Automatic Text Classification Based on Knowledge Tree

Lu Peng, Yibo Gao and Yiping Yang
Dept. of Integration Information System and Research Center
Institute of Automation Chinese Academy of Science
Beijing, China
lujupeng@gmail.com, yibo.gao@ia.ac.cn, yiping.yang@ia.ac.cn

***Abstract*—Automatic text classification is one of important fields in intelligent information process. Most researchers focus on statistic method (Rocchio, SVM, KNN etc.) which is based on Vector Space Model (VSM) representing text. On the basis of analyzing their disadvantages, a new method －automatic text classification based on background knowledge is proposed in this paper. This method is to simulate the classification process of human being. And it includes background knowledge and classification algorithm in order to make computer cognitive ability. It combines text semantic structure and background knowledge to activate relative branches of knowledge tree and decide which classification it belongs to by reasoning. The experiment indicates that the model has higher classification precision and recall.**

***Keywords*—Automatic Text Classification, Knowledge Tree, Cognitive Ability**

I. INTRODUCTION

As with the growth of internet, information has been explored. Automatic text classification as a branch of artificial intelligence is defined as the classification of text into predefined categories. And it has been applied to search engine, spam filter and so on. In this field, research has mainly focus on statistic methods such as Rocchio [1], Native Bayes [2], KNN [3], support vector machine (SVM) [4], neural network [5], etc.. In these methods, the representation of document is mainly concentrated on vector space model (VSM) which represents the document with character, words constructs a feature vector $(\omega_1, \omega_2, \ldots \omega_n)$ in which ω_i is the weight of feature i and n is the total number of features. Then the distance between vectors will classify text category.

Statistical methods are applied widely and achieve greatly, but they have many disadvantages as well as advantages that there is no semantic information. We present a new method based on knowledge tree. It simulates the classification process of human being and classifies documents from the view of cognitive. And our experimental result indicates that this method has higher classification precision and recall.

In this paper, we discuss a automatic text classification method based on knowledge tree. This paper proceeds as follows. Section 2 analyses the disadvantages of statistical methods. Section 3 shows original ideas of our algorithm and system architecture. Section 4 and 5 discuss the key techniques in detail. Section 6 describes the experiment result. Section 7 is conclusion.

II. ANALYSIS OF STATISTICAL CLASSIFICATION METHODS

Statistical text classification is mainly based on VSM, but it has many disadvantages as follow:

(1) Term relation: VSM is assumed to be independence of words in text [6]. In fact, text is made up of concepts and relations. Thus VSM can not be regarded as representation of text. (2) Term difference: There are few overlapped terms among some fields such as sport, politics, computer and so on. On the one hand, distances among these vectors are easy to be distinguished; on the other hand, there are same terms in some fields so that vectors of these categories are difficult to be distinguished because of its similarity. Thus it is difficult to classify these documents according to distances of categories in similar fields. (3) Unuseful terms: In general, text is made up of thousands of words, but many words such as ‘we’, ‘is’ and so on do not belong to any category. VSM takes these words as terms representing a certain category. (4) Other: VSM is based on words form not on the meanings of words. In fact, there are a lot of words with many meanings and a lot of the same meaning with many words forms. But VSM has processed these words equally.

Thus we present a new method for automatic text classification based on knowledge tree.

III. ORIGINAL IDEAS AND SYSTEM ARCHITECTURE

A. Original Ideas

In general, human classification can be divided into two processes. Firstly, we get the surface symbolic information of document, and understand shallowly at the same time. Secondly, information is processed with people’s experiment and knowledge so as to be classified, namely cognitive process. Further text category can be seen as a complicated cognitive process that consists of two aspects: cognitive ability and background knowledge. On the analogy of human classification, we represent a new method for automatic text classification based on knowledge. First, domain knowledge is constructed manually. Last, document is classified according to our classification algorithm. We let knowledge tree represent domain knowledge. In this paper, we regard document category as an integrated process of knowledge and semantic structure.

978-1-4244-1674-5/08 /$25.00 ©2008 IEEE

B. *System Architecture*

In this paper, automatic text classification is divided into three processes. Firstly, domain knowledge is constructed manually, namely, knowledge architecture in cognitive psychology is constructed. Secondly, text is processed such as segmentation to be convenience to be classified. Lastly, text classification algorithm is implemented on knowledge tree; likely, human cognitive ability is on his acquired knowledge. The system architecture is shown as figure 1.

