IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 3, March 2016.
www.ijiset.com

ISSN 2348 – 7968

Investigation and Application of Improved Text mining based on Bayes Algorithm

Tang Zhi-hang
School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China

Abstract*—Finding features in semi-structured or non-structured document data was not an easily thing in the past. With the development of text mining technique, it has become a widely used analysis tools to find out patterns or rules in hugely text data. The purpose of this study tries to discern some common patterns in the lyrics of music chart singles, which make these singles popular. We provide a walk-through of language identification fundamentals by first introducing the theoretical background, followed by a step-by-step guide on model building, according to the standard practice of data mining. We will put emphasis on understanding the problem, and solving it in an effective and costless manner within the open source environment RapidMiner. We will become familiar with RapidMiner's "Text Mining Extension", and learn how to create several simple and fast workflows for language identification. The methods we will use are generic, and applicable to any dataset, with the appropriate pre-processing. In addition, we will learn how to use the implemented models in custom projects, by exporting them as web services using Rapid Analytics, and deploying them in custom applications. We will go through a case study of a web application for language identification of web pages, using RapidMiner's "Web Mining Extension". Special emphasis will be given to the optimization of the exported workflow, to enable faster execution, and harmonious integration in other applications.

Keywords*—Bayes Algorithm, text *mining, web mining

1. Introduction

In the past, finding features in semi-structured or non-structured document data was not an easily thing. With the development of text mining technique, it has become a widely used analysis tools to find out patterns or rules in hugely text data. For example, Weng and Liu [1] proposed a mechanism based on text mining technology to reduce the load of replying customer service e-mails. de Oliveira et al[2] Used text mining technology to analyze and acquire the enterprise's competitive strategies. Aurora [3] proposed a topic discovery system to reveal the implicit knowledge with text mining technology. Although the topics are somewhat different, all of them showed that text mining is suitable for finding the information needed in the semi-structured or non-structured document data.

The purpose of this study is to find out the common features of famous singles using text mining technology. Base on the results of text mining, several analysis methods has been used to quantify the features in the lyrics and then proposed a quantitative measurement mechanism which can be used to predict whether a single might become attractive to audience.

2. Related work

2.1 Text Mining

Feldman and Dagan [4] was one of the pioneers to give much attention to KDT (Knowledge discovery from text) or text mining. They describe KDT as a process to find out the profitable and usable information in texts. Thus, text mining can be broadly defined as a knowledge discovery process in which an individual extracts the useful information from a text-based data by using analysis tools [5]. As compare with the data mining which is an automatically process to discover useful information from structured data stored in the database [6], the main objective of text mining is to discover valuable knowledge embedded in semi-structured or non-structured document data [7].

ISSN 2348 – 7968

Feldman and Sanger [5] indicate that the results of text mining usually represent the features of documents rather than the underlying documents themselves. Although the potential features of documents can be represented in various ways, the commonly types of feature used are: characters, words, terms, and concepts.

Nasukawa and Nagano [8] suggests that the entire text mining procedure can be divided into three components:

Concepts extraction. Since the same words or terms may sometimes express different meaning in the different document. As a result, the first part of text mining is concentrate on how to extract the meaningful concepts from the document.

Rules and patterns discovery. After extract the concepts from the documents, some quantitative analysis methods may be used to find out the hidden rules and patterns.

Display and interactive analysis. The final part of text mining is to visualization the results in a multi-dimension view and provide an interactive interface for user to interact with.

2.2 Application of Text Mining

The primary objectives of text mining are classification and prediction. Several methods were used to work for text mining in recent years. Among these methods, most of them are the widely studied and applied method of data mining. Weiss et al [9] pointed out that text mining technology can be used in five major areas:

Document classification. This area is about assigning documents to one of several predefined categories.

Information retrieval. Information retrieval is a common topic especially in web-based text mining. It allows researchers to gathering the documents on Internet as the clues to retrieve the specific document.

Clustering and organizing documents. Distinct from classification which assigning documents to the predefined categories, clustering technology is useful to group semi-structured or non-structured document data as clusters and assigning the labels needed for each cluster.

Information extraction. Information extraction is to extract useful information from a text-based data. As mentioned, it is more difficult to extract the useful information from semi-structured or non-structured document data as compare with the traditional data mining. Thus, finding useful information from semi-structured or non-structured document is more valuable and interesting to the researchers.

