SENTIMENT ANALYSIS OF WORDS USING FEATURE MAPPING BASED ON LABELS INFORMATION TECHNIQUE

A.SHAJI GEORGE[1], Dr.K.KRISHNAMOORTHY [2]

[1] Research Scholar,Faculty of Computer Science & Engineering, CMJ University,
Jorabat, Ri-Bhoi District, Meghalaya-793101, India
[2]Professor, Department of Computer Science & Engineering, Sudharsan Engineering College,
Pudukkottai-622003 ,Tamilnadu, India

ABSTRACT

In a feature representation, sentiment analysis plays an important role in text classification. It refers to natural language processing based techniques used to identify, extract or characterize subjective information and to classify the various topics into positive, negative or neutral categories. Using the proposed attribute weightage method, the attribute weightage is calculated and that value is passed to the Euclidean distance for clustering.The other feature, which is introduced in the modified bisecting K-Means algorithm, is the selection of cluster for splitting further. K-means algorithm to cluster the positive and negative samples of each label so as to extract the general characteristics. Feature mapping based on labels information (FM-BOLI) is an algorithm suitable for Binary Relevance. Compared with the conventional text representation, it makes the dimension of the text under control by means of word embedding. Then we use an AANN (Auto Associative Neural Networks) is a special kind of neural networks that is used to simulate associative process, which is connected through weighted connections. Results show that the FM-BOLI technique is more effective than other techniques in this application.

INTRODUCTION

Often text mining, also known as text data mining or text analytics, is confused with information retrieval, the correct definition of text mining is the “the process of deriving high-quality information from text”. Compared to data mining, which processes structured information and extracts useful information from data sets to transform them for further use, text mining takes care of unstructured information. Basically, it’s the process that allows identification of new and unexpected information from a collection of text.

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Text mining technology is now broadly applied to a wide variety of government, research, and business needs. All three groups may use text mining for records management and searching documents relevant to their daily activities. Legal professionals may use text mining for e-discovery. Governments and military groups use text mining for national security and intelligence purposes. Scientific researchers incorporate text mining approaches into efforts to organize large sets of text data (i.e., addressing the problem of unstructured data), to determine ideas communicated through text (e.g., sentiment analysis in social media) and to support scientific discovery in fields such as the life sciences and bioinformatics. In business, applications are used to support competitive intelligence and automated ad placement, among numerous other activities.

Using NLP, statistics, or machine learningmethods to extract, identify, or otherwisecharacterize the sentiment content of atext unit. Sometimes referred to as opinion mining,although the emphasis in this case is onextraction. Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation, affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication.

In many social networking services or e-commerce websites, users can provide text review, comment or feedback to the items. These user-generated text provide a rich source of user's sentiment opinions about numerous products and items. Potentially, for an item, such text can reveal both the related feature/aspects of the item and the users' sentiments on each feature. The item's feature/aspects described in the text play the same role with the meta-data in content-based filtering, but the former are more valuable for the recommender system. Since these features are broadly mentioned by users in their reviews, they can be seen as the most crucial features that can significantly influence the user's experience on the item, while the meta-data of the item (usually provided by the producers instead of consumers) may ignore features that are concerned by the users.

Clustering of Data with Weightage Analysis of Word

Data clustering is an important technique for data analysis which can be used to discover the similarity or dissimilarity between groups of items in a dataset such that items in one group are more similar than the other groups and vice versa. K-Means algorithm is the most popular partition based algorithm which is widely used in data clustering.

K-Means algorithm have been proposed for data clustering due to its simplicity, efficiency and ease of convergence. K-Means is one of the oldest and most commonly used clustering algorithms. It is a prototype based clustering technique defining the prototype in terms of a centroid which is considered to be the mean of a group of points and is applicable to objects in a continuous n-dimensional space.

Using the proposed attribute weightage method, the attribute weightage is calculated and that value is passed to the Euclidean distance for clustering. The other feature, which is introduced in the modified bisecting K-Means algorithm, is the selection of cluster for splitting further.

