IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INNOVATION AND APPLICATION OF INTELLIGENT PROCESSING OF DATA, INFORMATION AND KNOWLEDGE AS RESOURCES IN EDGE COMPUTING

Received October 6, 2019, accepted October 28, 2019, date of publication November 1, 2019, date of current version November 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950900

Orientation Analysis for Chinese News Based on Word Embedding and Syntax Rules

PENGWEI WANG[1], YUJUN LUO[2], ZHEN CHEN[1], LIANGHUA HE[2], AND ZHAOHUI ZHANG[1]
[1]School of Computer Science and Technology, Donghua University, Shanghai 201620, China
[2]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Corresponding author: Pengwei Wang (wangpengwei@dhu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities, in part by the National Natural Science Foundation of China (NSFC) under Grant 61602109, in part by the Natural Science Foundation of Shanghai under Grant 19ZR1401900, in part by the Shanghai Science and Technology Innovation Action Plan under Grant 19511101802, in part by the Shanghai Sailing Program under Grant 16YF1400300, and in part by the Donghua University (DHU) Distinguished Young Professor Program under Grant LZB2019003.

ABSTRACT In this era of Internet and big data, there is billions of news generated every day, and the traditional manual methods are insufficient for public opinion orientation analysis. Especially for Chinese, which has more complicated syntax and semantic structure, and there is no space between words as separator. This greatly increases the difficulty of analyzing opinion orientation. In this paper, a novel approach is proposed aiming at solving the problem of public opinion orientation analysis based on Chinese news. The approach combines word2vec, sentiment dictionaries and syntax rules, where the word2vec can map words into different vectors with finite dimensions. Through it we can calculate the cosine similarity between the words and sentiment dictionaries to get the orientation value of target words, which is helpful for calculating the orientation value of key sentences and full text. Specifically, the process consists of three steps. First, word2vec is used to train word embedding, and every word in corpus is mapped into a given vector space. Then, key sentences are extracted from news content. Finally, pre-defined syntax rules with word vector similarity are used to analyze document orientation based on key sentences. Several experiments are conducted on both closed and open datasets, and the results validate the effectiveness of the proposed approach.

INDEX TERMS Chinese news, orientation analysis, syntax rule, sentiment dictionary, key sentence.

I. INTRODUCTION

Online media such as microblog, portals, forums, mainstream news organizations, etc., are trying to publish various news at first time and the number of news is growing dramatically every day. A piece of news often conveys the author's opinions, and contains event's orientations during the process of its generating and disseminating. These positive or negative news orientations may affect the tendency of public opinions and the views of people. Orientation analysis is of high value in the area of monitoring Internet public opinion. The application background of this paper is monitoring the enterprise operation in Shanghai pilot free-trade zone (FTZ) on the Internet. Before it, we have done some relevant studies on the relationship extraction [1] and hot events clustering [2] for FTZ. Although these studies have achieved some good results, but there are still some deficiencies in public orientation opinion analysis. For regulators, it can help them to know the feedbacks of public events and supervise the development trend of enterprises under their jurisdiction. The new era of big data brings a great challenge to the original public opinion analysis. Currently, mainstream search engines only support to search news by keywords, and the search results do not have any further information such as classifications, orientations or opinions. Therefore, how to extract more valuable and correct information from large number of news by an effective way is a big challenge under big data surroundings.

The essence of public opinion orientation analysis is document-level sentiment analysis. As we all know, Chinese and English are the two most widely used languages in the world. The study of text orientation analysis in English context has been carried out for many years and has

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

achieved some good results. Compared to English language, the research on Chinese started late, and Chinese language has distinct linguistic characteristics which brings several unique challenges in text analysis: 1) Chinese does not segment words by spaces in sentences, so more complex preprocessing work is required; 2) Chinese shows a larger variety of word sense and syntactic dependency in sentences than English, and we can use these characteristics to help analyze the sentiment of words and sentences; 3) The length of news text is usually relatively long, and it is difficult to grasp the relationship between part and whole of context information in the process of analysis. The existing methods for text orientation analysis can be divided into two main categories. In the early stage, it is mainly based on rules, sentiment dictionaries and statistics tools to make the orientation analysis. Some early attempts of sentiment analysis predicted the polarity of a news document based on natural language processing characteristics. For example, Kim and Hovy [3] measured text sentiment polarity by using sentiment dictionaries to calculate weighted sum of words sentiment score directly, which will result in the calculation of full-text orientation values being affected by the noisy words in sentences. With the development of artificial intelligence and deep learning, some machine learning techniques have been widely adopted to solve the problem of sentiment orientation analysis due to their excellent performance. Some proposed supervised learning approaches can construct models by learning from labeled training datasets. For example, Pang *et al.* [4] first applied three machine-learning methods to predict sentiment of reviews and the results outperformed human-produced baselines. Supervised learning approaches require large training datasets with labels, and the process of labeling is very costly and time-consuming. As for the unsupervised learning approach, it requires a lot of predefined linguistic patterns as templates. However, the templates cannot cover all aspects. Moreover, most of these existing studies on sentiment analysis focus on articles written in English. The key contribution of this study is to try to propose a suitable approach for public opinion orientation analysis in Chinese. It proposed 1) a method of key sentence extraction to suppress the noise in long documents and 2) a novel approach which combines supervised-based and unsupervised-based methods to analyze the orientation in news. The approach consists of three main steps: word2vec training, key sentences extraction and orientation analysis. Experiments evaluate the effectiveness of the proposed method in the field of Chinese public opinion orientation analysis.

The rest of the paper is organized as follows: Section 2 reviews the existing main approaches of processing the orientation analysis issues. Section 3 describes the proposed method. Section 4 illustrates and discusses the experimental results. Finally, the conclusion is drawn in Section 5. It is important to mention that our work is especially based on Chinese language, thus the examples in the rest of the paper are all in Chinese, and we have translated them into English for easy understanding.