Prediction and evaluation. The ultimate goal of text mining is prediction. Based on the finding patterns in the input document data, researchers can use these patterns to form of some generalized rules that can be used to predict and evaluate something they care about.

3. Improved Text mining based on Bayes Algorithm

3.1 The Problem of Language Identification

The task of language identification can be simply described as discerning the language of a given text segment. For example, given a set of six sentences in different languages, the goal is to identify the language of each sentence. This task may be somewhat trivial for humans. We might be able to easily discern the language and possibly identify some of them, even though we might not speak any of those languages. The first sentence is easy, since we are reading this book written in English. We may say that the second sentence “feels” German, or that the third one “sounds” French. The alphabet of the fourth sentence looks completely different from the others, so we might guess it “looks like” Greek, while the fifth sentence might “lead to” Spanish. We might find the sixth sentence somewhat tricky since it looks different from the previous ones. If we are vaguely familiar with the Slavic family of languages we might say it “relates to” that family, since the sentence is written in Croatian. Finally, although we might find the last sentence “similar to” Spanish, it is, in fact, written in Portuguese.

ISSN 2348 – 7968

When identifying languages, we use our knowledge of languages, acquired either actively or passively. To design algorithms for achieving the same task, we first need to systematize the knowledge needed for language identification. There are several indicators we can rely on when identifying languages, without knowing those languages at all; Examples of several different alphabets are given in Figure 1.

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (1500 examples, 1 special attribute, 1 regular attribute) View Filter (1500 / 1500): all

Row No.	language	text
1	eng	Although we did not have a very positive exchange on Mr Graefe zu Baringdorf' s report earlier on and will certainly have to agree to disagree, things may l
2	eng	But as has been pointed out, the adoption of a rigorous definition of the precautionary principle is crucial.
3	eng	In Scandinavia, which perhaps has what is at present the most efficient eco-labelling system, there are now already a number of eco-labels complemen
4	eng	The first step towards achieving the emission ceilings is undoubtedly a dynamic implementation and enforcement of existing directives, such as the Aut
5	eng	There is the old case of visiting scientist B and how his contract came into being.
6	eng	Finally, I would like to underline our support for the creation of a European pension fund with the participation of the different economic sectors and the p
7	eng	Mr President, hooliganism is not a borderline phenomenon but clearly takes the form of a subculture and is, as such, difficult to control.
8	eng	The vote will take place at 5.30 p.m.
9	eng	Mr President, I shall be brief, because almost everything has been said already, while I was waiting my turn.
10	eng	It is very important that my colleague, Mrs Lucas, should have the opportunity in this House to raise the issue of whether her immunity extends to this cas
11	eng	We shall therefore vote in favour of this budget as an emergency measure.
12	eng	I would like to outline my view on the divided European reaction to the recent American/British action against Saddam Hussein' s regime in Iraq.
13	eng	This can and must be achieved while, at the same time, fully respecting the provisions of the Treaty, the rights of the institutions concerned and the existi
14	eng	They are committing more troops. They are making a greater effort.
15	eng	Some people - and I will tell you who, the British conservatives in particular - are calling for the Committee of the Regions to be abolished.
16	eng	I repeat: Galileo is a gateway to a European information society and I expect the Council of Transport Ministers to perceive it as such tomorrow.
17	eng	The euro currency does run a particular risk of being counterfeited at the time when it is introduced into circulation and during the period when it will circu
18	eng	However, I am not so sure whether the comments regarding the European Central Bank's first two years are exactly correct.
19	eng	The Letter of Amendment is another possibility to consider for the presentation of this document.

Figure 1 example set of English

3.2 Token-based Representation

The token-based representation is built on top of basic meaningful elements of text, called tokens. Tokens can be words separated by delimiters; logograms—signs and characters representing words or phrases, such as Chinese letters; idioms, or fixed expressions, such as named entities—names of places, people, companies, organisms, etc.

