Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 6, No. 2, 2023

Junk Information Recognition Based on Naive Bayes

Yuanyao Zhang[1, *]

[1] Colloge of overseas education, Nanjing Tech University, Nanjing, Jiangsu, 211816, China
[*]Corresponding Author's Email: 202021139077@njtech.edu.cn

Abstract: This work mainly aims to learn and improve the existing methods of identifying spam information. In this work, the fundamental central principle is applied to naive Bayes. The improvement part is to use naive Bayes to achieve automatic recognition and reporting of spam information in the system, as well as a refined classification of spam information. This paper aims to further deal with junk information (illegal information) through naive Bayes, so as to meet the needs of People's Daily life. In the simple junk information recognition, add life elements, so that junk information recognition more life, close to life, service human.

Keywords: Bayesian, Spam, Probability, Classification.

1. Introduction

In the rapid development of science and technology today, junk messages often appear in people's life, bring a lot of trouble to people's life, there may also be economic losses, so the design of a filter junk message classification algorithm has become a demand, the purpose is to automatically classify and screen junk messages, judge the junk information. Through these algorithms can realize the blocking of spam messages, so as to achieve the purpose of convenience. With the improvement of people's requirement for information function, junk information recognition gradually appears in people's life. At present, China's junk information identification and classification is in the stage of development, compared with some other developed countries in this industry, there is still some distance. At present, most literature reviews are based on the recognition and simple classification of junk information. However, it has not yet involved the identification of information after reporting and slightly complex information identification classification. Such as identifying the type of information (telecom fraud, advertising, etc.). In this research topic, the emphasis is on the identification of junk information after reporting and subdivision research. More in line with people's life needs, so that people to junk information at a glance. Subdivision can let people know the types of spam information and improve the awareness of prevention.

2. Bayes' Theorem

Bayes' theorem is a basic theory for dealing with spam and email. It uses probability to predict future events based on past events. It's about random events A and B. When the sample size is close to the population, the probability of the event occurring in the sample will be close to the probability of the event occurring in the population. In "Research and Implementation of Spam Filtering System Based on Bayesian Classification", the author filters spam information by Bayesian. This paper systematically expounds how to construct a spatial vector model and the application of Bayesian basic principles.[1] The construction of space vector model will also be involved in this study, which is a structural process. It does not cover the post-identification services of the garbage system (such as sorting and reporting).

Formula of Bayes' Theorem: P (A/B)=P (B/A)*P (A)/P (B)

2.1. Application of naive Bayes in junk information classification

The role of naive Bayes in junk information classification is to classify the acquired information with the constructed classification model. Firstly, through training the training set, the most frequent spam words are obtained. Then, the collected emails are tested, one is normal emails, the other is abnormal emails.

Suppose you receive an email X, without knowing whether it is a normal email or not. Firstly, the nouns in this email are constructed with a feature vector. Each of the elements in the eigenvector exists independently of each other. The mail discrimination method used in this study is as follows: the probability of normal mail and abnormal mail is calculated respectively. Spam is referred to here as Spam, which would normally be Ham. If P (Ham | X) > P (Spam | X), it is called a normal mail, the opposite is Spam.

P (Ham | X) and P (Spam | X) formula:

$$P(Ham|X)=\frac{P(Ham)P(X|Ham)}{P(X)} \quad (1)$$

$$P(Spam|X)=\frac{P(Spam)P(X|Spam)}{P(X)} \quad (2)$$

3. Data Set Establishment of High-frequency Words of Junk Information

In "Research on Spam Filtering Technology Based on Bayesian Classification", the tree structure idea and the improved algorithm of high-frequency word sensitivity are introduced emphatically. [2]The improvement of high frequency word sensitivity is very important for a spam filtering system. In a spam message, the sensitivity to high-frequency words determines whether the spam message can be successfully identified. In the research of junk information

processing, the concept of tree structure can help the system query keywords more quickly to achieve the purpose. With the improvement of high-frequency word sensitivity, a new idea is provided for this study to use high-frequency word sensitivity to complete a more refined classification of junk information.

3.1. Source of data set

In the Internet to find related spam, spam training, training a spam high frequency word vector set.

For amateurs to identify junk information, can directly download the corpus package from the Internet, directly test.

3.2. Capacity selection of training set and test set

Generally, the capacity of the training set is 200-300. In this study, the capacity of the training set is appropriately increased to improve the accuracy, so as to construct a high-precision high-frequency vocabulary data set. To a certain extent, the larger the capacity of training set and test set, the better.

3.3. Accuracy evaluation of high-frequency vocabulary data set

After the data set is built, part of the mails are collected for verification, and the judgment errors of normal mails and abnormal mails are calculated. In order to determine whether a suitable high-frequency vocabulary data set is established. The introduction of high-frequency words can effectively reduce the error of judgment. It is also a common way to judge whether the mail is spam by high-frequency words.