II. RELATED WORK

The existing document-level text orientation analysis methods can be divided into two groups. One group includes the syntax rule-based approaches. These approaches often require expert-defined syntax rules and a sentiment dictionary and analyze document orientation by matching a set of predefined linguistic patterns as templates. Turney [6] summarized several sequence patterns of part-of-speech (POS) tags to identify phrases, and estimate the semantic orientation of each extracted phrase based on pointwise mutual information (PMI). At last, they summed up of all semantic orientation to judge text classification. Popescu *et al.* [7] defined 10 types of rules to extract fine-grained features and associated opinions from reviews. By using these features, it is easy to evaluate the correlation between candidate objects and indicators in the related fields. The experimental results also show that the method achieves good results, but the domain words are very difficult to obtain. Regarding Chinese documents, Ye *et al.* [8] proposed an improved semantic approach based on pointwise mutual information and information retrieval (PMI-IR) for sentiment classification on movie reviews. Zhang *et al.* [9] proposed a text sentiment analysis method based on syntactic structure. Firstly, they found the dependence between sentence context words based on the syntactic tree, and calculated the sentiment values of the sentence. Finally, They accumulated the weighted values of sentences to calculate the sentiment value of full text and compared it with machine learning methods. All of the above studies predicted the whole text orientation from the single word information, which ignored the relation in context. So the result of orientation analysis is not very good.

Another group of studies focused on using machine learning techniques to predict classification of target documents. Pang *et al.* [4] employed Naive Bayes, maximum entropy classification, and support vector machines (SVM) methods to solve the sentiment classification problem. They found that standard machine learning techniques outperformed human-produced baselines but did not perform as well as traditional topic categorization tasks. Some studies have used SVM or combined SVM with other methods to solve the given task. For example, Wang *et al.* [16] integrated SVM with MaxEnt to make a comprehensive classification result. Zhang *et al.* [17] used word2vec as the input of SVM-perf to analyze Chinese comments sentiment classification. Wang and Manning [18] analyzed the characteristics of SVM and Naive Bayes classifier, and proposed a model called NBSVM which combines both of them. It performed well on both snippets and longer documents for sentiment, topic and subjectivity classification. However, although machine learning model overcomes the shortcomings of traditional analysis methods, such as relying on manual work and poor mobility, it is difficult to further improve the accuracy of machine learning model. The structure of Recurrent Neural Network (RNN) [10] and Recursive Neural Network in deep learning could fully consider the context information, which had a good effect on processing serialized data, such as text.

Zhou *et al.* [12] proposed an attention-based bilingual representation learning model, which learns the distributed semantics of documents in both the source and target languages. The Long Short Term Memory (LSTM) networks are used to model documents. And they proposed a hierarchical attention mechanism for the bilingual LSTM networks, which achieve good results on the benchmark data set. Socher *et al.* [13] used RNN to solve text classification problems. They introduced a Sentiment Treebank. The model outperformed all previous methods on several metrics when trained on the new treebank, which could accurately capture the effects of negation. Yang *et al.* [14] proposed a hierarchical text classification model based on the attention mechanism, which could effectively capture the key seniment information in the text. In addition, the Convolutional Neural Network (CNN) [15] structure could capture local information effectively, and some work using CNN to solve the problem of orientation analysis [25]–[28] also achieved good results. Although the accuracy of deep learning model is higher than that of machine learning model, the efficiency and accuracy still need to be improved under the condition of long text, and the corresponding model should be further optimized.

Generally speaking, key sentences can highly summarize the core ideas and viewpoints of the text. Therefore, we can intuitively obtain the theme of the article through key sentences, which is very helpful for analyzing the sentiment of an article. And the key sentences extraction process can effectively remove the redundant content and noise information in the paragraph, and greatly improve the convergence speed of the algorithm for calculating the orientation value of the full-text. Some studies [19]–[21] tried to extract some key sentences from the long document, and analyze the document orientation based on these sentences. In this paper, we also take this idea and propose a new key sentence extraction method to suppress the noise issue in long documents. Besides, Cao and Xie [5] proposed an algorithm for recognizing negative news of enterprises in Chinese language based on the simplified case grammar and this study is closely related to ours. As we know, words in the document contain a lot of information of sentiment, but this work did not take it into consideration. By contrast, in our framework, we combine word2vec and syntax rules from word-level to sentence-level to analyze the news orientation. As showed in experiments, the results of our approach outperformed both machine learning-based approaches and rules-based approaches on given datasets.

III. THE PROPOSED METHOD

A. OVERVIEW

The main work of this paper involves recognition of news orientation and calculation of sentiment intensity of opinions towards the subject. Fig. 1 shows an overview of the proposed approach. The structure of Chinese sentence is complex which contains a lot of redundant information. It is difficult to directly judge the orientation information from a large number of continuous sentences. So before processing, we need to prepare a core sentiment dictionary including positive and negative words which contains only a few but important sentiment words added manually and construct a set of grammar-based syntax rules, including negative words, degree adverbs and associated words, to reduce useless information in the sentences. Word2vec has been proven to be useful and efficient in many applications such as the word similarity computing, thus we choose it to train the word embedding. First, we preprocess the raw text data, which includes segmenting sentence by "。", "；", "！", etc., segmenting words and removing stopping words. After processing, a document will consist of a list of sentences, and a sentence is composed of a list of words. Then, we extract some key sentences from long documents to suppress noise information which may cause wrong orientation judgement. Finally, we analyze news text based on key sentences to determine the orientation of the news and calculate sentiment intensity of opinions. We use the word vector to calculate the word orientation information (WOS), and add up each WOS in the sentence to obtain the sentence orientation information (SOS) from the weighted syntax structure. Then, we can get the orientation value of the whole news based on the keyword voting mechanism.

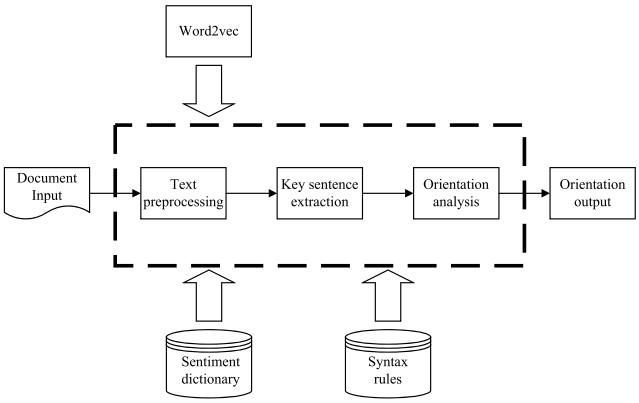


FIGURE 1. Overview of proposed approach.

