REFERENCE ALGORITHM OF TEXT CATEGORIZATION BASED ON FUZZY COGNITIVE MAPS

ZHANG Guiyun ,LIU Yang , ZHANG Weijuan,WANG Yuanyuan
Computer and Information Engineering College, Tianjin Normal University, Tianjin 300384, Email:dyxy1999@126.com, phone :013920736656

Abstract: This paper introduces the reference theory and algorithm of text categorization by using fuzzy cognitive map(FCM), which is based on value inference and can be able to infer by combing rule and statistics. This method is flexible and robust, and we do not need train the corpus time after time，it is suitable to the text categorization of insufficiency training, new subject and multi-classification.

Key words: text categorization; fuzzy cognitive map; reference algorithm

1. INTRODUCTION

The technology of text automatic categorization has gone through the rule-based technology, statistics-based technology, and to the combination of rule and statistics. Recently there are Rocchio classical algorithm, Naïve Bayes probability algorithm, decision tree matching algorithm, K-nearest neighbor method based on similarity, Support Vector Machine (SVM)suggested by Vapnik, Linear Least Square Fitting (LLSF), Neural Network，maximum entropy categorization method rough set method[1] [10] [12] [13] [14] [15] and so on. This paper introduces a reference algorithm of text categorization based on fuzzy cognitive map for making the text categorization become the result of FCM reference which is based on weight of text term, term and category, and category and relevancy.

2. CONSTRUCTION OF FCM IN TEXT CATEGORIZATION

The cognitive map (CM) is constituted by relations of concepts which are represented by nodes. The relation between concept is represented by an arc with arrow, its strength is represented by number value, namely, the weight of arc. FCM combines fuzzy logic and neural networks technology, the state space of an FCM is determined initially by an initial condition and then propagated automatically through the node function relative to a threshold until a static pattern is reached. A causal inference is achieved when the FCM reaches a stable limit cycle or fixed point.

The foundation of text term model and selection of categorization method are core problems. Now, although there are various categorization algorithms based on vector space model, most of which need training a large number of corpus. The method in this paper regards text term and classification as nodes of CM, the corresponding state values of node are weight values of terms, relevancy of term t_i and classification C_j and that of classification C_k and classification C_j are the weights of corresponding edges to realize the text categorization reference algorithm based on FCM.

Definition 1 A text categorization FCM is a quadruples ordered set U=(T,C,E,W),where T= $\{t_1,t_2,\ldots,t_n\}$ represents the term set in text, C= $\{C_1,C_2,\ldots,C_n\}$ represents classification set, E=$\{<t_i,C_j>,<C_k,C_j>|t_i \in T,\ C_k,C_j \in C\}$,directed arc $< t_i,C_j>$ represents that term t_i relates to classification C_j, $<C_k,C_j>$ represents that classification C_k relates to classification C_j,W= $\{W_{ij}, P_{kj} \mid W_{ij}$ is weight of directed arc $< t_i,C_j>$, P_{kj} is weight of directed arc $< C_k,C_j>\}$. $V_{ti}(0)$, $V_{Ck}(0)$ represent initial value of term t_i and classification C_k (weight value).The weight of corresponding edge is 0 if there is no any relevancy.

Therefore, the adjacency matrix of text categorization FCM can be simplified as a（n+m）×m matrix:

$$W_U = \begin{pmatrix} \mathrm{L} & \mathrm{L} & \mathrm{L} \\ \mathrm{L} & w_{ij} & \mathrm{L} \\ \mathrm{L} & \mathrm{L} & \mathrm{L} \\ \mathrm{L} & \mathrm{L} & \mathrm{L} \\ \mathrm{L} & P_{kj} & \mathrm{L} \\ \mathrm{L} & \mathrm{L} & \mathrm{L} \end{pmatrix} \qquad (1)$$

Where W_{ij} denotes the relevancy of node v_i and classification $C_{j,}$ P_{kj} denotes the relevancy of classification C_k and classification C_j.

The total input received by text categorization FCM at time t+1is

$$(v'_{c1}(t+1),...,v'_{cm}(t+1))=(v_{t1}(t),...,v_{tn}(t),v_{C1}(t),...,v_{Cm}(t))\times w_U \qquad (2)$$

Therefore, the output received by text categorization FCM at time t+1is

$$(v_{c1}(t+1),...,v_{cm}(t+1))=(f(v'_{c1}(t+1)),...,f(v'cm(t+1))). \qquad (3)$$

The input received by text categorization FCM at time t+1is determined by equ (4)as follows:

$$(v_{t1}(t),...,v_{tn}(t),v_{c1}(t+1),...,v_{cm}(t+1))=(v_{t1}(t),...,v_{tn}(t),f(v'_{c1}(t+1)),...,f(v'cm(t+1))). \qquad (4)$$

Namely, the weight of term is not changed, values of classification nodes are updated.