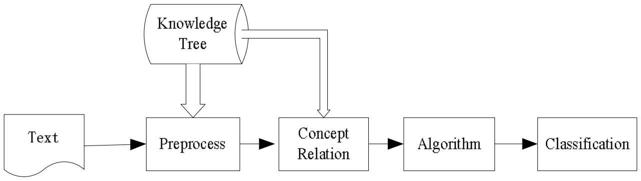


Figure 1. Schematic of the system architecture

IV. KNOWLEDGE TREE

We integrate text information into personal knowledge in order to classify texts, but without relevant knowledge, we can not identify text category because our knowledge system can not be activated. For example, some person can confuse wheat with leek because there is no relevant knowledge for the crucial distinction between wheat and leek in their background knowledge. Thus background knowledge plays the key role in the classifying process, and the integrated process of text information and personal knowledge is regarded as cognitive ability. According to cognitive psychology, personal knowledge is organized by a cognitive structure, namely knowledge structure [7, 8]. In general, knowledge structure consists of three parts: multi categories system, rules of distinguishing categories and relation of different categories. In this paper we divide this structure into two parts: concept and concept relation such as member relation, hyponymy relation and synonymy relation. Knowledge structure is denoted by knowledge tree. It is shown as figure 2.

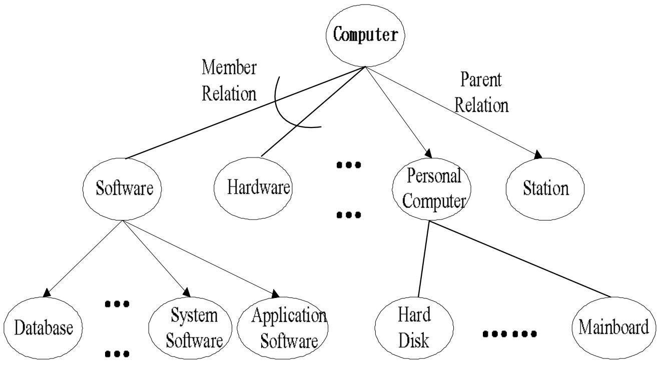
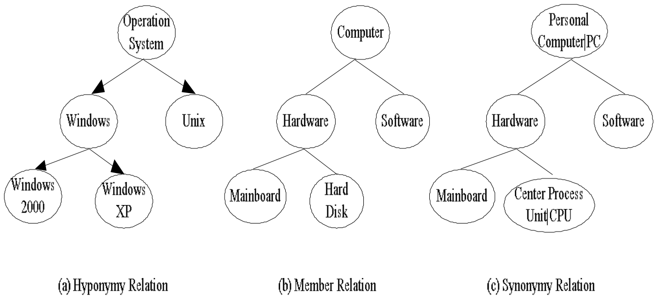


Figure 2. Knowledge Tree in Compute Science.

Knowledge tree is described in detail as follows:

Definition 1 (Node Concept) Node concept denotes concept on node in knowledge tree. Concept is a result that human being acquires objective things or a reflection of objective knowledge in the world [8]. Naturally, symbolic concept denotes the object and its meaning of objective world. In this paper, concept is denoted by nodes in knowledge tree. In general, after we read text title or a few line of text, document category can be judged to classify. In this process, we do not have to read the whole, further we depend on subject concept in order to decide to classify. Subject concept is word or phrase that can distinguish one category from the others. It is well known that the Chinese Library Classification and the Subject Terms List of Chinese Classification are the main basis of books. Actually they can be seen as expert knowledge. Our knowledge tree is constructed manually on the basis of these two books.

Definition 2 (Concept Relation) All things in the world are related to all other things, and some relations exit in concepts. Generally, text consists of concept information and concept relation information. We divide concept relation into three kinds: hyponymy relation, member relation and synonymy relation. Hyponymy relation have the qualities of inheritances, namely son concept can inherit all the features of its parent concept. In knowledge tree, the symbol '→' denotes this relation, for example, node concept 'person computer' and node concept 'computer'. It is shown as figure 3(a). Member relation which is called whole-part relation is a membership relationship between whole concept and its composed part concepts. It is denoted by the symbol '－', for example, 'computer' and 'main board' form member relation. It is shown as figure 3(b). Synonymy relation is a synonymy relationship between concepts. In the world, different things can be represented by different symbol, for example, 'Personal Computer' and 'PC' have synonymy relationship. The symbol '|' denotes this relation. It is shown as figure 3(c).



