2012 International Conference on Network and Computational Intelligence (ICNCI 2012)
IPCSIT vol. 46 (2012) © (2012) IACSIT Press, Singapore

A Domain Knowledge Based Weighting Approach in Protein Interaction Article Classification

Yifei Chen [1 +], Feng Liu [2] and Bernard Manderick [2]

[1] School of Information Science,Nanjing Audit University,86 Yushan Rd(W), 211815, Nanjing, P.R.China

[2] Computational Modeling Lab, Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussels, Belgium

Abstract. In this paper domain knowledge based feature representation and weighting approaches are proposed for interaction article classification (IAC) task. IAC is a specific text classification application in biological domain and tries to find out which articles describe protein interactions. However, the existing feature representation and weighting scheme commonly used for text mining are not well suited for IAC. We capture and use biological domain knowledge, i.e. gene mentions also known as protein or gene named entities, to address the problem. We put forward a new gene mention-based representation approach that highlights the important role of gene mentions in texts. Furthermore, we incorporate co-occurrences of gene mentions into a new feature weighting schema called gene mention-based term frequency (GMTF). It can indicate the potential interactions between proteins. Together with the extracted contextual features, our system performs better than other leading systems for the moment. It can classify biomedical literature with fairly high accuracy, which can achieve the precision of 78.55%, recall of 86.69% and balanced $F_{\beta=1}$ score of 82.42.

Keywords: interaction article classification, feature weighting, Biological Domain Knowledge

1. Introduction

In recent years together with the growing interest in biological research, especially the study of protein-protein interactions, an overwhelming amount of biological literature are published daily on the internet. To manage such an information overload problem, classification needs urgent settlement. It is essential to classify which articles are related to the knowledge we concerned. This makes protein-protein interaction extraction from biological literature more efficiently. Therefore study on automatic interaction article classification (IAC) has become a task with practical significance to the data mining in texts.

IAC is a text classification task in biological domain. In traditional text classification models, vector representations are premise and base to realize the automatic classification of texts. In most cases bag-of-word (BOW) approach [8] is used to transform texts into real value vectors. The basic idea is that different kinds of articles contain different kinds of words whose occurrences are clues for text classification. Using BOW, a text is represented by an entire vector and each component of the vector describes the value of one *feature* of the text, i.e. each distinct word that occurs in the training set is treated as one such feature. This method is easy to implement and efficient in computation.

The most straightforward approach of feature weighting is to assign binary weights, i.e. **1** for feature present and **0** for features absent in an input text. But in reality, different features have different importance in the text. In general, the tf-idf (term frequency-inverse document frequency) method is one of the most effective ways to calculate feature weighting. This method is improved by many literature [10,17,1].

+ Corresponding author. Tel.: 13770682463.
E-mail address: yifeichen91@nau.edu.cn.

However, all the weighting schemes are designed for generic purposes and hence may not be well suited for the biological domain.

This paper proposes new feature representation and feature weighting approaches based on the Gene Mentions (gene or protein names) that co-occur in the texts. Furthermore, the extracted feature can capture the contextual information in the text. Support vector machines are used as classifier models to implement a *Interaction Article Classifier* (*IACer*). Experiments show that the proposed method can pick out domain knowledge which makes great contributions to IAC and improve the classification performance.

2. Methods

2.1. Gene Mention-based Representation Schema

The traditional BOW representation treats gene mentions like other common English words. However, for the IAC application, gene mentions are very important because these gene mentions might be involved in the interactions to be discovered. Hence gene mention recognition is essential for *IACer*.

The study of the distribution of gene mentions in the training data shows that the average number of gene mentions per positive example (interactions are described) is 10.59, approximately twice as much as that per negative example (no interaction is described), 4.09. This is to be expected because the positive examples describe protein interactions and as a consequence should contain more gene mentions than the negative examples. Inspired by this noticing phenomenon that the gene mentions are subject to the skewed distribution, we are thinking that maybe adding the gene mention information to the traditional bag-of-word approach will lead to better performance.

To do this, the most straightforward approach is first to build a lexicon containing all the gene mentions appearing, and then replace every gene mention with its index in the lexicon. The resulting data set is called $\text{Text}_{\text{GmName}}$ because gene mentions are represented by the indices in the lexicon to their names.

