

# Discipline Decision Tree Classification Algorithm and Application based on Weighted Information Gain Ratio

Yan Xia<sup>1</sup>, Jian Shu<sup>2</sup>, Na Xu<sup>3</sup> and Hui Feng<sup>1</sup>

<sup>1</sup>Shanghai Joint Laboratory for Discipline Evaluation, Shanghai Education Evaluation Institute, Shanghai, China

<sup>2</sup>Shanghai General Motor, Shanghai, China

<sup>3</sup>Shanghai Municipal Education Examinations, Shanghai, China

**Keywords:** Data Mining, Information Entropy, Information Gain Ratio, Decision Tree, Discipline Classification, Discipline Evaluation.

**Abstract:** Discipline evaluation is an important part in higher education evaluation. It plays a significant role in discipline construction in universities and colleges. It is challenging how to use scientific discipline evaluation to classify disciplines, such as advantageous disciplines and newly-emerging ones. This paper proposes an algorithm of discipline decision tree classification based on weighted information gain ratio. It determines evaluation attributes and creates decision tree according to weighted information gain ratio. Discipline classification rules are deduced by decision tree. An automatic classification system is developed, implementing the algorithm and analysing data from universities and colleges in Shanghai. Experimental results show that our scheme can achieve about 83.33% accuracy in forecasts. It provides advice and guidance for discipline evaluation, and establishes foundation for discipline development strategy.

## 1 INTRODUCTION

Discipline is a basic unit of universities and colleges. Discipline construction is the core of constructions in universities and colleges, to improve talent cultivation quality and scientific research level, and to serve society. At present, most universities and colleges in China have completed the layout adjustment and structure scale of discipline. They come into a new stage of improving discipline construction quality, cultivating discipline characteristics, forming discipline advantages, and promoting discipline development (Han and Mei, 2011). At the stage, it is significant to develop advantageous disciplines and newly-emerging disciplines. Therefore, it is important how to use scientific method to carry out discipline evaluation, to classify disciplines to select advantageous disciplines and newly-emerging ones. Now discipline evaluation is usually carried out in a way which combines an objective calculation of data and peer review. The discipline evaluation indicator system mainly focuses on a university or college's teaching staff and resources, its scientific research level, its talent cultivation quality and its academic reputation. The data come from officially released

information for public use and the materials submitted by the universities and colleges for evaluation. Data mining is usually used to determine advantageous disciplines and newly-emerging ones. The widely used tool for data mining analysis in academic research and evaluation is bibliometric, which evaluates discipline according to indicators related to articles. Although it is objective and easy to operative, it is mainly focusing on scientific research level, and neglecting other perspectives. It is difficult to implement scientific and comprehensive discipline evaluation only from the perspective of bibliometric (Hood and Wilson, 2001). It has become a hotspot in higher education field to establish a discipline evaluation system based on objective data in order to make a scientific classification of disciplines in universities and colleges. With its help, the educational administrative department can easily understand the current situation of discipline development, and can promote the development of higher education in China healthily and rapidly.

This paper proposes a discipline decision tree classification algorithm based on weighted information gain ratio. It determines evaluation attributes and establishes decision tree according to

different weighted information gain ratio. Discipline classification rules are deduced by decision tree. An automatic classification system is implemented. It investigates the application of data mining technology in discipline classification, and provides advice for the discipline construction in universities and colleges.

## 2 RELATED WORK

Advantageous disciplines and newly-emerging disciplines are the basis of development of first-class discipline in the world. They play important roles in the development of the discipline cluster. At the same time, they take advantage of complementary disciplines to promote cooperative development among related disciplines. At present, discipline evaluation is commonly used to determine whether it is advantageous discipline, newly-emerging one or not. The research of discipline evaluation is composed of several categories. One kind of discipline evaluation is based on university ranking, such as Times Higher Education World University Rankings (THE) (Marijk, 2008), U.S. News College Rankings (USNWR) (Jamil and Alenoush, 2007), China Discipline Ranking (CDR) from China Academic Degrees and Graduate Education Development Center (CDGDC, 2013). One kind of discipline evaluation is based on scientific mapping, such as Bibliometric Rankings from the Centre for Science and Technology Studies (CWTS) in Laiden University of Holland (Moed, 2006). One kind of discipline evaluation is based on tendency research, such as discipline value evaluation in The National Center for Scientific Research (CNRS) in France. Another kind of discipline evaluation is based on scientific fund management, such as evaluation for the financing disciplines from Biotechnology and Biological Sciences Research Council (BBSRC) (Aghion et al., 2010). All the above discipline evaluation uses the traditional method, which is combined of subjective and objective evaluation, such as expert assessment and bibliometric. The evaluation process is complex. The evaluation result is easy to be influenced by the subjectivity and so on.

In order to solve the existing problems in discipline evaluation, this paper proposes a new discipline decision tree classification algorithm based on weighted information gain ratio. An automatic discipline classification system is implemented, verifying the algorithm and analyzing data from universities in Shanghai. It provides

advice and guidance for comprehensive discipline evaluation and developing strategy of disciplines.