(a) Hyponymy Relation (b) Member Relation (c) Synonymy Relation

Figure 3. Concept Relation

Knowledge tree is organized by hierarchical semantic structure in which different hierarchical concept and relation construct the whole background knowledge. We use Stias[1] to construct knowledge tree.

V. CLASSIFICATION ALGORITHM

In the process of intelligent action, some strategies are required to utilize as well as psychological activities are represented. Automatic text classification is similar to this

[1] Stias(Science Technology Information Analysis System) developed is a intelligence collection, processing and analysis system for science technology information in the Integration Information System and Research Center, Institute of Automation Chinese Academy of Science.

processing. On the basis of acquired knowledge and experience, we adapt a lot of strategies such as semantic, sequence, syntax and so on to process information. Acquired knowledge has influence over not only applications of strategies but also information integration including text and background knowledge. In other words, text content matches acquired knowledge to be activated so that document can be understood shallowly. Finally the text is classified.

Though the text written style is various, there are the same subject words and relations of words in a category. When we classify the text, the subject words and relations decide document to belong to a certain category. Thus automatic text classification is defined by the matching processing of subject words in text and node concepts on knowledge tree, namely we find the maximum probability of node concept matching the text information. Classification algorithm is described in detail as follow:

Definition 3(Initial Relevant Coefficient of Node Concept R_0) R_0 is defined as the ratio of the number of node concept S_j in the knowledge tree of the text to the number of all node concepts in the knowledge tree in the text. Thus R_0 is shown as formula (1).

$$R_0 = {}^{N_i}\!/_{M} \text{ ,} (N_i \geqslant 0, \quad \text{M} > 0) \qquad (1)$$

In which N_i is the number of S_j in the text, M is the number of all node concepts in the text.

Definition 4(Relevant Coefficient of Node Concepts) Relevant Coefficient of Node Concepts is defined as the relevancy of text to knowledge tree; this is mainly related to the initial relevancy of node concepts and correlation of the upper and lower node concepts. In this paper, there are three relations of nodes: hyponymy relation, member relation and synonymy relation. We only consider the impact of the lower nodes to the upper node, in other words, one node has no direct relationship with the other nodes except for its son nodes. In conclusion, relevant coefficient of non-leaf node concept is summed up three terms: its initial relevant coefficient R_0 , its inherited relevant coefficient R_h from hyponymy relation and its relevant coefficient R_c from member relation.

Inherited relevant coefficient R_h from hyponymy is show as formula (2)

$$R_h = \sum_n a_n \times R_n \qquad (2)$$

In which a_n is the weight of node concept and should meet $0 \le a_n < 1$ and $\sum_n a_n < 1$.

When one node has been activated, we think that its all sub-nodes have been activated. So each sub-node plays the same role to father node. R_h is rewritten by formula (3).

$$R_h = \sum_{n=1}^{N_h} \frac{R_n}{N_h} \qquad (3)$$

In which N_h is the number of its sub-nodes with hyponymy relation, R_n is the relevant coefficient of its sub-node n with hyponymy relation.

It is similar to member relevant coefficient R_c :

$$R_\text{c} = \sum_{m=1}^{N_M} \frac{R_m}{N_M} \qquad (4)$$

In which N_M is the number of its sub-nodes with member relation, R_m is sub-node m coefficient with member relation.

For synonymy relation, the concept is equal to node concept in the knowledge tree, so the concept affects initial relevant coefficient R'_0 .

$$R'_0 = \frac{N_i + N_\text{s}}{\text{M}} \qquad 5)$$

In which is N_i is the number of node concept S_j in the text, N_s is the number of synonymy concept of S_j .

For leaf nodes, their relevant coefficient is equal to their initial relevant coefficient coefficient.

Finally, we utilize independent probability formula shown as formula (6):

$$\begin{aligned} P(X \cup Y \cup Z) &= P(X) + P(Y) + P(Z) \\ &- P(X) \times P(Y) - P(Y) \times P(Z) \\ &- P(Z) \times P(X) + P(X) \times P(Y) \times P(Z) \end{aligned} \qquad (6)$$

So we compute the relevant subject category coefficient shown as formula (7):

$$\begin{aligned} R(S_j) &= R'_0 + R_h + R_c - R'_0 \times R_h \\ &- R_h \times R_c - R'_0 \times R_c + R_0 \times R_h \times R_c \end{aligned} \qquad (7)$$

For those texts not belonging to any category, node concepts may appear in these articles, but relations of these node concepts has no relevant. Thus we predefine to set rules that can remove those concepts in order to identify text categories. We set rules as follows:

(a) $\text{R}_0<0.01$ (b) node concept that there is no any other node concept in this branch is removed.