Using this representation approach, it is obvious that the same gene mention has the same index throughout the whole collection of articles. However, this representation scheme has an implicit drawback. Because there are so many different gene mentions, the extracted gene mentions will be sparsely distributed. Most of the gene mentions occur only once or twice in the data set $\text{Text}_{\text{norm}}$, e.g. 81.1% of the gene mentions occurs once, 9.0% occurs twice, 3.3% occurs three times and only 6.6% occurs more than three times.

Because of this distribution we consider an alternative that we call the *gene mention-based representation* scheme. In this scheme instead of substituting each gene mention by its index, we replace it by GM_i where i indicates the order of its first appearance in a given article. Now the same gene mention will have the same representation only within the same article. In different articles the same GM_i might represent different gene mentions.

The second approach highlights general protein interactions in the articles instead of interactions between the specific proteins. In the IAC application, it is only important to detect whether protein interactions are described in a certain article but it is not necessary to know which specific proteins are involved. We will denote texts represented using the second approach by $\text{Text}_{\text{GmOrder}}$.

2.2. Feature Extraction

As stated already, bag-of-word features are very useful for text classification applications. We extract more bag-of-word features to capture contextual information as following.

- $Feature_{BOW}$: This kind of features are the standard bag-of-word features.
- $Feature_{BObiW}$: This kind of features are the bag of bi-gram words, i.e., the current word and its previous neighbouring word.
- $Feature_{BOcW}$: This kind of features are the bag of contextual words of gene mentions.

We tried window sizes from 2 to 15 on the training data and found out that the optimal window size is 10, i.e. the 5 words to the left and to the right of each gene mention.

2.3. Gene Mention-based Feature Weighting Schema

In the text, different features have different importance, hence weights should be introduced to reflect the importance of these features. Next we will present all the feature weighting schemes [16] that will be investigated.

Binary weighting scheme: The features take on binary values, i.e. the value 1 if the feature occurs in the text or 0 if it does not occur.

Term frequency (TF) weighting scheme: We define a normalized term frequency to measure the importance of feature f_i within the particular text $\bar{t}$:

$$TF_i = \frac{n_i}{\sqrt{\sum_{k=1}^{N} n_k^2}} \tag{1}$$

Where n_i is the number of occurrences of the considered feature f_i in text $\bar{t}$ and the denominator normalizes this expression.

$TF \cdot IDF$ weighting scheme: Sometimes, term frequency factors alone do not ensure acceptable classification performance. Hence a collection-dependent factor is introduced that favours the terms concentrated in a few articles of the collection. The well-known inverse document frequency (IDF) factor performs this function. The definition of IDF is as follows:

$$IDF_i = \log \frac{M}{m_i} \tag{2}$$

where M: the total number of articles in the collection;

m_i: the number of articles where the feature f_i appears

By combining Equations 1 and 2, the $TF \cdot IDF$ weighting scheme is:

$$TF \cdot IDF_i = TF_i \times IDF_i = \frac{n_i}{\sqrt{\sum_{k=1}^{N} n_k^2}} \cdot \log \frac{M}{m_i} \tag{3}$$

$TF \cdot TRF$ weighting scheme: Since the $TF \cdot TRF$ weighting scheme does not consider the relevance properties of the articles, [2] and [14] proposed the *term relevance frequency (TRF)* weight to take into account this kind of information. In theory, *TRF* is defined as the proportion of relevant articles in which a feature occurs divided by the proportion of non-relevant articles in which it occurs. However, it is rather difficult to compute *TRF* directly without any knowledge of the occurrence properties of the features in the relevant and non-relevant subsets of the article collection. It has been shown that under well-defined conditions *TRF* can be simplified to an inverse document frequency factor of the form $\log((M - m)/m)$ [3]. In order to make sure that $TRF \geq 0$, we make the following modification to *TRF*:

$$TRF_i = \begin{cases} \log \dfrac{M - m_i}{m_i} & if\ \mathrm{M} > 2\mathrm{m}_i \\ 0 & otherwise \end{cases} \tag{4}$$