In this paper, we formulate intuitive properties that may allow a user to select an algorithm based on how it treats weighted data. Based on these properties we obtain a classification of clustering algorithms into various categories: those that are affected by weights on all data sets. Among the methods that always respond to weights are several well-known algorithms, such as k-means and k-median. On the other hand, algorithms such as Feature mapping based on labels information and Auto associative neural networks.

Auto-associative neural networks are feedforward nets trained to produce an approximation of the identity mapping between network inputs and outputs using backpropagation or similar learning procedures.

The key feature of an auto-associative network is a dimensional bottleneck between input and output.FM-BOLI is an algorithm suitable for Binary Relevance (positive and negative). Compared with the conventional text representation, it makes the dimension of the text under control by means of word embedding.Besides, it emphasizes the importance of specific label features more than common weighting algorithm. It extracts text features better so as to improve the performance of classification.The algorithm can still be improved in many aspects.

LITERATURE SURVEY

Nitesh Pradhan et al proposed a two types of similarity called lexical similarity and semantic similarity. The lexical similarity determines the similarity between word and text based on character by character matching. Nine algorithms were summarized; four of them are character based and remaining are statement based similarity. Semantic similarity determines the similarity between word and text based on their meaning. It can be used in biomedical ontology and can be applied to find similar geographic features.

Niphat Claypo et al proposed the clustering of customer opinions for restaurants by using unsupervised learning algorithm. In addition, the proposed method apply MRF feature selection technique for selecting relevant features. This can effectively reduce the number of features and computational times. Then, K-Means is adopted for clustering the reviews of restaurants into positive and negative groups. The experimental results showed that K-Means clustering is compatible with MRF feature selection since it can achieve the best performance in the clustering.

Ee-Peng Lim et alproposed scoring methods to measure the degree of spam for each reviewer and apply them on an Amazon review dataset. We then select a subset of highly suspicious reviewers for further scrutiny by our user evaluators with the help of a web based spammer evaluation software specially developed for user evaluation experiments. Our results show that our proposed ranking and supervised methods are effective in discovering spammers and outperform other baseline method based on helpfulness votes alone. We finally show that the detected spammers have more significant impact on ratings compared with the unhelpful reviewers.

Hao Lidong et al proposed a *Multi Mixed Convolutional Neural Network (MMCNN)* model to analyze the sentiment of online product comments. We mix the convolution and pooling features in mixed layer to enhance effectiveness of the online comments sentimental analysis. The skip-gram model is used to train the word vector. Because the length of each comment is not fixed, two new empirical matrix filling methods are designed which cyclic matrix filling and random matrix are filling.

Soujanya Poria et alproposeda method for assigning emotion labels to SenticNet concepts based on a semi-supervised classifier trained on WordNet-Affect emotion lists with features extracted from various lexical resources.

Yuxiang Bao et al proposeda brand new holistic system, which can deal with all the problems above simultaneously using aspect-based positive center similarity (ABPCS) model. We experiment our system on clothes and hotel domain, and the result shows considerable improvements over state-of-the-art baselines.

Yu-Hsun Lin purpose of the proposed method is to provide current filtering methods with another viewpoint in terms of discussion and analysis. Most important, the success of the method has been proven, both in the laboratory and onsite.

In terms of the results of onsite measurements, the method of noise suppression proposed by this study could improve the effectiveness of evaluating the insulation status of power equipment.

Zhenlong Sun proposed a novel combination forecasting model and applied the model to the prediction of Five Cities' election in Taiwan. Specifically, the exposure rate and approval rate which single models output were used for evaluating candidates and then added them to generate last forecasting result. The experiments showed that combination forecasting model based on machine learning successfully predicted four cities' election results of five cities'. The paper argued the methods of feature extraction and clustering.

Wang Yangproposed a statistical model based on these newcharacteristics. The model has the flexibility to deal with the "soft NLOS" and the "hard NLOS"environment, which is shown in measurements, by defining the polarity of a particular model parameter. Therefore, the channel impulse responses (CIR)generated by the proposed mode "resemble" the measured channel impulse responses better than SVIIEEE 802.15.3a model in terms of the cumulative distribution functions (CDFs) of the small-scale statistics, instead ofjust the average values.

