CHINESE-MINING PREPROCESSING TECHNOLOGY BASED ON TEXT TRAIT OPTIMIZING

[1] XIAOYONG WANG , [2]SIYOU XIAO， [3]FANG YUEFENG
[1]Intelligent Control Institute, Zhejang Wanli University, Ningbo 315101, Zhejiang, China
[2] Intelligent Control Institute, Zhejang Wanli University, Ningbo 315101, Zhejiang, China
[3] Department of Computer and information, Zhejang Wanli University, Ningbo 315000, Zhejiang, China

ABSTRACT

How to get the target text quickly becomes a technical limitation with the using of massive data. While obtaining the Chinese target information, the segmentation of the sentence is supposed to be the key according to research. To mine the segmentation of English text is relatively simple for the space is used as a interval, meanwhile the Chinese segmentation is much more difficult. So in this paper the reciprocal crossing segmentation algorithm and the trait-optimizing vector model are designed to improve the mining efficiency of Chinese information. Based on dictionary, an improved segmentation algorithm is adopted in text pretreatment processing, which is based on vector space module, to do experiments on the segmentation algorithm and to analyze the segmentation results. And that segmentation algorithm is already proved to be very effective in the text mining of text trait vector module.

Keywords: *Reciprocal Crossing Matching, Mutual Information, Information Gain, Expected Crossing Entropy, Ce2 Statistics*

1. INTRODUCTION

It’s most important for data-procurement to mine certain potential and useful data from the mass information. Eight percent of the information exists as a unstructured text form nowadays. So text data mining (TDM) technology has been a crucial research topic [1].

Since 1950s, researchers represented by H. P. Luhn of IBM have made a study on the classification of English text, and they have put forward a automatic classification ideology based on hash technique. The ideology is used in information searching and file-keeping field successfully. The automatic text classification and hash technique are now widely acknowledged, and are pretreated ad the key to improve the text mining application efficiency. Meanwhile, the great discrepancy between English and Chinese leads that the mining technique of English text which is a rather popular technology overseas can not be used in Chinese text field directly. With the wide use of Chinese and the rapid exploding of Chinese information, the research of mining technique becomes more and more important.

In this paper, a new matching algorithm is designed to optimize the Chinese words, then to vectorize the text based on the trait, and to pretreat the unstructured Chinese text. By this way, the efficiency of Chinese mining is improved considerably.

2. DESIGN OF RECIPROCAL CROSSING SEGMENTATION MATCHING

The text eigenvector must be chosen based on the effective segmentations of the text source library with the key of research of segmentation algorithm, among which the Maximum Matching(MM) method is most widely used. The MM method which is a mechanical segmentation method based on the string matching segments the text into words based on accepted and customized rules [2,3]. The MM method obeys the rule of longest segmentation matching. A character string contains 6~8 characters is chosen as the maximum character string and then matched with the word entry in the dictionary. If is can not match the word entry, a character will be delated to continue the match until the equivalent is found. The matching is done from right to left. Reverse maximum match method is opposite to MM method, that is from left to right. A series of experiments have that the reciprocal crossing matching (RCM) method is more effective than the MM method.

By experimenting, we find that the MM method has the following faults. On hand， several characters are matched with the entries in the dictionary from left to right. If the text string exists, it can be syncopated directly. Each step of syncopating depends on the dictionary a lot, which leads that the unlisted words can not be recognized well. On the other hand, According to lots of wrong syncopating cases, we find that the last character of the word entry to be matched and the n+1 character can be usually put together into word, that is to say they have crossing ambiguity. And it cannot be solved because of the unidirectional processing of the algorithm. The segmentation algorithm of this paper chooses the design proposal which considers both the forward and reverse direction. It solves the problem of un-identification of the new words and the fault of crossing ambiguity.

2.1 The Construction Of The Hash Segmentation Dictionary

Segmentation dictionary is an essential part of Chinese segmentation system, from which all information needed in automatic segmentation system can be found. According to the analysis of GB2312 Chinese Encoding and Character-making habits, the Hash segmentation dictionary is designed in this paper. The dictionary contains the following three parts: the lead-in hash list, the word index and the main body of the dictionary. The lead-in hash list is generated by the HASH function evaluation, that is to input an unfixed length character string to output a fixed one. The lead-in list generated by this way can realize the fast access of data. Each word in the dictionary has a unique correspondence in the word index. The main body of the dictionary is the words warehouse. During the matching process of MM method, much time is spent on matching the entries of the words warehouse. So a good words warehouse is a key technique to the improved matching method.

