ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Study on the Credibility Based on the Features of Agricultural Information

Yuyi Zhang [a)], Yayao Zuo, Xiaobang Chen

Faculty of Computer, Guangdong University of Technology, Guangzhou, 510006 China.

[a)] Corresponding author: 1540171202@qq.com

Abstract. This paper based on agricultural information with timeliness, regional and other complex factors, proposed a method based on the features of agricultural webpage information to classify the credibility of agricultural webpage information. By constructing the model of the relationship between the features of agricultural information and the credibility of information, and combining with the SVM algorithm of machine learning, the credibility classifier is trained to predict and assess the credibility of agricultural webpage information. Experiments show that this method is feasible, as well as to evaluate the credibility based on information sources, has some advantages.

Key words: agricultural webpage; SVM algorithm; credibility classifier.

INTRODUCTION

Nowadays, the study of the credibility of information is still in a period of growth, meanwhile, it is also a hot research field in the internet age. Scholars at home and abroad have also conducted valuable research from different perspectives. Based on the blog content, Juffinger A et al. Proposed to assess the credibility of the blog content by identifying the similarities and differences in the structure of blog content and news corpus content.1 Lee builds a database of trusted truths based on vast amounts of information on the Internet, and evaluates the credibility of suspicious information through trusted truth libraries.2 Based on the content credibility corpus, Adam achieves the goal of assessing the credibility of web pages by establishing a predictive model of the credibility of web content.3 Based on the military information, Zhang Tianyu proposed to take the credibility of information sources, information reviews and military information content into the idea of rank learning, and to obtain better ranking results of information credibility.4 Based on Internet medical information, G Eisenach helps consumers filter out harmful information and identify and select highly trusted information through the rating of its information. In response to the credibility of webpage information, domestic and foreign experts and scholars have conducted research on blog content, news content, military information and network medical information respectively. However, few scholars have conducted in-depth studies on the credibility of webpage information in agriculture field and agricultural web page information is an important source for agricultural industry-related personnel, especially farmers. Besides, Agricultural webpage information has such complicated factors as regionalism and timeliness. In view of this, the study of agricultural webpage information credibility is a challenging topic.

This paper makes a preliminary exploration on the credibility of agricultural webpage information, the main contribution which is constructing the model of relationship between the features of agricultural webpage content and information credibility and combine the SVM algorithm to train the credibility classifier to achieve the prediction of the credibility on agricultural webpage information. The experiment shows the method is feasible and have some advantages for the credibility assessment based on the information sources.

THE CREDIBLE MODEL OF AGRICULTURAL WEBPAGE INFORMATION FEATURE

The Construction of Agricultural Webpage Information Feature Library

The agricultural information feature database which constructed in this paper is mainly aimed at Guangdong Province. The construction of the agricultural information feature database is based on time and region, including the four seasons of the year, such as spring, summer, autumn and winter. Besides, other features that affect the credibility of agricultural webpages include climate, water, topography, soils, heat, light, temperature, precipitation and more. The high-relevancy feature is highly related to agricultural information and has a positive correlation with agricultural information. As the following sentence: in summer, the temperature in Guangdong Province is relatively high, farmers diligently harvest bitter gourds, gourd and other crops. In this sentence, the key words are "Summer" and "Guangdong Province", the temperature is relatively high, and the crops harvested by farmers are balsam pear and loofah. In fact, the climate of Guangdong Province in summer is characterized by sweltering heat, high rainfall, high temperature, and long duration of sunshine. The characteristics of abounding crops are mainly rice, bitter gourd, gourd, spinach, spinach, etc.Therefore,this sentence is credible. The low-relevancy feature is characterized by low relevance to agricultural information, such as the following terms: farmers, agriculture, climate, temperature, etc. These feature words have low impact on the reliability of information on agricultural web pages. The negative feature is mainly misleading people and negatively related to agricultural information. As the following sentence: in the summer, Guangdong Province is cold, with low temperatures and little rain. The real summer climate in Guangdong Province is sizzling, rainy, hot, and long hours of sunshine, and this shows that this sentence has misled people's understanding of the summer in Guangdong Province. Through the negative correlation of climate feature words in this sentence on the summer of Guangdong Province, which is cold; low temperature; less rain, we can immediately make an untrustworthy decision on this sentence. This is the function of the negative related feature words, which is an important indicator for us to judge the credibility of agricultural webpage information.

To sum up, this paper applies AHP to divide the construction of feature words into three levels:

The first level is the feature word of high relevance (word_h)

The second level is the feature word of low relevance (word_l)

The third level is the negative correlation of the feature words (word_n)

The feature words of agricultural web information determine the credibility of information. If there are many high-confidence feature words in the agricultural webpage information, the credibility of the agricultural web page information is determined to be high. The number of low-relevance features in agricultural webpage information also affects the credibility of web page information, though the influence index is relatively low. Negatively related feature words in agricultural webpage information has a negative correlation with the content of website information, the more negative feature words, the less credibility on the agricultural webpage information.