METHODOLOGY

Weight Calculation

Auto-associative neural networks are fully connected recurrentnetworks where all nodes are inter-connected except to themselves. The network is fed with a training vector $X(0)$ where each element corresponds to one node of the unique layer. The activation is spread through the connections and generates a new output vector Xi.

Then the synaptic weights W are adjusted using a Hebbian rule. The new output vector is reprocessed through the network and generates a second new output vector Xi. This cycle continues until two consecutive output vectors exhibit the same values.

This query stable state is used in a similarity calculation with each 'document' stable state to order the document's relevancy to the query. A standard hyperbolic tangent is used as the spreading activation function.

$$\mu_i(t+1) = \frac{1}{2}\left[1 + tanh\left(\left(\left(\sum_{j=i,j\neq i}^{N} w_{ij}\,\mu_j(t)\right) - \theta\right)/\lambda\right)\right]$$

$\mu_i(t)$is the activation level of node ***i*** at iteration ***t***,
w_{ij} is the connection weight from node ***i***to node ***j***,

λis the slope of the function,

θ is the activation threshold.

When processing a document vector, the output state stabilizes when $|\mu_i(t+1)$ - $\mu_i(t)| <$ **ε,** ∀ i. λand θ are two free parameters to be determined empirically. During the training phase, the weights of the synaptic matrix ***W*** are updated according to the following Hebbian learning rule.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha\left[x_i(t)x_j(t)\right] - \beta\left[x_i(t)x_j(t)\right]$$

Where w_{ij} (t) is the connection weight from node ***i***to node ***j*** at iteration ***t***,

$x_i(t)$is the output of node ***i***at iteration ***t***,

α is the learning rate parameter,

β is a forgetting rate parameter.

The first part of this rule is a regular Hebbian learning rule where the parameter α controls the learning rate. The rule states that the weights connecting the nodes ***i*** and ***j*** will getupdated by a portion (0 <α≤ 1) of the 'correlation' between the two nodes ($\boldsymbol{x_i \cdot x_j}$).

We used a weight representation for the documents and thequeries, where the weights are defined according to the *tf* × *idf*scheme.

$$w_{ij} = tf_{ij} \times idf_i = tf_{ij} \times \log\left(\frac{N}{n_i}\right)$$

w_{ij} is the weight of term ***i***for the document ***j***,
tf_{ij} is the frequency of term ***i*** in document ***j***,
$idf_i$is the inverse document frequency of term ***i***,
$n_i$is the number of documents that include term ***i***,
N is the total number of documents.

The *idf* factor captures the discriminating power of the term. If a term appears in too many documents, then it is not useful for discriminating those documents. This is reflected in the logarithm function. As the number of documents includingterm ***i*** (***n***$\boldsymbol{n_i}$) approaches the total number of documents in the collection (***N***), the *idf* factor approaches zero. The final weight of such term will approach zero, no matter how frequent it is in each document. On the contrary, if only one document includes a term, its *idf* factor will retain a maximumdiscriminating power.

The *tf* × *idf* weighting scheme states that a term is as important to represent a document, as it is frequent within the document and rare among all documents of the collection.

The feature mapping based on label information is compared with the conventional text representation, it makes the dimension of the text under control by means of word embedding.The FM-BOLI takes advantage of word embedding whose dimension is controllable and meaning can be measured by distance calculation. Besides, it emphasizes the importance of specific label features more than common weighting algorithm. It extracts text features better so as to improve the performance of classification.

FEATURE MAPPING BASED ON LABEL INFORMATION

Generally, features are categorized into related features,irrelevant features and redundant features. FM-BOLI is basedon the prior knowledge that the higher the similarity between aword and related features is, the word should be given a higherweight.

In Multi-Label Text Classification(MLTC), the positive samples of each label include general charactersof the label and the general characters of negative samples canbe regarded as the irrelevant features.

Accordingly, the implementation of the algorithm is divided into two steps. First, extract general features of positive samples and negative samples corresponding each label. In the case of unsupervised learning, clustering can obtain clusters with the similar characteristics. Secondly, map each text to aemphasizes the importance of label information and increases text information by extending dimension of text representation.

