IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 29, 2020, accepted June 10, 2020, date of publication July 14, 2020, date of current version August 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007763

News Text Summarization Based on Multi-Feature and Fuzzy Logic

YAN DU AND HUA HUO
School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China
Corresponding author: Hua Huo (pacific_hua@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 61672210.

ABSTRACT In the last 70 years, the automatic text summarization work has become more and more important because the amount of data on the Internet is increasing so fast, and automatic text summarization work can extract useful information and knowledge what user's need that could be easily handled by humans and used for many purposes. Especially in people's daily life, news text is the type of text most people are exposed to. In this study, a new automatic summarzation model for news text which based on fuzzy logic rules, multi-feature and Genetic algorithm (GA) is introduced. Firstly, the most important feature is word features, we score each word and extracted words that exceeded the preset score as keywords and because news text is a special kind of text, it contains many specific elements, such as time, place and characters, so sometimes these special news elements can be extracted directly as keywords. Second is sentence features, a linear combination of these features shows the importance of each sentence and each feature is weighted by Genetic algorithm. At last, we use fuzzy logic system to calculate the final score in order to get automatic summarization. The results of the proposed method was compared with other methods including Msword, System19, System21, System 31, SDS-NNGA, GCD, SOM and Ranking SVM by using ROUGE assessment method on DUC2002 dataset show that proposed method outperforms the aforementioned methods.

INDEX TERMS News text summarization, genetic algorithm, multi-feature, fuzzy logic system.

I. INTRODUCTION

Now in the era of big data, there is a large amount of data produced on the Internet every day, today's people feel is the most powerful social media data of explosive growth [1], such as our daily news from Web, WeChat, weibo, and various types of industry data. The volume of data is already much larger than existing storage, processing and analysis tools [2]. Among them, Web news has become one of the best media for people to get the latest information and the ever-changing current events. Facing these massive news, people do not have enough time to get the information they need by reading all the news online, especially for some enterprises and individuals in great demand for information. Therefore, in recent years, the research on automatic news summary [13] has become a focus of people's attention. On the one hand, it can solve the problem of information overload on the Internet and the other hand, it can simplify the information obtained by users [5].

The so-called news automatic summary is the process that the computer automatically extracts the corresponding content of the summary from the original news text and generates a concise and coherent short article that accurately and comprehensively reflects the central content of a news [3]. Some works use search engines to generate summarization for individual document on web pages [10]. In other works, text analysis method is used to extract the most important features on individual document [14], [15] and assigning labels to set up the web document cluster [16]. In recent years, Neural Networks and Support Vector Machine methods [6]–[8] have been used in text summarization due to their efficient performance in text data mining [19], [20].And some works use topic detection, tracking tasks [53] and literature summarization [54] to quickly mine the theme of the text. In recently, fuzzy logic is increasingly used in text data mining, [32]–[36], [38] and because of its characteristics of self-organization, self-adaptation and self-learning, evolutionary algorithm [41]–[45] is used in text summarization because it can effectively deal with complex problems that are difficult to be solved by traditional optimization algorithms without being restricted by the nature of the problem.

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

However, most of the researches only consider the overall word features and sentence features when extracting the features of news texts, and ignore the unique features of news articles, namely the three elements of time, place and person, which constitute the basic facts of a news. News as a kind of natural language processing (NLP) [11], [12] at the same time itself is the embodiment of the artificial intelligence, thinking, most studies based on calculating the importance of the sentences in the text to extract important sentences to form a summary, but these studies ignored the fuzziness of artificial intelligence, because different people will have different opinions for the same thing. Therefore, this paper proposes an automatic summarization method based on multi-feature extraction and fuzzy logic. It extracts the most important features by feature selection, the key words were selected by calculating the score of each word, and then the sentences were scored by combining the special three elements of the news text. Genetic algorithms assign optimised weights to each text feature, and then fuzzy logic system is used for sentence scoring and summarization extraction. The rest of this paper is organized as follows: Section II introduces current research status of this field. Section III introduces our proposal of news text features extaction and features weight calculation method. Section IV persents the experiments designed to evaluate this method. Section V displays the results of proposed experiments. Section VI discuss the results of proposed experiments. At the end, Section VII we make a conclusion of this paper.

II. RELATED WORK

Most studies calculate the score of each sentence [18] to get the most important sentences. In most cases, there are two steps to calculate the sentence score: word scoring [29] and sentences scoring [24]. These are summary generation models based on statistical principles [49]. In calculating the score of words, the score of each word is determined by the features it contains, and the score of a sentence is the sum of the scores of the words it contains. And some of the word features widely used in news text summarization works, such as capitalized nouns, title words, thematic words [27], [48] and indication words. Different works use different methods to determine the importance of the words such as word frequency [23] and TF/IDF [21], [22]. In calculating the score of sentences, sentence position [28], [30] and sentence length [28] are two most significant features in a sentence. LIN *et al.* [28] use sentence length, sentence position and similarity between sentence and title to improve the accuracy of summary sentences. Xiaojun Wan proposed a approach to simultaneous single-document and multi-document summarizations [47]. Naveen *et al.* apply the concept of multi-objective optimization to twitter summarization to produce high-quality summaries. And using the search capability of multi-objective differential evolution technology, different statistical quality measures were optimized at the same time, namely the length of tweets, TF-IDF score, anti-redundancy, and different aspects of the measurement summary [58]. LIU *et al.* [24] proposed four different weighting methods based on word frequency, sentence position, sentence length, sentence theme and title similarity, which solved the problem of automatic summarization based on microblog news texts. In recent years, Neural Networks [46], [57] and Support Vector Machine methods [50] have been used in text summarization due to their efficient performance in text data mining. Binwahlan *et al.* [44] proposed a fuzzy swarm diversity hybrid model for text summarization which includes three models: (i) Maximal Marginal Importance diversity, (ii) swarm diversity based method, and (iii) fuzzy swarm based method.

Lee *et al.* [25] proposed a method to apply fuzzy ontology to news summarization. Fuzzy ontology is the extension of domain ontology with fuzzy concept. Compared with domain ontology, domain knowledge is more helpful to solve uncertain reasoning problems. First, domain ontologies of various categories of news events are defined by language domain experts, documents are preprocessed according to the defined news corpus to generate meaningful terms, and meaningful terms are classified according to the nature of news events. The membership degree of each fuzzy concept in the fuzzy ontology is generated by using the fuzzy reasoning system. The final summary is generated through sentence extractors, sentence generators, and sentence filters.

In another work, A. Al-Radaideh *et al.* [26] proposed a hybrid, single-document text summarization approach, which includes three models: (i) domain knowledge method, (ii) statistical features method, and (iii) genetic algorithms based method. The first method mainly establishes a knowledge base manually in the professional domain corpus and summarizes the texts in the domain by using domain keywords. The second approach focus on segment extraction and ranking using heuristic methods that assign weighted scores to segments of text. The third approach treats the automatic text summary task as a classification problem. They use a machine-learning approach to categorization, using GA to produce a readable, cohesive, and good summary that is similar to the topic of the document, based on a set of attributes that describe the documents.

The model presented in this paper is single-text summary models. According to the characteristics of news text, a new text feature selection method combining the three news elements (time, location and person) was proposed, and the fuzzy logic system was introduced combined with genetic algorithm. The text features were weighted by genetic algorithm, and then the text features were adjusted twice by fuzzy logic system to get more accurate summary sentences, so as to generated a higher quality summary.