If meeting these two conditions, node concept S_j is not related to these subject nodes and will be removed.

VI. Experiment

Data set used in our experiment is from the internet and is manually divided into four categories: computer, automation, robot and sport. This data set consists of 1345 documents in total. So far, we have built a sound knowledge tree in automation, computer and robot fields. Our experiment is based on the knowledge tree to use automation, computers and robots corpus to test. And we can identity documents in other fields not to belong to the three fields.

In addition, we use KNN text classification algorithm to compare with. The result is shown as table I.

TABLE I
RESULTS OF TEXT CLASSIFICATION

Evaluation / Category	Our algorithm(%)			KNN(%)		
	P	R	F	P	R	F
Computers	96.9	96.9	96.9	85.3	86.1	85.8
Automation	95.0	94.0	89.7	80.2	82.0	81.1
Robots	92.6	92.2	92.4	79.6	81.0	80.3
Sports	98.2	99	98.6	95.8	97.0	96.4

P is precision, R is recall, and F is F-Value.

Precision= $\mathbf{n}_a$ / $\mathbf{n}_{c;}$ Recall= $\mathbf{n}_a$ / $\mathbf{n}_{b;}$ F-value is a function of recall R and precision P: $F=2P \cdot R/(P+R)$

In which $\mathbf{n}_a$ is the number of documents classified into a class correctly, $\mathbf{n}_b$ is the number of documents that belong to the class, $\mathbf{n}_c$ is the number of documents classified into the class.

From the experiment, we can draw a conclusion as follows:

(1) Our classification algorithm has achieved a preferable result.

(2) There is a good result in very different fields in which there are little same subject words. In my experiment, the sports result is best in four results.

Generally speaking, our algorithm based on knowledge tree has met our requirement and can be applied to the actual classification system. But this algorithm has a little inadequate aspects such as text title and text key words not to be processed.

VII. CONCLUSION

On the basis of the analysis of the statistical automatic text classification, we present a new classification method based on knowledge tree to simulation the process of human classification. This algorithm is based on text semantic structure to avoid disadvantages of VSM. From the experimental results, this algorithm is proved to be a practical one. Though the three fields are tested, knowledge tree has a good scalability to add and delete nodes. We can add any field conveniently by using Stias. In the future work, compositional concepts are processed in knowledge tree in theory; more and more knowledge in other field will be built into knowledge tree in practice.

ACKNOWLEDGMENT

This research is part of Research and Development of Modern Services Infrastructure Services (No. 2006BAH02A01) supported by Chinese National Key Technology R&D Program.

REFERENCES

[1] T. Joachims, “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization,” .In Proceedings of the 14th International Conference on Machine Learning, Nashville, Tennessee, USA, 1997, pp. 143-51.

[2] Susana Eyheramendy, David D. Lewis, David Madigan, **“**On the Naive Bayes Model for Text Categorization,” In the Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics,2003

[3] G.Guo, H.Wang, D.Bell, Y.Bi and K.Greer, “kNN Model-Based Approach in Classification,” In the Proc. of CoopIS/DOA/ODBASE 2003, pp. 986-996, 2003.

[4] T. Joachims, “Text categorization with Support vector machines,” In Proceedings of the 10th European Conference on Machine Learning, pages 137--142, Chemnitz, Germany, 1998.

[5] Pang Jian-feng, Bu Dong-bo, Bai Shuo, “Research and Implementation of Text Categorization System Based on VSM,” Application Research of Computers, 2001

[6] Li Xiao-li, Liu Ji-min, Shi Zhong-zhi,， The Concept-reasoning Network And its Application in Text Classification， Journal of Computer Research and Development， 2000

[7] Kazuhiro Morita , El-Sayed Atlam , Masao Fuketra , Kazuhiko Tsuda , etc., “Word classification and hierarchy using co-occurrence word information,” Information Processing and Management: an International Journal, 2004, pp. 957-972

[8] Wang Su, Wang An-Sheng, Cognitive Psychology. Beijing University Press, 1992, pp. 240-241, pp. 261-275