Clustering for positive and negative samples

To search for the general characters, FM-BOLI algorithm takes the method of text clustering. The text representation in the phase is assigning tf-idf weight to each word, which is the state-of-the-word method of text feature representation using word embedding. The specific method are as follows:

```
Algorithm: Clustering Implementation

for each label j do
PS (positive samples) = ($p_{1,}$ $p_2$, $p_{3,}$ …, $p_m$)

for each positive sample $p_i$ in PS do
        Split each text $p_i$ into word set {$s_1$,$s_{2,}$ $s_3$,…,$s_i$}
        The feature representation of $p_i$ is

                $x_i = \sum_{j=1}^{l} weight_{wj} \;*\; vector_{wj}$
end for

the positive sample set $Y_p$= {$y_{1,}$ $y_2$, $y_{3,}$ …, $y_m$}
    NS (negative samples) =  ($n_{1,}$ $n_2$, $n_{3,}$ …, $n_n$)
for each negative sample $n_i$ in NS do
        Split each text $n_i$ into word set {$s_1$,$s_{2,}$ $s_3$,…,$s_l$}

The feature representation of $n_i$ is
$x_i = \sum_{j=1}^{l} weight_{wj} \;*\; vector_{wj}$
end for
the negative sample set $Y_N$ = {$y_{1,}$ $y_2$, $y_{3,}$ …, $y_n$}
cluster for $Y_P$ and $Y_N$
end for
```

$weight_{wj}$ and $vector_{wj}$ represent tf-idf weight and word embedding of the word $w_j$respectively. The algorithm of clustering is K-means.

In this way, C_p = {$c_{1,}$ c_2, $c_{3,}$ …, c_P} (the set of positive sample's clustering centers) and C_N = {$c_{1,}$ c_2, $c_{3,}$ …, c_N} (the set of negative sample's clustering centers) can be obtained. Accordingly, the dimension of each clustering center is just the dimension of word embedding that we trained utilizing CBOW neural network.

Feature mapping algorithm

To emphasize the effect of each cluster center on the label characteristics, text feature representation would be extended to the dimension of the cluster center. Euclidean distance is chosen to measure the similarity between each word in a text and each clustering center.

Since the MLTC is transformed to q independent binary classification problems, feature mapping algorithm in only one binary classification problem will be described. The specific method are as follows:

```
Algorithm: Feature Mapping

for each $p_i$ in training samples do
    Split the text $p_i$ into word set $\{s_1,s_2, s_3,\ldots,s_i\}$
    for each $c_i$ in $C_p$do

            $\tau_i = \frac{1}{l}\sum_{j=1}^{l} \frac{1}{\alpha + d(vector_{wj}, c_i)} * vector_{wj}$

    end for
    The result of positive sample's feature mapping is
                    $[\tau_1, \tau_2, \tau_3, \ldots \tau_P]^T$
    for each $c_i$ in $C_N$do

            $\theta_i = \frac{1}{l}\sum_{j=1}^{l} d(vector_{wj}, c_j) * vector_{wj}$
    end for

The result of negative sample's feature mapping is
                    $[\theta_1, \theta_2, \theta_3, \ldots \theta_N]^T$
    Finally, the feature representation of the text $t_i$ is

                    $x_i = [\tau_1, \tau_2, \tau_3, \ldots \tau_P, \theta_1, \theta_2, \theta_3, \ldots \theta_N]^T$
end for
```

$\alpha$is a term of Laplacian Smoothing, which is preventing the denominator from being 0, where the value is 0. 00000001. $d\ (vector_{wj}, c_j)$ is the Euclidean distance between $vector_{wj}$ and c_j. x_i is a matrix of $(P+N)*d$, P and N is the number of positive and negative samples' clustering centers,d is the dimension of the word embedding.

Sentiment Analysis

Opinion mining (sometimes known as sentiment analysis or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine

Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event.

The attitude may be a judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).