Then, using Equations 1 and 4, we write the $TF \cdot TRF$ weighting scheme as:

$$TF \cdot TRF_i = TF_i \times TRF_i = \frac{n_i}{\sqrt{\sum_{k=1}^{N} n_k^2}} \cdot \max\left(0, \log \frac{M - m_i}{m_i}\right) \tag{5}$$

$TF \cdot RF$ weighting scheme: [11] proposed a new weighting method, $TF \cdot RF$, based on the idea that the higher the frequency of a frequency feature in the positive category as compared to its frequency in the negative category, the more important that frequency feature is to discriminate between positive and negative articles. *RF* stands for *relevance frequency* because only the relevant articles, i.e. those articles that contain this feature instead of the whole collection of articles, are considered. It is calculated as follows:

$$RF_i = \log \frac{NP_i}{NN_i} \tag{6}$$

Where NP_i: the number of relevant articles in the positive category

NN_i: the number of relevant articles in the negative category

In a similar way as with *TRF*, to make sure that $RF \geq 0$, we modify *RF* as:

$$RF_i = \log\left(\frac{NP_i}{NN_i} + 1\right) \tag{7}$$

Then, by combining Equations 1 and 7, the $TF \cdot RF$ weighting scheme becomes:

$$TF \cdot RF_i = TF_i \times RF_i = \frac{n_i}{\sqrt{\sum_{k=1}^{N} n_k^2}} \cdot \log\left(\frac{NP_i}{NN_i} + 1\right) \tag{8}$$

GMTF weighting scheme: All the weighting schemes above have one common drawback: they are designed for generic purposes and hence may not be well suited for the biological domain. In the interaction article classification application, the co-occurrences of gene mentions play an important role because protein interactions can only exist between genes mentioned. Hence, we think that for articles in the biological domain, the feature weight can be determined by two factors. The first one is the number of distinct gene mentions in a given article and the second one is the length of that article. The more distinct gene mentions occur in the article, the higher the weight the feature has. Similarly, the shorter the article, the higher weight the feature has. Inspired by Robertson's formula for *TF* [13,15], we propose the new feature weighting scheme, *gene mention-based term frequency (GMTF)*, defined as follows:

$$GMTF_i = \frac{TF_i}{TF_i + w_0 \cdot NGM + w_1 \cdot NAL} \tag{9}$$

where *NGM* is the *normalized* number of distinct gene mentions in a given article $\bar{t}$ and is defined as

$$NGM = \begin{cases} \text{Average \# GMs over all articles} & \text{(if no GM occurs in the article)} \\ \dfrac{\text{Average \# GMs over all articles}}{\text{\# GMs in a given article } \bar{t}} & \text{(otherwise)} \end{cases} \tag{10}$$

and *NAL* is the *normalized* article length of the given article $\bar{t}$ and is defined as:

$$NAL = \frac{\text{Total number of features in a given article } \bar{t}}{\text{Average number of features over all articles}} \tag{11}$$

The parameters w_0 and w_1 in Equation 9 have to be tuned. The parameter w_0, which can range between 0 and 5, weights the contribution of *NGM*. Higher values increase its contribution while $w_0 = 0$ eliminates the contribution of *NGM*. Trials on the training data showed that $w_0 = 0.1$ is the optimal value.

The parameter w_1, which can also ranges between 0 and 5, affects the contribution of *NAL*. Thus setting w_1 near 5, e.g. $w_1 = 4.5$, will increase its contribution. Trials on the training data showed that for our application $w_1 = 2.0$ is optimal. Therefore, we have set $w_0 = 0.1$ and $w_1 = 2.0$ during our experiments.

We can also combine *GMTF* in Equation 9 on the one hand with *IDF* in Equation 2, *TRF* in Equation 4 and *RF* in Equation 7 on the other hand to define several new weighting schemes:

$$GMTF \cdot IDF_i = GMTF_i \times IDF_i \tag{12}$$

$$GMTF \cdot TRF_i = GMTF_i \times TRF_i \tag{13}$$

$$GMTF \cdot RF_i = GMTF_i \times RF_i \tag{14}$$

3. Experiments and Results

In this section, we compare the fine-tuned *IACer* with other leading systems from the BioCreAtIvE II challenge. The data set, $Data_{BCII}$, used to build and evaluate all these systems is taken from interaction article task of the BioCreAtIvE II challenge [9]. The results are shown in Table 1. And the confidence intervals shown here are obtained by the *bootstrap resampling* method defined in [5] making use of 1,000 samples and for confidence level of $\alpha = 0.05$.

