Web Document Classification using Improved Graph based KNN Classification

Sana Irfan¹ Latika Sharma²
¹M.Tech. Student ²Assistant Professor
¹,²Department of Computer Science & Engineering
¹,²BIT, Meerut, UPTU

Abstract— this paper presents the classification of web document using improved graph based KNN. Most of the organization are facing problem of large amount of unorganized data. Large amount of data is present in the form text journals, email etc. there are many feature selection methods available like Mutual Information, Regularized mutual information, CHI Statics, term Strength, Document Frequency, Inverse Document Frequency, category Term Descriptor, Strong class information word. Feature selection method plays an important role in Text Categorization. The emphasis laid on the combination of standard method localized method i.e. weight of term and standard Dataset Reuters-21578 is used to verify the result.

Key words: Mutual Information, Weight of Term, KNN, Feature Selection, Data Mining

I. INTRODUCTION

Increased used of internet giving birth to the problem of managing the text data. A large amount of data is present in the form of Emails, journals etc. Text categorization is the part of data mining. Text categorization is the process of grouping of text into predefined categorized. Feature selection method plays an important role in the categorization process. The traditional text classification methods used VSM (Vector Space Model) in which each document was represented as a feature vector of the terms in the document. The similarity calculation between the documents was done using various similarity calculation methods like Jaccard method and Cosine angel method. However this VSM model ignored the structural information of the document like word order and co–occurrence between the words, which plays a vital role in identifying a document but in a graph based text categorization system whole information is captured in the form of edges and nodes. Combine the frequency of the term with the standard feature selection methods i.e. CHI square, mutual information (MI) and Regularized mutual information (RMI).

II. DATA MINING

Data mining is often defined as the finding hidden information in the database or it has been called exploratory data analysis, data driven discovery, and deductive learning.

III. TEXT CATEGORIZATION

The objective of text classification is to reduce the detail and diversity of the data and resulting information overload by grouping similar documents together. Text classification categorization may be viewed as assigning documents in a predefined categories. Text categorization has become an important task in text mining

IV. KNN

One of the most common classification scheme based on the use of distance measures is that of the K nearest neighbor (KNN). The KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. The K-NN is simple and is having better precision in classifying a document. Also this K-NN does not need any training resources or model to be built up and it categorizes on the fly .Therefore its cost is also less as no resources need to be trained and accuracy is also better than any other classifier.

V. SYSTEM DESIGN

A. Stop Word Filtering

Stop words such as conjunctions, articles, and even common words that occur frequently across all document are eliminated.
B. Steming
Steming is the process of reducing the words to its base/roots. For example eat, ate eaten to eat.

C. Feature Selection Methods
Yang And Pedersen gave a comparative study of various feature selection methods. The different methods describe were Document Frequency, Information Gain, Mutual Information, Chi Square Statistics and Term Strength. The results showed that IG, Chi and DF were reliable in selecting informative words. This process select irrelevant and redundant feature from data. After this process data is helpful in improving the performance of the algorithm.

1) Mutual Information:
It is used to measure the mutual dependence of the two terms in a paragraph or in a document. The formula used for mutual information to calculate mutual dependence between term t and category c is

\[ I(t,c) = \log \frac{P(t,c)}{P(t)P(c)} \]

2) CHI Square Statics
It is used to measure the lack of independence between the term w and the category c. If w and c are independent then the CHI will have a lowest value of 0. Its formula is

\[ \text{CHI} (w,c) = N \times \frac{(P(w,c) - P(w)P(c))^2}{P(w)c + P(w)c} \]

3) Regularized Mutual Information
Regularized mutual information measures the relevance of a term in the category. Its formula is

\[ \text{RMI} = 2MI(t,c) / H[t] + H[c] \]

4) Weight of Term
It is used to measure the weight of term appearing frequently as well as rarely in the document.

\[ WT = TF(t) \times MI(t,c) \]

VI. DATA SET
The dataset used for this project is REUTERS-21578. It the collection if articles related to different categories. The different categorized are health, education, science, sports, movie, business and travel.

<table>
<thead>
<tr>
<th>Category</th>
<th>No Of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>22</td>
</tr>
<tr>
<td>Health</td>
<td>29</td>
</tr>
<tr>
<td>Sports</td>
<td>29</td>
</tr>
<tr>
<td>Education</td>
<td>22</td>
</tr>
<tr>
<td>Science</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 1: corpus statics

VII. PERFORMANCE MEASURES
Performance measure for classification:
- \[ \text{RECALL} = \frac{a}{a + b} \]
- \[ \text{PRECISION} = \frac{a}{a + c} \]
- \[ F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

\[ a = \text{number of documents correctly assigned} \]
\[ b = \text{number of documents incorrectly assigned} \]
\[ c = \text{number of documents rejected incorrectly} \]

VIII. EXPERIMENTAL RESULTS
After combining WT with the standardized method the value of all the performance measure is increased.

Fig. 3: Results

IX. CONCLUSION
The graph clearly shows that WT increase the performance measures that are precision, recall and F1. WT has boosted the performance by using the fact that rarely occurring words are effective in classifying a document.

X. FUTURE SCOPE
Future work would be to combine the Weight of Term factor with the other classifier such as neural networks and support Vector Machine (SVM), more work are to be done on standard feature selection method. More work to be done to reduce the classification time.

REFERENCES
[7] Xiang Zhang, Mingquan Zhou, Guohua Geng, Na Ye (2009), “A Combined Feature Selection Method for Chinese Text Categorization”, College of Information Science and Technology, Northwest University, Xi’an, China, College of Information and Control Engineering, Xi’an University of Architecture and Technology, Xi’an, China, College of Information Science and Technology, Beijing Normal University, Beijing, China.


[14] Zonghu Wang, Zhijing Liu (2010) “Graph Based K-NN Text Classification “ School of Computer Science & Technology, Xi’an University, Xi’an, China.