Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/13437
DOI URL: http://dx.doi.org/10.21474/IJAR01/13437

RESEARCH ARTICLE

PERFECTIONOF CLASSIFICATION ACCURACY IN TEXT CATEGORIZATION

Dr. Rajeev Tripathi
Dept. of Computer Application Shri Ramswaroop Memorial Group of Professional Colleges Lucknow (India).

Manuscript Info

Manuscript History
Received: 25 July 2021
Final Accepted: 29 August 2021
Published: September 2021

Key words:-
Featureselection, Text Categorization, Categorization Accuracy, CHI Square Statistics, TFIDF

Abstract

Problems and strategies for text classification have already been known for a long time. They're widely utilised by companies like Google and Yahoo for email spam screening, sentiment analysis of Twitter data, and automatic news categories in Google alerts. We're still working on getting the findings to be as accurate as possible. When dealing with large amounts of text data, however, the model's performance and accuracy become a difficulty. The type of words utilised in the corpus and the type of features produced for classification have a big impact on the performance of a text classification model.

Introduction:-

With the rapid growth of online information, how to effectively handle large amounts of text has become a hot research issue, with text classification being one of the most important jobs. Text categorization is the process of assigning new documents to pre-existing categories, and it is frequently utilised in fields such as information retrieval, email classification, spam email screening, and subject spotting.

In recent years, the majority of research has been focused on developing new classification algorithms, with little attention paid to improving document representation models, which are utilised in information retrieval. There are three traditional models: vector space model [1], probabilistic model, and inference network model, vector space model is the most widely used.

In a vector space model, a feature is represented as a numerical weighting. There are several popular weighting methods, including Boolean weighting, frequency weighting, TF-IDF weighting, TFC weighting [2], LTC weighting [9], and entropy weighting, the most commonly used of which is TF-IDF weighting.

In this paper, we propose a vector space model enhancement of TF-IDF weighting. TF-IDF considers both term frequency and inverse document frequency; in this method, if the term frequency is high and the term only appears in a small portion of documents, the term has a very good differentiate ability, This method stresses the capacity to distinguish between classes more, while ignoring the fact that a phrase that appears frequently in documents belonging to the same class might better reflect that class's feature. So we created a new parameter to reflect the in-class characteristic, and then ran several tests to evaluate the effects, and the results show that this improvement is more accurate.

Basics of data streams

Stream data mining relies heavily on statistics, complexity, and computational theory in its computational techniques. The system's resource needs are considerable due to the real-time nature of data streams and their high arrival rates. Also, computational theory techniques have been implemented to achieve time and space competent solutions. Summarization is a common technique for extracting reasonably accurate information from databases. They combine approaches for data

Corresponding Author:- Dr. Rajeev Tripathi
Address:- Dept. of Computer Application Shri Ramswaroop Memorial Group of Professional Colleges Lucknow (India).

reduction and summary generation. Summarization is the process of converting data into a format suitable for stream analysis, which can be accomplished by condensing the entire data set or selecting a subset of the incoming stream to analyse. Techniques including sampling, drawing, and load shedding are employed to summarise the data set. Synopsis data structures and aggregation functions are used to pick a subset from the data stream [8][11].

TextCategorizationSteps
Document preprocessing, document representation, dimension reduction, model training, testing, and assessment are the five key processes in text organization [3].

DocumentPreprocessing
We remove html tags, uncommon words, halting words, and perhaps some stemming in this stage; this is straightforward in English, but complex in Chinese, Japanese, and other vernaculars.

DocumentRepresentation
Prior to classification, we must convert the documents into a format that a computer can understand; the most widely used approach is the vector space model (VSM). This model treats the document as a multi-dimensional vector with the dataset feature as a dimension.

DimensionReduction
Because there are tens of thousands of words in texts, selecting all of them as features would make categorization impossible, as the computer would be unable to comprehend such a large quantity of data. As a result, we must choose the most relevant and representative characteristics for classification, with CHI square statistics[4], information gain, mutual information, document frequency, and latent semantic analysis being the most often utilised methodologies.