In order to better evaluate the proportion of various feature words in agricultural webpage information, we define the ratio of each feature word as follows:

Among them, the definition of the high relevancy feature words in the webpage information is defined as follows:

$$H_{ratio} = \Sigma word\_h/sum\_word \tag{1}$$

The rate of low relevancy feature words in the webpage information is defined as follows:

$$L_{ratio} = \Sigma word\_l/sum\_word \tag{2}$$

The ratio of the negative-relevance feature words in the web page information is defined as follows:

$$N_{ratio} = \Sigma word\_n/sum\_word \tag{3}$$

In the above formula, word_h, word_l, and word_n indicate the number of corresponding feature words in the web page information, and sum_word indicates the total number of words in the agricultural web page information.

If the H_ratio of the feature words with high relevance in the webpage information is larger, the credibility of the webpage information is higher; on the contrary, the larger the N_ratio of the feature words of the negative relevance, the lower the credibility of the webpage information is, based on this, the credibility of information on agricultural websites is defined as follows:

$$P_{feature} = \alpha * H_{ratio} + \beta * L_{ratio} + \varepsilon * N_{ratio} \quad (4)$$

Among them, α, β, ε are the weight ratio of feature words with high relevancy degree, low relevancy degree and negative relevancy degree respectively. The relationship among the three is $1 > \alpha > \beta > \varepsilon$. The low correlation coefficient of feature words is positively correlated with the credibility of the webpage information, therefore, α, β values are larger than zero. While the negative correlation of feature-word ratio has negative correlation with the credibility of webpage information, therefore, the value of ε is less than zero.

The Credible Algorithm of Agricultural Webpage Information

The specific algorithm of the credibility of agricultural webpage information flow is shown in Figure 1.

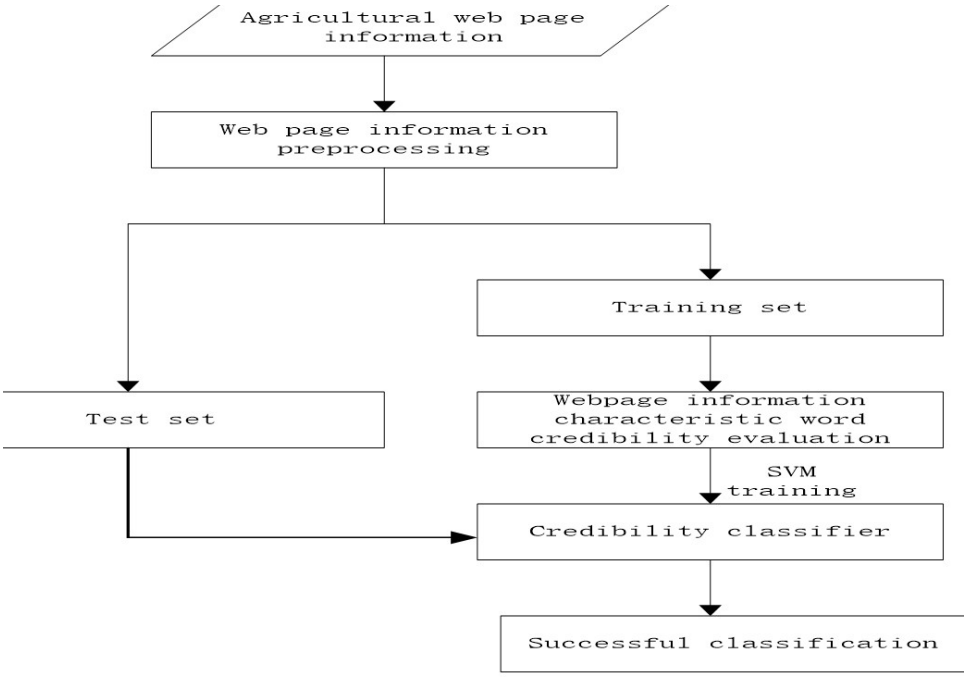

FIGURE 1. Agricultural web content credibility classification flowchart

This paper classified agricultural information credibility as shown above. First, the agricultural webpage information we have selected is recorded as Ci,i=1,2,3,...,n, and a series of preprocessing is performed to filter out noise information that does not reflect the credibility of the agricultural webpage information, which is marked as Cni.For example, various stop words and web page information that cannot reflect the credibility of agricultural, and a series of processing such as word segmentation of the pre-processed webpage information to obtain the agricultural webpage information. Then, the three-tiered feature modeling of the identified agricultural webpage information is modeled, which is word_h&word_l& word_n.Based on the relationship model between the feature data and information credibility that we have mastered, Combining with machine learning SVM algorithm, training generates agricultural webpage information credibility classifier.finally,we can get the initial credible or untrustworthy classification results.

