RESEARCH ARTICLE OPEN ACCESS

Text Categorization Using First Appearance And Distribution Of Words

M. Maharasi[1], P. Jeyabharathi[2], A. Sivasankari[3]
[1]Mca,M.Phil, M.E. Assistant Professor, Department of MCA, Dr. Sivanthi Aditanar College of Engineering Tiruchendur, Tuticorin Dist, India
[2]M.E. Assistant Professor, Department of MCA, Dr. Sivanthi Aditanar College of Engineering Tiruchendur, Tuticorin Dist, India
[3]M.E. Assistant Professor, Department of MCA, Dr. Sivanthi Aditanar College of Engineering Tiruchendur, Tuticorin Dist, India

ABSTRACT
Text categorization is the task of assigning predefined categories to natural language text. Previous researches usually assign a word with values that express whether this word appears in the document concerned or how frequently this word appears. These features are not enough for fully capturing the information contained in a document. This project extends a preliminary research 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. The proposed distributional features are exploited by a tfidf style equation, and different features are combined using ensemble learning techniques. The distributional features are especially useful when documents are long and the writing style is casual.

I. 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. 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, 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 and AdaBoost. 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. These values are not enough. 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.

II. EXISTING SYSTEM

A wide range of classifiers have been used ,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 useful for text categorization but these values are not enough for complete categorization so the distribution of a word is also important value. 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. 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. 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 were used unfortunately, the improvement of performance brought by these linguistic features was somewhat disappointing.

III. PROPOSED SYSTEM

In addition to frequency of appearance of word the proposed system has some distributional features they are:

1. **The compactness of the appearances of a word**
2. **The position of the first appearance of a word**

The compactness of the appearances of a word has 3 implementations

ComPactPartNum. The number of parts(index) a word appears can be used to measure the concept of compactness, 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.

ComPactPosVar. The variance of the positions of all appearances is used to measure the compactness.

Advantages of proposed system are:

1) Distributional features can help improve the performance, while requiring only a little additional cost.
2) Combining traditional term frequency with the distributional features results in improved performance.
3) The benefit of the distributional features is closely related to the length of documents and the writing style of documents.

IV. HOW TO EXTRACT DISTRIBUTIONAL FEATURES

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.

4.1 Modeling a Word's Distribution

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. 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.There are three types of passages used in information retrieval., 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 the following experiments. The discourse passage and window passages with different sizes are explored, respectively. 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].

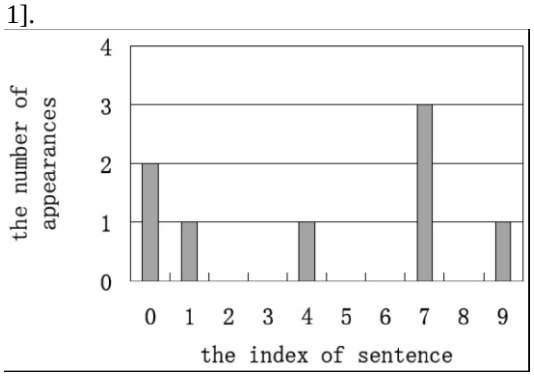

Fig.1.The distribution of "corn"

4.2 Extracting Distributional Features

Given a word's distribution, this section concentrated on implementing the two intuitively proposed distributional features. For the position of the first appearance, this feature can be extracted directly from the proposed word distribution model. For the compactness of the appearances of a word, three implementations are shown.

Suppose in a document d containing n sentences, the distributional array of the word t is array(t,d)=[c0,c1,…..cn-1]. Then the compactness of the appearances of the word t and the position of first appearance (FirstApp) of the word t are defined, respectively, as follows:

$$\text{FirstApp}(t,d)=\min_{i\in\{0....n-1\}} ci > 0?i : n, \quad (1)$$

$$\text{ComPactPartNum}(t,d)= \sum_{i=0}^{n-1} ci> 0 \ ?1:0, \quad (2)$$

$$\text{LastApp}(t,d)=\max_{i\in\{0....n-1\}} ci>0?i : -1 ; \quad (3)$$

$$\text{ComPactFLDist}(t,d)=\text{LastApp}(t,d)-\text{FirstApp}(t,d)$$

$$\text{count}(t,d)=\sum_{i=0}^{n-1} Ci,$$

$$centroid(t,d)=\frac{\sum_{i=0}^{n-1} Ci * i}{count(t,d)} \quad (4)$$

$$ComPactPosVar(t,d)=\frac{\sum_{i=0}^{n-1} Ci*i|i-centroid(t,d)|}{count(t,d)}$$

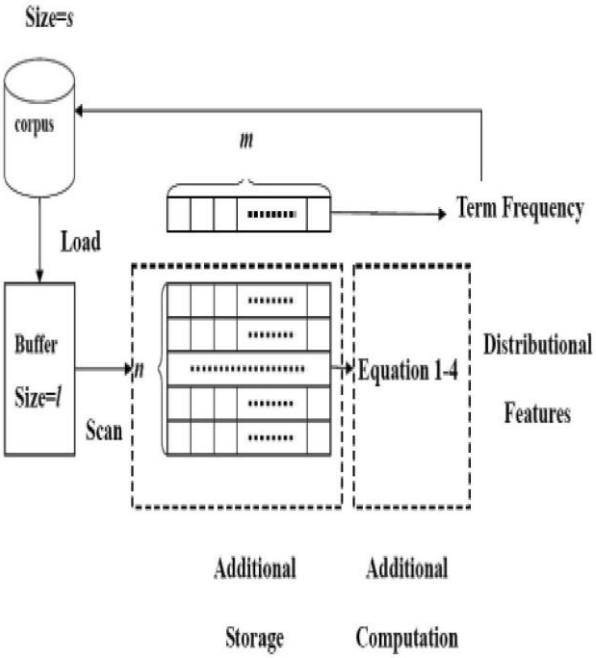

Fig.2. The process of extracting the term frequency and distributional features.