2.2 Process of Reciprocal Crossing Segmentation Matching

Reciprocal crossing segmentation, which is an improved MM method put forward in this paper, matches the meta segmentation sentence by reciprocal scanning method. By comparing the mutual information (IM), it enhances the new word identification and solves the crossing ambiguity. Thus the Chinese mining effect is much improved. The main process is shown as the following Figure 1.

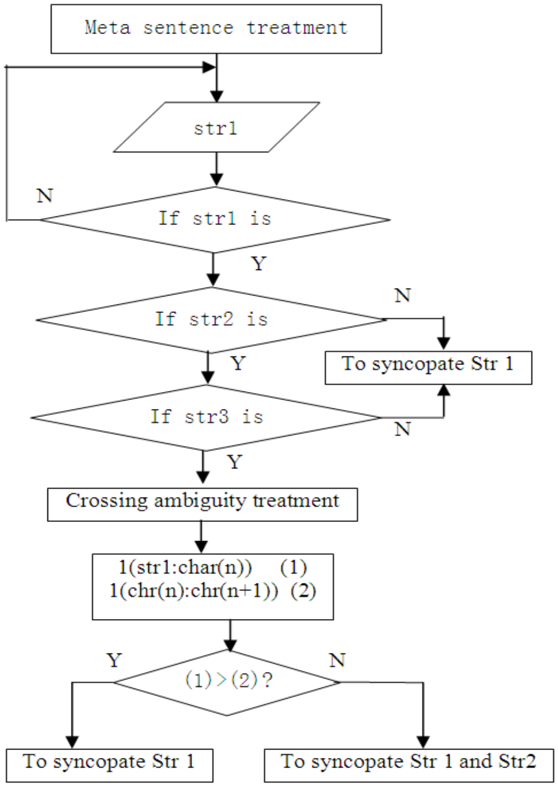

Figure 1: The process of improved segmentation matching algorithm

2.1.1 Primary segmentation treatment and construction of segmentation meta sentence

Do the first scanning to separate ，，、，；，！ and other mark from each other with a /. Then do the second scanning to add a / at each side of the keywords like punctuation, numbers, and other non-kanji symbols, and so on. The smallest unit between two /s is regarded as a meta sentence.

2.1.2 Forward scanning and segmentation-matching with the dictionary

Let the number of the characters of the longest phrase scanned in the dictionary be n. To the being-treated meta sentence X-text, n (6~8 usually) characters are read orderly from the first character. If X-text.length<n, the whole meta sentence will be chosen as str1, and str1 = X-text. sub-string (0, min (n, X-text. length)). If str1 is not contained in the dictionary, the n-1 string in the front of str1 will be chosen as str2. The str2 will be matched with the dictionary. If the matching is successful, it will be transferred into scrap processing. If the matching is unsuccessful, one character will be omitted to

continue the matching until the length of str2 is 1. If str1 is contained in the dictionary, the n-1 string in the front of str1 will be chosen as str2, and the length of str2 can be 1. Let str2=str1.Sub-string (0, n-1), then match str2 with the dictionary. If str2 is not contained in the segmentation dictionary, str1 will be syncopated out of the pending text. The redo the process. If str2 is contained in the dictionary, let str3= X-text.Sub-string (n-1, 2), and str3 will be matched with the words list. If str3 is not contained in the segmentation dictionary, str1 will be syncopated out of the pending meta sentence, and redo step 1. If str3 is contained, it will be transferred into the module of crossing ambiguity treatment.

2.1.3 Reciprocal matching and analyzing of the crossing ambiguity

The statistical analysis of mutual information str1and str2, if the former is larger than the later, str1 will be syncopated out, and step 2 will be redone. If the former is smaller than the later, str1 and str2 will be syncopated out, and step 2 will be redone. When crossing ambiguity appears, the module will choose the syncopating position according to the frequency of the two words used in current context. It is also good for str3 to identify the new words.