The process of token extraction is called tokenization. For most of the languages and uses, tokenization is done by splitting sentences over whitespaces and punctuation characters, sometimes ignoring numbers, or allowing specific punctuation in a word, e.g., wasn't, off-line. The frequent words approach identifies the most frequent words in the text, and uses them for the language identification. The most used words across different languages have similar meanings, and regarding the type of words, they are usually articles, prepositions, pronouns, and some frequently used verbs, and adjectives. Due to their high frequency of occurrence, they are suitable even for shorter texts. Figure 2 depicts main process of model frequent words

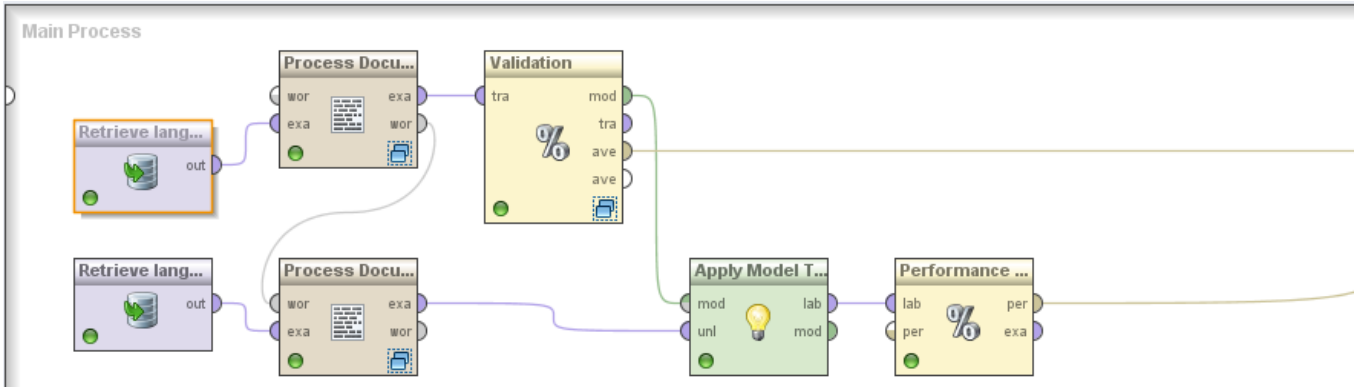


Figure 2 main processes of model frequent words

3.3 Examining the Word Vector

ISSN 2348 – 7968

A word vector is just a fancy name for a table, where each row is a document (SMS message in this case), and each column is a unique word in the corpus (all of the words in all of your documents). The values inside the table depend on the type of word vector you are creating. In this case we are using Term Occurrences, meaning that a value in a cell represents the number of times that word appeared in that document. You could also use the Binary Term Occurrences, meaning the value in the cell will be zero if the word did not appear in that document, and one if the word appeared one or more times in that document. It is always a good idea to examine your data, in order to "get a feel" for it, and to look for strange anomalies.

1. Click the Example Set tab to view the word vector. You will see that the word vector has 9755 attributes, meaning that there are 9755 unique words in the corpus. Equivalently, there are 9755 columns in the word vector.

2. Look at the Range column for the "text" role attribute. You will note that:

- "Sorry I'll call later" is the most common message.
- Below that, you can see that there are 4827 "ham" messages, and 747 "spam" messages.

3. Click the Data View button. You will see that document 13 has a "1" under the word "A", meaning that the word "A" appears one time in that document. Actually, "a" and "A" appear in that document, but we are considering the letter case of the words in this process, so they are counted as distinct words, Figure 3 depicts table of wordlist frequencies and Figure 4 depicts table of example set..

Result Overview | WordList (Process Documents from Data) | ExampleSet (Process Documents from Data)

Word	Attribute Name	Total Occurences	Document Occurences	ham	spam
yrs	yrs	12	11	5	7
ystrday	ystrday	1	1	1	0
ything	ything	1	1	1	0
yummy	yummy	2	2	2	0
yun	yun	2	2	2	0
yuo	yuo	4	2	4	0
yuou	yuou	1	1	1	0
zac	zac	1	1	1	0
zealand	zealand	1	1	1	0
zed	zed	6	6	0	6
zeros	zeros	1	1	1	0
zhong	zhong	1	1	1	0
zoom	zoom	1	1	1	0
zyada	zyada	1	1	1	0
É	É	1	1	1	0
Ü	Ü	53	50	53	0
èn	èn	1	1	1	0
é	é	3	1	3	0
ú	ú	1	1	0	1
ü	ü	120	97	120	0
鈥	鈥	1	1	1	0