Model Training
The most crucial aspect of text classification is this. It involves selecting a subset of texts from the corpus to form the training set, learning on the training set, and finally generating the model.

Testing and Evaluation
This phase takes the model created in step 4 and applies it to the testing set, then selects a suitable index to execute assessments

TF-IDF
The Word Frequency-Inverse Document Frequency (TF-IDF) [5,6] method gives a term weighting based on its inverse document frequency. The TF-IDF weighting approach, which was initially presented from information retrieval, is a frequently used weighting method in vector space models. It indicates that the more papers in which a phrase appears, the less essential that term is and the lower the weighting. It can be expressed as follows:

$$aij = \log(tfij + 1.0) * \log\left(\frac{N + 1.0}{nj}\right)$$

TF-IDF-CF
The TF-IDF score indicates the relative significance of a phrase in the text and the corpus as a whole. Term Frequency is abbreviated as TF, while Inverse Document Frequency is abbreviated as IDF:

In order to address the limitations of the TF-IDF, we propose a new parameter called class frequency, which estimates the word frequency in documents within a single class.The new weighting technique is therefore renamed TF-IDF-CF, and its formula is based on (2):

$$aij = \log(tfij + 1.0) * \log\left(\frac{N + 1.0}{nj}\right) * \frac{ncij}{Nci}$$

ncijrepresentsthe numberof documentswheretermjappears withinthe sameclasscdocument ibelongsto,
Nci represents thenumber of documents within the same classcdocumentibelongs to.

***TF–IDF* Vectors -** The TF-IDF score indicates the relative significance of a phrase in the text and the corpus as a whole. Term Frequency is abbreviated as TF, while Inverse Document Frequency is abbreviated as IDF:

$$TFIDF(t,d) = TF(t,d) * \log\left(\frac{N}{DF(t)}\right) \quad (4)$$

Being:
t: term(i.e. a word in a document)
d : document

TF(t) : term frequency (i.e. how many times the term t appers in the document d)
N : number of dicuments in the corpus
DF(t) : number of documents in the corpus cintaing the term t

The TF-IDF value increases in proportion to the number of times a word seems in the document and is offset by the number of documents in the corpus that include the term, which helps to compensate for the fact that certain words appear more frequently than others in general..

CHI square statistics

In text classification, CHI square statistics [4] is a helpful feature selection approach since it can quantify the connection between feature and class. Let A represent the times when both feature t and class c exist, B represent the times when feature t exists but class c doesn't, C represent the times when feature t doesn't exist but class c does, D represent the times when both feature t and class c don't exist, and N represent the total number of times when both feature t and class c don't exist.

$\chi^2(t,c)=$

$$\frac{N*(AD-BC)^2}{(A+C)*(B+D)*(A+B)*(C+D)}$$

TFCandLTC

TFC[2] is proposed to include the influence of the length of various documents on weighting, and it is actually the normalising of formula (1).

$$tf_{ij}*\log(^{N})$$

LTC[9] is a TF-IDF format that takes into account the limitations of tiny datasets and is basically a normalisation of formulas (2).

$$a_{ij}\text{-}\frac{\log(tf_{ij}+ \quad N \quad 1) \quad 1.0)*\log(\ )}{\sqrt{\sum^{M}[\log(tf_{ip}+1.0)*\log(\frac{N}{n_p})]^2}} \qquad (6)$$

Chi –Square (χ^2) based on the statistical theory: It measures the lock of independence among the terms and category.

$$\chi^2= \sum^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right] \quad i=1$$

The feature selection methods

The enormous dimensionality of feature space is the fundamental issue with text classification. A text document's feature set is a collection of unique terms or words that appear in all texts. Feature selection is a technique for reducing the amount of characteristics in a database. The benefit of lowering the attribute list is increased processing speed, which leads to improved performance.

