TEXT CATEGORIZATION

V.Shaik Parveen
Lecturer
Sethu Institute of Technology, Madurai,
Email: parveensh.81@gmail.com

Dr. C. Kavitha
Professor, Department of M.C.A,
K.S.R College of Engineering,
Triuchengode - 637215

ABSTRACT

Text categorization is the task of assigning predefined categories to natural language text. With the widely used "bag-of-word" representation, previous researches usually assign a word with values that express whether this word appears in the document concerned or how frequently this word appears. Although these values are useful for text categorization, they have not fully expressed the abundant information contained in the document. This paper explores the effect of other types of values, which express the distribution of a word in the document. These novel values assigned to a word are called distributional features, which include the compactness of the appearances of the word and the position of the first appearance of the word. The proposed distributional features are exploited by a tfidf style equation, and different features are combined using ensemble learning techniques. Experiments show that the distributional features are useful for text categorization. In contrast to using the traditional term frequency values solely, including the distributional features requires only a little additional cost, while the categorization performance can be significantly improved. Further analysis shows that the distributional features are especially useful when documents are long and the writing style is casual.

Keywords- text categorization, bag-of-word.

1. INTRODUCTION

In the last 10 years, content-based document management tasks have gained a prominent status in the information system field, due to the increased availability of documents in digital form and the ensuring need to access them in flexible ways [1] . Among such tasks, Text Categorization assigns predefined categories to natural language text according to its content. Text categorization has attracted more and more attention from researchers due to its wide applicability. Since this task can be naturally modeled as a supervised learning problem, many classifiers widely used in the Machine Learning (ML) community have been applied, such as Naı¨ve Bayes, Decision Tree, Neural Network, k Nearest Neighbor (kNN), Support Vector Machine (SVM), and AdaBoost.. Recently, some excellent results have been obtained by SVM [2] and AdaBoost [3] . While a wide range of classifiers have been used, virtually all of them were based on the same text representation, "bag of words," where a document is represented as a set of words appearing in this document. Values assigned to each word usually express whether the word appears in a document or how frequently this word appears. These values are indeed useful for text categorization. However, are these values enough? Considering the following example, "Here you are" and "You are here" are two sentences corresponding to the same vector using the frequency-related values, but their meanings are totally different. Although this is a somewhat extreme example, it clearly illustrates that besides the appearance and the frequency of appearances of a word, the distribution of a word is also important. Therefore, this paper attempts to design some distributional features to measure the characteristics of a word's distribution in a document. Note that the word "feature" in "distributional features" indicates the value assigned to a word, which is somewhat different from its usual meaning, i.e., the element used to characterize a document. The first consideration is the compactness of the appearances of a word. Here, the compactness measures whether the appearances of a word concentrate in a specific part of a document or spread over the whole document. In the former situation, the word is considered as compact, while in the latter situation, the word is considered as less compact. This consideration is motivated by the following facts. A document usually contains several parts. If the appearances of a word are less compact, the word is more likely to appear in different parts and more likely to be related to the theme of the document. For example, consider Document A and Document B in Reuters-21578. Document A talks about the debate on whether to expand the 0/92 program or to just limit this program on wheat. Obviously, this document belongs to the category "wheat." Document B talks about the US Agriculture Department's proposal on tighter federal standards about insect infections in grain shipments, and this document belongs to the

category “grain” but not to the category “wheat.” Let us consider the importance of the word “wheat” in both documents. Since the content of Document A is more closely related to wheat than Document B, the importance of the word “wheat” should be higher in Document A than in Document B. However, the frequency of this word is almost the same in both documents.1 Therefore; the frequency is not enough to distinguish this difference of importance. Here, the compactness of the appearances of a word could provide a different view. In Document A, since the document mostly discusses the 0/92 program on wheat, the word “wheat” appears in different parts of this document. In Document B, since the document mainly discusses the contents of the new standard on grain shipment and just one part of the new standard refers to wheat, the word “wheat” only appears in one paragraph of this document.