3.4. Using minimum risk Bayes to avoid misjudgment

In "Improved Bayesian spam filtering algorithm", the concept of the least risk Bayesian algorithm is introduced. The cost of a normal email misconstrued as spam is incalculable. So in the identification of spam should pay attention to the minimum risk decision. [3]The concept is worth learning from, and the algorithm is efficient. Minimal risk Bayes can be used in combination with mail identification, commercial use and security.

4. Precise Classification of Junk Information:

The junk information is divided into fraud information, advertising information, harassment information, etc. Convenient for customers at a glance, judge the type of information.

4.1. Construction of classification set

In "Research on Chinese Spam Filtering System based on Bayesian Algorithm", it provides ideas for this research. The purpose of this work is to analyze the recognition of Chinese spam messages, but English SMS messages are not involved. In this paper, text preprocessing and feature extraction are worthy of reference. In this paper, a new concept is proposed, namely 3-layer Bayesian network structure model. [4]This model can help to identify spam more accurately. The mail subdivision in this work can be optimized on this Bayesian network structure. The information to be obtained falls into several categories as required. The information is subdivided by categories. For example, scam messages are classified as Class A, advertising messages as class B, and harassing messages as class C. Collect A collection of A, B and C types of mail and train them separately. After training the collection of three features, the specific type is determined after the determination of spam. The number of mail training set, within a certain range, the more the higher the precision. The high-frequency words used to judge junk information are a large set, and the set of further information classification is a more accurate small set. Just like the idea of multi-layer Bayes, the total class is subdivided to achieve the purpose of efficiency.

4.2. Decision rules of A, B and C types of information

Based on A, B and C types of information, the classification of junk information is as follows: A, B, C, AB, AC, BC, ABC. A message has the probability of having all three characteristics.

The three features A,B and C exist independently. If the probability is greater than 50%, it is judged as the information. That is, an email with M:

P (A | M > 50%), P (B | M) < 50%, P (C | M) < 50%, the mail just for fraud

P (A | M > 50%), P (B | M) > 50%, P (C | M) < 50%, the email for fraud, spam

P (A | M > 50%), P (B | M) > 50%, P (C | M) > 50%, the mail fraud, advertising, E-mail harassment.

5. Report Junk Information:

Automatic reporting of spam messages is becoming increasingly popular. The report of spam information means that after the identification of spam information, the system itself carries out the report of relevant departments. Automatic reporting of junk information belongs to the late service after information identification. In “Research and Application of Spam processing Model”, the relationship between different reporting domains is involved, and the detailed network structure is described. This study has some inspirations for reporting within the system.[5]

5.1. Ways to report spam

At present, the country has launched relevant reporting APP, or in the mailbox system to report by oneself.

5.2. Implementation Process of spam reporting

Through the system to identify whether it is junk information and the type of information, write a program to let the system fill in the relevant information instead of the user report.

5.3. Service-type spam reporting

After the spam is identified, the user is asked whether they agree to report it or agree to report it. This feature is used in many mobile phone systems today. Following the user's opinion is the premise of this function.

5.4. Linkage between information reporting and information classification

Information reporting starts after information is classified into spam, and spam classification can better serve information reporting and inform the server of the basic information of spam.

6. Conclusion

An information filtering model is established based on naive Bayes principle to realize specific mail classification and report after spam identification. The specific classification of spam can help users see the type of mail at a glance, so as to prevent such things from happening. Tree model plays a very important role in the classification of spam, which can help classify the types of mail faster. Spam reporting is a service. By understanding the network structure and getting familiar with the relationship between different reporting domains, it can report the accounts that send spam to the relevant authorities after obtaining the consent of the users. These two improvements can make spam recognition better serve the user community. Convenient people's life, based on convenience. At the same time, also can effectively prevent the occurrence of fraud and other behaviors, reduce the user's property losses. To ensure the safety of use.

References

[1] Lin Wei.(2009). Research and Implementation of Spam Filtering System Based on Bayesian Classification (Master Dissertation, Xihua University). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2010&filename=2009199790.nh

[2] Wang Lu.(2020). Research on Spam Filtering Technology Based on Bayesian Classification (Master's thesis, Shanghai University of Engineering Science). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202101&filename=1021534561.nh

[3] Zhao Jinghui & Wei Zhengang.(2016). Improved Bayesian spam filtering algorithm. Computer Systems Applications (10),137-140. doi:10.15888/j.cnki.csa.005380.

[4] Liu Haoran, Ding Pan, Guo Changjiang & et al.(2018). Research on Chinese Spam Filtering System based on Bayesian Algorithm. Journal of Communications (12),151-159.

[5] Yu Muqing.(2010). Research and Application of Spam processing Model (Master Dissertation, Beijing University of Posts and Telecommunications). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2011&filename=2010222633.nh