Figure 3 A table of wordlist frequencies

ISSN 2348 – 7968

Result Overview | WordList (Process Documents from Data) | ExampleSet (Process Documents from Data)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (5574 examples, 2 special attributes, 9979 regular attributes) View Filter (5574 / 5574): all

Row No.	text	att1	A	ABOUT	ABOUTAS	ABTA	ACCEPT	ACL	ACTION	ACTUAI
5557	Well keep in mind I ve only got enough gas for one more rounc	ham	0	0	0	0	0	0	0	0
5558	Yeh Indians was nice Tho it did kane me off a bit he he We shu	ham	0	0	0	0	0	0	0	0
5559	Yes i have So that s why u texted Pshew missing you so much	ham	0	0	0	0	0	0	0	0
5560	No I meant the calculation is the same That It gt units at It gt Th	ham	0	0	0	0	0	0	0	0
5561	Sorry I ll call later	ham	0	0	0	0	0	0	0	0
5562	if you aren t here in the next It gt hours imma flip my shit	ham	0	0	0	0	0	0	0	0
5563	Anything lor Juz both of us lor	ham	0	0	0	0	0	0	0	0
5564	Get me out of this dump heap My mom decided to come to low	ham	0	0	0	0	0	0	0	0
5565	Ok lor Sony ericsson salesman I ask shuhui then she say quit	ham	0	0	0	0	0	0	0	0
5566	Ard like dat lor	ham	0	0	0	0	0	0	0	0
5567	Why don t you wait til at least wednesday to see if you get your	ham	0	0	0	0	0	0	0	0
5568	Huh y lei	ham	0	0	0	0	0	0	0	0
5569	REMINDER FROM O To get pounds free call credit and details	spam	0	0	0	0	0	0	0	0
5570	This is the nd time we have tried contact u U have won the Pou	spam	0	0	0	0	0	0	0	0
5571	Will ü b going to esplanade fr home	ham	0	0	0	0	0	0	0	0
5572	Pity was in mood for that So any other suggestions	ham	0	0	0	0	0	0	0	0
5573	The guy did some bitching but I acted like i d be interested in b	ham	0	0	0	0	0	0	0	0
5574	Rofl Its true to its name	ham	0	0	0	0	0	0	0	0

Figure 4 A table of example set

3.4 Validating the Model

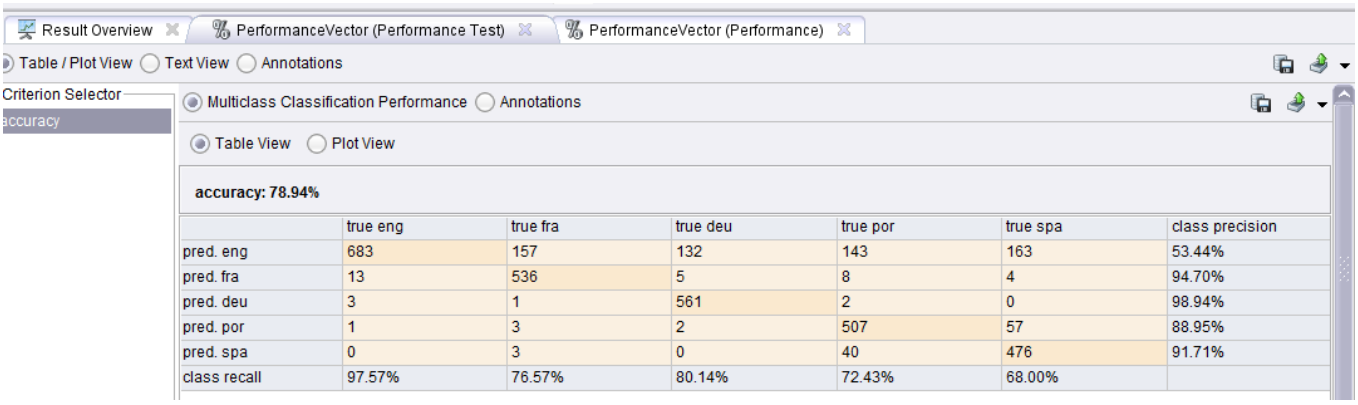
To build the Support Vector Machine model, add the Support Vector Machine operator after the Process Documents from Data operator. To find the model's predictive accuracy, we must apply the model to data, and then count how often its predictions are correct. The accuracy of a model is the number of correct predictions out of the total number of predictions. Add the Apply Model operator after the Support Vector Machine operator and connect their two nodes together. Add a Performance operator after the Apply Model operator and connect it to a res node.