III. PROPOSE METHOD

As Fig.1 shows, there are four main aspects of this work: (1) Use NLTK tools to preprocess the original news text, (2) Extract important features from news text, (3) Use genetic algorithm to assign appropriate weights to the extracted news text features and (4) Use fuzzy logic system to score the

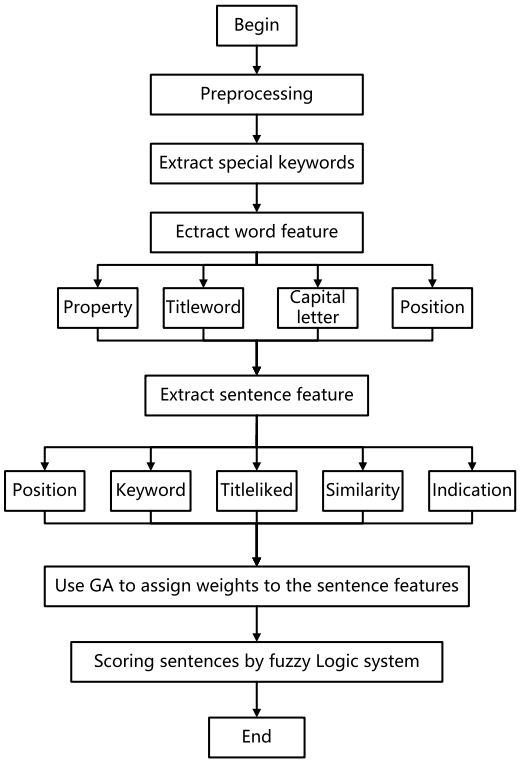

FIGURE 1. Automatic news text summarization model based on multi-feature, genetic algorithm and fuzzy logic.

sentences. In the step of extracting news text features, we will divide it into two parts. The first step is to extract features for each word, and the second step is to extract features for each sentence. The important features of the words we want to extract include five features: word frequency, word property, thematic word, capital letters and word position. And features of sentence we want extract contains five features: keyword, position, titleliked feature, similarity feature and indication feature.

A. PREPROCESSING

In this part, preprocessing process contains four parts: (1)Segmentation by NLTK [40], (2)Removing stop words by NLTK, (3)part-of-speech tokenization, (4)Stem extraction based on NLTK and porter algorithm.

B. NEWS TEXT FEATURE

In this part, we extract the most significant features from the preprocessed original news text. We can get the important sentences we want through these news text features. These features will be introduced as follows:

1) WORD FEATURES

Words are the basic unit of a sentence, so the importance of the words in a sentence determines the importance of the sentence in the news text. The features that determine the importance of a word as follows:

• Word frequency

Word frequency feature is the word occurrence frequency statistics in the article, which is the most used feature in all summarization systems. In general, a word appears more frequently in the article and spans more paragraphs, so it tends to be more important in the article.

• Word property

Content word features refer to the part of speech features of a word. From the perspective of natural language understanding, nouns and verbs constitute the core of a text, and their simple combination can be used as a simple expression of the whole text. In addition, adjectives and adverbs also play a certain role in the expression of the content of the text. Therefore, content words are usually nouns, verbs, idioms, adjectives and adverbs. In terms of the weight representation of content word features, we think that the weight of nouns, verbs and idioms is greater than that of adjectives and adverbs.

For the word t in the article, considering the property of word of t, if t is a noun and a verb, then $S_{word}(t) = 2$, if t is an adjective or adverb, then $S_{word}(t) = 1$, and if t is another part of word, then $S_{word}(t) = 0$.

• Title word

The feature of the inscription mainly considers whether the word appears in the headline of the news.The title is the soul of the news, is the first link in the process of news transmission and acceptance, is the eyes in the whole news transmission, the title has become a crucial factor to guide readers to read, it requires a high concentration of the main content of the message in just 10 or 20 words, to provide as much information as possible.The audience has formed a ''first dependence'' on the news title, which has become the first signal for the audience to identify the news content and judge the news value, and the first choice gateway for the audience to decide whether to ask for in-depth news information.Therefore, words in news headlines often play an important role in the expression of news content.

For the word t in an article, if in a news headline or subheading, $S_{title}(t) = 2$, otherwise $S_{title}(t) = 1$.

• Capital letters

When a sentence contains capitalized English words like APEC, IBM, etc, the sentence may contain important information. In news, capital letters appear especially frequently in science and finance news, and they often contain important information, often as key nominal features.

For the word t in an article, if the word is a capital letter, then $S_{eng}(t) = 2$, otherwise $S_{eng}(t) = 1$.

• Position

The first sentence in the news is called the ''leading sentence'', which usually clearly describes the content of the news and indicates the main idea of the news, while the following sentence usually contains a lot of redundant information in terms of information extraction and processing. The typical dominant sentence does not use subjective expressions or vague adjectives and verbs. Therefore, the words in the dominant sentence are also more important.

For the word t in an article, if the word appears in the leading sentence, then $S_{pos}(t) = 2$, otherwise $S_{pos}(t) = 1$.

The above characteristics are combined to comprehensively calculate the weight of the word. If the weight is greater than a preset value of min, it is extracted as a feature word. Then we improved the formula TF-IDF to calculate the weight of the words, the features of the above news texts are considered as well as the features of word frequency. The calculation formula is as follows:

$$W(t) = Stitle(t) \times Sword(t) \times Seng(t) \times Spos(t) \times \frac{tf(t) \times \log(\frac{N}{nt} + 0.01)}{\sqrt{\sum [tf(t) \times \log(\frac{N}{nt} + 0.01)]}} \quad (1)$$

In the formula above, W(t) is the weight of word t in the text, tf(t) is the word frequency of word t in the text, N is the number of paragraphs in the text, nt is the number of paragraphs with word t in the text, and the denominator is the normalized factor. $S_{title}(t)$,$S_{word}(t)$,$S_{eng}(t)$and $S_{pos}(t)$are the weighting coefficients, which are evaluated according to the above keyword definition.

2) EXTRACTION OF SPECIAL KEYWORDS FROM NEWS TEXTS

Before the extraction of the above feature words, we analyzed the composition of the summary of news articles. As long as the elements of news are taken as feature fields for extraction and indexing, a complete news fact can be formed, that is, the summary of news can be formed. Among these elements, time, place and character, as the basic elements, play an important role in the summaries, and their identification mainly depends on the result of word segmentation, which can be extracted directly. Therefore, we extract these three features as unique features of news and define them as follows:

a: TIME CHARACTERISTICS

Time words appear in the news (including numeric time and ''Christmas'', ''National Day'', ''eve'', etc.). According to the analysis of news features, news usually occurs in the first paragraph of an article. If the time characteristics does not appear in the first paragraph, this article is likely to be a kind of commentary news, so the time does not play a key role in the summary. Therefore, in the extraction of time features, we mainly consider the time in the first paragraph. If the time words appear in the first paragraph of news, they will be directly extracted as keywords.

b: PERSON CHARACTERISTICS

Names in the news. As the element of news, people's names appear in every paragraph of news. Therefore, the words marked with names in the news text are directly extracted as the feature words of the article.

c: SITE CHARACTERISTICS

Place names in the news. The words marked as place names in the news text are directly extracted as the feature words of the article.

