

Automatic Vocabulary Grouping and Deep Combination for News Credibility and Reliability Evaluation Corresponding to Specific Language

Ming-Shen Jian^{*}, Rong-Bin Deng^{**}, Chen-Wei Fang^{***}, Hua-Yu Wu^{****}, Wen-Hsiang Hsieh^{*****}

Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University

Yunlin County, Taiwan 632

jianms@nfu.edu.tw, 40743162@gm.nfu.edu.tw, 40743206@gm.nfu.edu.tw, 40743212@gm.nfu.edu.tw, 40743160@gm.nfu.edu.tw

Abstract—Considering the spread of the intentional false news, to be aware of the new news is important. In this research, the Automatic Deep Vocabulary Grouping and Combination for News Credibility and Reliability Evaluation is proposed which includes Key Vocabularies Merging Method and False News Warning Method. By merging different algorithms, the proposed system could cluster and group the fake news according to the found features of various fake news content. The proposed system could find and verify the features of the specific intentional false news according to the proposed deep combination evaluation function. After suitable verification and recursive checking, the minimum groups for clustering the fake news could be given. Based on the collected features of the various intentional false news, the new coming news can be classified as the false news due to the feature matching percentage. According to the verification, all the collected and proved fake news could be found and clustered into the corresponding fake news groups.

Keyword—False News, Cluster, Web Crawler, Cloud Computing

Manuscript received August 23, 2021, and a follow-up of the invited journal to the accepted & presented paper entitled "Automatic Vocabulary Grouping and Deep Combination for News Credibility and Reliability Evaluation Corresponding to Specific Language" of the 21th International Conference on Advanced Communication Technology (ICACT2021)." This work is Final File 20210125_finalpaper.pdf and was supported in part by the Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University.

Ming-Shen Jian is with the Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University, Taiwan (corresponding author to provide phone: 886-922-916612; fax: 886-563-60306; e-mail: jianms@nfu.edu.tw).

Rong-Bin Deng is with the Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University, Taiwan

Chen-Wei Fang is with the Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University, Taiwan

Hua-Yu Wu is with the Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University, Taiwan

Wen-Hsiang Hsieh is with the Cloud Computing and Intelligent System Lab., Dept. of CSIE, National Formosa University, Taiwan

I. INTRODUCTION

DUE to the online community or social media, various information can be exchanged quickly. Without verification, all the information broadcasted between people may be under misrepresentation [1]. Considering the anonymous using or none-recognition social media user, false news or fictitious information could be made for specific purpose. The false news could spread faster than the true story or clarification news. Although most news can be searched and verified through the search engine such as Google [2], Yahoo [3], etc., to find the correct information from various sources and various news is still difficult.

There were different methods proposed to detect the false news based on the supervision [4]. To find and detect the false new, the content of the news could be the important factor [5]. In addition, to take the background or profiles of the user into consider could be also the solution for false news finding [6]. Based on the crowd-sourcing approach, the false news could be verified according to the public annotations or replies. However, when the false news are intentionally given or spread, public news readers may not differentiate true or false news [7]. Furthermore, if the false news is written maliciously, it's more difficult to classify the news due to some organizations or groups would intentionally spread the false news.

In addition, some languages would make the verification of news more difficult. For example, the possible combination of Chinese vocabularies would make the meanings completely different [11]. Currently, by using the specific keywords, the search engines would find the possible information for verification according to the statics of keywords. Hence, considering the probability of the keywords included in the found information or web pages, different keywords could be combined as the new potential vocabularies [8-10]. In other words, by adaptively combining the key words, the features of the news can be described and defined.

Considering the rapidly changing of the online news and information spread through the online media or social community, the feature keywords or vocabularies are changed quickly. Furthermore, the content of false news

could be various and huge. To continuously collect the data from social media is needed. In addition, to manage big data from the online media is also required. Corresponding to the big data management, the Hadoop/MapReduce [10,15,19,20] and Spark were proposed [25]. All the vocabularies are treated as the key. Then, the total among of the keys (vocabularies) appeared in the article or web page can be evaluated and counted as the value. Hence, each online article or news can be represented as (key, value) corresponding to all the vocabularies appeared.

According to the proposed research, by adaptively combining the found vocabularies, the feature of the similar online information can be found [10]. However, different false news would be various in different content. Some false news may appear only few hours, some may cause several days spreading with huge data. Therefore, to collect the huge data online for feature of the intentionally spread false news is an important issue.

In this research, the first step is to collect the proved false news from the third party impartial agency. To collect the verified data or information, the automatic data collection based on the online crawler is needed. Corresponding to the vocabulary combination procedure, the suitable length of keywords for the same false news could be found. In other words, the suitable features of the same feature can be found. By comparing different types of false news and the evaluation based on the dynamic multiple dimension K-mean algorithm, the features or groups which intentionally spread the false news could be found.

This paper is presented as follows. Section 2 gives the related works. Then, the proposed procedure for news credibility and reliability prediction is given in section 3. Section 4 provides the conclusion.

