A Text Categorization Based on Summarization Technique

Sue J. Ker
Department of Computer Science,
Soochow University
Taipei 100, Taiwan,
ksj@cis.scu.edu.tw

Jen-Nan Chen
Department of Management,
Ming Chuan University
Taipei 111, Taiwan,
jnchen@mcu.edu.tw

Abstract

We propose a new approach to text categorization based upon the ideas of summarization. It combines word-based frequency and position method to get categorization knowledge from the title field only. Experimental results indicate that summarization-based categorization can achieve acceptable performance on Reuters news corpus.

Introduction

With the current explosive growth of Internet usage, the demand for fast and useful access to online data is increasing. An efficient categorization system should provide accurate information quickly. There are many applications for text categorization, including information retrieval, text routing, text filtering and text understanding systems.

The text categorization systems use predefined categories to label new documents. Many different approaches have been applied to this task, including nearest neighbor classifiers (Masand, Linoff and Waltz, 1992; Yang, 1994; Lam and Ho, 1998; Yang, 1999), Bayesian independence classifiers (Lewis and Ringuette, 1994; Baker and McCallum, 1998; McCallum and Nigam, 1998), decision trees (Fuhr et al., 1991; Lewis and Ringuette, 1994; Apte et al., 1998), induction rule learning (Apte et al., 1994; Cohen and Singer, 1996; Mouilinier et al., 1996), neural networks (Wiener, Pedersen and Weigend, 1995; Ng, Goh and Low, 1997), and support vector machines (Joachims, 1998). These categorization algorithms have been applied to many different subject domains, usually news stories (Apte et al., 1994; Lewis and Ringuette, 1994; Wiener, Pedersen and Weigend, 1995; Yang, 1999), but also physics abstracts (Fuhr et al., 1991), and medical texts (Yang and Chute, 1994).

In this research to resolve the task of text categorization we apply a method of text summarization, that is, combining word-based frequency and position method to get categorization knowledge from the title field only. Experimental results indicate that summarization-based categorization can achieve acceptable performance on Reuters news corpus. Additionally, the computation time for the title field is very short. Thus, this system is appropriate for online document classifier.

Following is a description of the organization of this paper. Section 2 describes the previous work of summarization. Summarization-based algorithms for text categorization are outlined in Section 3. The experiments we undertook to assess the performance of these algorithms are the topic of Section 4. Quantitative experimental results are also summarized. Finally, concluding remarks and recommendation for future work is made.

1 Text Summarization

The task of summarization is to identify informative evidence from a given document, which are most relevant to its content and create a shorter version of summary of the document from this information. The informative evidence associated with techniques used in summarization may also provide clues for text categorization to determine the appropriate category of the document.

Several techniques for text summarization have been reported in the literature, including methods based on position (Edmundson, 1969; Hovy and Lin, 1997; Teufel and Moens, 1997), cue phrase (McKeown and Radev, 1995; Mahesh, 1997), word frequency (Teufel and Moens, 1997), and discourse segmentation (Boguraev and Kennedy, 1997).

Of the above approaches, both word frequency and position methods are easy to implement. In this research we combine these two approaches to investigate the efforts for categorization. In regard to the position method, Hovy and Lin (1997) considered the title is the most likely to bear topics. They claim words in titles are positively relevant to summarization. Teufel and Moens (1997) also confirmed this viewpoint; they mentioned that words in the title are good candidates for document specific concepts. They showed 21.7% recall and precision, when the title method is used alone, with an increased performance of 3%, when combined with other methods.

Furthermore, from observation of the TREC evaluation during recent years, it has been shown that there is no significant difference between short and long query. It seems reasonable to acquire informative clues from the title, still not degrading the categorization performance severely.

2 Methods

This section describes a series of algorithms based on the title summarization technique for text categorization.

2.1 Preprocessing and Feature Selection

We divide the corpus texts into words, delineate by white space and punctuation. All characters are lower-case and stop words are removed. After the words are stemmed, we call them terms. These terms are then used as features.

2.2 Term Weighting

Weights are now assigned to the surviving features in each category. We design several different formulas for term weighting. In each formula, we associate a weight, W(f, c), with each surviving feature, f, in category c, in the same way weights can be obtained in information retrieval when assigning them to index terms. In addition, we normalize the value of term frequency, tf, between categories. The probability of category is also taken into account. We define W(f, c) as equations 1 through 3.

$$W(f,c) = \frac{tf_{f,c}}{Max_c} \times idf_f \qquad \text{(Eq. 1-a)}$$

$$W(f,c) = p(c) \times \frac{tf_{f,c}}{Max_c} \times idf_f \qquad \text{(Eq. 1-b)}$$

$$W(f,c) = p(c) \times tf_{f,c} \times idf_f \qquad \text{(Eq. 1-c)}$$

$$W(f,c) = tf_{f,c} \times idf_f \qquad \text{(Eq. 1-d)}$$

$$idf_f = \frac{T}{df_f} \qquad \text{(Eq. 2)}$$

$$p(c) = \frac{N_c}{\sum_c N_c} \qquad \text{(Eq. 3)}$$

where $tf_{f,c}$ = the frequency of the feature f appearing in the category c,
T = the number of categories,
df_f = the number of categories that contain the feature f,
Max_c = the maximum frequency of any feature in category c,
N_c = the document numbers belonging category c in training sets.