As the basic elements of news, the above three features play an important role in the news summary, and can be directly identified accurately after text preprocessing. Therefore, before the extraction of feature words in the text, the above three unique features of news were firstly extracted, and then the selection and extraction of feature words in the whole article were conducted. The whole input feature words in the news text were the sum of the features extracted in the two stages.

3) SENTENCE FEATURE

After extracting the keywords of the article, we can calculate the weight of each sentence in the article, so as to quantitatively determine the importance of each sentence in the document, and extract the sentences that constitute the summary according to the weight of the sentence, so as to provide a quantitative standard for summary sentence extraction. First, sentence features should be extracted, and the selected features should fully reflect the function of the sentence in the original document, that is, the degree to which the sentence expresses the main content of the original text. According to the sentence weight calculation function, the sentence weight obtained after text preprocessing is calculated. The heavier the sentence is, the more important it is considered to be, and the more it can represent the content of the article, and the more likely it is to be used as the topic sentence. The extraction of the sentence output as the summary is based on its weight value.

• Position

In the feature selection of words, we consider the feature of word position. Similarly, for a sentence, its importance in the article is also closely related to its position. In general, because sentences are usually organized hierarchically, sentences at the beginning and end of a paragraph are more likely to contain material useful for the summary, especially the first paragraph of a news text.

The weight of the statement position is divided into 7 levels, from 1 to 6. Where 0 means not summary sentence, 1 to 6 means the different importance of sentences, 1 means least important, and 6 means most important. Specific division method is by expert rules [31]:

- ▶ The first sentence of the article: 6
- ▶ The last sentence of the article: 5
- ▶ First sentence per paragraph: 4
- ▶ Last sentence per paragraph: 3
- ▶ The second sentence of each paragraph: 2
- ▶ The last second sentence of each paragraph: 1
- ▶ Others: 0

For each statement S_i in the file, its position feature is calculated as a formula:

$$Score_{pos}(s_i) = \frac{Position_1}{6} \quad (2)$$

In the formula:

s_i - a sentence in an article

$Position_1$ - The position weight of sentence s_i in the article

• Keyword

The more important keywords a statement has, the more likely it is to belong to the summary sentence. In addition, the weight of sentences containing keywords is unified with the weight of keywords. The weight of keywords is different. In the case of sentences containing different keywords and the same number of keywords, the weight of sentences is determined by the importance of keywords.

For each sentence S_i in the article, the calculation method of sentence keyword features is like a formula:

$$Score_{Key(s_i)} = \sum_{k=1}^{j} c_k \times w_k \tag{3}$$

In the formula:s_i - a sentence in an article

j - The number of keywords that appear in sentence s_i

c_k - The number of times the k_{th} keyword appears in sentence s_i

w_k - The weight of the k_{th} keyword appearing in sentence s_i

• Titleliked

This feature considers how similar each statement is to the title of the article. In the above key feature extraction of words, we analyze the important role of news headlines on their subject. Similarly, if the sentence in the article is similar to the title, that is, close to the meaning expressed by the title, then the sentence will be more relevant to the topic of the article.

The degree of similarity between two sentences is expressed by calculating the cosine similarity of two sentences. When getting the sentence vector, first we used Google word2vec kit [55], [56] to get the corresponding Vector for each word, then add all the vectors and average them, so we can get the sentence vector, and then calculate the cosine of the angle between the sentence and title. For a statement s in the article, the feature calculation method of sentence and title similarity is as followed formula:

$$Score_{Title}(s) = \frac{\frac{\sum_{i=1}^{ns} wvsi}{ns} \times \frac{\sum_{i=1}^{ns} wvti}{nt}}{\left\| \frac{\sum_{i=1}^{ns} wvsi}{ns} \right\| \left\| \frac{\sum_{i=1}^{ns} wvti}{nt} \right\|} \tag{4}$$

In the formula:

s - a sentence in an article

n_s - The number of words in sentence s

n_t - The number of words in title

wv_{si} - The vector of the ith word in sentence s

wv_{ti} - The vector of the ith word in title

• Similarity feature

This feature mainly considers the degree of similarity between the sentences in the article. The cumulative similarity of a sentence reflects the degree of correlation between the sentence and other sentences in the article. The greater the cumulative similarity of a statement, the more sentences related to the statement in the article, the more the statement can reflect the main content of the article. We believe that sentences with greater cumulative similarity are more likely to be part of the summary. Cosine similarity is also used to calculate sentence similarity.

The similarity calculation method between two statements is as the formula:

$$Sim(s_i, s_j) = \frac{\frac{\sum_{x=1}^{nsi} wvsix}{nsi} \times \frac{\sum_{x=1}^{nsj} wvsjx}{nsj}}{\left\| \frac{\sum_{x=1}^{nsi} wvsix}{nsi} \right\| \left\| \frac{\sum_{x=1}^{nsj} wvsjx}{nsj} \right\|} \tag{5}$$

In the formula:

s_i,s_j - a sentence in an article

ns_i - The number of words in sentence s_i

ns_j - The number of words in sentence s_j

wvs_{ix} - The vector of the xth word in sentence s_i

wvs_{jx} - The vector of the xth word in sentence s_j

Then, the cumulative similarity calculation method of statement s_i is as the formula:

$$Score_{sim}(s_i) = \sum_{j=1(j\neq i)}^{m} Sim(s_i, s_j) \tag{6}$$

In the formula:

s_i - a sentence in an article

m - the number of sentences in the file

• Indication

The purpose of news is to tell readers the latest information as soon as possible, so news writing is more focused. Compared with other types of articles, news articles are more dependent on some indicative words or strings to identify the key points of articles, so as to facilitate readers to read quickly. Subject items are different for each story, whereas indicators are generic, and these words have the same effect on each story.

News text is a very special type of text, the main purpose of news text is to objectively state clearly the news facts, so the language writing of news text is very rigorous, the use of words is very fastidious. Through the observation of news text, we find that the use of indication words is relatively single. Therefore, we built a small corpus to include these indication words commonly used in news texts. According to the importance of deixis, deixis are divided into three levels.

1) Topic deixis (Itit): this kind of deixis is also called topic prompt string. The sentences containing this kind of deixis are called topic prompt sentences. They are the preferred candidate sentences for summary. Such as ''the author thinks, the final summary'' and so on. Such indicators should be given a high weight.

2) Topic deixis of paragraph (Iseg): the sentence containing this kind of deixis is usually the central sentence of the

paragraph. The topic of the paragraph should be given a certain weight, such as ''central task'', ''core idea'', etc

3) Subject deixis (Isen): these words usually modify a sentence to emphasize something or an event in the sentence. Make the thing or event the subject of the sentence or the focus of the sentence, such as ''very, special''. In a sentence, the stressed part is often more important than the general sentence. Therefore, this kind of word can be given a smaller weight value to reflect the emphasis.

For a statement S_i in the article, the calculation method of sentence indicator features is as followed formula:

$$Score_{Indic}(s_i) = \frac{3 \times N_{Int} + 2 \times N_{Iseg} + N_{Isen}}{3} \quad (7)$$

In the formula: s_i - a sentence in an article

N_{Init} - the number of topic indicator in a sentence

N_{Iseg} - the number of paragraph topic indicator in a sentence

N_{Isen} - the number of sentence topic indicator in a sentence

According to the features of news texts, the importance of each feature in a sentence varies. In view of this, it is assumed that the weight of sentence position is W_1, the weight of keyword feature is W_2, the weight of similarity feature with title is W_3, the weight of cumulative similarity feature is W_4, and the weight of indicator feature is W_5. These features are combined to calculate the comprehensive weight of sentence.