Algorithm flow is as follows:

Enter: C1, C2, ..., Cn // web collection

Output: True, False // Web information classification results, that is, credible or not credible

1) filtering the noise information Cni based on the input web pages C1, C2, ..., Cn to obtain a cleaned web page Cr;

2) Calculate the confidence in Cr based on information features Pfeature = α * Hratio + β * Lratio + ε * Nratio, where Hratio = Σword_h / sum word, Lratio = Σword_l / sum_word, Nratio = Σword_n / sum_word, Reliability, which is Ccre (i), where i = 1,2,3 ... n;

3) Based on the above obtained webpage credible Ccre (i) and SVM algorithm, trained SVM classifier;

4) the web page to be tested to achieve classification, the output results;

5) End.

EXPERIMENT AND ANALYSIS

In order to verify the feasibility of the proposed method for the credibility of information, this paper designs the following experiment:

The experimental platform for this experiment is the Linux system Centos 6.5, which is mainly implemented based on python. The data is captured using pyspider. In the experiment process, the body pipe is extracted using the boiler pipe, and the word segmentation tool is jieba. The training set and test set used in the experiment were mainly based on the agricultural information-based Internet. They mainly included 8 agricultural websites, such as: China Agricultural Information Network, China Fisheries Information Network, China Fishery Information Network, etc. There are also 6 forums, such as: Aquaculture Technology Forum, Planting Technology Forum, etc. Among them, 1750 positive cases and 250 counterexamples were used, with total of 2000 webpages. And based on the above characteristics of agricultural webpage information content and the SVM classification algorithm used in this paper, the performance of SVM, Naive Bayes and Logistic Regression three classification algorithms in machine learning is compared and analyzed. The experimental results are as follows:

TABLE 1. Classification Algorithm Experimental Results Table

category	A (%)	P (%)	R (%)
SVM	90.5	89	91
Naive Bayes	85.2	81.5	81.3
Logistic	81.5	80.2	82.3

It can be seen from the figure above that the SVM classification algorithm is obviously better than Naive Bayes and Logistic in this experiment, and the number of sample training is not large; under the condition of high data feature dimension, it further shows that SVM has unique superiority as a classification algorithm.

CONCLUSION

This paper studies the relationship between information content and information credibility based on agricultural web pages. By modeling the characteristics and information credibility of agricultural information content, combined with the machine learning SVM algorithm, the reliability of agricultural web page information Played a very good role of judge. Through experiments, this method is feasible for the credibility of agricultural web information. However, at the same time, subject to the quality and scale of training corpus, the results of this paper's agricultural webpage information credibility classifier are more general and there is still much room for improvement. In addition, the construction of the feature database of agricultural webpage information needs to be further Perfect, the above questions are the next research direction of this article.

ACKNOWLEDGMENTS

At the time of this thesis, I would like to express my deep appreciation to all those who have taken care and help in learning and living during the Master's degree.

First of all, we would like to thank Professor Zuo. Can be successfully completed the writing of the paper, all embodies the teacher's effort and sweat. The teacher in the paper topics, research programs to determine and the

specific implementation process have given careful guidance, their rigorous attitude and systematic research ideas I benefit for life.

REFERENCES

1. Juffinger A,Granitzer M, Lex E. Blog Credibility Ranking by Exploiting Verified Content[C]//Proc of the 3rd Workshop on Information Credibility on the Web.2009:51-58.
2. LEE R, KITAY AMA D, SUMIYA K.Web-based evidence excavation to explore the authenticity of local events[C]//Proceedings of WICOW 2008, California:ACM, 2008:63-66.
3. Adam Wierzbicki, Michal Kakol, Radoslaw Nielek. Understanding And Predicting Web Content Credibility Using The Content Credibility Corpus,2017, Volume 53,Issue 5,Pages 1043-1061.
4. Zhang Tianyu, Lin Hongfei. Research on the credibility of military information based on multiple representation. Computer Engineering and Science, 2011, 33(9):109-116
5. G Eysenbach,G Yihune,K Lampe,P Cross,D Brickley.(2000)Quality Management, Certification and Rating of Health Information on the Net with MedCERTAIN: Using a medPICS/RDF/XML metadata structure for implementing eHealth ethics and creating trust globally.J. Med. Internet Res.2 (suppl.), article el.
6. Dong XL, Gabrilovich E, Murphy K, Dang V,Horn W, Lugaresi C ,Sun S, Zhang W. Knowledge-based trust: Estimating the trustworthiness of web sources. In: Proc. Of the VLDB.2015.938-949.