Table. 1: Comparison of IACer with other interaction article classifiers.

Method	Precision	Recall	$F_{\beta=1}$
TaiWan system	68.90%	85.07%	76.13
$TF \cdot RF$ system	-	-	77.75
GeneTeam system	75.00%	51.20%	61.00
p-spectrum system	73.52%	82.93%	77.94
IACer	78.55%	86.69%	82.42 ± 2.36

The TaiWan system [4] makes use of the SVM algorithm and the $TF \cdot IDF$ weighting scheme. Besides bag-of-word features, it also uses contextual bag-of-word features and extracts likely positive and likely negative data to enhance its performance.

The $TF \cdot RF$ system [12] proposes a new weighting scheme $TF \cdot RF$ to represent how much the bag-of-word feature contributes to the semantics of a document. This SVM algorithm together with feature ranking techniques achieved the second best performance in the BioCreAtIvE II challenge.

Unlike the two systems above, the GeneTeam system [6] does not select the words in the articles as features. Instead it makes use of MeSH (Medical Subject Headings) terms, interaction verbs and the number of proteins to build the system. Reported results indicate that this kind of features does not work well for the IAC application.

The *p*-spectrum system [7] got the best performance in the BioCreAtIvE II challenge. It differs from the the feature-based SVMs above in the sense that it does not extract features but uses the *p*-spectrum kernel to compare strings. As a consequence each article is treated as a string and for every two articles all common substrings of length *p* are computed using the *p*-spectrum kernel function. The best result is achieved for substring length $p = 7$.

IACer uses a SVM classifier with the polynomial kernel with *degree* 2 and *coefficient* 1 which is tuned on the training data using 5-fold cross validation to get the optimal parameters $C = 8500$ for the box constraint. The selected features considered are $Feature_{BOW}$, $Feature_{BObiW}$ and $Feature_{BOcW}$. *IACer* outperforms all the other systems shown in Table 1 with the differences significant for confidence level $\alpha = 0.05$ and it is the only system that obtains a 80+ $F_{\beta=1}$ measure. The best $F_{\beta=1}$ measure of *IACer*, i.e. 82.42, is achieved by the following procedures:

1. First we used a preprocessor on the original input $Text_{original}$ to generate *normalized* and *gene mention-based* texts $Text_{normGmOrder}$.
2. We not only considered $Feature_{BOW}$ but also other features, such as $Feature_{BObiW}$ and $Feature_{BOcW}$.
3. The *GMTF* feature weighting scheme is used to weight the features.

4. Discussion

This section is organized as follows. In Section 4.1 describes the contribution of the gene mention-based representation approach and Section 4.2 discusses the contribution of the different types of features together with the different weighting schemes.

4.1. Contribution of GM-based Representation Approach

In this section, we motivate our choice for $Text_{normGmOrder}$ in ICAer based on an experimental comparison. Hence, we have four different representations: $Text_{original}$ that is the original text, $Text_{norm}$ that is the preprocessed text, $Text_{normGmName}$ that combines $Text_{norm}$ and $Text_{GmName}$, and finally $Text_{normGmOrder}$ that combines $Text_{norm}$ and $Text_{GmOrder}$. We use $Feature_{BOW}$ with either the binary or the standard *TF* weighting scheme and the results are listed in Tables 2 and 3, respectively.

Table. 2: Contribution of different representations based on IACer with FeatureBOW and the binary weighting scheme.

	Precision	Recall	$F_{\beta=1}$
$Text_{original}$	67.85%	84.91%	75.43 ± 2.65
$Text_{norm}$	68.71%	86.39%	76.54 ± 2.57
$Text_{normGmName}$	71.50%	83.14%	76.88 ± 2.62
$Text_{normGmOrder}$	69.62%	86.95%	76.98 ± 2.60

Table. 3: Contribution of different representations based on IACer with TF and the binary weighting scheme.