B. WORD EMBEDDING

The most common ways to represent words are one-hot encoding. However, the simple features have many disadvantages. For example, taxi and cab are the same thing, but the similarity of their one-hot encoding representations is 0.

Word2vec is a tool of word distributed representations based on deep learning techniques [22]. Every word is mapped into a high-dimensional feature space, whose length is less than dictionary size, thus it can suppress the drawback of traditional word representations that are high dimension and sparse. The word vector has the semantic relationships between words in the document. In this paper, we choose continuous bag-of-words [22] architecture as training model and use Chinese Wikipedia corpus and some recent news from the Internet to train word vector. The Wikipedia corpus

TABLE 1. Similar words extraction result.

word	most similar words
捷豹/Jaguar	雪佛兰/Chevrolet(0.8052), 菲亚特/Fabbrica(0.7983), 雪铁龙/Citroen(0.7947), 积架/Jaguar(0.7944), 玛莎拉蒂/Maserati(0.7878), 沃尔沃/Volvo(0.7819)
收购/ purchase	并购/merge(0.7833), 购并/merge(0.6940), 买下/buy(0.6606), 母公司/parent company(0.6597), 注资/funding(0.6490), 入股/stack in(0.6464), 股权/equity(0.6440)
慈善/charity	筹款/fundraising(0.7446), 公益/Public welfare(0.7334), 助学/aid(0.6793), 义演/Charity performance(0.6775), 募款/Fundraising (0.6600)
重组/restructuring	重整/reforming(0.6538), 改组/reorganization(0.6268), 私有化/privatisation(0.5988), 分拆/spin-off(0.5625), 整并/Integral union(0.5462), 合并/merge(0.5307)

supplies enough training corpus and recent news fills the gap of missing new words. After training, words are represented by vectors which size are 400. Formula (1) is used to calculate the similarity of two words:

$$sim_{ij} = cos(w_i, w_j) = \frac{(w_{i1}, \cdots, w_{in}) \cdot (w_{j1}, \cdots, w_{jn})}{\sqrt{(w_{j1}, \cdots, w_{jn})^2} * \sqrt{(w_{j1}, \cdots, w_{jn})^2}} \tag{1}$$

This work uses the Gensim [11] toolkit to train word vectors. Gensim is a deep learning toolkit based on Python language, which contains the training module of word2vec, in which the word vector can be trained by custom parameters, such as window size, word vector dimension, training times and so on. Table 1 lists the most similar synonyms calculated by the cosine similarity. The value in the brackets after the synonym is the similarity. The closer to 1, the higher similarity between two words. We can see that word2vec works well in the task of word similarity computing.

C. KEY SENTENCES EXTRACTION

News often contains a large amount of information, including the description of the reported event, the opinions of media journalists and the knowledge related to the event. However, not all the information is useful for orientation analysis, and even some redundant contents may affect the judgment of orientation. The purpose of key sentences extraction is to obtain sentences from news that are really helpful for clarifying main ideas and filtering out the irrelevant information as much as possible. In general, keywords are highly condensed news topics. Therefore, this work proposes a key sentence extraction method based on keywords, and the extraction results will be used as the input for orientation analysis. In order to filter the deceitful and misleading information in long documents, some key sentences that make big contributions to illustrate the theme or help infer the orientation are extracted as the summary of a document, and then the document orientation can be inferred from them. The key method to weight the importance of sentences is computing the similarity of words in sentences with keywords. Moreover, title is called ''the eyes of news'' in press which means its importance, thus the title is added into the key sentences set automatically.

Document can be segmented and represented by a sentence list: document $= \{s_1, s_2, \ldots, s_n\}$ where s_i is the i-th sentence in the document. For each sentence, segmenting words and removing stopping words are used to translate a sentence into a list of words before processing.

The keywords can summarize the subject of a document. We choose TextRank algorithm [23] to extract keywords from a document. The core idea of TextRank comes from Google's PageRank, which is a keyword extraction algorithm based on the dictionary model. Keywords contain key information related to the topic of the news, which is a high degree of abstraction of the content described in the document. Therefore, the model measures the importance score of words according to the cosine similarity between words and keywords in the news. For each word, the similarity with each keyword is calculated, and the maximum value of all values is finally taken as the importance score of the word. A sentence is made up of a list of words, so the importance of a sentence can be inferred from the importance of its included words. In our approach, keywords are used to estimate the importance score of each sentence.

For a sentence, we first calculate the importance score of each included words. The definition of word importance score is defined as below.

Definition 1: Word importance score (WIS): It means the importance of a word in the document, which is calculated by formula 2. The higher the word importance score, the more important the word. Then we calculate the importance score of each sentence, and its definition is shown as below.

Definition 2: Sentence importance score (SIS): It means the importance of a sentence in the document (As shown in formula 3, L is the number of words contained in sentences). The higher the sentence importance score, the more important the sentence.

$$WIS(W) = MAX(SIM(W, W^k)), \quad W^k \in KeyWords \tag{2}$$

In formula 2, MAX is the maximum operation, and $SIM(\cdot, \cdot)$ indicates the degree of similarity between two words, which uses the cosine similarity described in formula 1. The W^k means any word in the keyword set.

$$SIS(S) = \frac{\sum_{i=1}^{L} WIS(W_i)}{L} \tag{3}$$

Finally, we sort the sentence list ordered by importance score, and extract top k sentences as key sentence list: keysentence $= \{sen_0, sen_1, \ldots, sen_k\}$, where sen_i is the i-th key sentence in the list.

The overall algorithm of key sentence extraction is shown as Algorihm 1. In the initial stage, we use TextRank algorithm to extract 10 keywords from the document to generate the keyword set and create an empty keyword set. After that, the title of the article should be added to the key sentence set since that it can highly condense the content of the article.

