A SVM-Based Text Classification System for Knowledge Organization Method of Crop Cultivation

Laiqing Ji[1],Xinrong Cheng[1],Li Kang [1],Daoliang Li[1] ,Daiyi Li[1],Kaiyi Wang[2],Yingyi Chen[1,※]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083;

[2] Beijing Research Center for Information Technology in Agricultural, Beijing 10097

※Yingyi Chen, Corresponding author, Tel: +86-10-62738489, Fax: +86-10-62737741, Email: chyingyi@126.com

Abstract. The organization of crop cultivation practices is still far from completion, and Web Resources are not used adequately. This paper proposed a method, based on SVM, to organize the knowledge of crop cultivation practices efficiently from Web Resources. The knowledge organization method of crop cultivation was proposed with Good Agricultural Practices (GAP) in the application of the crop cultivation practices. It is that how to organize the existing crop cultivation knowledge, according to the requirements of crop cultivation practices. It mainly includes a text classification method and a search strategy on the knowledge of crop cultivation. For the text classification method, it used a text classification method based on SVM Decision Tree; for the search strategy, it used a strategy, organized by Ontology and custom knowledge bases. The experiment shows that performance of the proposed text classification method and the knowledge organization method with wheat, is workable and feasible.

Key words: Support Vector Machine (SVM); Text Classification; organization method; crop

1 Introduction

The amount of information about the knowledge of crop cultivation in Web pages is huge and the information is disorganized, so that it is a big workload to organize the related knowledge and not sufficient to use it, resulting in a number of waste of the useful resources. So it is a meaningful thing to organize the related knowledge [1].

To solve these problems, a lot of methods were proposed both at home and abroad. However, these methods [2] only collect the content from the Web Resource, not classifying the collected content and getting what we truly need. The search engine is a retrieval tool for the network information, but the performance of traditional search engine is unsatisfactory [3]. By analyzing Web documents to get the related knowledge, Text mining, as an important branch of data mining, can help have a better use of Web knowledge, based on information retrieval, data mining and knowledge management and so on.

This paper proposed a method on how to organize the knowledge for more effectively organizing and analyzing the documents about the related knowledge. This method involved in how to classify the documents and set the search strategies on the knowledge of crop cultivation.

2 The knowledge organization method

2.1 Knowledge characteristics analysis

The complexity of crop cultivation brings about the complexity of the related knowledge organization, and there is not a very clear knowledge or practice to guide the process of crop cultivation. However, the current research for crop cultivation practices, mainly combined with Good Agricultural Practices (GAP), has also had some progress. In this paper, it mainly referred the document "Combinable crops control points and compliance criteria" in GAP as required.

With the analysis of GAP as a prerequisite, it divided the process of crop cultivation into three phases: pre-production, in-production and post-production. The pre-production mainly consisted of land selection and seed selection. The land selection mainly referred the land type, the time of sowing and so on, and the seed selection mainly referred varietal characteristics, suitable areas for planting, output, resistance, price and so on. The in-production mainly referred sowing and seedling raising, fertilization, irrigation, pest and disease control and so on. For different crops in different areas, the detailed operating of crop cultivation may be different. For example, some crops just need seedling raising, not sowing; the crop cultivation in outdoor and greenhouse is different. The post-production mainly consisted of agricultural harvest, storage and so on.

2.2 The knowledge organization model

Although the knowledge of crop cultivation is numerous and messy with the above analysis, it has certain law so that we can organize it with the knowledge of GAP. The following graphic shows this method.