V. 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 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 the position-of-the-first-appearance-based features are implemented to reflect different considerations. The distributional features are useful for text categorization, especially when they are combined with term frequency or combined together. The effect of the distributional features is obvious when the documents are long and when the writing style is informal.

REFERENCES

[1] L. D. Baker and A.K. Mc Callum," Distributional Clustering ofWords for Text Classification," Proc. ACM SIGIR '98, pp. 96-103, 1998.

[2] 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.

[3] D. Cai, S.-P. Yu, J.-R. Wen, and W.-Y. Ma, "VIPS: A Vision-Based Page Segmentation Algorithm," Technical Report MSR-TR-2003-79, Microsoft, Seattle, Washington, 2003.

[4] J.P. Callan, "Passage Retrieval Evidence in Document Retrieval," Proc. ACM SIGIR '94, pp. 302-310, 1994.

[5] M.F. Caropreso, S. Matwin, and F. Sebastiani, "A Learner- Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization," Text Databases and Document Management: Theory and Practice, A.G. Chin, ed., pp. 78-102, Idea Group Publishing, 2001.

[6] M. Craven, D. DiPasquo, D. Freitag, A.K. McCallum, T.M. Mitchell, K. Nigam, and S. Slattery, "Learning to Extract Symbolic Knowledge from the World Wide Web," Proc. 15th Nat'l Conf. for Artificial Intelligence, pp. 509-516, 1998.

[7] F. Debole and F. Sebastiani, "Supervised Term Weighting for Automated Text Categorization," Proc. 18th ACM Symp. Applied Computing (SAC '03), pp. 784-788, 2003.

[8] T.G. Dietterich, "Machine Learning Research: Four Current Directions," AI Magazine, vol. 18, no. 4, pp. 97-136, 1997.

[9] S.T. Dumais, J.C. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and representations for Text Categorization," Proc. Seventh Int'l Conf. Information and Knowledge Management (CIKM '98), pp. 148-155, 1998.[10] C. Fellbaum, WordNet: An Electronic Lexical Database. MIT Press, 1998.

[11] J. Fu¨ rnkranz, "A Study Using n-Gram Features for Text Categorization," Technical Report OEFAI-TR-98-30, Austrian Inst., for Artificial Intelligence, Vienna, Austria, 1998.

[12] J. Fu¨ rnkranz, T. Mitchell, and E. Riloff, "A Case Study in Using Linguistic Phrases for Text Categorization on the WWW," Proc. First AAAI Workshop Learning for Text Categorization, pp. 5-12, 1998.

[13] 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.

[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.

[15] 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.

[16] 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.

[17] K. Lang, "Newsweeder: Learning to Filter Netnews," Proc. 12th Int'l Conf. Machine Learning (ICML '95), pp. 331-339, 1995.

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

[19] D. Lewis, Reuters-21578 Text Categorization Test Collection, Dist. 1.0, 1997.

[20] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. ACM SIGIR '92, pp. 37-50, 1992.

[21] F. Li and Y. Yang, "A Loss Function Analysis for Classification Methods in Text Categorization," Proc. 20th Int'l Conf. Machine Learning (ICML '03), pp. 472-479, 2003.

[22] 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.

[23] A. Moschitti and R. Basili, "Complex Linguistic Features for Text Classification: A Comprehensive Study," Proc. 26th European Conf. IR Research (ECIR '04), pp. 181-196, 2004.

ABOUT THE AUTHORS

M.Maharasi She is presently working as a Assistant Professor in Dr.Sivanthi Aditanar College of Engineering, Tiruchendur. She has done her M.E (CSE) in Dr.Sivanthi Aditanar College of Engineering, Anna University @ Tiruchendur in 2010. She received her M.Phil degree from Manonmaniam Sundaranar University at Tirunelveli in 2004. She received her MCA degree in Sri Saratha College for women @ Bharathidasan University, Karur in 1996. She received her B.sc (computer Science) degree in Cauvery College for women Bharathidasan University @ Trichy in 1993. She has 12 years of teaching experience in this field. She had presented papers in national and International Conferences.

P.Jeyabharathi She is presently working as a Assistant Professor in Dr.Sivanthi Aditanar College of Engineering, Tiruchendur. She has done her M.E (CSE) in Francis Xavier Engineering College, Anna University @ Tirunelveli in 2011. She received her B.E degree from Dr.Sivanthi Aditanar College of Engineering. Anna University @ Chennai in 2009. She had presented papers in national and International Conferences.

A.Sivasankari She is presently working as a Assistant Professor in Dr.Sivanthi Aditanar College of Engineering, Tiruchendur. She has done her M.E (CSE) in Pavendhar Bharathidasan college of Engg. & Technology, Anna University @ Trichy in 2011. She received her B.E degree from Dr.Sivanthi Aditanar College of Engineering. Anna University @ Chennai in 2009. She had presented papers in national and International Conferences.