C. USE GA TO ASSIGN WEIGHTS TO NEWS TEXT FEATURES

After extracting the text features from original news text, the importance of each feature is determined by allocating weight to the text features [9].For different news articles, it is difficult to find the most suitable feature weight value.In this paper, Genetic algorithm is used to adjust the weight value to get the best function suitable for news articles as algorithm 1.

Algorithm 1 The Adjust Weights to the Features

1.Input document D: D = { $s_1, s_2, s_3, \ldots s_m$}

2.Calculating Features: For each sentence(S_i), there are a set of features F, F = {f_{Pos}, f_{Key}, f_{Title}, f_{Sim}, f_{Indic}}.

3.Use GA to score weights of the features: There are weights w_j correspond to each features.

4.WF:Assign a fair weight to each feature and store it in a set WF, WF = {w_1f_{Pos}, w_2f_{Key}, w_3f_{Title}, w_4f_{Sim}, $\ldots$ w_5f_{Indic}}.

Genetic Algorithm, also known as evolutionary Algorithm. Genetic algorithm (GA) is a heuristic search algorithm inspired by Darwin's theory of evolution. Its main characteristic is directly on the object structure, so different from other algorithms for solving the optimal solution, the genetic algorithm and continuity of function limit, there is no derivative method of probability optimization method, do not need to make sure the rules can automatically acquire and to guide the optimization of search space, adaptively adjust the search direction. It has the advantages that GA does not have too many mathematical requirements for the optimization

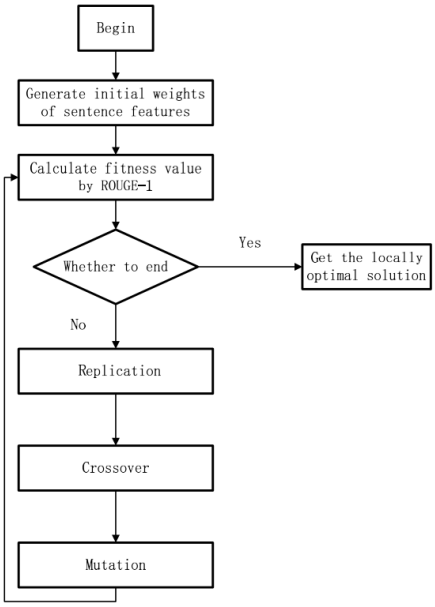

FIGURE 2. Use GA to find most suitable weights of the news text features.

problem. And because of its evolutionary characteristics, the search process does not need the inherent properties of the problem, for any form of objective function and constraints, whether linear or nonlinear, discrete or continuous can be processed. However, since the optimization process of GA is stochastic convergence, the optimized result of GA is likely to be a local optimal solution rather than a global optimal solution. For this, we use the restart method to start the genetic algorithm several times and the evaluation method (ROUGE-1) is used to evaluate the quality of the summary. Then select the weight distribution with the best effect as the optimal solution.

We selected 300 documents from the DUC2002 as training to form the dataset. The evaluation method (ROUGE-1) is used fitness function to show whether the weight used is the best weight. In addition, the parameters of the algorithm (crossover and mutation) are set as follows: crossover rate is 70''%'', variation is 0.2. And also, the multi start GA is used so that GA was run five times. Eq(9) is the adjusted combination.

$$ROUGE-1 = \frac{\sum_{s\in\{ReferemceSummaries\}} \sum_{gram_1\in S} Count_{match}(gram_1)}{\sum_{s\in\{ReferemceSummaries\}} \sum_{gram_1\in S} Count(gram_1)} \quad (8)$$

$$GA - weight(s_i) = \sum_{j=1}^{5} w_j \times score_{feature}(s_i) \quad (9)$$

In the Eq(9), GA weight(s_i) is the score of the sentence s_i, W_j is the weight of the sentence feature j, j = 1-5 is the number of sentence features and $score_{feature}(s_i)$ is the

score of the feature j which is given to sentence s_i. Now, optimized weights of $W = W_1, W_2, W_3, W_4, W_5$ is used to adjust the feature scores of F = Position, Keyword, Titleliked, Similarity, Indication. In the end, a set of adjusted feature scores of $WF = W_1 * Score_{Pos}(s_i) + W_2 * Score_{Key}(s_i) + W_3 * Score_{Title}(s_i) + W_4 * Score_{Sim}(s_i) + W_5 * Score_{Indic}(s_i)$ is obtained. The formula for calculating the score of a sentence s_i is Eq.(10).

$$\begin{aligned} Score(s_i) = {} & W_1 \times Score_{Pos}(s_i) + W_2 \times Score_{Key}(s_i) \\ & + W_3 \times Score_{Title}(s_i) + W_4 \times Score_{Sim}(s_i) \\ & + W_5 \times Score_{Indic}(s_i) \end{aligned} \tag{10}$$

D. FUZZY LOGIC ON NEWS TEXT SUMMARIZATION

Automatic summarization, just as which name impling is the process to make the machine automatically generated by imitating human based, is one of the earliest artificial intelligence thought way. At the same time, fuzzy logic refers to imitating the thinking mode of human brain's uncertain concept judgment and reasoning, Fuzzy thinking is a uniquely human way of thinking, so automatically in the process of generation, human not only prepared machine learning algorithms, but also can determine the generation of the importance of each sentence. This method, known as the fuzzy method, prevents conflicting training data. As we say in section 3.2, genetic algorithms was used to train the data to find the optimal weight for each feature. After the training, we assign optimal and fair weights to each feature as $W = W_1, W_2, W_3, W_4, W_5$ and then the assigned weights are used to calculate the scores for these features $WF = W_1 * Score_{Pos}(s_i) + W_2 * Score_{Key}(s_i) + W_3 * Score_{Title}(s_i) + W_4 * Score_{Sim}(s_i) + W_5 * Score_{Indic}(s_i)$as the input to fuzzy logic systems, and more accurate sentences can be obtained, resulting in high-quality summaries. In addition, human fuzzy logic expressions are used in if-then fashion.This paper adopts Mamdani fuzzy inference method to implement fuzzy rules in Matlab fuzzy logic toolbox [37].

Furthermore, the trapezoidal membership function is a nonnegative continuous function and is used to determine the degree and input values that belong to each appropriate fuzzy set. We use high, medium and low fuzzy sets to represent different degrees of fuzzy. The trapezoidal membership function, as its name implies, is expressed as trapezoid, so the function consists of four parameters(a,b,c,d) to determine the trapezoid breakpoint and two parameters(i,j) to determine the fuzzy set as Eq.(11). The parameter a,b,c and d determine the positions of two upper angles and two base angles of the trapezoidal membership function. As Fig.3 shows, a whole fuzzy controller system contains a fuzzyfication interface, a knowledge base (fuzzy data-base and fuzzy rule-base), an fuzzy inference engine and a defuzzyfication interface. Given the input and output variables of the fuzzy logic control system, fuzzification converts the input value of the fuzzy controller into the membership value of the corresponding fuzzy set as the result of fuzzification.In addition, the output of the trapezoid membership function is the fuzzy

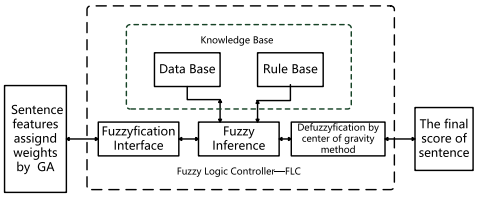

FIGURE 3. Fuzzy logic controller.