2.1.4 Defragment

Several meta sentences will be generated after the above syncopating, and the number of the characters between the two /s will be calculated. If the meta sentence X-text contains n words after syncopating, the character array will be [c_1, c_2...c_n]. The vowel whose length is 1 will be chosen from right to left from the first word, then continue reading starting from left. If two numbers which are both 2 exist continuously, that is the string satisfies the /vowel 1/ vowel 2 / structure, make sure if the two vowels exist in the dictionary. If they exist in the dictionary, they will be merged as /vowel 1 vowel 2 /. If there exist m (2< m< maximal word length in the dictionary) numbers which is 1, that is the string satisfies the /vowel 1/ vowel 2 / vowel 3 / vowel 4 / structure, make sure if the vowels exist in the dictionary. If they exist in the dictionary, they will be merged as /vowel 1 vowel 2 vowel 3 vowel 4/. If the don’t exist, a vowel will be omitted to make sure if the m-1 vowels are regarded as words. If not, redo the process.

3. MINING PREPROCESSING BASED ON THE TEXT EIGENVECTOR

The research of Helsinki University break the bottleneck of the effective English text mining technique. By pre-doing the format conversion and uniform treatment to all text, it adopts corresponding module to mine them uniformly, which improves the text mining efficiency a lot [4]. Recently, scholars at home and abroad have put forward some representative text retrieval models, such as Boolean Model, Vector Space Model, Probability Model, and so on. Those models have done researches on the questions such as processing trait weighting, category learning, similarity algorithm, and so on. They have quite good effects on English text treatment, but little on Chinese text. Based on theory aforementioned, the paper projects the vector model into Chinese Text pretreatment.

While transferring the traditional vector model into Chinese mining research, dimension of the corresponding vector space of the text is very larger, even can be hundreds thousands [5]. If we search and query with the original dimension of the text, we will pay much time, which makes the algorithm meaningless. The vector model based on text trait designed in this paper reduces the vector space dimension of the text greatly. Vector space model of the text trait consists of a set of basic linearly independent vectors. The dimensions of the vector agrees with that of the vector space and can describe the vector space with the trait. It contains the following details: Documentation D, which is a documentation or a passage in it. trait item t, which is a basic linguistic unit appears in the text and can represent the file properties. trait weight wt, which shows the importance of the trait in the file. Similarity s, which is the correlation coefficients between the contents of two files. If the text is described by the vector, it can be measured by the length between the vectors of the two files, and is usually calculated by the inner product or the cosine of the angle. The larger the cosine value is, the greater the similarity of the files is. The paper improve the text trait vector technique to do pretreatment to the text classification. And the pretreatment can be realized by constructing the following trait selecting methods.

First, Information Gain, which is usually used in machine learning field, calculates the trait information content according to whether the trait appears in the text.

$$InfoGain(t) = P(t)\sum_i P(c_i \mid t)\log_2 \frac{P(c_i \mid t)}{P(c_i)} + P(\bar{t})\sum_i P(c_i \mid \bar{t})\log_2 \frac{P(c_i \mid \bar{t})}{P(c_i)}$$

(1)

Second, Expected Cross Entropy, which has the only difference between which and the Information Grain is that it discounts the disappearance of words.

$$CrossEntry(t) = P(t)\sum_i P(C_i \mid t)\log_2 \frac{P(Ci \mid t)}{P(C_i)} \quad （2）$$

Third, Mutual Information, which discounts the disappearing frequency of the words, comparing to Expected Cross Entropy. It is a disadvantage, but also makes Mutual Information measurement be used to mine the infrequent words.

$$MutualInfo(t) = \sum_i P(C_i \mid t)\log_2 \frac{P(Ci \mid t)}{P(C_i)} \quad (3)$$

Fourth, Ce2 Statistics. On the original test platform, A is used to calculate the simultaneous of trait t and category c. B is used to calculate the times that trait t occurs but category c doesn't occur. C is used to calculate the times that trait t doesn't occur but category c occurs. D is used to calculate the times that both trait t and category c don't occur. N is the text set number. Then the value of Ce2 between trait t and category c can be calculated as the following.

$$Ce^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} (4)$$

3.1 The Trait Item Determined by The Mutual Information Content

The trait vector optimization in this paper obeys the justification based on the fixed trait item of the word and the mutual information content of its category. The algorithm process is as follows.

At the initial situation, all words appear in the category are contained in the trait database.