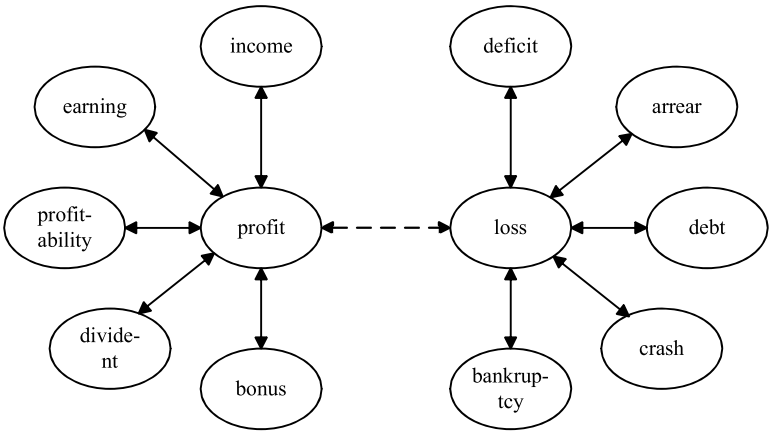


FIGURE 2. An example of adding new words into the dictionary.

Algorithm 1 Key Sentence Extraction

Require: *Set* < *sentences* >: *sentence set*
Ensure: *Set* < *sentences* >: *K key sentence set*
1: *Initialize key sentence set* ← {}, *keyword set* ← {}
2: Extract 10 keywords using TextRank algorithm and add into keyword set *Keyword set* ← { $kw_1, \ldots, kw_n$ }
3: Add news title as sen0 into key sentences set *Keysentenceset*.*add*(*title*)
4: Extract k key sentences from content and add into sentence set
5: **for** *each sentence* to *Set* < *sentences* > **do**
6: **for** *each word* to *sentence* **do**
7: *WIS* ← *MAX*(*SIMILARITY*(*word*, *keyword*))
8: **end for**
9: *SIS* ← *SUM*(*WIS*)
10: **end for**
11: Extract k sentences with top SIS and add into sentence set
12: Key sentence set.add(sen)
13: **return** K key sentence set { $sen_0, sen_1, \ldots, sen_k$ }

For each sentence in the article, we calculate the similarity between each word and the key words, and then choose the maximum value as the word importance score (WIS), so we can get the sentence importance score (SIS) by calculating the average value of all these WIS. Finally, all sentences are sorted in descending order according to the sentence important scores. We extract top k sentences and add them into the key sentence set, which is our final target. Since the number of keywords is only 10, the complexity of calculating a single sentence importance score is just O(n) (Line 7 of algorithm 1), and the time complexity of algorithm 1 is O(mn) in the worst case, where m represents the number of sentences in an article and n represents the number of words in the longest sentence.

D. ORIENTATION ANALYSIS

As described in the part of introduction and related work, news texts are relatively long, so we should consider the context information during the analysis process. Because directly adding the orientation values of words will lose the semantic characteristics of the sentences, and the final result is not good enough. A sentiment dictionary is needed and defined as criterion to distinguish the positive or negative orientation. Sentiment dictionary is closely related to the application field and its quality will affect the recognition results directly. Even though there are many dictionary sources, researchers usually build their own suitable dictionaries to solve a specific problem. We build our specific sentiment dictionary by a semi-automatic process. First, some positive and negative words are defined manually as core sentiment words. These words can represent the syntax features of each orientation well and distinguish from the others. Then we extract synonyms of core words from existing dictionaries to expand the core dictionary. Finally, both of words extracted from existing dictionaries and core words are added to build the new dictionary.

As shown in Fig.2, at first, there is a pair of opposite words ''profit'' and ''loss'', where ''profit'' is in the positive word set and ''loss'' is in negative word set. After calculating the cosine value of word vector, we add the most similar words to ''profit'' such as ''bonus'', ''divident'' into positive word set. Similarly, some similar words as ''deficit'', ''arrear'', ''debt'' are added into negative word set.

In order to get sentence orientation, we should analyze its word orientation first. The existing studies usually judged word orientation by searching whether a word is in sentiment dictionaries or not. This simple strategy has a big disadvantage that some synonyms of sentiment words may be missed during judgement. For example, when we get a word of ''abundant'', but there is only word ''fruitful'' in the dictionary, we will miss the sentiment word by just searching the dictionary. To overcome this problem, word vector is used in our approach to judge a word with orientation information or not. After calculation, the similarity score of the two words is up to 0.8072, thus the positive orientation score will be assigned to ''abundant''.

Definition 3: Word orientation score (WOS): It means the orientation of a word. The positive value means the word has

TABLE 2. Examples of associated words.

Associated words sample	
虽然/虽是/尽管(0.3)... 但/但是/可是/不过(0.7)	Similar to "but" in English
不是(0.0)… 而是(1.0), 是(1.0)… 不是(0.0)	Similar to "not" in English
不管/无论/任凭(0.2)… 还/还是/总/总是/都(0.8)	Similar to "no matter" in English

TABLE 3. Examples of associated words.

虽然 发布会 火爆，但 业内 并不 看好	
0.51	(-1) * 0.46
The product announcement is hot, but industry insiders do not think highly of it.	

positive orientation and the negative value means the word has negative orientation. The bigger absolute value means the sentiment intensity of a word is stronger. The basic calculation method is shown below, in which $Dict^O$ is the sentiment dictionary and made up of the sentiment word W^O.

$$WOS = \tau \cdot MAX\left(SIM\left(W, W^O\right)\right), \quad W^O \in Dict^O \tag{4}$$

$$\tau = \begin{cases} 1 & (W^O \text{ is a positive word}) \\ -1 & (W^O \text{ is a negative word}) \end{cases} \tag{5}$$

In orientation analysis, negative words play an important role in judging. A negative word such as ''no'', ''not'' and ''otherwise'' can completely reverse the orientation of a sentence. Formula (6) is used to assign weight to sentiment words, and n is the number of negative words before a given word.

$$WOS = \begin{cases} WOS & (n = 0, 2, \ldots) \\ -1 * WOS & (n = 1, 3, \ldots) \end{cases} \tag{6}$$

In Chinese, a compound sentence is made up of more than one simple sentences, and each simple sentence is joined by conjunctions, prepositions or articles in a compound sentence. The simplest form of computing sentence orientation score is to calculate the average value of all WOS. Unfortunately, different conjunctions may have different particular emphasis of expression, but the average method cannot reflect it. Table 2 shows samples of associated words with its weights defined in our approach.

