ISSN 2229-5518

Graph Based Text Classification

*Lakshmi Kumari,

ABSTRACT

In this paper, a graph based classifications model is proposed on the basis of the word semantic space. This model can solve the problems of vector space model, such as the order or words, the boundary between sentences and phrases, etc. Different feature selection methods are also explained in this paper.

Keywords – Feature Selection, Graph based model, K-NN algorithm, Preprocessing, Semantic Space, Text Classification, and Vector Space Model

1.INTRODUCTION

With the development of internet, a large amount of data in any organization needs efficient classification. The traditional text classification techniques have been based on vector space model which ignored the structural information of the document that is word order and co-occurrence of the words in the document. Therefore the paper has used graph based technique which takes into account the structural information of the document.

The Graph based technique has been recently developed by Schenker A. and Zhaotao et al. More work on graph based technique has been done by Wei Jin, Rohini k and Maual and Alxander. Word semantic space has been proposed by Zhao et al [1] in 2010 Graph based text classification has been introduced by Whang and Liu [2] in 2010.

2.FEATURE SELECTION AND PREPROCESSING

The text classification begins with preprocessing which includes the stop word removal and stemming of the document. The stemming algorithm that the paper has used is porter stemmer. After preprocessing the next phase is feature selection. The feature selection that the paper has used are MI[3] with Chi, RMI[4] with Chi [5] and WT .

RMI

Regularized mutual information measures therelevance of a term in a category. It is effective than mutual information and do not takes into account the numerical values.

RMI = 2MI (t, c) / H[t] +H[c]

Weight Of Terms

It is formed by replacing the IDF [inverse document frequency] in TF-IDF. It is used to measure the weight of terms appearing frequently as well as rarely in the document.

WT = TF (t).MI (t, c)/MI

It is used to measure the mutual dependence of the two terms in a paragraph or in whole document.

I (t, c) = log P (t, c) / P(t/c) *P(t)

I (t, c) = log P (t/c) – log P (t)

Where P (t, c) is the probability of the term t in the category c, P (t) is the probability of the term t and P(c) is the probability of the category c.

Chi Square Statistics

It is used to measure the lack of independence between the term t and the category c.

$CHI(w,c) = N*(P(w,c)* P(\overline{w}, \overline{c}) - P(w,\overline{c})*P(\overline{w},c))/P(w)* P(\overline{w})*P(c) * P(\overline{c})$

P(w) is the probability of w in the document d and P(c) is the probability when the text belong to category c. $P(\overline{w}, c)$ is the probability that word do not occur in the category, $P(\overline{w}, \overline{c})$ is the probability that word w and category do not appear .similarly the meaning of rest of the terms can be known.

3. GRAPH BASED TEXT CLASSIFICATION

After the feature selection step the whole text converted into graph based on the features selected.

3.1. Graph Based Text Representation Model [7]

A graph is 3 tuple G= (V, E, F, W, M), where V is a set of nodes, E is a collection of weighted edges connecting nodes. FWM (Feature Weight Matrix) [8] is defined as the feature weight matrix of the edges.

- Node:

Unique feature terms obtained from the train set using feature selection methods.

- Edges:

Constructed based on order and co-occurrences relationship between feature words.

- Feature Weight Matrix:

Here every document is represented as incidence matrix. The weight w of the edge indicates the degree of constraint between the two features related to the edge. The weight between the two features is semantic measure which is defined as

$$W_{AB} = 1/ (\text{num}(B) - \text{num}(A))$$

Where num (B) is the order of the feature A in the document, num (A) is the order number of the feature B in the document. If the two feature terms appear one after another in the document and A appears before B, then in the picture G there is directed edge from A to B and the weight is 1. this phenomenon is called A directly restrict B or B is directly restricted by A. if the feature A and B are not adjacent and A appears before B, then in the graph G there is a directed edge, and its weight can be computed through the formula (5).This is called A indirectly restrict B or B is indirectly restricted by A.