3. REFERENCE ALGORITHM OF TEXT CATEGORIZATION BASED ON FUZZY COGNITIVE MAP

3.1 Decision of term weight and edge weight in text categorization FCM

Many weight functions about term weight such as Boolean weight function, TF-IDF weight function, ITC weight function, Okapi weight function, the algorithm of TF·IDF·IG (information gain) come out. In addition, the algorithm[10] by assigning weight value for the regions of term words is considered. Term frequency*inverse document frequency(TF·IDF) is a basic one.Assume that the term frequency t_i in document d_j is $tf_{ij}=freq_{ij}$, inverse document frequency $idf_i=log(N/n_i)$, where N is the number of texts in data corpus, n_i is the sum of texts which comprise term t_i, and the base-number of log can be 10,e or 2. Initially, the weight of term t_i in document d_j is:

Vt_i（0）= tf_{ij}• idf_i（5）, then normalize it, the basic way is maximum normalization (others see paper[6]): $tf_{ij} = \frac{freq_{ij}}{\max_k \{freq_{kj}\}}$ (6).The relevancy of term t_i and classification C_j is weight W_{ij} in text categorization cognitive map. The common methods are Mutual Information, IG, and Expected Cross Entropy etc. Many researches show that Mutual Information algorithm is much better than others[12] .The mutual information of term t_i and classification C_j is: $MI(ti,Cj)=\log(\frac{P(ti/Cj)}{P(ti)})$ (7), where $P(ti/Cj)=\frac{1+\sum_{k=1}^{N}tf_{ik}}{|V|+\sum_{l=1}^{|V|}\sum_{k=1}^{N}tf_{lk}}$ and $P(t_i)$ denote the specific weight of term t_i in classification C_j and word frequency in corpus, |V| and N denote sum of all term and the amount of documents, respectively.

3.2 Reference algorithm of text categorization based on fuzzy cognitive map

The reference algorithm of text categorization based on fuzzy cognitive map is as follows:

Input: weight of term, relevancy of term and text classification, relevancy among classifications.

Output: classification of text

Step1 Calculate weight of term t_i through equ(5), and normalize it e by using equ(6);

Step2 Calculate relevancy of term t_i and classification $C_{j,}$ W_{ij}through equ(7), read relevancy of classification C_k and C_j which are specified by experts as weight P_{kj},and then decide the adjacency matrix through equ (1).

Step3 Calculate the output of C_j at time t+1 through equ(2) and equ(3),mostly f is a sigmoid function: $f(x)=1/(1+e^{-cx})$;

Step4 Whether $Vc_j \geq P_T$ (threshold),if yes, output C_j, and if there are many S_j ,then output the maximum; if no, goto step1(or terminate iterated algorithm by limiting its degree) .The output C_j is text classification.

4. EXPERIMENTS AND ANALYSIS

Recall and precision are classical performance evaluate standards of text classification, where the precision reflects the proportion of correct text classification. We randomly choose 30,50,100,150,200,250,300,500 pieces of documents concerning economy, politic, computer, physical, education and law to train and carry out experiments from corpus in Fudan university, disk edition of *People Daily* corpus in 1999 and web. tf_{ij} is calculated by using the simplest word frequency, C is 0.5, C_j is the biggest output weight after 300 iterative. Table 1 indicates relationships weights between classifications.

Table2-1 and table2-2 describe the result of test, and the recall and precision of different pieces of texts，the calculated formulas are seed by reference 1.

Table1 Relationships weights between classifications

	economy	politic	computer	physical	education	law
economy	1	0.7	0.4	0.5	0.5	0.7
politic	0.7	1	0.2	0.3	0.6	0.8
computer	0.4	0.2	1	0.1	0.6	0.2
physical	0.5	0.3	0.1	1	0.3	0.3
education	0.5	0.6	0.6	0.3	1	0.5
law	0.7	0.8	0.2	0.3	0.5	1