Implication for implementation proposed methodology

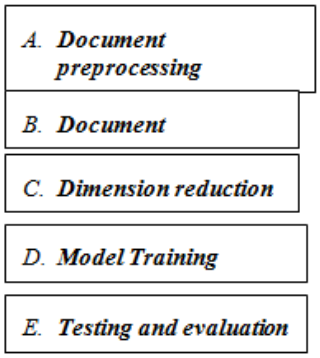
In the context of implementation the text classification system can be divided into three main steps:

1. **Pre-processing step:** Where the punctuation marks, stop words, and diacritics and non meaningful words are separated.
2. **Features selection step:** In this step the relevant features are chosen from the original text. They present the next that is input into the learning step.
3. **Learning steps:** Numerous techniques have been deployed to teach systems how to segregate text documents into different categories**.**

Improvement of TFIDF

TF-IDF is extensively used weighting method in text categorization. As we discuss Text categorization steps.
It includes 5 main steps:

(5)



Extracting features from text files:

Text files are just a collection of words (ordered). We must transform text files into numerical feature vectors in order to execute machine learning algorithms. As an example, we'll use the bag of words model. To summarise, we divide each text file into words (for English dividing by space), count the number of times each word appears in each document, and then give an integer id to each word.

Loading the data set

```
from sklearn.datasets import collegedata
cdata = collegedata(subset='student', shuffle=True)
```

TF: Just counting the number of words in each document has one issue it will give more weightage to longer documents than shorter documents. To avoid this, we can use frequency (TF - Term Frequencies) i.e. #count(word) / #whole words, in each document.

TF-IDF: Finally, we can even diminish the weightage of more common words like (it,the, is, an etc.) which transpires in all document. This is known as TF-IDF i.e Term Frequency times inverse document frequency.

```
from sklearn.feature_extraction.text import CountVectorizer
count_data = CountVectorizer()
mstudent = count_data.fit_transform(collegedata.data)
mstudent.shape
```

Conclusion:-

Text categorization is a hot issue in today's information retrieval research, and it's an important area of data mining and retrieval. In order to tackle this challenge, significant research has been done to develop new classifiers that would increase the accuracy, whereas this work aims to improve the accuracy by providing an improvement on the old technique. We can see from the trials that this modification improves accuracy substantially, therefore we believe it is promising. The technique is a rigorous, methodical, and non-arbitrary method for determining whether or not any data differs considerably from a previously compiled frameset, as well as the size of that difference. This model, like other statistical models, is most useful in situations where huge volumes of data must be analysed and summarised for significant choices to be made.

References:-

1. Abualigah, L. M. Q., & Hanandeh, E. S. (2015). Applying genetic algorithms to information retrieval using vector space model. International Journal of Computer Science, Engineering and Applications, 5(1), 19.
2. Boyer, C., & Dolamic, L. (2015). Automated detection of HON code website conformity compared to manual detection: an evaluation. Journal of medical Internet research, 17(6), e3831.
3. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv: 1607.01759.
4. Yang y, pedersen j q. A comparative study on feature selection in text categorization. Proceedings of the14thinternationalconferenceonmachine learning(icml),1997:2-3.
5. Salton g, Wong a, yang c s. A vector space model for automated indexing. Communications of the acm,1975:1-8.
6. Saltong,mcgillc.Anintroductiontomoderninformationretrieval.Mcgrawhill,1983.
7. *N.D. Gagunashvili*, "Chi-square tests for comparing weighted histograms", 2009
8. *R Tripathi, S. Dwivedi*. A Quick Review of Data Stream Mining Algorithms.IJIR,Vol-2, Issue-7, 2016
9. Chirawichitchai, N., Sanguansat, P., & Meesad, P. (2010, November). Developing an effective Thai Document Categorization Framework base on term relevance frequency weighting. In 2010 Eighth International Conference on ICT and Knowledge Engineering (pp. 19-23). IEEE.
10. *R Tripathi, S. Dwivedi.* Resolution of E-Commerce Market Trend Using Text Mining. IJSRCSE,Vol.8, Issue.1, pp.01-05, February (2020)
11. Mirza, B., Lin, Z., & Liu, N. (2015). Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift. Neurocomputing, 149, 316-329.