. Voge, K. A. Bachman, M. A. Park, H. Kita, I. T. Croghan, and H. Liu, "Application of a natural language processing algorithm to asthma ascertainment. An automated chart review," *Amer. J. Respiratory Crit. Care Med.*, vol. 196, no. 4, pp. 430–437, 2017.

[23] N. Afzal, S. Sohn, S. Abram, C. G. Scott, R. Chaudhry, H. Liu, I. J. Kullo, and A. M. Arruda-Olson, "Mining peripheral arterial disease cases from narrative clinical notes using natural language processing," *J. Vascular Surg.*, vol. 65, no. 6, pp. 1753–1761, 2017.

[24] S. Yazdani, S. Fallet, and J. M. Vesin, "A novel short-term event extraction algorithm for biomedical signals,"IEEE Transactions on Biomedical Engineering, vol. 65, no. 4, pp. 754–762, 2018.

[25] C. D. Manning, C. D. Manning, and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[26] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proc. 12th Int. Conf. World Wide Web*, May 2003, pp. 519–528.

[27] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report," in *Proc. DARPA Broadcast News Transcription Understand. Workshop*, 1998, pp. 194–218.

[28] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2003, pp. 330–337.

[29] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2004, pp. 297–304.

[30] C. Wang, M. Zhang, L. Ru, and S. Ma, "An automatic online news topic keyphrase extraction system," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, vol. 1, Dec. 2008, pp. 214–219.

[31] N. Stokes and J. Carthy, "Combining semantic and syntactic document classifiers to improve first story detection," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Sep. 2001, pp. 424–425.

[32] D. Trieschnigg and W. Kraaij, "Hierarchical topic detection in large digital news archives," in *Proc. 5th Dutch Belgian Inf. Retr. Workshop*, Jan. 2005, pp. 55–62.

[33] Z. Li, B. Wang, M. Li, and W. Y. Ma, "A probabilistic model for retrospective news event detection," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2005, pp. 106–113.

[34] K. Zhang, J. Zi, and L. G. Wu, "New event detection based on indexing-tree and named entity," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 215–222.

[35] C. C. Chen, Y. T. Chen, Y. Sun, and M. C. Chen, "Life cycle modeling of news events using aging theory," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, Sep. 2003, pp. 47–59.

[36] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," *Knowl.-Based Syst.*, vol. 117, pp. 56–69, Feb. 2017.

[37] X. Cui, P. Zhu, X. Yang, K. Li, and C. Ji, "Optimized big data K-means clustering using MapReduce," *J. Supercomput.*, vol. 70, no. 3, pp. 1249–1259, 2014.

[38] A. George, "Efficient high dimension data clustering using constraint-partitioning k-means algorithm," *Int. Arab J. Inf. Technol.*, vol. 10, no. 5, pp. 467–476, 2013.

[39] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[40] M. S. G. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. KDD Workshop Text Mining*, vol. 400, no. 1, Aug. 2000, pp. 525–526.

[41] A. D. Amensisa, S. Patil, and P. Agrawal, "A survey on text document categorization using enhanced sentence vector space model and bi-gram text representation model based on novel fusion techniques," in *Proc. 2nd Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2018, pp. 218–225.

[42] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[43] A. L. Garrido, O. Gomez, S. Ilarri, and E. Mena, "NASS: News annotation semantic system," in *Proc. IEEE 23rd Int. Conf. Tools Artif. Intell.*, Nov. 2011, pp. 904–905.

[44] A. L. Garrido, M. G. Buey, S. Ilarri, and E. Mena, "GEO-NASS: A semantic tagging experience from geographical data on the media," in *Proc. East Eur. Conf. Adv. Databases Inf. Syst.* Berlin, Germany: Springer, Sep. 2013, pp. 56–69.

[45] A. L. Garrido, M. G. Buey, S. Escudero, A. Peiro, S. Ilarri, and E. Mena, "The GENIE project—A semantic pipeline for automatic document categorisation," in *Proc. WEBIST*, vol. 2, 2014, pp. 161–171.