	Precision	Recall	$F_{\beta=1}$
$Text_{original}$	63.35%	90.53%	74.54 ± 2.61
$Text_{norm}$	70.29%	86.09%	77.39 ± 2.55
$Text_{normGmName}$	69.05%	88.46%	77.56 ± 2.45
$Text_{normGmOrder}$	72.20%	87.57%	79.14 ± 2.47

From Tables 2 and 3, it can be seen that 1) $Text_{normGmOrder}$ is the best representation for both the binary and *TF* weighting scheme. 2) both gene mention-based representations considered have positive effects on the performance. Moreover, it can be seen that both $Text_{normGmName}$ and $Text_{normGmOrder}$ give significantly better results than $Text_{original}$ for confidence level $\alpha = 0.05$. And $Text_{normGmOrder}$ overcomes the problem of $Text_{normGmName}$ caused by the sparse distribution of gene mentions although this difference is not significant.

Therefore we conclude that $Text_{normGmOrder}$ is the best representation for *IACer*.

4.2. Contribution of Different Features and Weighting Schemes

In this section, we discuss in detail the contributions of 1) the different types of features and 2) the different weighting schemes.

Figures 1 shows the $F_{\beta=1}$ measures for the different features types and the different weighting schemes. We can draw the following conclusions:

1. $Feature_{BOcW}$ is critical for *IACer*. After it is combined with $Feature_{BOW}$, all the performances are improved, which implies that the context words of gene mentions play an important role. However, adding $Feature_{BObiW}$ does not lead to consistent improvement. In most cases, the $F_{\beta=1}$ measures are decreased when introducing $Feature_{BObiW}$. But there still are several exceptions, e.g. for all three types of features $Feature_{BOW}$+$Feature_{BObiW}$+ $Feature_{BOcW}$ are combined. This has a positive influence on the performance only for binary, *TF*, *GMTF*, $TF \cdot IDF$ and $GMTF \cdot IDF$ weighting. The best result is achieved when all types of features are combined with *GMTF*. Therefore, it can be concluded that a carefully designed weighting scheme, *GMTF* using $Feature_{BOW}$+$Feature_{BObiW}$+ $Feature_{BOcW}$ leads to the best performance.

2. Choosing the most appropriate feature weighting method depends on a number of factors such as the extracted feature sets.
    a. It can be seen immediately that the widely used *IDF*, *TRF* and *RF* do not improve the performance. And $TF \cdot IDF$, $TF \cdot TRF$, $TF \cdot RF$, $GMTF \cdot IDF$, $GMTF \cdot TRF$ and $GMTF \cdot RF$ perform even much worse than binary weighting. Therefore, document-related frequencies (*IDF*, *TRF* and *RF*) are not helpful at all.
    b. The best three results are achieved by using *GMTF* (80.17), *TF* (79.58) and binary (79.57) weighting schemes respectively. This means that these pure term frequencies (binary, *TF* and especially *GMTF*) are very beneficial for *IACer*. Moreover, because the traditional weighting schemes are not tailored for the biological domain, they are not the most suitable choice. The new feature weighting method *GMTF* capture the co-occurrences of gene mentions to revise the weights of features and hence is better adapted for the IAC application.
3. To summarize, the proposed *GMTF* weighting scheme together with the set consisting of the features $Feature_{BOW}$, $Feature_{BObiW}$ and $Feature_{BOcW}$, accomplishes the best performance because they can capture and use the domain specific information which are essential for the IAC application.