II. RELATED WORKS

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

These days, due to the online media including video streaming, social media, or online forum, people could receive various information more quickly. However, some intentional false news would cause damage from the health to the vote. Some research were proposed to differentiate the true and false news. In addition, corresponding to the various languages, the complete vocabulary in the sentence or article should be taken into consideration [10]. There were procedures proposed based on the language corpus [9,11,12]. According to the vocabularies stored in the corpus, the various vocabularies could be identified and recorded as the text-pattern. However, the information or new vocabulary comes every day. To automatically collect the information online is important.

Hence, in this research, the web crawler [13,14,20] is used to collect the online information included in the web pages. Based on the web crawler, the application could continuously collect the data from the Internet. From different search engines or web pages, the information or potential news can be gathered.

After obtaining the data from the Internet by crawler, the data or vocabularies should be evaluated. Since the total

amount of the collected data could be huge, to parallel process the big data based on the distributed system is needed. Hadoop/MapReduce was the data analysis application proposed by Apache or Google [10,15,19,20]. Since the data size from the Internet could be huge, the Hadoop Distributed File System or Google File System was proposed for dividing the huge data into blocks and distributed storing [16-19]. All the corresponding information about data storage can be managed based on the HBase, the database based on the cloud environment. The HBase is the noSQL type database. Hence, the length of each data can be various instead of the traditional fixed data length in database. In other words, although the size of information or data from different network sources could be different, the HBase on cloud could still provide the storage space and record it with better performance.

Suppose that to analyze the complete collected data is called Job. Then, the data will be divided into several partitions for parallel processing, called Tasks. Each job could be divided into multiple tasks. These tasks are mapped to multiple computing nodes or virtual machines [26]. After the computing and evaluation, each report of the task from the computing node could be sent back and summarized. Based on the Hadoop/MapReduce, each found vocabulary from the collected information or data will be represented as 'key'. The total amount of occurrences corresponding to this 'key' vocabulary can be counted and represented as the 'value'. In other words, the report related to the data after analysis could be represented as the set: {key, value}. Finally, according to the reduce procedure, the value with the same key will be added up. At last, only one report corresponding to each information collected from Internet could be given.

Since the content of the collected information from the Internet could be various, the data size after analysis would be different and huge. In addition, to execute the Hadoop/MapReduce program also requires huge computing resource. Hence, considering the computing ability, the cloud computing environment is needed. Via network connection and virtualization technology, the virtualized computing resource such as CPU, memory, and storage space, could be merged into the computing resource pool. When a user needs the computing resource, the computing resource in the resource pool can be integrated as the virtual machine for the user. If no more using the virtual machine, the computing resource can be returned to the resource pool for further using. Therefore, the computing resource for services such as IaaS, PaaS, and SaaS, can be rapidly used [27]. In other words, if the configuration and environment of the virtual machine is on demand defined, the user could obtain the cloud services such as SaaS more directly and quickly. Therefore, the cloud computing could be the solution to repeatedly and continuously analyze the information gathered from the Internet.

After analysis related to the collected information, the feature of the content should be found. In addition, there are various news, information, and data from Internet. The features between different information or data could be different. Therefore, to group the similar information or data from these collected data is important. Then, the possible features of this grouped data such as the false news makers or spreaders could be found. To find the possible relationship

between different data, the clustering algorithms based on K-means were proposed. According to the group analysis with vector quantization, different groups with different features could be found and divided by the following equation:

$$J = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (1)$$

Suppose that there are total k randomly selected group centres from μ_1 to μ_k . The all found features x are measured the distance between x and group centres. Finally, all the x which is included in the group S_i with the shortest distance between x and μ_i are called members of group i . In other words, when the minimized J is found, all the x could be individually classified to its nearest group with the centre μ .

III. PROPOSED SYSTEM

In this research, to continuously collect the proved news, especially proved false news, is very important for training the proposed system. Therefore, some specific online websites or news proof data sources are on demand recorded in the proposed web crawler procedure which is based on the HTTP protocol [20,21]. By using the selenium [22], Google search Engine [2], or Taiwan FactCheck Center [28], the corresponding news (proved false news) could be collected. Then, the false news spread through the websites, social media, forum, etc., could be further collected. In other words, the false spreading paths, media, content, etc., can be searched and found.

The false news studied in this research is spreading in Chinese language. Due to the characteristics of the Chinese, the vocabularies in the content should be divided according to the ambiguity resolution rules such as Maximum Matching [23] (mmsegj). Each chunk which includes the neighbour vocabularies would follow the rules: largest average and smallest variance of word length, largest degree of the summation related to morphemic freedom corresponding to the one-character word, and maximum matching [24]. Therefore, each independent data or information from web page or social media could be divided into multiple key vocabularies.

However, to find the key vocabularies, the huge data from the Internet should be managed. Hence, in this research, the Hadoop/MapReduce procedure is implemented based on the cloud environment. By looped mmsegj and Hadoop/MapReduce procedures, the largest average length of key vocabularies could be found. However, considering the intentionally spread false news today, new vocabularies, names, or new homophonic or metaphor vocabularies would appear. Therefore, in this research, the Key Vocabularies Merging method is proposed. By sorting, grouping and combining different found vocabularies according to the on demand defined limitations and considerations, the key vocabularies could be deeply searched through the verification via search engines. The flowchart can be shown as follows.