4. IMPROVED KNN CLASSIFICATION BASED ON GRAPH

As a graph consists of nodes, edges and the weight of the edges, we can define the similarity measure of two graphs by those elements. Three different algorithms [8] have been used for classification. First has been used to convert the text into graph and then two improved classification measures have been used in calculating the similarity between two graphs and finally classifying the document.

First of all the document will be converted into graph based representation with the help of following matrix

$T=[t_1,t_2,\ldots t_n]$

$$M=\begin{pmatrix} a_{11} & a_{12} & .. & .. & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & .. & .. & a_{2,n-2} & a_{2n} \\ \ldots & \ldots & .. & .. & \ldots & .. \\ a_{n-1,1} & a_{n-2,2} & .. & .. & a_{n-1,an-1} & a_{n-1,n} \\ a1n & an2 & \ldots & .. & a_{n,n-1} & a_{nn} \end{pmatrix}$$

In the above, T is the set of features; t_i is the Feature for i = 1, 2,, n. M is the incidence matrix of the features; a_{ij} is the relevance degree between the features t_i and t_j ($1< i <j < n$).If some word A restricts another word B several times, then the nearest constraint (the maximal constraint) between them is considered. According to the definition 1, the maximal constraint is 1 and the matrix U is obtained:

$$U=\begin{pmatrix} u_{ll} & u_{12} & u_{l,n-l} & u_{ln} \\ u_{21} & u_{22} & u_{2,n-l} & u_{2n} \\ u_{n-I,l} & u_{n-I,2} & u_{n-l,n-l} & u_{n-I,n} \\ u_{nl} & u_{n2} & u_{n,n-l} & u_{nn} \end{pmatrix}$$

The Matrix U need to be normalized.

Let $W_{ij} = U_{ij} / \sum_k \sum_l U_{kl}$

where I,j,k,l = 1,2,….n. Then normalized matrix w is as follows

$$W=\begin{pmatrix} w_{11} & w_{12} & .. & .. & w_{1,n-1} & w_{1n} \\ w_{21} & w_{22} & .. & .. & w_{2,n-2} & w_{2n} \\ \ldots & \ldots & .. & .. & \ldots & \ldots \\ w_{n-1,1} & w_{n-1,2} & .. & .. & w_{n-1,n-1} & w_{n-1.n} \end{pmatrix}$$

ISSN 2229-5518

W_{n1} W_{n2} .. .. $W_{n,n-1}$ $W_{n,n}$

Input :

Training set $D=\{d_1,d_2\ldots d_i\ldots d_n\}$ D_i is a text after segment and stop words filtering $D_i=\{f_1,f_2,..f_i..f_m\}$, f_i is the i-th word of text f_i = Feature selected, N_i=Node w_i = Weight

Output :

Training set $G=\{g_1,g_2,\ldots g_i,\ldots g_n\}$ g_i is the i-th text represented by graph

Procedure:

1. For each d_i in D
2. Initialize the node set N_i , edge set E_i and Feature Weight
Matrix FWM_i to be empty.
3. For each f_i in d_i
4. If($f_i \in N_i$)
5. create a new node n_i representing f_i,
6. add n_i to N_i,set w_{ii}=1 // w_{ii} is defined in (3)
7. End If
8. End for
9. For each f_i in d_i
10. Create a new edge e_i connecting f_i and f_{i+1}
11. find the node n_k which representing f_i
12. If($e_i \in E$)
13. add e_i to Ei, set weight e_i =1
14 w_{kk++}; // w_{kk} represents the frequency of n_k
15. Else If($e_i \in E_i$)
16. weight e_i ++;
17. w_{kk++}
18. End If
19. End For

Algorithm1. Text to graph conversion

FW(feature weight): It describes the similarity between two graphs by weight of both nodes and edges appear in both two graphs. It can be calculated as follows.