To be able to predict a model's accuracy on unseen data, we must hide some of the data from the model, and then test the model on that unseen data. One way to do this is to use K-fold Cross-Validation.

When using, say, 10-fold Cross-Validation, we would hide 1/10th of the data from the model, build the model on the remaining 9/10ths of the data, and then test the model on the whole dataset, calculating its accuracy. We would do this again, hiding a different $1/10^{th}$ of the data from the model, and test again. We would do this 10 times in total, and take the average of the accuracies. This provided a better idea of how the model will perform on data that it has not seen before, Figure 5 depicts accuracy of performance vector and Figure 6 depicts classification error of performance vector.

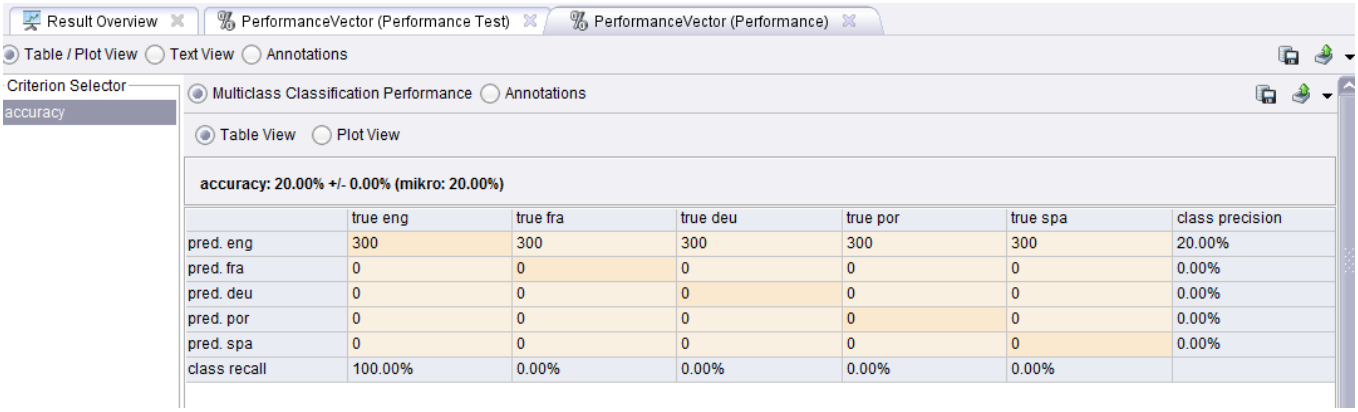
1. Remove the Support Vector Machine, Apply Model, and Performance operators from the Main Process window.
2. Connect an X-Validation operator to the Process Documents from Data operator, and connect its ave (for average performance) node to a res node.
3. Double-click the X-Validation operator. Put a Support Vector Machine operator in the left side of this inner process, and an Apply Model operator and a Performance operator in the right side of the process. Connect all required nodes.

ISSN 2348 – 7968



	true eng	true fra	true deu	true por	true spa	class precision
pred. eng	683	157	132	143	163	53.44%
pred. fra	13	536	5	8	4	94.70%
pred. deu	3	1	561	2	0	98.94%
pred. por	1	3	2	507	57	88.95%
pred. spa	0	3	0	40	476	91.71%
class recall	97.57%	76.57%	80.14%	72.43%	68.00%	

Figure 5 accuracy of performance vector



	true eng	true fra	true deu	true por	true spa	class precision
pred. eng	300	300	300	300	300	20.00%
pred. fra	0	0	0	0	0	0.00%
pred. deu	0	0	0	0	0	0.00%
pred. por	0	0	0	0	0	0.00%
pred. spa	0	0	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	0.00%	0.00%	