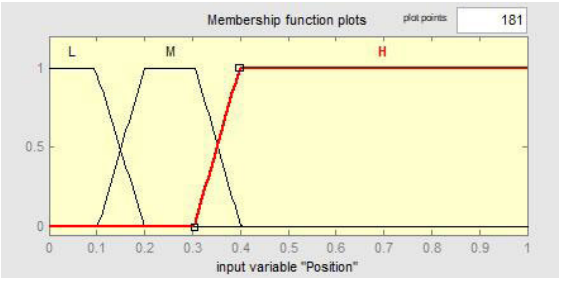

FIGURE 4. The sentence feature (Position) in trapezoidal membership functions.

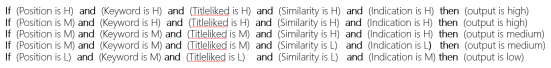

FIGURE 5. Some of the fuzzy rules.

membership of the fuzzy set (within [0-1]). Fig.4 shows the input value of sentence feature (Position) in trapezoidal membership functions.

$$A_{ij}(x_j) \begin{cases} \dfrac{x_j - a_{ij}}{b_{ij} - a_{ij}} & if\ a_{ij} \le x_j < b_{ij} \\ 1 & if\ b_{ij} \le x_j < c_{ij} \\ \dfrac{d_{ij} - x_j}{d_{ij} - c_{ij}} & if\ c_{ij} \le x_j < d_{ij} \\ 0 & if\ d_{ij} \le x_j \end{cases} \tag{11}$$

In that $a \leqslant b \leqslant c \leqslant d$ must hold.

At the beginning, fuzzy logic system is used in the control system, which is also known as the expert system in the control system. More importantly, it builds the rule base based on the expert experience and knowledge. The fuzzy logic rule library of 150 IF-THEN rules in this paper was built with the help of my tutor based on his experience and knowledge in the research of natural language processing language processing. And we did the experimental by modifying the rules several times to calculate the quality of the summary generated to achieve the best effect. The Fig. 5 show some of these rules. Fig. 6 persents the output in trapezoidal membership functions. Defuzzification is to convert the fuzzy value in fuzzy logic system into the final score of sentence by using centriod method [39] in Eq.(12) which takes the barycenter of the area enclosed by the curve of membership function and abscissa as the final output value of fuzzy reasoning (a fragile number). In Eq.(12), k is the number of samples used by the function, v represents the output fuzzy set, u_v represents

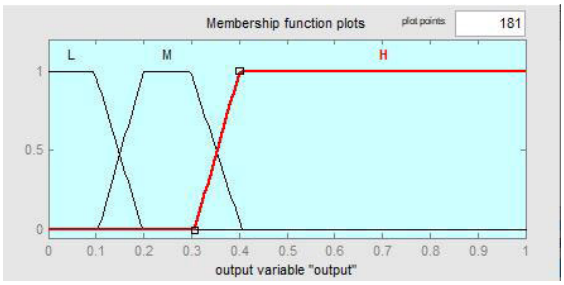

FIGURE 6. The output in trapezoidal membership functions.

the membership function on the fuzzy set and then V_0 is the value of the sentence applying fuzzy rules. We show news text summarization based on Fuzzy-GA algorithm as algorithm 2.

$$V_0 = \frac{\int_k v u_v(v) dv}{\int_k u_v(v) dv} \tag{12}$$

Algorithm 2 The Fuzzy-GA Text Summarization Algorithm

Input a news text T: Take the text T as a input
1. Preprocessing:
1.1 Division and integration by NLTK [40].
1.2 Use NLTK to eliminate stop words.
1.3 Stem extraction based on NLTK and porter algorithms.
2. Features Extraction: extract word features and sentence feature.
3. Assign weights: allocating suitable weights to each text feature by GA.
4. Calculate sentence scores: use fuzzy logic system to calculate the score of each sentence. This step is divided into three parts:
4.1 (1) Take the input fuzzification, (2) Fuzzy Inference and (3) Take the output defuzzification
5. Summary generation:
5.1 Sort the sentences by fraction: Choose the n sentences with the highest score.
5.2 Arrange the selected sentences in the order of the original text.

IV. EXPERIMENTAL DESIGN

A. DATASET

In order to evaluate the proposed method, selected documents are the Document Understanding Conference (DUC) for 2002. The DUC dataset contains many different text types, including 30 sets of single newswire/newspaper document. These 60 sets of single newswire/newspaper document as the dataset and divide them into two groups: (1) training dataset and (2) test dataset. The training dataset contains 30 sets and each set includes nearly 10 documents meanwhile each document is written summaries by 5 human experts. The test dataset is also consisted of 30 document sets.

B. PREPROCESSING

Preprocessing is the process of cleaning up and preparing news texts before generating summaries. The key information we need in online web news is often surrounded by a large amount of noise information, so it is a very important process to remove the useless information that has no effect on the original news text.The news text preprocessing process is mainly had four steps: division, integration, cleaning and stem extraction. In this paper, we use Natural Language Toolkit (NLTK) to preprocess news text.

Separating words and sentences is the first step in preprocessing news text. In this step, division and interation are done by NLTK. NLTK allows word segmentation of news text through whitespace. Therefore, NLTK can identify the edge between sentences by various punctuation marks, such as commas, periods, etc.

C. EVALUATION METHOD

In this part, we use ROUGE measure to evaluate the quality of the summary generated by the proposed method. The full name of ROUGE is Recall-Oriented Understudy for Gisting Evaluation which proposed by Chin-Yem Lin in 2004. The core idea of the ROUGE method is that the summaries of original text are wrote by several human experts as a standard set of summaries. The quality of the summary is evaluated by comparing the automatic summary generated by the system with the standard summary generated by human experts, and counting the number of overlapping basic units (n-gram grammar, word sequence and word pairs) between the two.The stability and robustness of the evaluation system can be improved by comparing the artificial summarization of several experts. This method has become one of the general criteria for evaluating techniques. The main criteria for ROUGE are ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W, ROUGE-SU, etc. Eq.(13) presents the formula of this method(ROUGE-N).

$$ROUGE-N = \frac{\sum_{s\in\{ReferemceSummaries\}} \sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{s\in\{ReferemceSummaries\}} \sum_{gram_n\in S} Count(gram_n)} \tag{13}$$

In the equation, n on behalf of the length of the n-gram, $Count_match(gram_n)$ represents the number of n-gram that appear in both the candidate summaries and reference summaries, $Count(gram_n)$ represents the number of n-gram in the reference summaries, ROUGE-N is the method based on recall rate, so the denominator is the number of all n-gram in the reference summary set.

D. COMPARISON WITH OTHER METHODS

- Msword [44]: Msword is a basic summary generation tool, and as a benchmark summary generation method, most researchers used for comparing with their proposed method.

• SDS-NNGA [46]: This work presentd a method of single document text summary generation method based on artificial neural network and genetic algorithm.

• System: System19 [47], System21 [48] and System31 [49] are three systems selected from participating systems for each task.System19 proposes a method for extracting and abstracting summary generators from single and multiple documents in GISTEXTER. System 31 is an automatic English informatve titles generation method based on hidden Markova (HM) model.