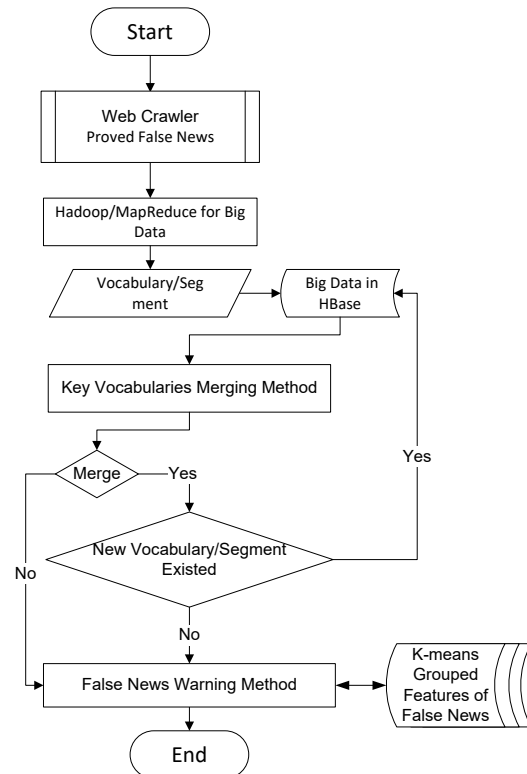


Fig. 1. The flowchart of the proved false news system training.

In this research, the proved false news in Chinese language could be collected from Taiwan FactCheck Center [28]. Since the news was already proved as the false news, the corresponding data or information could be the training data of the proposed system. According to the mmsegj and Hadoop/MapReduce procedures, the largest average length of vocabularies or frequently appeared specific vocabularies could be found and sorted. Then, the clustered and found vocabularies should be proved and verified as the features of the specific false news. In this research, the similar content of the proved false news from different news sources could be collected and analysed together. Therefore, these similar intentionally spread false news could be used to train the proposed system and cluster the possible features of the collected false news. To find the available and possible features of the similar false news, the Key Vocabularies Merging Method is proposed.

A. Key Vocabularies Merging Method

In this research, the Key Vocabularies Merging Method is proposed for extending the found key vocabularies. Based on the mmsegj algorithm for Chinese vocabulary recognition, the basic segments or simple vocabularies can be found. Considering the intentionally spread false news, the feature vocabularies may be combined from various found vocabularies. For example, the vocabulary could be combined from the name, location, food, even foreign language. Hence, in this research, the Key Vocabularies Merging Method is proposed for further extended key vocabulary searching.

After analysing the similar proved false news, the initial key vocabularies which represented and sorted according to value of the set, $\{\text{key}, \text{value}\}$, can be obtained. Then, suppose that there are total v found initial vocabularies or segments according to the mmsegj algorithm. The occurrences of the i^{th}

vocabulary which represented as v_i can be shown as $O_c(v_i) =$ value. Hence, the Key Vocabularies Merging could evaluate and decide to merge two vocabularies by following the decision function based on difference percentage shown as follows:

$$\text{Merge} = \begin{cases} \text{true} & \frac{|O_c(v_i) - O_c(v_j)|}{O_c(v_i)} \leq \alpha \\ \text{false} & \text{otherwise} \end{cases} \quad (2)$$

where $i \neq j$ and α is the on demand given value. If the result of the merge evaluation is true, the i^{th} vocabulary and the j^{th} vocabulary will be merged as the new vocabulary.

If there are new vocabulary or segment merged according to the proposed decision function, these new merged vocabulary or segment should be verified by the following two checking steps. First step, these new vocabulary will be tested through the search engines online including different language vocabulary checking. If the new merged vocabulary is existed in the general articles, then this new merged vocabulary should occur frequently. Second step, the new merged vocabulary should also occur in the all original proved similar false news. In other words, the occurrence of the new merged vocabulary should be satisfied the function shown as follows:

$$O_c(v_i) \geq O_c(\text{new merged vocabulary}) \geq O_c(v_j) \quad (3)$$

where the $O_c(v_i) \geq O_c(v_j)$. If these two steps can be satisfied, this new merged vocabulary or segment can be recorded. To find the longest feature vocabulary or segment, the found merged vocabulary and segment will be checked according to the proposed Key Vocabularies Merging Method again. When no new merged vocabulary can satisfy the two checking steps, the Key Vocabularies Merging Method stops.

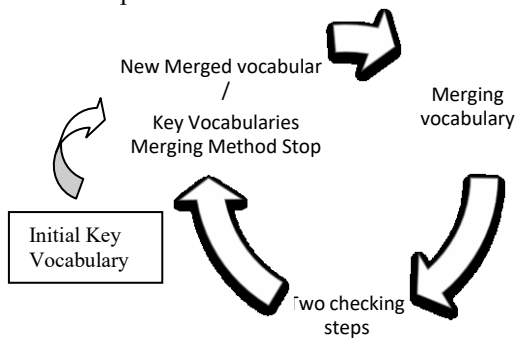


Fig. 2. The cycle of the proposed Key Vocabularies Merging Method.