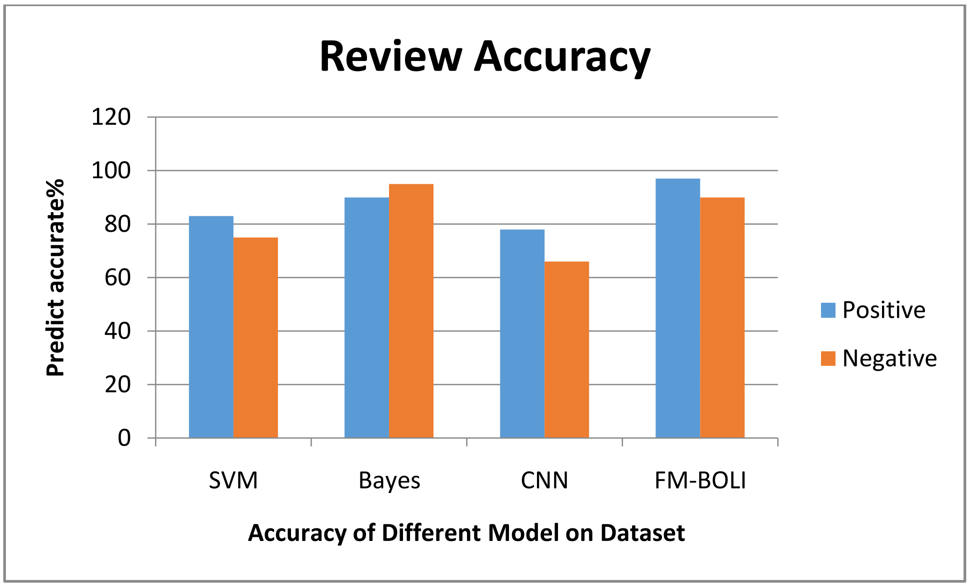
It normally involves the classification of text into categories such as “positive”, “negative” and in some cases “neutral”. The main reason sentiment analysis so difficult is that words often take different meanings and are associated with distinct emotions depending on the domain in which they are being used.

Efficiency Analysis

To evaluate the performance of our model, we adopt the Accuracy metric, which is defined as:

$$\text{Acc}=\frac{T}{N}$$

where T is the number of correctly predicted samples, N is the total number of samples. Accuracy measures the percentage ofcorrect predicted sample in all samples.



The above graph shows the Accuracy of FM-BOLI model based method, classical convolutional neural network based methods and SVM and Naive Bayes based method on Dataset respectively.

The dataset is the online comments about infant milk power sold in Jindong Mall (www.jd.com). 46,000 comments generated from 2015 to 2017 are crawled. We labelled the comments with 4 stars or 5 stars as positive comments. The comments with I star and 2 stars are labelled as negative comments. The comments with 3 stars are ignored. There are 23000 positive comments and 23000 negative comments.

In our hyper parameters Settings, we compare the Accuracy of FM-BOLI model based method with SVM, Naive Bayes, and classical convolutional neural network based methods. *Chi-Square* feature selection is applied in SVM, Naive Bayes methods. Classical convolutional neural network's *minibatch*is set to 100 and dropout is set to 0.5. 3x3 convolution layer is employed. lxl, 3x3, 5x5 convolutional layer and 2 x2 max-pooling layer are used in FM-BOLI. Other parameters are set as that ofCNN.

Precision:

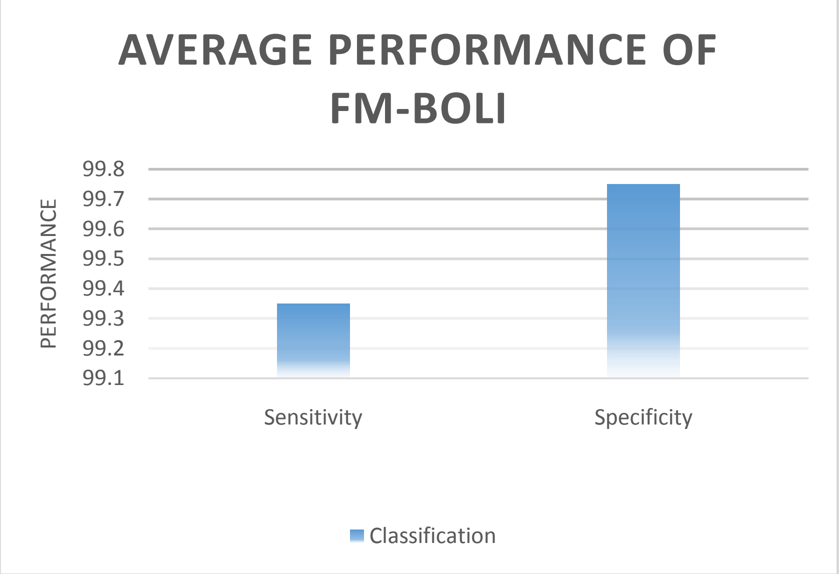
Precision is measured over the total predictions of the model. It is the ratio between the correct predictions and the total predictions. In other words, precision indicates how good the model at whatever it predicted is.