Thus, the compactness of the appearances of the word “wheat” is lower in Document A than in Document B, which well expresses the importance of this word. The second consideration is the position of the first appearance of a word. This consideration is based on an intuition that the author naturally mentions the important contents in the earlier parts of a document. Therefore, if a word first appears in the earlier parts of a document, this word is more likely to be important. Let us consider Document A and Document B in Reuters-21578. Document A belongs to the category “grain” and talks about the heavy rain in Argentine grain area. Document B belongs to the category “cotton” and discusses that China is trying to increase cotton output. Obviously, the word “grain” should be more important in Document A than in Document B. Unfortunately, the frequency of the word “grain” is even lower in Document A than in Document B. Now, let us consider the position of the first appearance of the word “grain.” In Document A, it first appears in the title. It is not strange, since this document mainly talks about Argentine grain area.

In Document B, the word “grain” first appears at the end of the document. It is not strange either. Since the theme of this document is about increasing cotton output, the suggestion that the production of cotton be coordinated with other crops such as grain is indirectly related to this theme, so the author naturally mentioned this suggestion at the end of the document. Obviously, the position of the first appearance of a word could express the importance of this word to some extent. Above all, when the frequency of a word expresses the intuition that the more frequently a word appears, the more important this word is, the compactness of the appearances of a word shows that the less compactly a word appears, the more important this word is and the position of the first appearance of a word shows that the earlier a word is mentioned, the more important this word is.

The contribution of this paper is the following:

- Distributional features for text categorization are designed. Using these features can help improve the performance, while requiring only a little additional cost.
- How to use the distributional features is answered. Combining traditional term frequency with the distributional features results in improved performance.
- The factors affecting the performance of the distributional features are discussed. The benefit of the distributional features is closely related to the length of documents in a corpus and the writing style of documents.

2. RELATED WORKS

When the features for text categorization are mentioned, the word “feature” usually have two different but closely related meanings. One refers to which unit is used to represent a document or to index a document, while the other focuses on how to assign an appropriate weight to a given feature. Consider “bag of words” as an example. Using the former meaning, the feature is a single word, while tfidf weighting is the feature given the latter meaning. This section will focus on previous researches about the features used for text categorization based on these two meanings. Other topics about text categorization can be found in a review paper [1] . For the first meaning, besides the single word, syntactic phrases have been explored by many researchers[4] . A syntactic phrase is extracted according to language grammars. In general, experiments showed that syntactic phrases were not able to improve the performance of standard “bag-of-word” indexing. Statistical phrases have also attracted much attention from researchers. A statistical phrase is composed of a sequence of words that occur contiguously in text in a statistically interesting way, which is usually called n-gram. Here, n is the number of words in the sequence. When statistical phrases were used to enrich the text representation of the single word, better performance has been reported with the help of a feature selection mechanism. Researchers also indicated that the short statistical phrase was more helpful than the long one. In addition to phrases, other linguistic features such as POS-tag, word-senses, and the synonym and

hypernym relations in WordNet [2] were tried by researchers. Unfortunately, the improvement of performance brought by these linguistic features was somewhat disappointing. Word cluster was another promising feature for the first meaning [1], [2] . A word's distribution on different categories was used to characterize a word [1], [2] . The clustering methods used by researchers included the agglomerative approach [1] and the recently proposed Information Bottleneck [2] . Experiments showed that the word-cluster-based representation outperformed the singleword-based representation sometimes.