Therefore, based on the proposed Key Vocabularies Merging Method, the key vocabulary can be extended. Suppose that the false news intentionally broadcasted would be written corresponding to the specific person, specific fake news or sentences, specific organizations, etc. Then, the probabilities of individual word or vocabularies should be very closer to another frequently appearing word or vocabulary. For example, the personal name "韓國瑜" in Chinese could be divided as two vocabularies: "韓國" and "瑜". The first vocabulary is the proper nouns. However, initially the complete personal name is not defined as the proper nouns in the database. Therefore, after several cycles of the proposed Key Vocabularies Merging Method, these two independent vocabularies, "韓國" and "瑜", could be found that the probability of appearance corresponding to

these two vocabularies are almost the same since originally it is a personal name. According to Key Vocabularies Merging Method, the key vocabulary "韓國" can be extended with another key vocabulary "瑜" as "韓國瑜". Especially, according to two checking steps of Key Vocabularies Merging Method, the new found key vocabulary can be proved as the new merged and useful key vocabulary as the feature of the false news. In addition, the new merged key vocabulary can be recorded into the Chinese proper nouns database.

Since it is difficult to find the new potential vocabularies excluding in the database, in this research, additional method: jieba algorithm, is used for extending the vocabularies. Jieba algorithm is used to find the possible and potential vocabulary segmentation of Chinese language [29].

Jieba algorithm is the method to find the vocabulary segmentation in Chinese [29]. In this research, to find the possible segment of Chinese words, two possible methods are used: 1) Rule Segmentation and 2) Statistical Segmentation.

Based on jieba algorithm, a vocabulary dictionary tree would be used for words comparing and matching by rule segmentation. In other words, if the segments are already recorded in the dictionary, the results of matching could identify the found segments. Considering the vocabularies used in the social community media, new created words are frequently appeared. It means that only matching the vocabulary tree by rule segmentation is not enough. Therefore, the Statistical Segmentation is also included for the new vocabularies or segments finding. By counting the appearance frequency of the potential new vocabularies, the potential new created vocabularies could be found due to the frequent using in the text content from various sources. In other words, with the statistical results, the high frequently appeared words could be identified and recognized as the new vocabulary. The new created new vocabulary will be recorded into the database for matching by Rule Segmentation in the future. Hence, the new created vocabularies used in the fake news or social media could be found and extended more automatically. In this research, the MapReduce for evaluating the various key segments is used. Since the fake news spread today would include the personal names or some homophonic vocabularies created by social media, the found potential new vocabularies would be various with huge amount. Therefore, to evaluate and obtain the statistical results of different text content from various sources would request huge computing resource. In this research, by following the MapReduce procedure, the statistical results could be obtained through counting the {Key, Value} format which Key is the potential vocabulary and the Value indicate the appearance times.

In addition to the Statistical Segmentation, the directed acyclic graph is used to structure the possible vocabularies according to the word graph scanning based on the prefix dictionary. The maximum probability of the dynamic programming path could be found. By using the Hidden Markov Models (HMM), the words excluded in the dictionary database could be learned and found. Hence, even the new words or vocabularies are created according to the pronounce or other non-traditional grammar method, the statistical results of the similar content from various news sources could still be used to identify and recognize the new

feature vocabularies.

After recursively processing the proposed Key Vocabularies Merging Method and the jieba algorithm with Hidden Markov Models (HMM), the hidden vocabularies and txt segmentations could be found and identified. The possible feature key vocabularies could be merged and extended by different segments.

B. Clustering Method

After finding the vocabularies and counting the appearance of each vocabulary, the features of the specific fake news should be found and defined. Since the similar fake news would have similar key vocabularies, these vocabularies could be used to cluster or group the fake news. In this research, term frequency-inverse document frequency (TF-IDF) method is used. Term frequency-inverse document frequency is a statistical method for information exploration and text mining through the weighting techniques [30,31]. Each found vocabulary will be evaluated for its importance of the text content or corpus. Suppose that the term frequency (TF) of the key vocabulary is high if this specific vocabulary is important in the fake news. In other words, if the words are not important in these fake news, the appearance frequency of the words is decrease in inverse proportion called inverse document frequency (IDF).

Due to the term frequency-inverse document frequency method, the important vocabularies could be filtered with less common words. In other words, after filtering multiple similar fake news, the specific key vocabularies could be possibly found since these specific key vocabularies would be important in the same or similar fake news. In other words, the similar fake news would include the similar specific key vocabularies. In this research, these similar or the same specific key vocabularies are called features of the specific fake news. Suppose that the features of different fake news related to various topics would be also different. Then, after term frequency-inverse document frequency method, the fake news with the similar key vocabularies or features should be grouped or clustered as the same topic or purpose of the fake news. In this research, the Scikit-learn [32,33] is used for enhance the accuracy of classification and clustering. Scikit-learn is the machine learning method which uses Numpy to efficiently perform the linear algebra and array operations. Scikit-learn provides various classification, regression and clustering algorithms, such as support vector machine, gradient boosting, term frequency-inverse document frequency, k-means clustering, DBSCAN, etc. In this research, the value k for the k-means clustering in eq. (1) should be assigned and given. To find the suitable value k, the Elbow Method [34] is used. Based on Elbow Method, the elbow point is used to find the minimized value of k which to be used for k-means algorithm. Hence, by recursively executing the Scikit-learn method and Elbow Method, the fake news which were already verified could be clustered and grouped automatically with less manual operation according to the found k. In other words, the when the value k is defined, the total numbers or the fake news groups or clusters is also defined. In addition, when the numbers or the fake news groups is defined, the key vocabularies related to each groups could be called as the features of the specific fake news group. In other words, after Scikit-learn method and Elbow Method,

the groups and the corresponding features are defined and found.