• Ranking SVM [50]: The semantic features between words and sentences in news texts are calculated by mutual information [51], the sentences are divided into topics according to their correlation degree, and the topic sentences are given higher weight. At the same time, a variety of combined features are extracted from the text, and the sentences are sorted by SVM, so as to get the automatic summary.

• GCD [57]: Through introduced the notion of a general context combined with RNN and propose a model for summarization based on it.

• SOM [58]: This method employ the concept of multiobjective optimization in microblog summarization to produce good quality summaries. And uses the search capability of multi-objective differential evolution technology, different statistical quality measures were optimized at the same time, namely the length of tweets, TF-IDF score, anti-attenuation, and different aspects of the measurement summary.

V. EXPERIMENTAL RESULTS

In this paper, the performance of this method was evaluated by ROUGE-1 and ROUGE-2.

Table 1 shows the comparison between this method and other methods in ROUGE-1 and ROUGE-2, where rank1 and rank2 represent the ranking of these methods in ROUGE-1 and ROUGE-2. From the results shown in table 1, we can see that MFG shows higher performance than other methods in both ROUGE-1 and ROUGE-2. In ROUGE-1 assessment method, GCD, SOM, SDS-NNGA, System21, System31, System19, Ranking SVM and Msword ranked second, third, fourth, fifth, sixth, seventh and eighth respectively. In ROUGE-2 assessment method, SOM, GCD, SDS-NNGA, System21, System19, Ranking SVM, System31 and Msword ranked second, third, fourth, fifth,sixth, seventh and eighth respectively.

Eq.(14) is to calculate the improvement rates of MFG over other methods in ROUGE-1 and ROUGE-2, in that, RM represents the recommended method and OM represents the other methods. We discribed the improvement rates in Table 2. As Table 2 shows, MFG improves performance in Msword, GCD, SOM, System19, System31, SDS-NNGA, System21 and RankingSVM on ROUGE-1 by 18.28%, 1.3%, 3%, 13.17%, 5.28%, 4.73%, 7.17% and 13.18% respectively. MFG improves performance in Msword, GCD, SOM, System19, System31, SDS-NNGA, System21 and RankingSVM on ROUGE-2 by 100.69%, 5.46%, 2.87%, 11.39%, 10.14%, 9.04%, 20.76% and 18.35% respectively. Fig. 7 shows the same results of ROUGE-1 and ROUGE-2 in Table 2 respectively.

TABLE 1. The results of proposed method compared with other methods by using ROUGE-1 and ROUGE-2.

Method	ROUGE-1	Rank1	ROUGE-2	Rank2
MFG	0.48932	1	0.21969	1
GCD	0.48303	2	0.20832	3
SOM	0.47506	3	0.21357	2
Msword	0.41369	9	0.10947	9
System19	0.43237	7	0.19722	6
System21	0.46477	5	0.19946	5
SDS-NNGA	0.46722	4	0.20147	4
System31	0.45657	6	0.18167	8
Ranking SVM	0.43235	8	0.18563	7

TABLE 2. The improvement rates of recommended method(MFG)(%).

Method	ROUGE-1	ROUGE-2
Msword	18.28	100.69
GCD	1.3	5.46
SOM	3	2.87
System19	13.17	11.39
System21	5.28	10.14
SDS-NNGA	4.73	9.04
System31	7.17	20.76
Ranking SVM	13.18	18.35

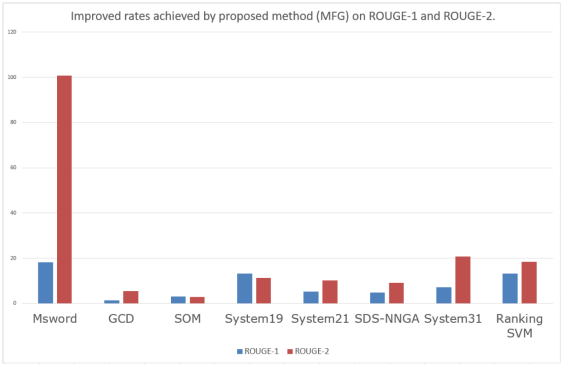

FIGURE 7. The improve rates of proposed method(MFG) on ROUGE-1 and ROUGE-2.

$$\frac{RM - OM}{OM} \times 100 \tag{14}$$

However, Tables 1 and Table 2 do not specify which approach achieves the best performance with ROUGE-1 and ROUGE-2 on dataset. Therefore, we use Eq.(15) to calculate the cumulative result ranking of these methods, where n represents the total number of all methods(n = 9) and t is the number of times this method has ranked i(1-9)th. Table 3 shows the ranking matrix for each method and table 4 shows the cumulative result ranking calculated by using the ranking matrix and Eq.(15) between MFG and other methods.

$$CumulativeResult = \sum_{i=1}^{n} \frac{(n-i+1)t}{n} \tag{15}$$

Based on Table 4, 8 levels were created for these methods, and we divided them into 4 groups based on the limitations of

TABLE 3. The ranking matrix for each method.

Method	rank								
	1	2	3	4	5	6	7	8	9
MFG	2	0	0	0	0	0	0	0	0
Msword	0	0	0	0	0	0	0	0	2
GCD	0	1	1	0	0	0	0	0	0
SOM	0	1	1	0	0	0	0	0	0
System19	0	0	0	0	0	1	1	0	0
System21	0	0	0	0	2	0	0	0	0
SDS-NNGA	0	0	0	2	0	0	0	0	0
System31	0	0	0	0	0	1	0	1	0
Ranking SVM	0	0	0	0	0	0	1	1	0

TABLE 4. Cumulative results rank calculated by Eq.(15).

Method	comparison rank	rank
MFG	2	1
Msword	0.22	8
GCD	1.67	2
SOM	1.67	2
System19	0.78	5
System21	1.11	4
SDS-NNGA	1.33	3
System31	0.67	6
Ranking SVM	0.56	7

TABLE 5. The results rank calculated by Eq.(15).

Set	Limit	Method
1	$1.5<RR\leq 2.0$	MFG,GCD,SOM
2	$1.0<RR\leq 1.5$	System21,SDS-NNGA
3	$0.5<RR\leq 1.0$	System31,System19,Ranking SVM
4	$0.0<RR\leq 0.5$	Msword

the result levels in Table 5. According to the results in table 5, the following situations can be observed:

• Set 1:In this set, MFG, GCD and SOM are three best performing methods of all, but the MFG method is superior to GCD and SOM in both ROUGE-1 and ROUGE-2. As shown in Table 1, in ROUGE-1, GCD, SOM ranks second and third, in ROUGE-2, SOM, GCD ranks second and third. GCD is the method which based on recurrent neural network combined with the nation of geberal context. SOM is the method based on multiobjective optimization and mul- tiobjective differential evolution technique.

• Set 2: In this set, the System21 and SDS-NNGA methods have similar result levels and have similar performance to each other. However, SDS-NNGA performs better than System21. SDS-NNGA is the method which based on artificial neural network and genetic algorithm. System21 is one of the top methods of the system. As shown in Table 1, in ROUGE-1 and ROUGE-2, SDS-NNGA ranks fourth and Systen21 ranks sixth.

• Set 3: In this set, the performance of the System19, System31 and Ranking SVM result grades is roughly the same. However, System19 has a better score than System31 and Ranking SVM. As shown in Table 1, System19 ranks seventh in ROUGE-1 and sixth in ROUGE-2. System31 ranks sixth in ROUGE-1 and eighth in ROUGE-2. Ranking SVM ranks eighth in ROUGE-1 and seventh in ROUGE-2. System19 proposes method in GISTEXTER to extract and abstract summarizer in both single and multiple documents. System 31 is a method based on Hidden Markova (HM) model to automatic generation of informative headlines for English texts. Ranking SVM is the method based on mutual information and ranking SVM.