Recently, Sauban and Phahringer [7] have proposed a new text representation method, which explicitly exploited the information of word sequence. In their work, a discriminative score for every word was first calculated. Then, with every word input in sequence, a document was shown as a curve depicting the change of the accumulated scores. This curve was called "Document Profiling." Two different methods were used to turn a profile into a constant number of features. One was to sample from the profile with a fixed gap, while the other was to get some high-level summary information from the profile. Comparable results with the "bag-of-word" representation were achieved with a lower computational cost. For the second meaning, the weight assigned to a given feature comes from two sources: intradocument and interdocument. The intradocument-based weight uses information within a document, while the interdocument-based weight uses information in the corpus. For tfidf, the tf part can be regarded as a weight from an intradocument source, while the idf part is a weight from an interdocument source. There were relatively few researches on the intradocument- based weight. Several variants of tf, such as the logarithmic frequency and the inverse frequency, were used by researchers [8], [9] . The logarithmic frequency reflected that the intuition that the importance of a word should increase logarithmically instead of linearly with the increase of its frequency. The inverse frequency was derived in order to distribute term frequencies evenly on the interval from 0 to 1 [9] . Ko et al. [10] used the importance of each sentence to weight the term frequency. Specifically, the importance of a sentence was measured by two methods. One was to calculate the similarity between the title and a given sentence, while the other one summed the importance of all words appearing in this sentence as the final importance.

Given the importance of a sentence, for a word, a weighted term frequency was used to replace the original tf, where each appearance was weighted by the importance of the sentence where this appearance occurred. For the interdocument-based weight, researchers tried to improve the idf from both the unsupervised view and the supervised view. Researches from the unsupervised view did not use the category information in the training set. Leopold and Kindermann [9] proposed the Redundancy to measure the importance of a word, which quantifies the skewness of the distribution of this word's frequency in different documents. Lan et al. [8] used the term relevance weight in their comparative study. The term relevance weight used the number of documents containing a word to divide the number of documents without this word, instead of the total number of documents in idf. In contrast, many researchers believed that the idf derived directly from text retrieval was not well suited for text categorization where the categories of training documents were available. In order to focus on the categorization task at hand, a lot of supervised weights were proposed. Shankar and Karypis [11] used a measure similar to the Gini Index to calculate the discriminating power of each word. Debole and Sebastiani [12] modified the idf using some feature scoring functions widely used for feature selection, such as Chi-square, Information Gain, and Gain Ratio. The best finding was for Gain Ratio, a variant of Information Gain. Soucy and Mineau used a weighting method based on statistical confidence intervals. This method had an advantage of performing feature selection implicitly. In their work, a significant improvement over the standard tfidf method was reported on benchmarks.

After talking about the related work in this area, a relatively accurate position can be found for our proposed distributional features. These features, which are the compactness of the appearances of a word and the position of the first appearance of a word, could be considered as a new weighting method using the information within a document.

3 HOW TO EXTRACT DISTRIBUTIONAL FEATURES

Recall that the definitions of the two proposed distributional features are both based on the analysis of a word's distribution; thus, modeling a word's distribution becomes the prerequisite for extracting the required features.

3.1 Modeling a Word's Distribution

In this paper, a word's distribution is modeled by two steps: first, a document is divided into several parts; then, distribution of a word is modeled as an array where each element records the number of appearances of this word in the corresponding part. The length of this

array is the total number of the parts. For the above model, how to define a part becomes a basic problem. According to Callan [13], there are three types of passages used in information retrieval.Kim and Kim [14] discussed the advantages and disadvantages of these three types of passages. The discourse passage is based on logic components of documents such as sentences and paragraphs. The discourse passage is intuitive, but it has two problems: the length of passages is inconsistent, and sometimes, no passage decoration is provided for documents. The semantic passage is partitioned according to contents. This type of passage is more accurate, since each passage corresponds to a topic or subtopic, but its performance is heavily influenced by the effect of the partition algorithm. The window passage is simply a sequence of words. The window passage is simple to implement, but it may break a sentence, and the length of window is hard to choose.

Considering efficiency, the semantic passage is not used in . ComPactFLDist. The distance between a word's first and last appearance is used to measure the compactness. It is motivated by the fact that, for a less compact word, the distance between the first mention and the last mention should be long. A slightly extreme example is the word that the author first mentions at the beginning of the document and then mentions again at the end of the document. . ComPactPosV ar. The variance of the positions of all appearances is used to measure the compactness. It is a natural implementation of the idea of compactness using the language of statistics. The mean position of all appearances is first calculated, and then, the mean distance between the position of each appearance and the mean position is calculated as the position variance. For the position of the first appearance, this feature can be extracted directly from the proposed word distribution model.