[46] J. R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson,, "FASTUS: A cascaded finite-state transducer for extracting information from natural-language text," in *Finite-State Language Processing*. Cambridge, MA, USA: MIT Press, 1997, p. 482.

[47] A. McCallum, "Information extraction: Distilling structured data from unstructured text," *Queue Social Comput.*, vol. 3, no. 9, p. 4, 2005.

[48] H. Nanba, R. Saito, A. Ishino, and T. Takezawa, "Automatic extraction of event information from newspaper articles and Web pages," in *Proc. Int. Conf. Asian Digit. Libraries*, Cham, Switzerland: Springer, Dec. 2013, pp. 171–175.

[49] E. Lefever and V. Hoste, "A classification-based approach to economic event detection in dutch news text," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 330–335.

[50] G. Jacobs, E. Lefever, and V. Hoste, "Economic event detection in company-specific news text," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–10.

[51] Z. Yang, Q. Li, L. Wenyin, and J. Lv, "Shared multi-view data representation for multi-domain event detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[52] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An overview of graph-based keyword extraction methods and approaches," *J. Inf. Organizational Sci.*, vol. 39, no. 1, pp. 1–20, 2015.

[53] C. Zhang, "Automatic keyword extraction from documents using conditional random fields," *J. Comput. Inf. Syst.*, vol. 4, no. 3, pp. 1169–1180, 2008.

[54] H. Zhao and Q. Zeng, "Micro-blog keyword extraction method based on graph model and semantic space," *J. Multimedia*, vol. 8, no. 5, pp. 611–618, 2013.

[55] H. Hromic, N. Prangnawarat, I. Hulpuş, M. Karnstedt, and C. Hayes, "Graph-based methods for clustering topics of interest in Twitter," in *Proc. Int. Conf. Web Eng.*, Cham, Switzerland: Springer, Jun. 2015, pp. 701–704.
[56] L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. M. De Matos, J. Neto, and J. Carbonell, "Automatic keyword extraction on Twitter," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2015, pp. 637–643.
[57] D. Kim, D. Kim, S. Rho, and E. J. Hwang, "Detecting trend and bursty keywords using characteristics of Twitter stream data," *Int. J. Smart Home*, vol. 7, no. 1, pp. 209–220, 2013.
[58] M. Zimniewicz, K. Kurowski, and J. Węglarz, "Scheduling aspects in keyword extraction problem," *Int. Trans. Oper. Res.*, vol. 25, no. 2, pp. 507–522, 2018.
[59] S. Duari and V. Bhatnagar, "sCAKE: Semantic connectivity aware keyword extraction," *Inf. Sci.*, vol. 477, pp. 100–117, Mar. 2019.
[60] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings," in *Proc. 11th Int. Conf. Inf. Knowl. Manage.*, Nov. 2002, pp. 600–607.
[61] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

JIAN WAN received the Ph.D. degree in computer technology from Zhejiang University. He serves as a Professor with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, China. He is also currently a Professor with Hangzhou Dianzi University. His research interests include virtual computing, grid computing, service computing, and embedded systems. He is a member of the Association for Computing Machinery (ACM) and the China Computer Federation (CCF).

PIAOPIAO ZHENG is currently pursuing the M.S. degree with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, China. Her research interests include data mining, machine learning, and natural language processing.

HUAYOU SI received the M.S. and Ph.D. degrees in computer science from Peking University, in 2004 and 2012, respectively. He is a Lecturer with the School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include P2P network, service-oriented computing, and the semantic Web. He has published more than 20 academic papers in these areas. In addition, he has served on the technical program committee of several international conferences.