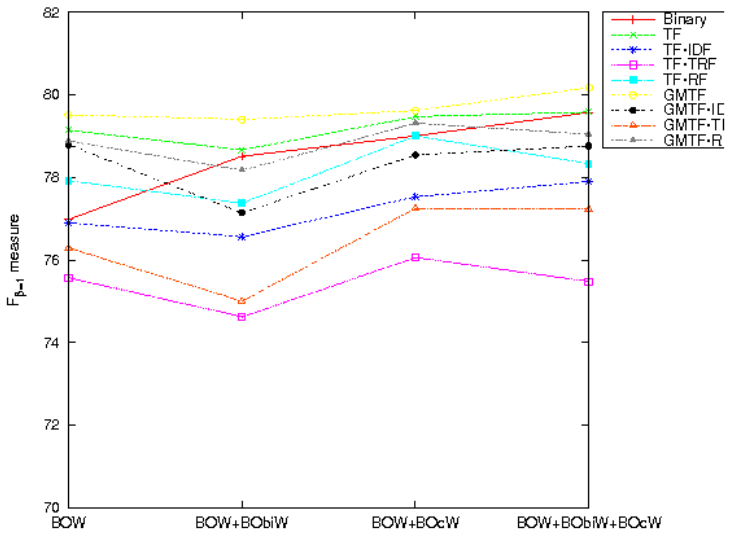

Fig. 1: Comparison based on the $F_{\beta=1}$ measures for the different features types and the different weighting schemes. "BOW": the results achieved using $Feature_{BOW}$; "BOW+BobiW": the results achieved using $Feature_{BOW}$ and $Feature_{BObiW}$; "BOW+BocW": the results achieved making use of $Feature_{BOW}$ and $Feature_{BOcW}$; "BOW+BObiW+BocW": the results achieved by using $Feature_{BOW}$, $Feature_{BObiW}$ and $Feature_{BOcW}$.

5. Conclusion

In this paper, biological domain knowledge was used to improve the performance of the *IACer* successfully. However, here only a small portion of the available external biological resources are used. In the future, we will introduce more background information on genes and proteins, including chromosomal locations, families, diseases, functions and biological processes related to the genes. This will provide more detailed and correlative information like where the gene is located on a chromosomal band, which family it belongs to, which diseases and functions it is related to, which kind of biological processes it takes part in.

6. Acknowledgements

The authors wish to acknowledge the support of the project 11KJB520007 of Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China.

7. References

[1] Z. Aihua, J. Hongfang, W. Bin, and X. Yan. Reseach on effects of term weighting factors for text categorization. Journal of Chinese Information Processing, pp. 97-104, 2010.

[2] W. S. Cooper and M. E. Maron. Foundations of probabilistic and utility-theoretic indexing. Journal of the ACM, 25(1):67-80, 1978.

[3] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35(4):285-295, 1975.

[4] H. Dai, H. Hung, R. T. Tsai, and W. Hsu. Iasl systems in the gene mention tagging task and protein interaction article sub-task. Proceedings of the Second BioCreative Challenge Evaluation Workshop, pp. 69-75, 2007.

[5] B. Efron and R. J. Tibshirani. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1994.

[6] F. Ehrler, J. Gobeill, I. Tbahriti, and P. Ruch. Geneteam site report for biocreative ii: Customizing a simple toolkit for text mining in molecular biology. Proceedings of the Second BioCreative Challenge Evaluation Workshop, pp. 199-207, 2007.

[7] M. Huang, S. Ding, H. Wang, and X. Zhu. Mining physical protein-protein interactions by exploiting abundant features. Proceedings of the Second BioCreative Challenge Evaluation Workshop, pp. 237-245, 2007.

[8] T. Joachims. Learning to Classify Text using Support Vector Machines. Kluwer Academic Publishers/Springer, 2002.

[9] M. Krallinger, F. Leitner, and A. Valencia. Assessment of the second biocreative ppi task: Automatic extraction of protein-protein interactions. Proceedings of the Second BioCreative Challenge Evaluation Workshop, pp. 41-54, 2007.

[10] Q. Kuang and X. Xu. Improvement and application of tfidf method based on text classification. 2010 International Conference on Internet Technology and Applications, pp. 1-4, 2010.

[11] M. Lan, C. Tan, and H. Low. Proposing a new term weighting scheme for text categorization. Proceedings of the Twenty First National Conference on Artificial Intelligence, 1:763-768, 2006.

[12] M. Lan, C. Tan, and J. Su. A term investgation and majority voting for protein interaction article sub-task 1 (ias). pp. 183-185, 2007.

[13] S. E. Robertson and K. S. Jones. Simple, proven approaches to text retrieval. Technical Report 356, Computer Laboratory, University of Cambridge, 1997.

[14] S. E. Robertson and S. K. Jones. Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3):129-146, 1976.

[15] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. Proceedings of the Third Text REtrieval Conference (TREC-3), pp. 109-126, 1996.

[16] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513-523, 1988.

[17] M. Zhanguo, F. Jing, C. Liang, H. Xiangyi, and S. Yanqin. An improved approach to terms weighting in text classification 2011 International Conference on Computer and Management (CAMAN), pp. 1-4, 2011.