C. False News Warning Method

After Key Vocabularies Merging Method, the features of the proved similar false news can be found. By analysing various types of the proved false news, different features related to these various false news could be obtained. Therefore, based on the K-means algorithm, the intentional false news even the news sources can be grouped. After K-means grouping, the possible types of false news and possible spreading news sources could be found.

Therefore, based on the classification of the various intentional false news groups, the features of each group could be collected. In other words, these features could be used to represent or indicate the similar false news. Finally, through comparing and matching the features, the new news can be analysed and predicted.

Suppose that there are total f_g features of the g^{th} false news group. In addition, by K-means algorithm, the new news would be classified and grouped to the g^{th} false news group. Then, comparing the found features of the new coming news, there would be f_n features of the new news that included or the same in the g^{th} false news group. Therefore, in this research, the new news could be evaluated as the possible intentional false news by following function:

$$\text{False News Warning} = \begin{cases} \text{true} & \frac{f_n}{f_g} \times 100\% \geq \beta \\ \text{false} & \text{otherwise} \end{cases} \quad (4)$$

where the β is the on demand given value for the intentional false news warning.

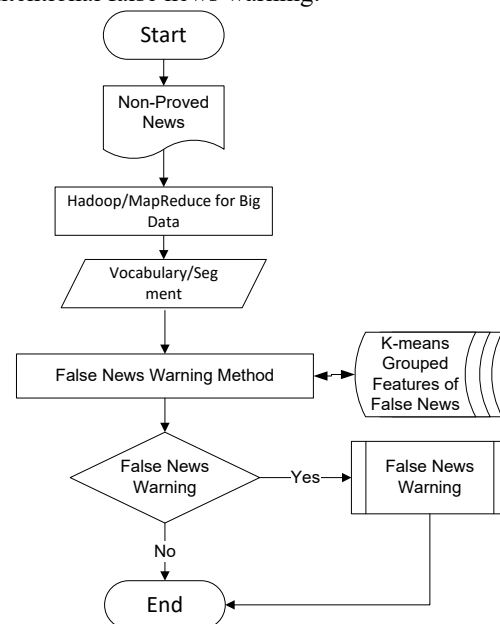


Fig. 3. The flowchart of the false news warning.

IV. VERIFICATION AND RESULTS

Following is the implementation example. Based on mmsegj algorithm for Chinese vocabulary recognition, the Chinese vocabularies presented in the website or web page could be separated and recognized as Figure 4. According to the {key, value} structure, the initially found key vocabularies and the corresponding appearance probability are checked. Some proper nouns are already recognized.

However, many key vocabularies are recognized as the single Chinese word. Based on the proposed Key Vocabularies Merging Method, the two vocabularies could be merged according to equation 1. In figure 4, for example, the appearance probabilities of "西" and "坑" are almost the same. According to equation 1, these two vocabularies could be merged as "西坑" or "坑西". However, after next cycle of the Key Vocabularies Merging Method, due to that the "西坑" is the independent proper nouns, and "坑西" is not the independent proper nouns, the appearance probability differences between "西坑" and "坑西" will be very huge. In other words, only the "西坑" can be reserved and survived for the further feature searching.

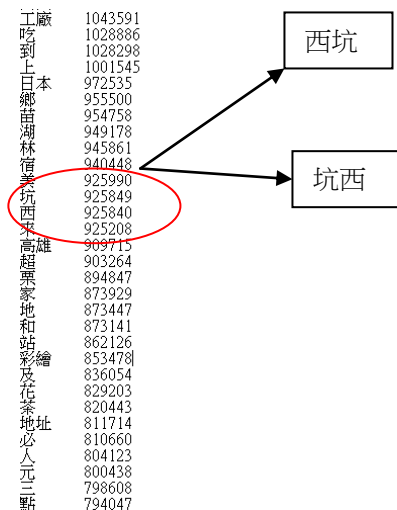


Fig. 4. The example of the implemented Key Vocabularies Merging Method.

In this research, to cluster and group the fake news which are collected by the web crawler, the scikit-learn for and the jieba algorithm are connected by the python program. The collected fake news and the related features are stored in the noSQL database on cloud. To provide the visual presentation and interface, the python matplotlib is also used. In the beginning, the specific topics of news from the Taiwan Fact Check Center are all already proved and recognized as the fake news. According to this recognized topic, the normal search engines such as Google or Yahoo are used to collect the similar news. Based on the proposed Key Vocabularies Merging Method, jieba algorithm, and term frequency-inverse document frequency (TF-IDF) method, the total found feature vocabularies or segmentations are various and huge. To cluster and group the collected fake news which are corresponding to the fake news proved by Taiwan Fact Check Center, the K-means algorithm integrated with the Elbow methods is recursively executed. According to Elbow Method, elbow point could be found around 4 with less differences of summation of squared distance when the value k is increased. The result is shown as Figure 5.