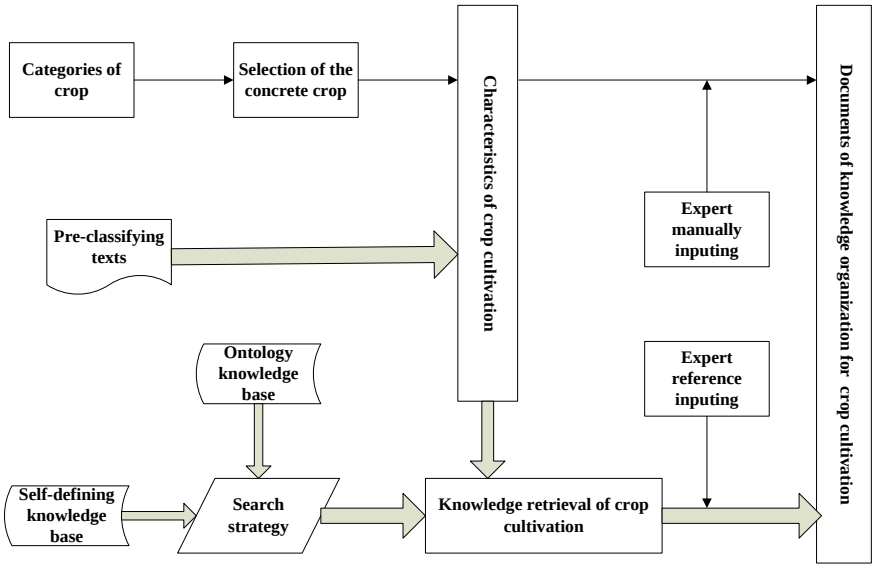

Fig.1. Knowledge organization model of crop cultivation

As shown in Fig.1，the kernel of this model is the text classification method and a search strategy on the related knowledge of crop cultivation. For the strategy in this model, it mainly includes the retrieve key provided by Ontology knowledge database and self-defining knowledge database. For the search strategy for Ontology knowledge database, it is domain Ontology built in accordance with the request

of GAP; for the search strategy for self-defining knowledge database, its setting rule is that the key in the self-defining knowledge database can stand for the most common characteristics in some category.

Taking the wheat as an example, the key in the search strategy of pest and disease control stands for the following: the symptom and their solutions. These solutions mainly refer the irrigation time, the irrigation methods and so on. Because the setting of the search strategy directly affects the accuracy rate of the related knowledge retrieve, we should analyze a lot of related texts, and summarize the most common characteristics of these texts, then make the related search strategy, using the key which can best represent the characteristics.

3 Text classification method of crop cultivation

This paper proposed the text classification method with actual demand and text classification algorithm, based on the knowledge characteristics of crop cultivation. Firstly, collect the sample texts on the knowledge on crop cultivation; then get the training model of the text classification through training the samples; Finally classify the related texts and get the categories of the related knowledge. The following graphic shows the knowledge classification model of crop cultivation.

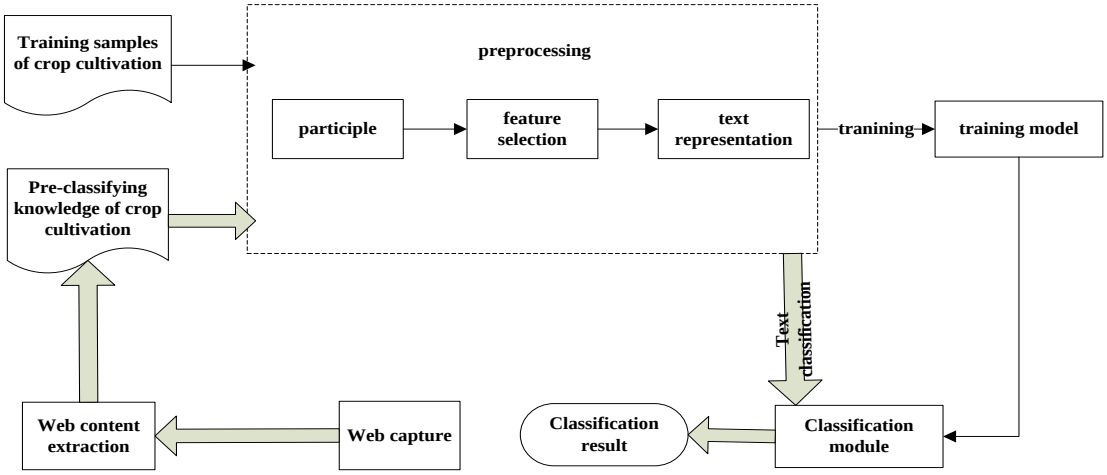

Fig.2. Knowledge classification model of crop cultivation

3.1 Support Vector Machine (SVM)

Based on the minimization principle to structure risk, SVM is a non-linear method. It maps the limited training data(Input Vector) to high dimension feature space, resulting in transforming the problem of the non-linear separability in the sample spaces into the problem of linear separability in the characteristic spaces. In other words, it is how to search the best linear classification Hyperplane. For the finding this Hyperplane, SVM is realized by the Support Vector (SV) and the decision boundary.