Table 2-1 The recall (%)and precision(%) of different pieces of texts

evaluate	50 pieces		100 pieces		150 pieces		200 pieces	
classification	Pi	Ri	Pi	Ri	Pi	Ri	Pi	Ri
economy	60	100	70	100	69	91	72.3	97.1
politic	85.7	85.7	100	80	98.3	84.6	91.7	91.7
computer	100	100	100	93.3	100	88	100	87.9
physical	100	75	100	77.8	88.9	76.2	96	78.1
education	100	71.4	90.5	90.5	93.3	90.3	88.2	85.7
law	100	87.5	81.3	86.7	88	91.7	93.1	90

Table 2-2 The recall (%)and precision(%) of different pieces of texts

evaluate	250 pieces		300 pieces		500 pieces	
classification	Pi	Ri	Pi	Ri	Pi	Ri
economy	74.6	97.6	74.2	98	75.4	98.9
politic	93.5	89.6	95	89.8	92.4	90.1
computer	77.6	86.4	98	87.3	96.6	89.4
physical	96.9	81.6	97.2	83.3	100	81.4
education	88.1	90.2	92	92	86.7	87.8
law	88.0	86.5	91.1	89.1	90.8	83.1

5. CONCLUSION

Text categorization is the basis of passage-chapter level text process, but different information demands will produce different categorization requirements. This paper suggests a reference theory and algorithm of text categorization based on fuzzy cognitive map which is derived from the weight of text term , the relevancy of term and classification and the relevancy of classification and classification. Although it is a new attempt, the results indicate its effect. The merits of using FCM to categorize text are:①It is a number value reference based on iterative calculation.②This method emphasizes feedback so that it is suitable to insufficiently training or new subject classification.③Considering the relevancy between classifications and the relevancy between terms.④Merging statistics and number value reference , so it overcomes the shortcoming of depending on experts' knowledge. ⑤When FCM reaches stable, a unitary classification is received, while when FCM converges a limit cycle, then multi-classification is received, so it is suitable to the classification of cross science and synthetical science.⑥The method is open. It can be added, deleted or combined, and it's suitable for real-time different requirement.

REFERENCE

1. PANG Jianfeng , BU Dongbo, BAI Shuo Research and Implementation of Text Categorization System Based on VSM Application research of computers 2001,9:23-26
2. CHEN Ruifen Chinese text categorization algorithm combined with feedback Computer Applications 2005 Vol. 25 No. 12
3. Axelrod R. Structure of Decision: the Cognitive Maps of Political Elites. Princeton, NJ:Princeton University Press, 1976.
4. Kosko B. Fuzzy cognitive maps. Int. J. Man-machine Studies, 1986, 24: 65-75
5. Kosko B. Adaptive inference in fuzzy knowledge networks. In: Proc. 1st Int. Conf. Neural Networks. 1987. 2: 261-268
6. Wang Xiaolong, Guan Yi Compter Natural Language Processing Tsinghua Press 2005 ,pp.146-154
7. YE Hao, WANG Mingwen, ZENG Xueqiang Automatic text multi-classification model based on latent semantic Tsinghua Univ (Sci & Tech), 2005, Vol. 45, No. S1: 1818-1822
8. Luo Xiangfeng, Cognitive map theory and its applications in image analysis and understanding ,dissertation of Ph. D,Hefei University of Technology, P. R. China, April 2003
9. LuSong LiXiaoli ,BaiShuo, Wang Shi An Improved Approach to Weighting Terms in Text JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 14 No. 6:8-13,20
10. Zhang Dong li, Wang Dongsheng, Zheng Weimin Chinese text classification system based on VSM J T singhua Univ (Sci & Tech) ,2003, Vo l. 43, No. 9:1288-1291
11. Liu Yunfeng , QiHuan , Xiangen Hu , Zhiqiang Cai A Modif ied Weight Function in Latent Semantic Analysis JOURNAL OF CHINESE INFORMATION PROCESSING Vol119 No16:64-69
12. Chen Zhigang He Pilian Research and Implementation of Text Categorization System Based on VSM Computer Applications 2004,Vol . 24.277-279
13. LU Jiaoli , ZHENGJiaheng The Research of Text Categorization Based on Rough Set JOURNAL OF CHINESE INFORMATION PROCESSING Vol119 No12:66-69
14. Sheng Xiaowei Jiang Minghu Automatic Classification of Chinnese Documents Based on Rough Set and Improved Quick-Reduce Algorithm Journal of Electronics & Information Technology 2005, Vol27 No7: 1047-1052
15. Demetrius A.Georgiou, Despina Makry. A Learner' s Style and Profile Recognition via Fuzzy Cognitive Map. Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04) 2004 0-7695-2181-9/04