• Set 4:In this set, the Msword obtains the lower rank(9) in all 9 methods. It gets rank 9 in ROUGE-1 and ROUGE-2. Msword is a benchmark method.

VI. DISCUSSION

The method mentioned in this paper is either a recently proposed summary generation method or a representative method in the development of summary generation technology. Table 1 shows the results of ROUGE-1 and ROUGE-2 for these methods on the dataset in both cases. Later, Table 2 shows the improvement rates of MFG for other methods. In addition, MFG has made appropriate improvements to the methods of Msword, GCD, SOM, System19, System31, SDS-NNGA, System21 and RankingSVM by 13.66%, 1.3%, 3%, 10.63%, 1.1%, 0.48%, 3.81% and 11.67% in the measure of ROUGE-1 respectively. Moreover, it has significant improvement by 27.26%, 5.46%, 2.87%, 11.87%, 2.55%, 0.04%, 12% and 110.18% in Msword, GCD, SOM, System19, System31, SDS-NNGA, System21 and Ranking SVM, in ROUGE-2, respectively. As we have seen, by extracting the key features of the news text and combining the fuzzy logic system with the genetic algorithm to assign more optimized weights to the text features, more accurate summary sentences can be obtained to generate higher quality summaries. The performance of MFG is significantly higher than that of other methods.

VII. CONCLUSION

In this paper,we proposed a new model based on Multi-feature, genetic algorithm and fuzzy logic for news text summarization. First, extract the most important features (word features and sentence features) and use a linear combination of these features to identify important sentences. In this step, we choosed a new text feature extraction method based on the characteristics of news text. Firstly, each word is extracted with features. Considering capital letters and words appearing in news headlines can reflect the central content of the article, these two features are added to the word features for extraction, and each word is graded according to the different features.At the same time, according to the characteristics of news text writing, the three news elements are directly extracted as news keywords. Then, each sentence is graded on this basis. When selecting sentence features, the sentence similarity features and the similarity with the title are taken into account. Meanwhile, the deixis features contained in the sentence are added to make the feature selection more in line with the writing features of news text. Then, fuzzy logic system introduced combined with the genetic algorithm is used to generate appropriate and fair weights for text features according to their importance. In GA, fitness values

are calculated by using ROUGE-1. In fact, ROUGE-1 is a fitness finction that evaluating the quality of the generated summaries by counting the number of base units that overlap each other, as opposed to the standard summaries generated manually. Meanwhile we start the genetic algorithm several times to select the optimal weight allocation to avoid falling into the local optimal result. Finally, the sentence feature vector that allocates the weight given to fuzzy inference system as inputs, so we are able to get more precise sentence which can result in creation of high quality summaries. In fuzzy logic, three steps are adopted: (1) fuzzification, (2) inference and (3) defuzzification. In addition, in reasoning proccess, about 150 if-then rules are defined by human experts and three fuzzy sets are used: low, medium, and high. The input of fuzzy logic system is a linear combination of extracted news text features and their corresponding reasonable weights obtained by the genetic algorithm. After the scores of all sentences are obtained through the fuzzy reasoning system, they are arranged in descending order according to the scores of sentences. Finally, choose the first n sentences as the summarization, where n is equal to the compression rate. The performance of this method and other methods was evaluated by experiments. Therefore, by using ROUGE-1 and ROUGE-2, proposed method was compared to other methods (Msword, GCD, SOM, System19, System31, SDS-NNGA, System21, and Ranking SVM) on dataset. As we can see from Table 2, experimental results show that this method performs better than other methods in the quality of news summary generation.

REFERENCES

[1] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[2] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016.

[3] R. Nikhil *et al.*, "A survey on text mining and sentiment analysis for unstructured Web data," Tech. Rep., 2015.

[4] A. Porselvi and S. Gunasundari, "Survey on webpage visual summarization," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, pp. 26–32, 2016.

[5] I. Mani, *Automatic Summarization*. Amsterdam, The Netherlands: JohnBenjamins Publishing Company, 2001.

[6] A. S. Nengroo and K. S. Kuppusamy, "Machine learning based heterogeneous Web advertisements detection using a diverse feature set," *Future Gener. Comput. Syst.*, vol. 89, pp. 68–77, Dec. 2018.

[7] A. Sinha, A. Yadav, and A. Gahlot, "Extractive text summarization using neural networks," Tech. Rep., 2018.

[8] Y. Chali, S. A. Hasan, and S. R. Joty, "A SVM-based ensemble approach to multi-document summarization," in *Advances in Artificial Intelligence*. Kelowna, BC, Canada: Springer-Verlag, May 2009.

[9] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, Dec. 2004.

[10] A. Porselvi and S. Gunasundari, "Survey on Web page visual summarization," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 1, pp. 26–32, 2013.

[11] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research [review article]," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, May 2014.

[12] A. K. Joshi, "Natural language processing," *Science*, vol. 253, no. 5025, pp. 1242–1249, 1991.

[13] M. A. F. F. Ren, "Automatic text summarization," *Digithum*, vol. 4, no. 3, pp. 82–83, 2008.

[14] V. E. Abramov and N. N. Abramova, "Automated summarization of the volumineous documents (exemplified by dissertation materials)," in *Proc. 22nd Int. Crimean Conf. Microw. Telecommun. Technol.*, Sep. 2012, pp. 425–426.

[15] T. Nasukawa, "Text analysis and knowledge mining," in *Proc. 8th Int. Symp. Natural Lang. Process.*, Oct. 2009, pp. 967–984.

[16] C. Carpineto, S. Osiáski, G. Romano, and D. Weiss, "A survey of Web clustering engines," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–38, Jul. 2009.

[17] F. Kyoomarsi, H. Khosravi, E. Eslami, P. K. Dehkordy, and A. Tajoddin, "Optimizing text summarization based on fuzzy logic," in *Proc. 7th IEEE/ACIS Int. Conf. Comput. Inf. Sci. (ICIS)*, May 2008, pp. 347–352.

[18] E. Lloret and M. Palomar, "A gradual combination of features for building automatic summarisation systems," in *Proc. 12th Int. Conf. Text, Speech Dialogue*. Berlin, Germany: Springer-Verlag, 2009, pp. 16–23.

[19] S. M. Weiss, N. Indurkhya, and T. Zhang, "Text mining," Tech. Rep., 2010.

[20] "Text mining," *Encyclopedia of Social Network Analysis & Mining*, vol. 31. pp. 91–99.

[21] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *Int. J. Artif. Intell. Tools*, vol. 13, no. 1, pp. 157–169, Mar. 2004.

[22] C.-H. Chen, "Improved TFIDF in big news retrieval: An empirical study," *Pattern Recognit. Lett.*, vol. 93, pp. 113–122, Jul. 2017.

[23] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, "Text summarization features selection method using pseudo genetic-based model," in *Proc. Int. Conf. Inf. Retr. Knowl. Manage.*, Mar. 2012, pp. 193–197.

[24] M. F. Liu, L. Wang, and L. Q. Nie, "Weibo-oriented Chinese news summarization via multi-feature combination," in *Proc. Conf. Natural Lang. Process. Chin. Comput.*, Nanchang, China, 2015, pp. 581–589.