NEAL N. XIONG received the first Ph.D. degree in sensor system engineering from Wuhan University and the second Ph.D. degree in dependable sensor networks with the Japan Advanced Institute of Science and Technology. He is currently an Associate Professor (3rd year) with the Department of Mathematics and Computer Science, Northeastern State University, OK, USA. Before he went to Northeastern State University, he worked at Georgia State University, the Wentworth Technology Institution, and Colorado Technical University (a Full Professor approximately 5 years) for 10 years. His research interests include cloud computing, security and dependability, parallel and distributed computing, networks, and optimization theory.

He has published over 280 international journal papers and over 120 international conference papers. Some of his works were published in the IEEE JSAC, IEEE, or ACM Transactions, ACM Sigcomm Workshop, IEEE INFOCOM, ICDCS, and IPDPS. He has been the General Chair, the Program Chair, the Publicity Chair, a PC member, and an OC member of over 100 international conferences and a Reviewer of approximately 100 international journals, including the IEEE JSAC, IEEE SMC (Park: A/B/C), the IEEE Transactions on Communications, the IEEE Transactions on Mobile Computing, and the IEEE Transactions on Parallel and Distributed Systems. He is serving as the Editor-in-Chief, an Associate Editor, or an Editor Member for over 10 international journals (including as an Associate Editor for the IEEE Transaction on Systems, Man, and Cybernetics: Systems, an Associate Editor for *Information Science*, the Editor-in-Chief for the *Journal of Internet Technology*, and the Editor-in-Chief for the *Journal of Parallel and Cloud Computing*) and a Guest Editor for over ten international journals, including *Sensor Journal*, WINET, and MONET. He was a recipient of the Best Paper Award in the 10th IEEE International Conference on High Performance Computing and Communications (HPCC-08) and the Best Student Paper Award at the 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009).

Dr. Xiong is the Chair of "Trusted Cloud Computing" Task Force, IEEE Computational Intelligence Society (CIS), and the Industry System Applications Technical Committee. He is a senior member of the IEEE Computer Society.

WEI ZHANG received the B.E. degree from the School of Information Science and Engineering, Wuhan University of Science and Technology, China, in 2000, and the M.Ec. and Ph.D. degrees from the Computer School, Wuhan University, China, in 2004 and 2008, respectively. He is currently a Lecturer with the School of Computer Science and Technology, Hangzhou Dianzi University, China. His research interests include wireless sensor networks and intelligent computing. He is a member of the Association for Computing Machinery (ACM) and China Computer Federation (CCF).

ATHANASIOS V. VASILAKOS is currently a Professor with the Department of Computer and Telecommunications Engineering, University of Western Macedonia, Greece, and a Visiting Professor with the Graduate Program, Department of Electrical and Computer Engineering, National Technical University of Athens (NTUA). He has coauthored (with W. Pedrycz) the books *Computational Intelligence in Telecommunications Networks* (CRC Press, USA, 2001), *Ambient Intelligence, Wireless Networking, Ubiquitous Computing* (Artech House, USA, 2006); (with M. Parashar, S. Karnouskos, and W. Pedrycz) *Autonomic Communications* (Springer), and *Arts and Technologies* (MIT Press); and (with M. Anastasopoulos) *Game Theory in Communication Systems* (IGI, Inc., USA). He has published more than 150 articles in top international journals and conferences. He is the Editor-in-Chief of the *International Journal of Adaptive and Autonomous Communications Systems* (Inderscience Publishers), the *International Journal of Arts and Technology*. He has been on the Editorial Board of more than 20 international journals, including the *IEEE Communications Magazine* from 1999 to 2002 and since 2008, the IEEE Transactions on Systems, Man, and Cybernetics–Part b: Cybernetics, since 2007 (SMC), the IEEE Transactions on Wireless Communications (invited), and *ACM Transactions on Autonomous and Adaptive Systems* (invited). He is the Chairman of the Telecommunications Task Force of the Intelligent Systems Applications Technical Committee (ISATC) of the IEEE Computational Intelligence Society (CIS). He is the Senior Deputy Secretary-General and a Fellow of the International Society of Intelligent Biological Medicine. He is a member of ACM.