To each word, it and the mutual information content of its category are calculated as $\log_2(\frac{P(W \mid C_i)}{P(W)})$

Among which, $P(W \mid C_i) = \frac{1+\sum_{i=1}^{|D|} N(W,d_i)}{|V| + \sum_{i=1}^{|V|}\sum_{i=1}^{|D|} N(W_i,d_i)}$, and $P(W \mid C_i)$ is the probability of occurrence of W in c_i . |D| is the number of the training text in the category. $N(W,d_{i)}$ is the weight frequency of W in d_i . |V| is the total weight. $\sum_{i=1}^{|V|}\sum_{i=1}^{|D|} N(W_i,d_i)$ is the weight frequency of all words in the category.

All words in the category are ordered according to the mutual information content calculated above.

A certain number of words are selected as the trait item. The exact dimension number of the selected trait item cannot still be made sure presently. The paper chooses determined initial value firstly, and then makes sure the optimal value by experimentally testing and statistics

The dimension compression is done to vector set operation according to the trait item selected in all the training texts, so that the vector representation is condensed.

This strategy ponders the case that low-frequency words contain information. The mutual information of low-frequency words is more than that of high-frequency words.

3.2 The Weight of Text Trait Vector

Statistical technique is now the most popular weight-determining method. that is to calculate the trait weight according to the text statistics which mainly concerns about the word frequency. The weight calculation formulas TF-IDF which is widely accepted is $W_{ik} = tf_{ik} \times idf_k$. In this formulas, tfik(Str Frequency) represents the in-text frequency of tk in text D_i . Idfk(Inverse Document Frequency) is the controlling-text frequency of tk, They have various computing methods, among which the popular equation is $W_{ik} = if_{ik} \cdot \log_2(\frac{N}{n_k} + 0.01)$.

In this equation if_{ik} represents the times that tk appears in the text D_i . N is the number of the total texts. n_k shows the number that tk-texts appear in the training text. Each component in the vector space can be described as this. Each component determines the ability that item tk distinguishes the sub-text contents attributes. The wider the width of the item appears in the text set is, the lower its ability to distinguishes the text contents attributes is. The high its appearing frequency in a certain text is, the higher its ability to distinguishes the text contents attributes is. Because of influence of the text length on weight, the term weight equation should be normalized. The weights of each item should be limited between [0, 1].

$$W_{ik} = \frac{if_{ik} \cdot \log_2(\frac{N}{n_k} + 0.01)}{\sqrt{\sum_{i=1}^{N} (if_{ik})^2 \cdot if_{ik} \cdot \log_2(\frac{N}{n_k} + 0.01)}} \qquad (5)$$

Additionally, the function of the marks is to provide the information (title, beginning, paragraph, and so on) and the format(bold, italics, and so on) of the text structure. While extracting the webpage feature, the paper calculates the feature lemmas of those marks and their appearing frequency, and makes them have higher weight. Experiment has shown that TF-IDF is an effective text processing tools. It not only is applied in information retrieval, but also has great effects in other fields like information dissemination, information filtration and text classification [6]. Being segmented by the segmenting program, the text deletes the inactive words first and merges the words like numbers and names. Then it calculates the term frequency. And at last, it fixes the trait vector of the text.

3.3 The Realization of Text Ming Pretreatment

The Chinese mining pretreatment in this paper is designed according to the cooperation of frequency that the document appears and the mutual information. The experiment indicates that the co-pretreatment to Chinese mining.

The document frequency (DF) shows the number of the documents with this trait appear in the corpus. The former n biggest trait words are selected according to DF in this design. This strategy has the following advantages. It omits the low-frequency words and reduces the dimension of the trait space. It has simple algorithms and little computations. There are still some disadvantages. The design regards the low-frequency as information-less, meanwhile it may has much information. And that has direct effects on the classification when the low-frequency words are omitted. The concrete steps are as following. DF of the trait dimension is calculated to get the trait dimension T first. Then the former M biggest trait words are selected by ordering DF value to get the new trait dimension T1. And the IM of each word in T1 is calculated. At last, the former N biggest trait words are selected by ordering M value to get the new trait dimension T2. The experiment shows that the text mining system based on B/S model of J2EE is realized in WINXP operating systems, with the database of DB2. All ideals of text pretreatments are applied in the experiment. And the text corpus of electrical category is designed by the microelectronic design, among which seven categories are chosen as the training sets. In the760 texts, the ratio of the training sets and testing sets of each category chosen is 10：3. The results of the experiment are shown as Table 1and Table 2.