[25] C.-S. Lee, Z.-W. Jian, and L.-K. Huang, "A fuzzy ontology and its application to news summarization," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 35, no. 5, pp. 859–880, Oct. 2005.

[26] Q. A. Al-Radaideh and D. Q. Bataineh, "A hybrid approach for Arabic text summarization using domain knowledge and genetic algorithms," *Cognit. Comput.*, vol. 10, no. 4, pp. 651–669, Aug. 2018.

[27] S. Malhotra and A. Dixit, "An effective approach for news article summarization," *Int. J. Comput. Appl.*, vol. 76, no. 16, pp. 5–10, Aug. 2013.

[28] C. Y. Lin and E. Hovy, "Identifing topics by position," in *Proc. 5th Appl. Natural Lang. Process. Conf.* NJ, USA, 1997, pp. 283–290.

[29] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Develop.*, vol. 2, no. 2, pp. 159–165, Apr. 1958.

[30] H. P. Edmundson, "New methods in automatic extracting," *J. ACM*, vol. 16, no. 2, pp. 264–285, Apr. 1969.

[31] K. S. Leung and W. Lam, "Fuzzy concepts in expert systems," *Computer*, vol. 21, no. 9, pp. 43–56, Sep. 1988.

[32] P. Angelov, *Evolving Fuzzy Systems*. New York, NY, USA: Springer, 2009.

[33] P. P. Angelov and X. Zhou, "Evolving fuzzy-rule-based classifiers from data streams," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 6, pp. 1462–1475, Dec. 2008.

[34] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion, "A text summarization method based on fuzzy rules and applicable to automated assessment," *Expert Syst. Appl.*, vol. 115, pp. 264–275, Jan. 2019.

[35] F. Kyoomarsi, H. Khosravi, E. Eslami, P. K. Dehkordy, and A. Tajoddin, "Optimizing text summarization based on fuzzy logic," in *Proc. 7th IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, May 2008, pp. 347–352.

[36] Z. Xia, Y. Yixin, X. Mingzhu, and Z. Hailong, "Research of text clustering based on fuzzy granular computing," in *Proc. 2nd IEEE Int. Conf. Comput. Sci. Inf. Technol.*, Aug. 2009, vol. 46, no. 13, pp. 288–291.

[37] S. N. Sivanandam, S. Sumathi, and S. N. Deepa, *Introduction to Fuzzy Logic Using MATLAB*, 1st ed. New York, NY, USA: Springer-Verlag, 2006.

[38] M. Li, L. Liu, and C.-B. Li, "An approach to expert recommendation based on fuzzy linguistic method and fuzzy text classification in knowledge management systems," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8586–8596, Jul. 2011.

[39] F. Kyoomarsi, F. Rahmani, E. Eslami, P. K. Dehkordy, and A. Tajoddin, "Text summarization based on cellular automata," in *Proc. Int. Conf. Inf. Commun. Manage.*, Singapore, 2011, pp. 1–7.

[40] S. Bird and E. Loper, "NLTK: The natural language toolkit," in *Proc. Acl-Workshop Effective Tools Methodologies Teach. Natural Lang. Process. Comput. Linguistics*. Stroudsburg, PA, USA: Association Computational Linguistics, 2004, pp. 69–72.

[41] Y. K. Meena and D. Gopalani, "Evolutionary algorithms for extractive automatic text summarization," *Procedia Comput. Sci.*, vol. 48, pp. 244–249, 2015.

[42] A. Abuobieda, N. Salim, Y. J. Kumar, and A. H. Osman, "An improved evolutionary algorithm for extractive text summarization," in *Intelligent Information and Database Systems*. Berlin, Germany: Springer, 2013.
[43] A. Abuobieda, N. Salim, Y. J. Kumar, and A. H. Osman, "Opposition differential evolution based method for text summarization," in *Proc. 5th Asian Conf. Intell. Inf. Database Syst.* Berlin, Germany: Springer, 2013, pp. 487–496.
[44] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy swarm diversity hybrid model for text summarization," *Inf. Process. Manage.*, vol. 46, no. 5, pp. 571–588, Sep. 2010.
[45] S. Dutta, V. Chandra, K. Mehra, A. K. Das, T. Chakraborty, and S. Ghosh, "Ensemble algorithms for microblog summarization," *IEEE Intell. Syst.*, vol. 33, no. 3, pp. 4–14, May 2018.
[46] N. Chatterjee, G. Jain, and G. S. Bajwa, "Single document extractive text summarization using neural networks and genetic algorithm," Tech. Rep., 2018.
[47] S. M. Harabagiu and F. Lacatusu, "Generating single and multi-document summaries with GISTEXTER," in *Proc. Workshop Text Summarization*, Philadelphia, PA, USA, 2002, pp. 1–9.
[48] X. Wan, "Towards a unified approach to simultaneous single-document and multi-document summarizations," in *Proc. 23rd Int. Conf. Comput. Linguistics Coling*, Beijing, China, 2010, pp. 1137–1145.
[49] D. Zajic, B. Dorr, and R. Schwartz, "Automatic headline generation for newspaper stories," in *Proc. Workshop Text Summarization*, 2002, pp. 11–12.
[50] L. I. Mengshuang, Z. Hongying, and J. Huizhen, "Micro-blog-oriented chinese news summarization based on multi-feature and ranking SVM algorithm," *J. Zhengzhou Univ. (Natural Sci. Ed.)*, vol. 2, p. 8, 2017.
[51] H. Huo and X. H. Liu "Automatic summarization based on mutual information," *Appl. Mech. Mater.*, vol. 513, pp. 1994–1997, 2014.
[52] P. Yang, W. Li, and G. Zhao, "Language model-driven topic clustering and summarization for news articles," *IEEE Access*, vol. 7, pp. 185506–185519, 2019.
[53] Y. Ma, P. Zhang, and J. Ma, "An ontology driven knowledge block summarization approach for Chinese judgment document classification," *IEEE Access*, vol. 6, pp. 71327–71338, 2018.
[54] H. Xu, Z. Wang, and X. Weng, "Scientific literature summarization using document structure and hierarchical attention model," *IEEE Access*, vol. 7, pp. 185290–185300, 2019.
[55] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," *Comput. Sci.*, 2013.
[56] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*. [Online]. Available: http://arxiv.org/abs/1310.4546
[57] H. Kim and S. Lee, "Document summarization model based on general context in RNN," *J. Inf. Process. Syst.*, vol. 15, no. 6, pp. 1378–1391, 2019, doi: 10.3745/JIPS.02.0123.
[58] N. Saini, S. Saha, and P. Bhattacharyya, "Multiobjective-based approach for microblog summarization," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 6, pp. 1219–1231, Dec. 2019, doi: 10.1109/TCSS.2019.2945172.

YAN DU is currently pursuing the M.S. degree with the School of Information Engineering, Henan University of Science and Technology, China. His current research interests include pattern recognition, machine learning, and natural language processing.

HUA HUO received the Ph.D. degree from the Laboratory of Intelligent Computing and Application Technology for Big Data, School of Software, Information Engineering College, Henan University of Science and Technology. He is currently a Professor with the Laboratory of Intelligent Computing and Application Technology for Big Data, School of Software, Information Engineering College, Henan University of Science and Technology. His research interests are in intelligent information processing and video semantic extraction.