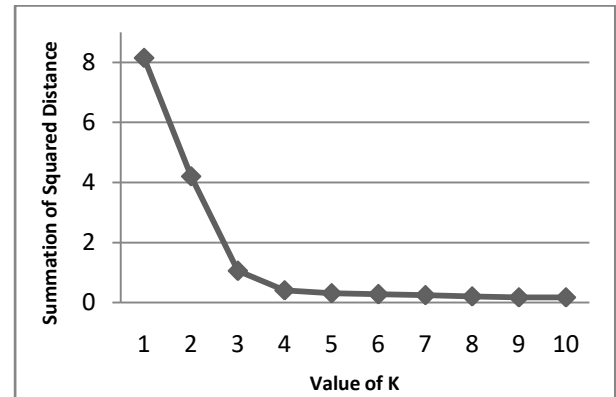


Fig. 5. The found value K based on Elbow Method according to the corresponding squared distance.

In the beginning of the verification, there are four different topics of fake news from the Taiwan Fact Check Center are collected. Since according to the proposed methods the final groups could be clustered into correct 4 groups, the related news from the search engine could be also recognized and clustered. In this research, there are total 34 fake news from various sources corresponding to the four main fake news topics are collected as the data for verifying the classification. There are 9 fake news are proved similar or the same as fake news group 1; 10 fake news are proved similar or the same as fake news group 2; 9 fake news are proved similar or the same as fake news group 3; and 6 fake news are proved similar or the same as fake news group 4. To verify the proposed system, these all 34 fake news are initially given without grouping and definitions. The proposed system receives these fake news without understanding the numbers of topics and the numbers of fake news in each fake news groups.

After the recursively processing the proposed Key Vocabularies Merging Method, jieba algorithm, and term frequency-inverse document frequency (TF-IDF) method with Elbow Method, the related fake news could be clustered into the same group by evaluation the square distance between each fake news and the group center. In addition, due to the Elbow Method, the minimum numbers of the groups could be also decided according to the obtained square distance. Finally, the related fake news could be correctly clustered into the corresponding groups. Figure 6 presents the average distance between each fake news group and the different group centers. The average distance of the 9 fake news which clustered into the fake news group 1 is only about 0.1 square distance from the cluster center 1 but more than 0.4 square distance to any other cluster centers. The average distance of the 10 fake news which clustered into the fake news group 2 is only about 0.1 square distance from the cluster center 2 but more than 0.6 square distance to any other cluster centers. The average distance of the 9 fake news which clustered into the fake news group 3 is only about 0.13 square distance from the cluster center 3 but more than 0.58 square distance to any other cluster centers. The average distance of the 6 fake news which clustered into the fake news group 4 is only about 0.05 square distance from the cluster center 4 but more than 0.47 square distance to any other cluster centers. In other words, the proposed algorithm and processes could really recognize and cluster the fake news correctly.

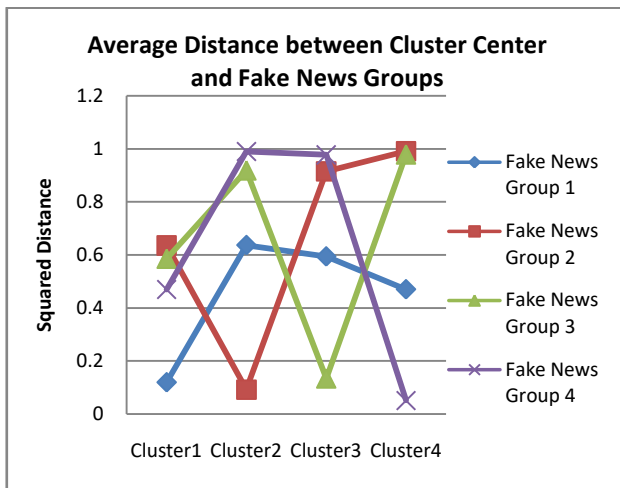


Fig. 6. The average square distance between the fake news and the cluster center.

Therefore, according to the proposed Key Vocabularies Merging Method, jieba algorithm, term frequency-inverse document frequency (TF-IDF) method with Elbow Method, the trained fake news data could be correctly recognized. Considering the purposes of the malicious organizations or users, the fake news would be frequently announced or created related to some specific person or topics. In other words, the rapidly announced fake news from the same organizations or groups could be recognized and found.

V. DISCUSSION

Although according to the proposed system the proved fake news could be recognized and clustered, to predict the new coming news is still difficult. First, the proposed system can recognized and cluster the existed and proved fake news. Based on the collected huge amount of the proved fake news, the recognition and clustering could be correct. However, the accuracy of the prediction should depend on huge data found from the Internet. It means that the prediction is difficult when there is only a few data collected from the Internet. In addition, since the new coming news is not identified and verified, the proposed system could only differentiate this news according to the related information collected and announced in the past. In other words, if the most of the news disseminated by the social media or organization is true news, then according to the proposed system the new coming news would be easily recognized as true news. In opposition, if most of the news disseminated by the social media or organization is fake news, the new coming news would be identified and clustered as fake news even it is really true. The found features would result in the bias when cluster and recognize the news from specific media or group. In other words, the history of the media or organization itself would affect the prediction result. The media or group with more proved fake news in the past would be easily recognized and clustered as the fake news in the future.