Testing set graphs $G=\{g_1,g_2,\ldots g_i,\ldots g_n\}$
Training set graphs $CG=\{cg_1,cg_2,\ldots cg_i,\ldots cg_n\}$ w_{ij} = weight of the edge

Fw= Feature Weight

Procedure:

1. For each edge in g_i
2. If edge in cg_i
3. If($w_{ij}(g_i)$>= w_{ij} (cg_i) // w_{ij} is the weight of edge
4. If(j>i)
5. Fw+=α w_{ij} (cg_i)
6. Else if(j=i)
7. Fw+= w_{ij} (cg_i)
8. End if
9. Else If(w_{ij} (g_i)< w_{ij} (cg_i)
10. If(j>i)
11. Fw+=α w_{ij} (g_i)
12. Else if(j=i)
13. Fw+= w_{ij} (g_i)
14. End if
15. End if
16. End if
17. End for

Algorithm 2. Calculation of feature weight

The following algorithm has been used for the final classification of the document into its category. **Input:**
Testing set graphs $G=\{g_1,g_2,\ldots g_i,\ldots g_n\}$, value k =5 Training set graphs $CG=\{cg_1,cg_2,\ldots cg_i,\ldots cg_n\}$
Nfp= Node Fit Percent
Efp= Edge Fit Percent
F_w = Featured Weight
Output :
Result set $R=\{r_1,r_2,\ldots r_i,\ldots r_n\}$

Procedure:

1 For each g_i in G
2. Initial List RL to store F_w and text category (length is K)
3 For each cg_i in CG
4. If Nfp(g_i,cg_i)>α && Efp (g_i,cg_i)> α
5. Calculate Feature weight F_w (g_i,cg_i)
6 If RL is not full
7 Add F_w (g_i,cg_i) and category of cg_i to RL
8 Else If RL is full
9 If F_w (g_i,cg_i) >min (F_{wi} in RL)
10 Replace F_{wi} in RL with F_w(g_i,cg_i)
11 End if
12 End if
13 End if
14 . End For
15 the category of g_i is the category appears most in RL
16 add the category of g_i to the Result Set R.
17 End For

Algorithm 3. Classification of document

5. Conclusion

In this paper semantic space method had been proposed with the graph based method so as to have a better and efficient classification. This can be further combined with different classifier to have efficient classification.

References

[1] Faguo Zhao, Fan Zhang ,Bingru Yang ,2010 .Graph Based Text Representation model and its Realization. IEEE.

[2] Zonghu Wang , Zhijing Liu ,2010 .Graph-based Chinese Text Categorization. *In Proc. Of Seventh International Conference On Fuzzy systems and Knowledge Discovery*. pp 2363- 2366.

[3] Xiang Zhang, Mingquan Zhou, Guohua Geng, Na Ye, 2009. A Combined Feature Selection Method for Chinese Text Categorization. *In Proc. Of International Conference Information Engineering and Computer Science,* Wuhan, pp. 1-4.

[4] Yao- Tsung, Chen, Meng Chang Chen, 2011. Using Chi-Square Statistics To Measure Similarities For Text Categorization. Taiwan. pp – 3085-3090. 14.

[5] Zonghu Wang, Zhijing Liu, 2010. Graph Based K-NN Text Classification**.** *In Proc. Of International Conference on Electrical and control engineering.* pp 1092-1095.

[6] Zhou, Fan Zhang, Bingru Yang ,2005. Towards Graph-based Text Representation Model and Its Realization. *In Proc. Of International Conference on Natural Language and Knowledge Engineering , CNLP-KE*, Beijing ,vol.19, pp. 1-8 .

[7] Zonghu Wang , Zhijing Liu ,2010 .Graph-based Chinese Text Categorization. *In Proc. Of Seventh International Conference On Fuzzy systems and Knowledge Discovery*. pp 2363- 2366.

Lakshmi Kumari , Mtech, Working as Assistant Professor in HMR Institute of Technology Management , GGSIPU , New Delhi.

Emailid- singhlakshmik@gmail.com