Definition 4: Sentence orientation score (SOS): It means the orientation of a sentence. The positive value means the sentence has positive orientation and the negative value means the word has negative orientation. The bigger absolute value means the sentiment intensity of a sentence is stronger.

We propose to add up each WOS in a sentence based on weighted syntax structure to recognize orientation in compound sentences. First, the sentence is matched with associated words and the weight is assigned to each clause. Then, WOS multiplied by weight to get clause orientation score. Finally, formula (7) is used to aggregate every clause orientation score to get SOS, in which $weight_i$ is the corresponding weight of the conjunctions contained in the current clause, which can be found from the pre-defined conjunction list.

$$SOS = \sum_{i=1}^{n} (weight_i * WOS) \tag{7}$$

In traditional method, it is common to directly add the orientation values of each word to obtain the average orientation score of sentences. Such method is simple and fast in calculation, but in many cases it will output wrong results. For example, we use this method to calculate orientation score of the sentence shown in Table 3, which consists of two clauses joined by adversative conjunctions, ''although'' and ''but'', and the result of calculation is $0.3 * 0.51 + 0.7 * (-1) * 0.46 = -0.169$. The orientation is expressed on the clause after ''but'', so it is reasonable to get a negative value of the sentence. But if using average method, we will get $0.51+(-1)*0.46 = 0.05$, which cannot reflect the orientation correctly.

In a piece of news, it is common and reasonable for several individual sentences to express different orientations due to the need of narrations or quotations. However, it is likely to obtain opinions of sentences that are completely opposite to the news during orientation analysis. A piece of news usually expresses one point of view. The sentences with different polarities may cause meaningless conflicts when calculating the orientation of news. Therefore, the model proposes a mechanism based on the key sentence voting to calculate the news orientation. After getting every key sentence orientation value, we then compose all of them to get document orientation value. If the number of positive SOS greater than the negative, the orientation is attributed to the positive, and all positive sentiment orientation scores are averaged to get the result. For the opposite case, the orientation of document is attributed to the negative and the rest processes are the same. The formula is shown as below:

$$orientation(document) = \begin{cases} \dfrac{\sum SOS_{pos}}{posnum} & (posnum > negnum) \\ \dfrac{\sum SOS_{neg}}{negnum} & (posnum < negnum) \end{cases} \tag{8}$$

The overall algorithm of orientation analysis is shown as Algorithm 2. On the basis of algorithm 1, algorithm 2 needs to

Algorithm 2 Orientation Analysis

Require: *Set* $<$ *sentences* $>$: *key sentences set*, *Dict* $<$ *sen* $-$ *words* $>$: *Sentiment dictionary*, *List* $<$ *rules* $>$: *Associated rules*, *Dict* $<$ *negativewords* $>$: *Negative words*

Ensure: *Orientation Score* $\in [-1, 1]$

```
Calculate sentences orientation score
for each sentence to Set < sentences > do
    for each word to sentence do
        WOS ← τ · MAX(SIMILARITY(word, sen − word))
        Match negative words with Dict < negativewords >
    end for
    Match compound sentence with List < rules >
    SOS ← SUM(WOS)
end for
Calculate news orientation
if posnum > negnum then
    Orientation Score ← AVG(positive score)
else
    Orientation Score ← AVG(negative score)
end if
return Orientation Score
```

prepare another three dictionaries or sets, namely, associated rules list, sentiment dictionary and negative words dictionary. For each sentence in the key sentence set obtained by algorithm 1, the method of calculating the word orientation score (WOS) of each word is similar to that of calculating the WIS, which regards the maximum cosine similarity between the target word and sentiment words as the result. The main difference is that we need to consider the polarity of the selected sentiment word to correct our target WOS. If the sentiment word is negative, the intermediate result WOS needs to be multiplied by -1. In addition, it is necessary to count the number of negative words before our target word and record it as n, so that we can determine the positive and negative of the final WOS according to the parity of n. When calculating the orientation value of the whole sentence, we need to pay attention to the importance of the conjunction, because it can greatly affect the meaning of the sentence, so we should multiply each WOS by the weight which corresponding to the conjunctions in the current clause. Finally, in the key sentence voting stage, we compare the number of positive and negative key sentences and average all the SOS of the target set that has the bigger number to calculate the orientation of document.

IV. EXPERIMENTS AND ANALYSIS

In this section, three comparative experiments are conducted to evaluate our proposed approach for public opinion orientation analysis based on two datasets. Experiment 1 evaluates the effectiveness of word2vec in word orientation judgement by comparing with searching word directly. Experiment 2 determines that the word vector dimension of subsequent experiments is 200, in which case the model has the highest accuracy. Experiment 3 proves the proposed key sentence extraction algorithm is beneficial for document orientation analysis. Experiment 4 compares the overall approach with Naive Bayes, KNN and SVM classification algorithm, and it shows that the evaluation metrics of our approach is better than the others.

A. DATASET

Since there is no standard dataset for public opinion orientation analysis, the following two datasets are pre-processed to perform the experiments.

1) CLOSED DATASET

We use web crawler to fetch near 10000 news about enterprise business activities in Shanghai FTZ from the Internet. Then we select 1511 news and label them with 1 or -1 manually. There are 715 news with positive label and 796 news with negative label.

2) OPEN DATASET

The third Chinese Opinion Analysis Evaluation (COAE2011) corpus [24] is an open dataset and we select 1121 news in finance domain which have 687 positive news and 434 negative news. The length of news is shorter than the closed one, and every document has at least one sentence.

Two dictionaries are used in the experiments. The small one is a core dictionary defined by ourselves which contains 81 positive words and 154 negative words. The big one is based on core dictionary and select other words from the dictionary set up by Li and Sun [29]. There are 1552 positive words and 2941 negative words.

B. EXPERIMENTS

The value of document orientation is decimal between -1 and 1 after processing. In order to evaluate the result conveniently, we simplify the negative value to -1 and the positive value to 1. In the following experiments, we use standard evaluation metrics such as accuracy, precision, recall and F-measures to evaluate the performance.