2.3 Category Ranking

We now have an index suitable for use in the category ranking process. The index contains features and a weighted value, W(f, c), associated with each feature f in each category c. Given a document, d, a rank can be associated with each category with respect to d. Let F_c is the set of features, f, in category c. The ranking of category c with respect to document d, R(c, d), is defined as equation 4.

$$R(c,d) = \sum_{f \in F_c \cap d} tf_{f,d} \times W(f,c) \qquad \text{(Eq. 4)}$$

where $tf_{f,d}$ = the frequency of the feature f appearing in the document d,
F_c = the set of features f in category c.

3 Experiments

To assess the proposed method's effectiveness, we apply the algorithms described in the previous section and conduct a series of experiments. Tests are performed on the Reuters corpus. A general description of the materials used in these experiments follows. Finally, the success rates are quantitatively evaluated.

3.1 The Reuters Corpus

To make our effectiveness comparable to other researchers' results in text categorization, we chose the commonly used Reuters news story corpus for the data. This corpus has many different versions. Yang (1999) points out

there are at least five versions of the Reuters corpus, depending on how the training/test sets are divided and the scope of categories or documents used for evaluation. In this paper, we select the Reuters version 3 (a formatted version is currently available at Yang's homepage http://moscow.mt.cs.cmu.edu: 8081/reuters_21450/apte), constructed by Apte et al., as our data set.

This version contains 7,789 training and 3,309 test documents within 93 categories. The distribution of category number is tabulated in Table 1. Most of these documents have only a single category, but some documents are multicategory. The average numbers of categories per document are 1.23 and 1.24 on training and test sets, respectively. The number of training documents per category varies widely, from 2 (dfl, fishmeal, …, etc.) to a maximum of 2,877 (earn). Tables 2 and 3 show the top ten most frequent categories and ten least frequent categories on the training sets. The average length of title field and whole document are 7.4 and 126.9 words per document, respectively.

3.2 Experimental Design

In this paper, we only use TITLE field as the scope of texts. In our first experiment, the variable is variant term weighting formulas that are described in Section 3. We want to see the effects on categorization performance, when probability of category and normalized process of term frequency are used. The first experiment is summarized in Table 4.

A second experiment is to locate the most preferred threshold value of minimum term frequency. For the number of features in our experiment, the values 10, 20, 50, 100, 150, 200, 300 and 900 are tested.

3.3 Experimental Results and Discussion

We survey the effectiveness of our algorithms by using the conventional 11-point average precision (Salton and McGill, 1983; Yang 1999).

We first investigate a suitable term weighted formula by doing a set of initial categorization from Method 1 through 4. Threshold of minimum term frequency is fixed at 3. The results are tabulated in Table 5. It can be seen that Method 4 appears to perform well in our measure. The average 11-point evaluation can achieve 82.7% precision for Method 4 (tf×idf).

Table 1 The distribution of category number on corpus.

Category No.	Training sets		Test sets	
	Doc #	Percentage	Doc #	Percentage
1	6586	84.6%	2823	85.3%
2	878	11.3%	347	10.5%
3	188	2.4%	65	2.0%
4	61	0.8%	36	1.1%
5	39	0.5%	21	0.6%
Above 5	37	0.5%	17	0.5%

Table 2 The ten most frequent categories in the training sets.

Topic Name	Document No.	
	Training sets	Test sets
earn	2877	1176
acq	1651	776
money-fx	538	207
grain	433	168
crude	388	197
trade	369	135
interest	347	150
wheat	212	81
ship	198	92
corn	176	64

Table 3 The ten least frequent categories in the training sets.

Topic Name	Document No.	
	Training sets	Test sets
cornglutenfeed	2	0
dfl	2	1
fishmeal	2	0
linseed	2	0
naphtha	2	4
nzdlr	2	1
palladium	2	1
palmkernel	2	1
rand	2	1
wool	2	0

Table 4 The choice of term-weighting formulas in the first experiment.

Method Id.	1	2	3	4
Formula Id.	1-a	1-b	1-c	1-d
Prob. used	∨	∨	×	×
Max_c used	∨	×	∨	×

It seems to point out that small text size (only TITLE field is used) is not bad for text categorization, when compared with kNN's 93% and LLSF's 92% for full texts (Yang, 1999).

The other experimental variable is the number of chosen features. Table 5 shows the large feature sets earn the better result when probability is absent.

In the next experiment, with the term weighting formula fixed at Eq. 1-d (Method 4), we vary the minimum number of term frequency from 1 to 3. Table 6 indicates that there are no significant differences among judgements, but shows a little improvement for those small threshold values. The data also shows that the information contained in the title field is almost come together and very little noise. Thus, it seems to have no effects for the processing of sparse data.

Table 5 The 11-point average precision scores of the first experiment. For the minimum term frequency, value 3 was used.