$$\text{Sensitivity (SE)} = 100.\frac{TP}{TP+FN}$$

Recall:

Recall is the Ratio of the correct predictions and the total number of correct items in the set. It is expressed as % of the total correct(positive) items correctly predicted by the model. In other words, recall indicates how good the model at picking the correct items is.

$$\text{Specificity (SP)} = 100.\frac{TN}{TN+FP}$$



CONCLUSION

We study the behavior of clustering algorithms on weighted data, presenting three fundamental categories that describe how such algorithms respond to weights and classifying several well-known algorithms according to these categories. Our results are summarized in the above graphs. Our analysis also reveals the following interesting phenomenon: algorithms that are known to perform well in practice, tend to be more responsive to weights. For example, k-means is highly responsive to weights while single linkage, which often performs the weight robust.

The FM-BOLI takes advantage of word embedding whose dimension is controllable and meaning can be measured by distance calculation. Besides, it emphasizes the importance of specific label features more than common weighting algorithm. It extracts text features better so as to improve the performance of classification.The algorithm can still be improved in many aspects. For example, extract features of positive and negative samples in other way instead of clustering. Possibly, more methods are adopted to measure the distance between two vectors.

REFERENCES

[1] Madjarov G, Kocev D, Gjorgjevikj D and Deroski S, An extensive experimental comparison of methods for multi- label learning, Pattern Recognition, 2012, 45(9): 30843104.

[2] Min-Ling Zhang, Zhi- Hua Zhou, A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, 2014, 26 (8) :1819-1837.

[3] BoutellMR, Luo J, Shen X, Brown CM, Learning multi- label scene classification. Pattern Recognition, 2004,37(9):1757-1771.

[4] J Rnkranz, E Llermeier, L Menc, A Eneldo and K Brinker classification via calibrated label rankin Machine Learning, vol. 73, no. 2, pp. 133153, 2008.

[5] G Tsoumakas, I Katakis and I Vlahavas -Labelsetsfo r multi IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 7, pp. 10791089, 2011.

[6] -KNN: A lazy learning approach to multi- Recognition., vol. 40, no. 7, pp. 2038 2048, 2007.

[7] Clare A, King RD. Knowledge discovery in multi-label phenotype data. In: Raedt LD, Siebes A, eds. LNCS 2168. Berlin: Springer-Verlag, 2001. 42-53.

[8] Elisseeff A, Weston J. A kernel method for multi- labelled classification. In: Dietterich TG, Becker S, Ghahramani Z, eds. Proc. of the Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2002. 681-687.

[9] Algorithm based on VSM and its application in Question Answering IEEE International Conference on IEEE, 2010, pp. 368 - 371.

[10] Weizhu Chen, Jun Yan, Benyu Zhang, Zheng Chen and Qiang Yang, Doc umentTra ns fo r mat io n fo r Multilabel Feature Selection in Text Categorization, IEEE International Conference on Data Mining, 2007, 80 (s 13) :451-456

[11] Bengio, Yoshua, Ducharme, Vincent, Pascal, Christian, et al. "A Neural Probabilistic Language Model. " Journal of Machine Learning Research, 3.6(2003), pp. 11371155.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient Estimation of Word Representation in Vector Space, arXiv:1301.3781v3 [cs.CL] 2013.

[13] Q Le, T Mikolov, Distributed Representations of Sentences and Documents, ICML'14 Proceed ings of the 31st International Conference on International Conference on Machine Learning, 2014, pp. 1188-1196.

[14] Liu Wanli, Liu Sanyang, Wang Jinyan, The adjustment for unbalanced support vector machine. computer science, 2009, 36(3):148-150.