3.2 The text classification ideology and algorithm

Through analyzing the related knowledge characteristic of crop cultivation, we know that the knowledge classification is the multi-class issue. As we know, SVM can only solve the two-class classification, so that we need graft many SVMS onto one to realize the knowledge classification of

crop cultivation.

Through analyzing the flaws of algorithm and the related knowledge characteristics, this paper proposed a method, combined binary decision tree with SVM, to realize the related knowledge classification of crop cultivation. Every non-leaf node refers to a SVM classifier and every leaf node refers to the final classified category. The text classification algorithm of crop cultivation, and the algorithm were described as Fig.3 and Fig.4 [11-12]:

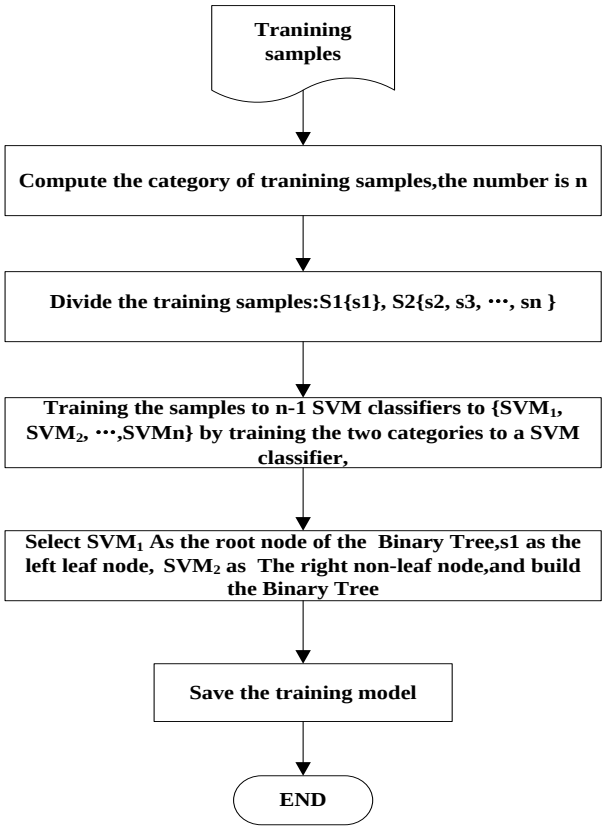

Fig.3. Training Algorithm（n>2）

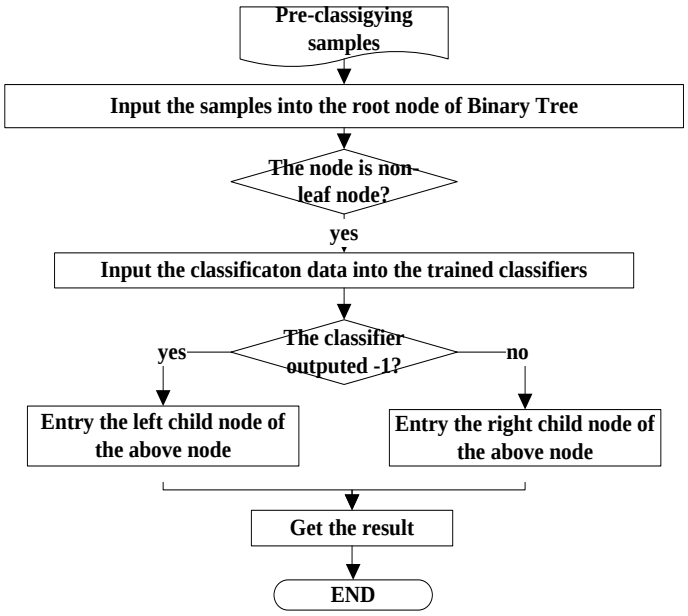

Fig.4. Classification Algorithm

4 Results and discussions

The text classification of crop cultivation is the kernel of the related knowledge organization. According to the actual requirement of the related document arrangement, this paper divided the crop cultivation into six critical control points: land selection, seed selection(seedling raising),fertilization, irrigation, pest and disease control ,harvest. It extracted the related content, using the knowledge of Web crawling and Web analysis, finally classified the content, and made these documents be the training samples in experiment.