Now, an example is given. For a document d with 10 sentences, the distribution of the word "corn" is depicted in Fig. 1; then, the distributional array for "corn" is [2, 1, 0, 0, 1, 0, 0, 3, 0, 1].

3.2 Extracting Distributional Features

Given a word's distribution, this section concentrated on implementing the two intuitively proposed distributional features. For the compactness of the appearances of a word, three implementations are shown as follows (note that under the word distribution model mentioned above, the position of a word's appearance is just the index of the corresponding part):

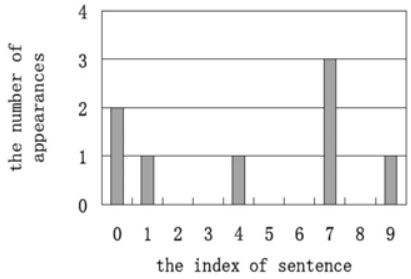

Figure 1. The Distribution of "Corn."

- ComPactPartNum. The number of parts where a word appears can be used to measure the concept of compactness. This is a natural implementation of the idea proposed in the introduction part. As what is mentioned, a word is less compact if it appears in different parts of a document.

- ComPactFLDist. The distance between a word's first and last appearance is used to measure the compactness. It is motivated by the fact that, for a less compact word, the distance between the first mention and the last mention should be long. A slightly extreme example is the word that the author first mentions at the beginning of the document and then mentions again at the end of the document.

- ComPactPosV ar. The variance of the positions of all appearances is used to measure the compactness. It is a natural implementation of the idea of compactness using the language of statistics. The mean position of all appearances is first calculated, and then, the mean distance between the position of each appearance and the mean position is calculated as the position variance.

4. FACTORS INFLUENCING THE PERFORMANCE OF DISTRIBUTIONAL FEATURES

As observed, when the distributional features are introduced, there is no obvious improvement on Reuters but a significant improvement on 20 Newsgroup and WebKB. Thus, the second question arises: what factors will influence the performance of distributional features? Recall that when the compactness of the appearances of a word is introduced, it is assumed that a document contains several parts and the word that only appears in one part is not closely related to the theme of the document. Also, when the position of the first appearance of a word is introduced, it is assumed that the word mentioned late by the author is not closely related to the theme of the document. Intuitively,

these two assumptions are more likely to be satisfied when a document contains some loosely related content. Then, the following question is: in what situation may a document contain the loosely related content.

The first exploration is about the length of a document. This exploration is based on human's habit of writing. When the length of a document is limited, the author will concentrate on the most related content, such as when writing the abstract of a paper. When there is no limit for the length, the author may write some indirectly related content, such as when writing the body of a paper. The mean length of documents of the three data sets used is reported. Here, the length of a document is measured by its number of words. It seems that the improvement brought by the distributional features is closely related to the mean length of documents. In order to further verify this idea, each of these three data sets is split into two new data sets, i.e., the Short data set and the Long data set, according to the length of documents. For each data set, the Short data set contains documents with length no more than 100, and the Long data set contains documents with length more than 100.

5. CONCLUSION

Previous researches on text categorization usually use the appearance or the frequency of appearance to characterize a word. These features are not enough for fully capturing the information contained in a document. The research reported here extends a preliminary research [33] that advocates using distributional features of a word in text categorization. The distributional features encode a word's distribution from some aspects. In detail, the compactness of the appearances of a word and the position of the first appearance of a word are used. Three types of compactness-based features and four position-of-the-first-appearance-based features are implemented to reflect different considerations. A tfidf-style equation is constructed, and the ensemble learning technique is used to utilize these distributional features. Experiments show that the distributional features are useful for text categorization, especially when they are combined with term frequency or combined together. Further analysis reveals that the effect of the distributional features is obvious when the documents are long and when the writing style is informal.