Accuracy can reflect the overall discriminant ability of the model, which is defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

TP is the number of texts which are predicted to have the positive orientation, and these texts are indeed the positive one. TN denotes the number of texts predicted to have the negative orientation, and they indeed have the positive one. FP is the number of texts predicted to be the positive sentiment word but are the negative one. FN is the number of texts predicted to have the negative orientation but have the positive one.

With these definitions, it is easy to define the precision, recall and F-measure. They are used to measure the discriminant ability of the model for each type of data. The definition of precision is the ratio of the number of texts correctly discriminated under their respective category to the total number

of texts discriminated into their respective category. The definition of recall is Number of texts correctly discriminated under their respective category to total number of texts under each category. There is a close relationship between accuracy and recall rate. When accuracy is low, recall is usually high, whereas when accuracy is high, recall will be usually low. Therefore, F-measure value is usually introduced to find a balance between them in practical application. Here are their definitions:

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{12}$$

Experiment 1: In order to evaluate the effectiveness of word2vec in extracting word orientation information, especially without a complete sentiment dictionary, we compare two methods of judging WOS. One method is using word2vec to compute the similarity of given word and sentiment words which is proposed in this work, the other is the simple method of searching the given word in dictionaries directly. We use the small sentiment dictionary to validate the conclusion. Fig. 3 shows the changes of accuracy according to the number of news varies from 200 to 1500 performed on closed dataset. The accuracy of proposed method is always higher than the simple searching one. With the increase in the number of news, the accuracy of using word2vec remains high, and the simple method continues to decline. The experimental results on open dataset are shown in Fig. 4, our approach also performs better. The reason is that as the amount of data increases, the coverage of the dictionary will gradually decrease. The direct comparison method may not be able to match the corresponding words in the dictionary. Take the words ''abundant'' and ''fruitful'' for example, if the direct comparison method is adopted, since the word ''abundant'' does not exist in the dictionary, the search algorithm will ignore the word. In comparison, by word embedding method, you can find that the similarity between ''abundant'' and ''fruitful'' is 0.8072, which can also be considered as a positive word.

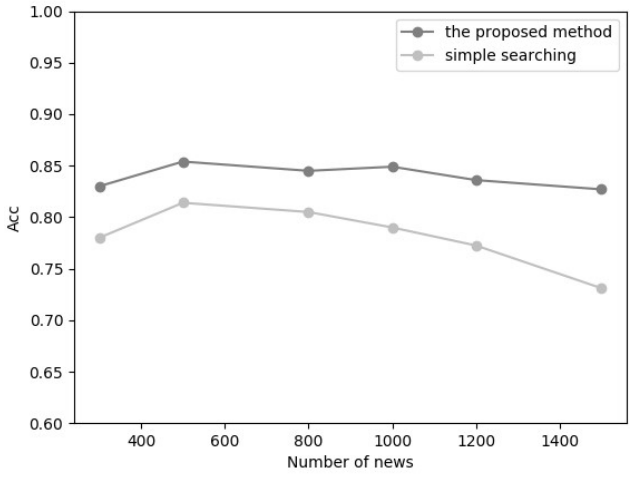


FIGURE 3. Comparison of different WOS calculation method on closed dataset.

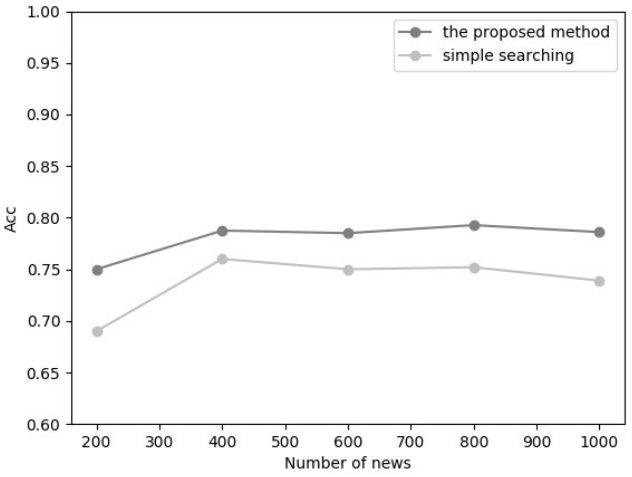


FIGURE 4. Comparison of different WOS calculation method on open dataset.

In many cases, syntax features in different background are different from each other, so researchers need to re-define suitable sentiment dictionary for given fields every time. This Experiment shows that even if there is no complete sentiment dictionary, using word2vec to vectorize the text can also improve the flexibility of feature extraction, which plays a positive role in improving the accuracy of the orientation analysis and maintaining good sentiment-oriented analysis performance.

Experiment 2: The dimension of word vector is an important super parameter of text representation. How to choose the dimension is one of the key factors in training high quality word vector. If the dimension is too low, the features of each dimension cannot be well distinguished. If the dimension is too high, features may become too sparse. Therefore, finding an appropriate dimension can maximize the text expression ability of word vector.

In order to evaluate the influence of the word vector dimension on the orientation analysis model, we conducted an experiment on the above two datasets. Firstly, word vector models with dimensions of 50, 100, 150, 200, 250 and 300 were pre-trained by word2vec. Then, we use these models in orientation analysis steps and record the accuracy rate with different word vector dimensions, which is shown in Fig. 5. It is found that when the word vector dimension is 200, the method achieves the highest accuracy on both datasets. Therefore, we adopt the word vector model with dimension 200 for the subsequent experiment.

Experiment 3: In order to verify that the key sentence extraction steps can improve the performance of orientation analysis model, we design a set of comparative experiments.

TABLE 4. Result of orientation analysis based on different content.

Dataset	Metrics	Key sentence		Random		Full-text	
		Pos	Neg	Pos	Neg	Pos	Neg
Closed	P	0.834	0.912	0.733	0.836	0.780	0.890
	R	0.910	0.837	0.844	0.724	0.893	0.773
	F	0.870	0.873	0.785	0.777	0.833	0.828
Closed	P	0.809	0.786	0.762	0.720	0.803	0.774
	R	0.874	0.691	0.842	0.606	0.867	0.681
	F	0.840	0.735	0.800	0.658	0.834	0.725

TABLE 5. Results of orientation analysis based on different methods.