Second, due to the verification and the proof of each news, some information or messages hidden in the news would need time for searching and finding. Some inside information could be only be proved and recognized after long time. Therefore, to prove the news true or fake itself is an difficult

issue. In addition, only more information found then the current news could be identified as the true or fake news in the future. Therefore, to directly verify and recognize a news is true or fake is not easy especially when the related information is not enough. Finally, the credibility of the media or organization still plays the main and important role.

However, due to some political purposes, the fake news would be created and broadcasted by specific malicious media organization and groups. These malicious parties would actively spread the fake news. According to the proposed system, these groups, individual person, even media organization could be found and evaluated. In the future, some more tags could be found and defined for these malicious fake news spreaders. In other words, the credibility of these malicious fake news spreaders could be evaluated and given the warning according to the new spreading history.

VI. CONCLUSION

In this research, the proposed Key Vocabularies Merging Method could find the features of the specific intentional false news. Based on the term frequency-inverse document frequency (TF-IDF) method with Elbow Method, the proved false news or fake news could be correctly clustered into the corresponding groups. In addition, the numbers of the groups related to the false news could be automatically increased. In other words, according to the analysis of the historic proved news, the repeatedly found news could be recognized and verified.

Based on the proposed False News Warning Method, different intentional false news could be grouped. According to the comparing with the various groups of the intentional false news, the false news warning could be given corresponding to the new news.

ACKNOWLEDGMENT

Thanks for the support of Cloud Computing and Intelligent System Lab. (CCIS Lab.) of National Formosa University.

REFERENCES

- [1] Soroush Vosoughi, Deb Roy, Sinan Aral, " The spread of true and false news online," *Science*, Vol. 359, Issue 6380, pp. 1146-1151, 2018. DOI: 10.1126/science.aap9559
- [2] <http://www.google.com>
- [3] <http://tw.yahoo.com>
- [4] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu. "Unsupervised fake news detection on social media: A generative approach." *Proc. of 33rd AAAI Conference on Artificial Intelligence*, 2019. DOI: <https://doi.org/10.1609/aaai.v33i01.33015644>
- [5] W. Y Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*. 2017.
- [6] C. Castillo, M. Mendoza, B. Poblete, "Information credibility on twitter." *In Proc. of the 20th international conference on World wide web*, 675–684. ACM, 2011.
- [7] C. F. Bond Jr, B. M. DePaulo, "Accuracy of deception judgments." *Personality and social psychology Review* 10(3) pp.214–234, 2006.
- [8] J. Kim, B. Tabibian, A. Oh, B.S cholkopf, M. Gomez Rodriguez, "Leveraging the crowd to detect and reduce the spread of fake news and misinformation." *In Proc. of the 11th ACM International Conference on Web Search and Data Mining*, pp.324–332. ACM 2018.
- [9] K.J. Chen & Ming-Hong Bai, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," *International Journal of Computational linguistics and Chinese Language Processing*, 1998, Vol.3, No.1, pages 27-44