Table 1: The Classification Result Of Different Trait Selection Algorithms In Different Dimensions

Trait Dimension (TD)	100	200	500	1000
Information Gain(IG)	0.4326	0.4870	0.5375	0.5732
Ce^2 Statistics	0.4179	0.4607	0.5514	0.5799
Mutual Information (IM)	0.5312	0.6021	0.6301	0.6355
Cooperation of DF and IM	0.6211	0.6301	0.6555	0.6612

Table 2: The Classification Result Of Different Trait Selection Algorithms In Different Dimensions

Trait Dimension (TD)	1500	2000	3000	4000
Information Gain(IG)	0.5731	0.5637	0.5721	0.5644
Ce^2 Statistics	0.5909	0.5902	0.5798	0.5890
Mutual Information (IM)	0.6339	0.6299	0.6408	0.6192
Cooperation of DF and IM	0.7021	0.6913	0.6899	0.6711

By results analyzing, we can get the accuracies of the four Chinese mining pretreatment algorithms . Ce2 is from 0.4179 to 0.5909. IG is from 0.4326 to 0.5731. IM is from 0.5312 to 0.6336. And The Cooperation of DF and IM is from 0.6211 to 0.7021. Form that we can get the following facts.

First, the best accuracies of each algorithms occur under the dimension of 1500. That is to say the accuracy of text mining increases not monotonously with the increase of the trait dimension but with that of trait quantity. The great amount of meaningless traits occur in the trait dimension does not increase the accuracy, but decrease it. When the trait dimension is below 1500, the accuracy increases unidirectional. that is because the identifying ability of system increases with that more and more information input in the trait vector.

Second, the cooperating algorithm accuracy of DF and IM is the highest. The disadvantage of IM is it ends to low-frequency words, which leads ineffective classification. And DF mainly is inclined to the negative effect of low-frequency words to classification. The cooperation of these two algorithms integrates their advantages and makes up their shortages, and gets good results.

4. CONCLUSION

According to the vector space module, the paper designs the text trait vector to do pretreatment to text mining processing. The establishment of text trait vector can be divided into two steps. They are to choose the text trait item and to calculate its weight. An improved segmentation algorithm based on dictionary is introduced to text pretreatment processing based on vector space module to do experiment on the segmentation algorithm and analyze the segmentation results. And the segmentation algorithm is proved to have great effects on the text mining of text trait vector module.

ACKNOWLEDGMENTS

This research was supported by the special significant sub - items of Core Electronic Devices, High-end Universal Chip and Basic Software of the Ministry of Industry and Information Technology. No. 2009ZX01039-002-001-04-06.

REFERENCES:

[1] Broder A. Fontoura M, Gabrilovich E, and etc, “Robust Classification of Rare Queries Using Web”, *Proceedings of the 30th annual international ACM SIGIR*, 2007, pp. 231-238.

[2] Backstrom L, Caruana R. C2FS, “An Algorithm for Feature Selection in Cascade Neural Networks”, *International Joint Conference*, 2006, pp. 4750-4752.

[3] Mobasher B, Cooly R, Srivastara J, “Automation personalization based on web usage mining”, *Communication of the ACM*, Vol. 43, No. 8, 2000, pp. 143-150.

[4] R. Kosla, H. Blockeel, “Web mining research a survey”, *SIG KDD Explorations*, 2000, February, pp. 1-15.

[5] Hanan A, Mohamed K, “Topic discovery of firm text using aggregation of different clustering methods”, *Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*, LNCS 2338, Springer_Verlag, 2002, pp. 159-174.

[6] Sebastian F, “Machine leaning in automated text categorization”, *ACM Computing Suvrvey*, Vol. 34, No. 1, 2002, pp. 1-47.

[7] Schapire Y, “A boosting_ based system for text categorization”, *Machine Learning*, 2000, pp. 39(2/3),136-167.

[8] Chakrabartis, DomBE, KumarSR, et al, “Mining the Web's Link Structure”, *Computer*, Vol. 32, No. 8, 1999, pp. 60-67.

[9]Y. Yang, “An evaluation of statistical approaches to text categorization”, *Journal of Information Retrieval*, 1999, pp. 67-79.

[10] Tseng, Y. H, “Automatic Thesaurus Generation For Chinese Documents”, *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 13, 2002, pp.1132-1136.

[11] Lu H, et al, “Effective data mining using neural networks”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, 1996, pp.958-960.