Since no specific combination of TF, CP, and FA consistently shows the best performance on different data sets from current experiments, how to find the optimal combination for different tasks is an important practical issue. In addition, designing the specific idf term for the distributional features is a promising direction. It is also interesting to test the effect of the distributional features on the blog data set in the future.

REFERENCES

[1] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[2] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. 10th European Conf. Machine Learning (ECML '98), pp. 137-142, 1998.

[3] R.E. Schapire and Y. Singer, "Boostexter: A Boosting-Based System for Text Categorization," Machine Learning, vol. 39, nos. 2/3, pp. 135-168, 2000.

[4] D. Mladenic and M. Globelnik, "Word sequences as Features in Text Learning," Proc. 17th Electrotechnical and Computer Science Conf. (ERK '98), pp. 145-148, 1998.

[5]L.D. BakerandA.K.McCallum,"Distributional Clustering ofWords for Text Classification," Proc. ACM SIGIR '98, pp. 96-103, 1998.

[6] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional Word Clusters versus Words for Text Categorization," J. Machine Learning Research, vol. 3, pp. 1182-1208, 2003.

[7] M. Sauban and B. Pfahringer, "Text categorization Using Document Profiling," Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD '03), pp. 411-422, 2003.

[8] M. Lan, S.Y. Sung, H.B. Low, and C.L. Tan, "A Comparative Study on Term Weighting Schemes for Text Categorization," Proc. Int'l Joint Conf. Neural Networks (IJCNN '05), pp. 546-551, 2005.

[9] E. Leopold and J. Kingermann, "Text ategorization with Support Vector Machines: How to Represent Text in Input Space?" Machine Learning, vol. 46, nos. 1-3, pp. 423-444, 2002.

[10] Y. Ko, J. Park, and J. Seo, "Improving Text Categorization Using the Importance of Sentences," Information Processing and Management, vol. 40, no. 1, pp. 65-79, 2004.

[11] S. Shankar and G. Karypis, "A Feature Weight Adjustment Algorithm for Document Classification," Proc. SIGKDD '00 Workshop Text Mining, 2000.

[12] F. Debole and F. Sebastiani, "Supervised Term Weighting for Automated Text Categorization," Proc. 18th ACM Symp.

Applied Computing (SAC '03), pp. 784-788, 2003.
[13] J.P. Callan, "Passage Retrieval Evidence in Document Retrieval," Proc. ACM SIGIR '94, pp. 302-310, 1994.
[14] J. Kim and M.H. Kim, "An Evaluation of Passage-Based Text Categorization," J. Intelligent Information Systems, vol. 23, no. 1, pp. 47-65, 2004.

BIOGRAPHY

Mrs.V. Shaik Parveen received the B.Sc in computer Science from Fatima College, Madurai in 2002, M.Sc in Information Technology from Gandhigram Deemed University in 2004 and M.Phil in computer Science from Madurai Kamaraj University in 2006 respectively. Currently, she is a PhD student in the Research and Development, Bharathiar University, Coimbatore. Her research interests include information retrieval, machine learning and text mining. Currently, she works as lecturer in Sethu Institute of Technology, Madurai.

Dr.C.Kavitha received the B.Sc and M.Sc., Degree from Bharathidasan University and M.C.A., Ph.D., from Periyar University. She has got more than 15 years of teaching and 10 years of research experience. She has published 3 international journals, 2 national journals in her research area and also presented paper at 4 international conferences and 16 national conferences. She is working as the Head of the Department of M.C.A at K.S.R College of Engineering, Tiruchengode. She is guiding more than 11 research candidates in various areas. Her areas of interest are Computer Networks, Network Protocols, Digital Image Processing, Data Mining. She is a life member of ISTE, CSI. She is also a review member for IANENG International journal of Computer Science.