- [10] Ming-Shen Jian, Wei-Cheng Hong, Sheng-Che Tsai, Yu-Wei Chen, Chih-Ling Fan "Based on Automatic Correlation Keyword Grouping and Combination Based Deep Information Search Corresponding to Specific Language Big Data -- Case of Leisure Recreation " *ICACT*, pp.372-377, Feb. 2020.
- [11] Wei-Yun Ma & K.J. Chen, "A bottom-up Merging Algorithm for Chinese Unknown Word Extraction," *Proc. of ACL workshop on Chinese Language Processing*, pages 31-38, 2003.
- [12] <http://120.108.221.55/profchwu/ml/%E6%95%99%E6%9D%90/%E6%96%87%E4%BB%B6%E5%88%86%E9%A1%9E/%E4%B8%AD%E7%A0%94%E9%99%A2%E8%A9%9E%E6%80%A7%E5%AE%9A%E7%BE%A9.pdf>
- [13] Chandni Saini, Vinay Arora, "Information retrieval in web crawling: A survey," *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016.
- [14] M. S. Jian, et al, "Cloud Based Agriculture Safety Inspection with Multiple Standard Sources," *Proc. of IEEE ICACT*, pp. 201-206, Feb. 2018.
- [15] Yu-Chen Wang, "The Structure of the Big Data Based on the Hadoop Platform", *thesis*, National Chi Nan University, Taiwan, 2015.
- [16] Zhi-Kai Liao, "Fault-Tolerant Management Framework for Hadoop Distributed File System", *thesis*, Tamkang University, Xinbei, Taiwan, 2013.
- [17] Jia-Chun Lin, "Study of Job Execution Performance, Reliability, Energy Consumption, and Fault Tolerance in the MapReduce Framework", *thesis*, National Chiao Tung University, Hsinchu, Taiwan, 2015.
- [18] Man-Ning Wu, "Data Analysis Results Storage Design in HBase, Chung Yuan Christian University", *thesis*, Taoyuan, Taiwan, 2015.
- [19] Hadoop – Introduction, Available online : https://www.tutorialspoint.com/hadoop/hadoop_introduction.htm
- [20] Ming-Shen Jian, Yi-Chi Fang , Yu-Kai Wang, Chih Cheng, " Big Data Analysis in Hotel Customer Response and Evaluation based on Cloud," *Proc. of IEEE ICACT*, pp.791-795, Feb. 2017.
- [21] Fredrik Erlandsson, Roozbeh Nia, Martin Boldt, Henric Johnson, S. Felix Wu, " Crawling Online Social Networks," *Proc. of 2015 2nd ENIC*, pp.9-16, Sept. 2015
- [22] <https://www.seleniumhq.org/>
- [23] Chih-Hao Tsai, "MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm," <http://technology.chtsai.org/mmseg/>
- [24] <https://www.itread01.com/articles/1476609057.html>
- [25] <https://spark.apache.org/>
- [26] Ming-Shen Jian, Ming-Sian You, "Cloud Based Hybrid Evolution Algorithm for NP-Complete Pattern in Nurse Scheduling Problem," *International Journal of Innovation, Management and Technology*, Vol. 7, No. 5, pp.234-237, 2016
- [27] Ming-Shen Jian, "Intelligent System Based on Cloud and IOT," Mar. 2019, EHGBooks Publishing, (ISBN: 978-1-62503-503-5)
- [28] <https://tfc-taiwan.org.tw/articles/report>
- [29] Zhang, X.; Wu, P.; Cai, J.; Wang, K. A Contrastive Study of Chinese Text Segmentation Tools in Marketing Notification Texts. *Journal of Physics: Conference Series*. Vol. 1302, No.2, 2010. doi: 10.1088/1742-6596/1302/2/022010
- [30] Hakim, A. A.; Erwin, A.; Eng, K. I.; Galinium, M.; Muliady, W. Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*. pp.1-4, 2014. doi: 10.1109/ICITEE.2014.7007894.
- [31] Havrillant, L.; Kreinovich, V. A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (tf-idf) Heuristic (and Variations Motivated by This Explanation). *International Journal of General Systems*. Vol. 46, No.1, pp.27-36, 2017. doi: 10.1080/03081079.2017.1291635
- [32] Hishamuddin, M. N. F.; Hassan, M. F.; Tran, D. C.; Mokhtar, A. A. Improving Classification Accuracy of Scikit-learn Classifiers with Discrete Fuzzy Interval Values. *2020 International Conference on Computational Intelligence (ICCI)*, pp.163-166, 2020. doi: 10.1109/ICCI51257.2020.9247696.
- [33] Brites, D. and Wei, M. PhishFry - A Proactive Approach to Classify Phishing Sites Using SCIKIT Learn. *2019 IEEE Globecom Workshops (GC Wkshps)*, pp.1-6, 2019. doi: 10.1109/GCWkshps45667.2019.9024428.
- [34] Syakur, M. A.; Khotimah, B. K.; Rochman, E. M. S.; Satoto, B. D. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *2018 IOP Conf. Ser.: Mater. Sci. Eng.* 336, 2017.



Ming-Shen Jian was born in Kaohsiung City, Taiwan in 1978. He received the B.S. from the National Chiao Tung University, HsinChu, and Ph.D degrees in Computer Science and Engineering from the National Sun Yat-sen University, Kaohsiung, Taiwan in 2007.

From 2018, he was an Associate Professor and director with the National Formosa University Cloud Computing and Intelligent System Laboratory. Currently he is also an IEEE Senior Member. Since 2009, he has been an Assistant Professor with the Computer Science and Information Engineering Department, National Formosa University. He is the author of four books, more than 50 articles, and at least 15 invention patents. His research interests include IOT development and application, Big Data, Optimal Solution, Intelligent System, and Cloud Computing. He was a Secretary of the Taiwan Association of Cloud Computing. Dr. Jian was a recipient of the IEEE sponsored international conference Paper Award in 2016, 2017, and 2018.



Rong-Bin Deng was born in 2000, Taiwan. Currently he is an B.S. degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Web Service, and Cloud Computing. He joins the Cloud Computing and Intelligent System Lab. (CCIS Lab.) from 2019.



Chen-Wei Fang was born in 1999, Taiwan. Currently he is an B.S. degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to IoT, and Cloud Computing. He joins the Cloud Computing and Intelligent System Lab. (CCIS Lab.) from 2019.



Hua-Yu Wu was born in 1999, Taiwan. Currently he is an B.S. degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Big Data, and Cloud Computing. He joins the Cloud Computing and Intelligent System Lab. (CCIS Lab.) from 2019.



Wen-Hsiang Hsieh was born in 2000, Taiwan. Currently he is an B.S. degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to Web Service, and Cloud Computing. He joins the Cloud Computing and Intelligent System Lab. (CCIS Lab.) from 2019.