It mainly adopt the recall ratio(r) and the precision ratio(p) to evaluate the quality of text classification, and it also refers to the comprehensive evaluation indexes of the recall ratio and precision ratio[10]:F1,its mathematical formula is: $F1=\frac{r*p*2}{r+p}$ [10].

Taking the pest and disease control of wheat as an example. For the search strategy of the pest and disease control, it mainly includes the symptoms of pest and disease control, the name of the medicines, the usage and dosage of the medicines and so on. Let the number of the related texts about the pest and disease control be *num2*; Let the number of the related texts about the pest and disease control of wheat be *num2*; Let the number of the texts including all of the search strategies on the pest and disease control of wheat be *num3*.

Let M be the reference values of the specialists, and the specific definition of M is that it is the ratio of the number of the text, including all of the search strategy in a category, and the number of the related texts in a category. It is shown by the definition of M , M =*num3* / *num1*.It shows the result of the experiment about the text classification method of crop cultivation and the knowledge organization method of the pest and disease control to wheat in the following table1.

Table1. Result of the text classification and the knowledge organization method of the pest and disease control to wheat

categories	Training samples	p	r	F1	num1	num2	num3	M
Land selection and preparation	205	87.3%	86.5%	86.9%	×	×	×	×
Seed selection and seedling raising	170	93.2%	93.0%	93.1%	×	×	×	×
Fertilization	205	95.8%	96.1%	95.9%		×	×	×
Irrigation	160	90.6%	89.7%	90.1%	×	×	×	×
Pest and disease control	263	91.5%	89.5%	90.5%	113	40	12	0.107
Harvest	200	94.1%	95.4%	94.7%	×	×	×	×

5 Conclusions

On the premise of the related knowledge characteristics analysis of crop cultivation, this paper proposed the knowledge organization model of the crop cultivation. In this model, it mainly described the principle how to set the search strategy and the SVM-Based text classification ideology and algorithm.This paper had the related test. The experiment result shows that the organization method is workable and feasible, and it can effectively provide the data for the document arrangement of crop cultivation practices. The experiment result also shows that it needs a lot of data to achieve the

effectiveness, or the effectiveness is not apparent.

Acknowledgements

Funding for this research was provided by National Key Technology R&D Program. The project name is Study on the comprehensive evaluation system of crop cultivation practices. The project number is 2009BADB6B02-01.

References

1. Usama MF. Data Mining and knowledge discovery: making sense out of data. IEEE Expert, 1996,11(5):20-25
2. Hand DJ. Intelligent Data Analysis: Issue and Opportunities. Proc. Advances in Intelligent Data Analysis, London,1998,2:67-79
3. Guowu Jiang, Xinrong Cheng, Li Kang, etc. Building Knowledge base for Consulting System on Agricultural Practical Techniques.IEEE proceedings of the 2009 International Conference on Computer and Computing Technology Applications in Agriculture.2009
4. O. L. Mangasarian and D.R.Musicant. Active set support vector machine classification. In Advances in Neural Information Processing Systems,2000,577-583
5. T. Joachim's. Making large-scale support vector machine learning practical. In Advances in Kernel Methods: Support Vector Machines,1999
6. Qiang Niu，Zhixiao Wang, Dai Chen, etc. The method of the Web text classification based on Support Vector Machine. Microelectronics and compute,2006, 23(9): 102-104
7. Hanbin Zou. Application of Support Vector Machine in the text classification.2006
8. C.W.Hsu and C.J.Lin. A comparison on methods for multi-class Support Vector Machine. Technical report,Department of Computer Science and Information Engineering,2001
9. Knerr S,et al.Single-layer learning revisited:A stepwise procedure for building and training a neural network. In Fogclman-Soulie et al.(ed.),Neuro-computing:Algorithms,Architectures and Applications,NATO ASI.Springer,1990
10. WenBi Rao，HuiYan Ke. Research and implementation of Web Text Classification. The computer technology and development, 2006, 16(3):116-118
11. E.Osuna, R.Freund, and F.Girosi. Training support vector machines: An application to face detection. In, editor, Proceedings CVPR'97,1997b
12. E.Osuna, R.Freund, and F. Girosi. Training support vector machines: An application to face detection. In Proceedings of CVPR'97,1997