Dataset	Our approach	Naive Bayes	KNN	SVM
Closed	0.8723	0.8297	0.7917	0.8612
Open	0.8013	0.6768	0.6402	0.622

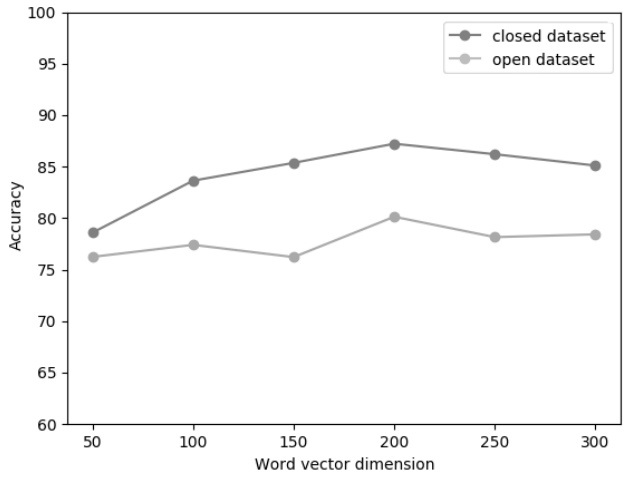


FIGURE 5. The influence of word vector dimension on accuracy.

The compared experiment replaces the key sentence extraction algorithm with the random extraction step. Compared with the key sentence extraction algorithm, the random extraction method extracts the same amount of texts. This experiment was also carried out on the above two datasets, using accuracy rate, recall rate and F value as evaluation metrics, and paying attention to the positive and negative two classifications at the same time. As shown in Table 4, the method based on key sentence has obvious improvements compared with the method based on full document and random extraction. It is worth nothing that since the length of most documents in open dataset is relatively short, the improvement of using key sentence is not so significant than that using full document. The experiment also demonstrates that our key sentence extraction method has a good performance in long document analysis.

Experiment 4: In order to validate the overall proposed method has a better performance on public opinion orientation analysis, the other three machine learning-based methods, Naive Bayes classifier, KNN classifier and SVM, are also implemented on the same dataset. In Table 5, we can see that our approach significantly outperformed the Naive Bayes and KNN classifier on both two datasets. It has nearly the same accuracy with SVM on the closed dataset. However, on the open dataset, the performance of SVM has a significant decrease and our method can keep a good performance. The reason is that the size of open dataset is smaller than that of closed one. For a supervised machine learning algorithm such as SVM, a certain amount of training corpus is required to obtain a better model. And it is found that the model proposed in this paper is generally higher than the other three models in the accuracy rate, which proves the effectiveness of the model.

Our approach not only focus on local orientation information, but also from the overall structure to predict orientation by considering the contribution of key sentences. Orientation scores of individual key sentences are aggregated to predict document orientation and the effectiveness of approach has been validated in the experiments.

V. CONCLUSION

In this study, we attempt to utilize the novel approach to deal with the problem of Chinese public opinion orientation analysis. There are three main steps in the framework, including word embedding training, key sentence extraction and orientation analysis. Experimental results show that our approach outperform both machine learning-based and ruled-based approaches on given datasets. Besides, it has been used in Shanghai FTZ successfully and has a good performance. However, there are still a lot of rooms for performance improvement. In further work, we want to find a way to combine word vector with traditional rule-based approach deeper to make full use of their own advantages. Furthermore, the proposed method will be applied to more datasets to examine the effectiveness and the improvement will be pursued. Big data as a service (BDaaS) and service selection and recommendation [30]–[40] is another direction in this big data and service-based era.

REFERENCES

[1] T. Zhuang, P. Wang, Y. Zhang, and Z. Zhang, "A novel method for open relation extraction from public announcements of Chinese listed companies," in *Proc. IEEE 5th Int. Conf. Adv. Cloud Big Data*, Shanghai, China, Aug. 2017, pp. 200–205.

[2] X. Sun, P. Wang, Y. Lei, W. Liu, L. Yang, and Z. Zhang, "A method for discovering and obtaining company hot events from Internet news," in *Proc. Int. Conf. Prog. Inform. Comput.*, Suzhou, China, Dec. 2018, pp. 26–32.

[3] S.-M. Kim and E. Hovy, "Automatic detection of opinion bearing words and sentences," in *Proc. 2nd Joint Int. Conf. Natural Lang. Process.*, 2005, pp. 61–66.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process. Assoc. Comput. Linguistics*, vol. 10, 2002, pp. 79–86.

[5] H. Cao and X. Xie, "Research and application of an algorithm about negative news judgment," (in Chinese), *J. Chin. Comput. Syst.*, vol. 36, no. 5, pp. 1047–1051, 2015.

[6] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics Assoc. Comput. Linguistics*, 2002, pp. 417–424.

[7] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural Language Processing and Text Mining*, A. Kao and S. R. Poteet, Eds. London, U.K.: Springer, 2007, pp. 9–28.

[8] Q. Ye, W. Shi, and Y. Li, "Sentiment classification for movie reviews in Chinese by improved semantic oriented approach," in *Proc. 39th Annu. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Kauia, HI, USA, Jan. 2006, p. 53b.

[9] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo, "Sentiment analysis of Chinese documents: From sentence to document level," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 12, pp. 2474–2487, 2009.

[10] A. Karpathy, J. Johnson, and F.-F. Li, "Visualizing and understanding recurrent networks," 2015, *arXiv:1506.02078*. [Online]. Available: https://arxiv.org/abs/1506.02078

[11] K. Khosrovian, D. Pfahl, and V. Garousi, "GENSIM 2.0: A customizable process simulation model for software process evaluation," in *Proc. Int. Conf. Softw. Process*, 2008, pp. 294–306.

[12] X. Zhou, X. Wan, and J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 247–256.

[13] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.

[14] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[15] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[16] P. Wang, H. Zhang, Y.-F. Wu, B. Xu, and H.-W. Hao, "A robust framework for short text categorization based on topic model and integrated classifier," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 3534–3539.

[17] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVM^{perf}," *Expert Syst. Appl.*, vol. 42, pp. 1857–1863, Mar. 2015.

[18] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2012, pp. 90–94.

[19] L. W. Ku, Y. T. Liang, and H. H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in *Proc. AAAI*, 2006, pp. 100–107.

[20] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, 2004, Art. no. 271.