Feature No.	Method Id.			
	1	2	3	4
10	71.9%	69.2%	70.1%	72.1%
20	76.3%	74.1%	73.8%	77.1%
50	78.8%	77.4%	74.3%	80.2%
100	80.2%	79.2%	74.5%	81.9%
150	80.1%	79.8%	73.8%	82.4%
200	80.2%	80.2%	73.3%	82.6%
300	80.0%	80.5%	73.0%	82.6%
900	79.6%	80.9%	72.2%	82.7%

Table 6 The 11-point average precision scores of Method 4.

Feature #	Tf>=3	Tf>=2	Tf >=1
10	72.1%	72.2%	72.3%
20	77.1%	77.2%	77.3%
50	80.2%	80.3%	80.5%
100	81.9%	82.1%	82.3%
150	82.4%	82.7%	82.9%
200	82.6%	82.9%	83.1%
300	82.6%	82.9%	83.2%
900	82.7%	83.0%	83.2%

Conclusion

In this paper, we apply the most popular methods, in text summarization, position and word frequency, to resolve the task of text categorization. We use a word-based term weighted technique from the title field, which is informative but short in length, to process categorization. The results show short title field will reduce execution time, and provide acceptable performance. Thus, this system would be appropriate for an online document classifier.

Previous work shows the hybrid approach for the text categorization and summarization is more efficient than a single scheme. Thus, we will try to combine several schemes in the future. In addition, in the position method, we could use hybrid structure to consider the title and some specific position in the document, for instance, the first sentence in the first paragraph or the first sentence in the second paragraph. When there is insufficient information in title field, it is helpful to proceed to the next position.

Acknowledgements

This research is partially supported by the ROC NSC grants 88-2213-E-031-003 and 89-2213-E-031-004. We would like to thank the anonymous referees for their valuable comments. Any errors that remain are solely the responsibility of the authors.

References

William J. Hutchins (1986) *Machine Translation : Past, Present, Future.* Ellis Horwood, John Wiley & Sons, Chichester, England, 382 p.

Apte C., Damerau F. and Weiss S. (1994) Towards language independent automated learning of text categorization models. In 17th *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'94*), pp. 23-30.

Apte C., Damerau F. and Weiss S. (1998) Text mining with decision rules and decisions trees. In *Proceedings of the Conference on Automated Learning and discovery, Workshop 6: Learning from text and Web*.

Baker L.D. and McCallum A.K. (1998) Distributional clustering of words for text categorization. In *21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'98*), pp. 96-103.

Boguraev B. and Kennedy C. (1997) Salience-based content characterisation of text documents. In *Proceedings of ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 2-9.

Cohen W.W. and Singer Y. (1996) Context – sensitive learning methods for text categorization. In *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'96*), pp. 307-315.

Edmundson H.P. (1969) New methods in automatic extracting. *Journal of ACM*, 16(2): 264-285.

Fuhr N., Hartmanna S., Lustig G., Schwantner M. and Tzeras K. (1991) Air/X – a rule-based multistage indexing systems for large subject fields. In *Proceedings of RIAO'91*, pp. 606-623.

Hovy E. and Lin C.Y. (1997) Automated text summarization in SUMMARIST, In *Proceedings of ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 18-24.

Joachioms Thorsten (1998) Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning* (*ECML*).

Lam W. and Ho C.Y. (1998) Using a generalized instance set for automatic text categorization. In *21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'98*), pp. 81-89.

Lewis D.D. and Ringuette M. (1994) Comparison of two learning algorithms for text categorization. In *proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval* (*SDAIR'94*), pp. 81-93.

Mahesh K. (1997) Hypertext summary extraction for fast document browsing. In *Proceedings of AAAI Spring Symposium: NLP for WWW*, pp. 95-104.

Masand M., Linoff G. and Waltz D. (1992) Classifying news stories using memory based reasoning. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'92*), pp. 59-64.

McCallum A. and Nigam K. (1998) A comparison of event models for Naive Bayes text categorization. In *AAAI-98 Workshop on Learning for Text Categorization*.

McKeown K. and Radev D. (1995) In *18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'95*), pp. 74-82.

Mouilinier I., Raskinis G. and Ganascia J. (1996) Text categorization: A symbolic approach. In *proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval* (*SDAIR'96*).

Ng, H.T., Goh W.B. and Low K.L. (1997) Feature selection, perceptron learning, and a usability case study for text categorization. In *20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'97*), pp. 67-73.

Salton G. and McGill M.J. (1983) *Introduction to modern information retrieval*. McGraw-Hill Computer Science Series. McGraw-Hill, New York.

Teufel S. and Moens M. (1997) Sentence extraction as a classification task, In *Proceedings of ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 58-65.

Wiener E., Pedersen J.O. and Weigend A.S. (1995) A neural network approach to topic spotting. In *proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval* (*SDAIR'95*).

Yang Y. (1994) Expert network: Effective and efficient learning from human decision in text categorization and retrieval. In *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'94*), pp. 13-22.

Yang Y. (1999) An evaluation of statistical approaches to text categorization, *Information Retrieval*. Vol. 1, pp. 69-90.

Yang Y. and Chute C.G. (1994) An application of expert network to clinical classification and MEDLINE indexing. In *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, pp. 157-161.