Figure 6 classification error of performance vector

4. Conclusions

Language identification is an important pre-processing step in many text mining applications, and is in its basic form easy to implement in RapidMiner. In this chapter, we demonstrated three different basic approaches to the problem. The frequent words method identified languages over the most frequent words in the corpus, per language. We saw that this method yields relatively high accuracy, but is unable to correctly classify sentences which do not contain frequent words, since the number of words in a language is practically unlimited. The character n-gram method, on the other hand, uses character n-grams instead of whole words in order to do the language identification. This method yields higher accuracy than the frequent words method, and possibly results in a smaller and faster model. The similarity-based method presented in the end, simplifies the resulting model and the computation, though at a cost of a lower accuracy. We implemented and tested these methods using RapidMiner, on an easily obtainable dataset. In the end, we showed how to use the similarity-based workflow in order to create a workflow for web page language identification, and export it as a web service using Rapid Analytics. Having a language identification workflow as a web service enabled us to apply it in various custom applications by simply issuing an HTTP call.

By getting familiar with the problem, its possible solutions, and using the solution in a real-world problem, we hope you mastered language identification and are ready for further challenges. Further challenges might include experimentation with method improvement, optimization of accuracy, and computational performance. Though it might seem that language identification is a solved problem, there is still room for further research and improvement, especially in the application areas like tweet, multilingual document, or language-variant language differentiation. We hope that we successfully gave you the ground knowledge, and inspired you to tackle such challenges and venture deeper into the new horizons of language identification.

ACKNOWLEDGEMENTS:

ISSN 2348 – 7968

A Project Supported by Scientific Research Fund of Hunan Provincial Education Department（15A043）

REFRENCES:

[1]. Weng, S.S. & Liu, C.K. 2004. Using Text Classification and Multiple Concepts to Answer E-mails. Expert Systems with Applications, 26(4), 529–543.
[2]. de Oliveira, J. P. M., Loh, S., Wives, L. K., Scarinci, R. G., Musa, D. L., Silva, L., & Zambenedetti, C. 2004. Applying Text Mining on Electronic Messages for Competitive Intelligence, In Proceeding of the 5th International Conference on Electronic Commerce and Web Technologies, Spain: Zaragoza.
[3]. Aurora, P. P., Rafael, B. L., & José, R. S. 2006. Topic Discovery Based on Text Mining Techniques. Information Processing and Management, 43(3), 752–768.
[4]. Feldman, R. & Dagan, I. 1995. KDT- Knowledge Discovery in Texts, In Proceeding of the First International Conference on Knowledge Discovery and Data Mining (KDD), Canada: Montreal.
[5]. Feldman, R. & Sanger, J. 2007. The Text Mining Handbook-Advanced Approaches in Analyzing Unstructured Data, USA: New York.
[6]. Tan, A. H. 1999. Text Mining: The State of the Art and the Challenges. In Proceedings of the 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining, China: Beijing.
[7]. Losiewicz, P. B., Oard, D. W., & Kostoff, R. N. 2000. Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information System, 15(2), 99–119.
[8]. Nasukawa, T. & Nagano, T. 2001. Text Analysis and Knowledge Mining System. IBM Systems Journal, 40(4), 967–984.
[9]. Weiss, S. M., Indurkhya N., Zhang T., & Damerau F. J. 2005. Text Mining-Predictive Methods for Analyzing Unstructured Information, USA: New York.

< Zhi-hang Tang >, <1974-08-08>, <hunan, China>

Current position, **Doctor of Hunan Institute of Engineering**

University studies: **control theory and control engineering in donghua University**

Scientific interest: intelligent decision and knowledge management

Publications <number or main>: 40 Papers

Experience: Zhihang TANG was born in Shaoyang, China, in 1974. He earned the M.S. degrees in control theory and control engineering from zhejiang University of techonlogy, in 2003 and Ph.D. from donghua University China in 2009. At the same time ,he is a teacher in department of computer and communication, Hunan Institute of Engineering(Xiangtan, China) from 2003.Chaired the 49th China Postdoctoral Science Foundation grant, presided over science and technology projects in Hunan Province in 2010, presided over the Education Department of Hunan Province in 2010 Outstanding Youth Project, as the first author more than 30 papers were published.His current research interests include intelligent decision and knowledge management.