[21] Z. Lin, S. Tan, and X. Cheng, "Using key sentence to improve sentiment classification," in *Proc. Asia Inf. Retr. Symp.*, 2011, pp. 422–433.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 3111–3119.

[23] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 404–411.

[24] X. Liao, H. Xu, L. Sun, and T. Yao, "Construction and analysis of the third chinese opinion analysis evaluation (COAE2011) corpus," *J. Chin. Inf. Process.*, vol. 1, no. 27, pp. 56–63, 2013.

[25] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: https://arxiv.org/abs/1408.5882

[26] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 649–657.

[27] Y. Xie, W. Rao, P. Duan, and Z. Chen, "A Chinese POS tagging approach using CNN and LSTM-based hybrid model," (in Chinese), *J. Wuhan Univ. (Natural Sci. Ed.)*, vol. 63, no. 3, pp. 246–250, 2017.

[28] M.-C. Hong, K. Zhang, J. Tang, and J.-Z. Li, "A Chinese part-of-speech tagging approach using conditional random fields," (in Chinese), *Comput. Sci.*, vol. 33, no. 10, pp. 148–151, 2006.

[29] J. Li and M. Sun, "Experimental study on sentiment classification of Chinese review using machine learning techniques," in *Proc. NLP-KE Int. Conf. Natural Lang. Process. Knowl. Eng.*, vol. 12, Aug./Sep. 2007, pp. 393–400.

[30] Y. Yin, L. Chen, Y. Xu, and J. Wan, "Location-aware service recommendation with enhanced probabilistic matrix factorization," *IEEE Access*, vol. 6, pp. 62815–62825, 2018.

[31] Y. Yin, Y. Xu, W. Xu, M. Gao, L. Yu, and Y. Pei, "Collaborative service selection via ensemble learning in mixed mobile network environments," *Entropy*, vol. 19, no. 7, p. 358, 2017.

[32] Y. Yin, W. Xu, Y. Xu, H. Li, and L. Yu, "Collaborative QoS prediction for mobile service with data filtering and SlopeOne model," *Mobile Inf. Syst.*, vol. 2017, Jun. 2017, Art. no. 7356213.

[33] Y. Yin, F. Yu, Y. Xu, L. Yu, and J. Mu, "Network location-aware service recommendation with random walk in cyber-physical systems," *Sensors*, vol. 17, no. 9, p. 2059, 2017.

[34] Y. Yin, S. Aihua, G. Min, X. Yueshen, and W. Shuoping, "QoS prediction for Web service recommendation with network location-aware neighbor selection," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 26, no. 4, pp. 611–632, 2016.

[35] H. Gao, H. Miao, L. Liu, J. Kai, and K. Zhao, "Automated quantitative verification for service-based system design: A visualization transform tool perspective," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 28, no. 10, pp. 1369–1397, 2018.

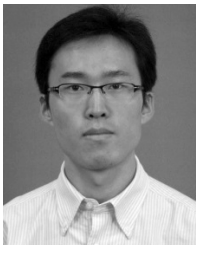
[36] H. Gao, W. Huang, X. Yang, Y. Duan, and Y. Yin, "Toward service selection for workflow reconfiguration: An interface-based computing solution," *Future Gener. Comput. Syst.*, vol. 87, pp. 298–311, Oct. 2018.

[37] H. Gao, S. Mao, W. Huang, and X. Yang, "Applying probabilistic model checking to financial production risk evaluation and control: A case study of Alibaba's Yu'e Bao," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 3, pp. 785–795, Sep. 2018.

[38] H. Gao, Y. Duan, H. Miao, and Y. Yin, "An approach to data consistency checking for the dynamic replacement of service process," *IEEE Access*, vol. 5, pp. 11700–11711, 2017.

[39] H. Gao, D. Chu, Y. Duan, and Y. Yin, "Probabilistic model checking-based service selection method for business process modeling," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 27, no. 6, pp. 897–923, Feb. 2017.

[40] Y. Yin, L. Chen, Y. Xu, J. Wan, H. Zhang, and Z. Mai, "QoS prediction for service recommendation with deep feature learning in edge computing environment," *Mobile Netw. Appl.*, pp. 1–11, Apr. 2019.



PENGWEI WANG received the B.S. and M.S. degrees from the Shandong University of Science and Technology, Qingdao, China, in 2005 and 2008, respectively, and the Ph.D. degree from Tongji University, Shanghai, China, in 2013, all in computer science.

He finished his Postdoctoral Research work at the Department of Computer Science, University of Pisa, Italy, in 2015. He is currently an Associate Professor with the School of Computer Science and Technology, Donghua University, Shanghai. His research interests include cloud computing, data mining, and service computing.

YUJUN LUO received the B.S. degree in computer science and technology from East China Normal University, Shanghai, China, in 2015, and the M.S. degree in computer science from Tongji University, Shanghai, in 2018. Her research interests include data mining and machine learning.

ZHEN CHEN received the B.S. degree in software engineering from Donghua University, Shanghai, China, in 2018, where he is currently pursuing the master's degree in computer science and technology. His current research interests include cloud computing, crowd intelligence, and machine learning.

LIANGHUA HE received the B.S. degree in surveying and mapping from the Wuhan Technology University of Surveying and Mapping, Wuhan, China, in 1999, the M.S. degree in surveying and mapping from Wuhan University, Wuhan, in 2002, and the Ph.D. degree in electronic engineering from Southeast University, Nanjing, China, in 2005.

From 2005 to 2007, he was a Postdoctoral Researcher with Tongji University, Shanghai, China, where he is currently a Professor. Since 2007, he has been with the Department of Computer Science and Technology, Tongji University. His current research interests include signal processing, pattern analysis, and machine learning.

ZHAOHUI ZHANG received the B.S. degree in computer science from Anhui Normal University, Wuhu, China, in 1994, the master's degree from the University of Science and Technology of China, in 2000, and the Ph.D. degree in computer science from Tongji University, Shanghai, China, in 2007. He became a Teacher at Anhui Normal University. He was a Professor with Anhui Normal University, in July 2015. He is currently a Professor with the School of Computer Science and Technology, Donghua University, Shanghai. His research interests include network information services, service computing, and cloud computing.