## 3 DISCIPLINE DECISION TREE CLASSIFICATION ALGORITHM

The discipline decision tree classification algorithm determines evaluation attributes according to weighted information gain ratio and correlation between them. Then it establishes decision tree. The decision tree is a directed graph to classify items. It consists of a root node (a node in the graph to which no other node points), internal nodes (nodes that are pointed at and to other nodes), and leaves (nodes that don't point to other nodes) (Han et al., 2011). The classified item travels from the root to one of the leaves, where classification is made. Discipline classification rules can be deduced by decision tree.

### 3.1 Basic Definitions

Definition 1 Let S be the set of training samples, and  $C_i$  the set of all classification attributes. Let  $S_{C_i}$  be the subset of S. Probability mass function  $P_i$  is defined as (1).  $T(S)$  is the cardinal number of S.

$$P_i = T(S_{C_i}) / T(S) \quad (1)$$

Definition 2 The entropy of S relative to  $C_i$  is defined as (2). Accum is accumulation of  $P_i$ ,  $i=1 \dots m$ .

$$I(S) = - \text{accum} (P_i * \text{Log}_2 P_i) \quad (2)$$

Definition 3 Let D be the subset of S, which contains several different evaluation attributes,  $\{A_1, A_2, \dots, A_n\}$ . If each  $A_i$  has  $k_i$  corresponding characteristics, D can be divided into  $k_i$  subsets according to evaluation attribute  $A_i$ . The entropy of D relative to  $A_i$  is defined as (3).

$$E(D, A_i) = - \text{accum} \{ [T(D_{ij}) / \text{accum} T(D_{ij})] * I(D_{ij}) \} \quad (3)$$

Definition 4 The information entropy of D relative to evaluation attribute  $A_i$  is defined as (4).

$$G(D, A_i) = I(D) - E(D, A_i) \quad (4)$$

Definition 5 The information entropy of evaluation attribute  $A_i$  is defined as (5).

$$S\text{Info}(D, A_i) = - \text{accum} \{ [T(D_{ij}) / \text{accum} T(D_{ij})] * \text{Log}_2 [T(D_{ij}) / \text{accum} T(D_{ij})] \} \quad (5)$$

Definition 6 Considering that the relationship between evaluation attributes has effect on

classification prediction, the information entropy of evaluation attribute  $A_i$  will be adjusted. If  $A_i$  has  $m_i$  non-associated characteristics, the mean value of non-associated information entropy of evaluation attribute  $A_i$  is defined as (6).

$$SInfo(D, A_i) = -\text{accum} \{ [T(D_{ij}) / \text{accum} T(D_{ij})] * \text{Log}_2 [T(D_{ij}) / \text{accum} T(D_{ij})] \} / m_i \quad (6)$$

Definition 7 The information gain ratio of evaluation attribute  $A_i$  is defined as (7).

$$GRatio(D, A_i) = r * G(D, A_i) / [SInfo(D, A_i) + t * SInfo(D, A_i)] \quad (7)$$

The  $r$  is set as the weight of evaluation attribute  $A_i$ .  $GRatio$  increases as far as  $r$  increases. So the corresponding evaluation attribute is more probably used as partition. On condition that there is same number of items in the evaluation attributes,  $SInfo$  increases while  $GRatio$  decreases if a special evaluation attribute has more classification attributes. Thus the corresponding evaluation attribute is less probably used as partition.  $t$  is set as the correlation coefficient, and  $t$  belongs to  $[0,1]$ . The value of  $t$  can be adjusted in the process of algorithm optimization.

### 3.2 Workflow

The work flow of Discipline Decision Tree Classification Algorithm is shown as algorithm 1 according to the above definition.

Algorithm 1: DTCA-WIGR (Discipline Decision Tree Classification Algorithm based on Weighted Information Gain Ratio)

Input: Samples (the set of training samples), Attributes (the set of evaluation attributes)

Output: The decision tree of discipline classification

Work flow:

```

DTCA-WIGR (Samples, Attributes)
Begin
Create root node R-Node;
If (All items in Samples belong to
same classification attribute C) then
Return R-Node as leaf. Mark it as C;
If (There are no extended evaluation
attributes in Attribute) then
Return R-Node as leaf. Mark it as
classification attribute, which most
items in Samples belongs to;
Else // R-Node is non-leaf.
Continue to classify them.
a. Calculate the information gain
ratio GRatio for each evaluation
attribute  $A_i$  ;

```

```

b. Select the evaluation attribute
Amax, which has the maximum GRatio;
c. Mark R-Node as evaluation
attribute Amax ;
d. Samples are divided into kmax
subsets, according to the number of
characteristics in Amax;
e. For each (  $S_i$  )
e.1. A branch with condition  $k_i$  is
generated according to R-Node;
e.2. If (There are no items in  $S_i$ )
then
Return  $S_i$  as leaf. Mark it as
classification attribute, which most
items in Samples belongs to;
e.3. Else
DTCA - WIGR ( $S_i$ , Attributes - Amax),
End

```

The recursive step of algorithm DTCA-WIGR stops when it meets with following conditions.

1. All items in Samples belong to same classification attribute.
2. There are no extended evaluation attributes in Attribute for partition.
3. There are no items in  $S_i$ .

The time complexity of algorithm DTCA-WIGR is  $O[\text{cardinal}(\text{Attributes}) * \text{cardinal}(\text{Samples}) * \text{Log}_2 \text{cardinal}(\text{Attributes})]$ .  $\text{cardinal}(\text{Attributes})$  is the cardinal number of the set of evaluation attributes.  $\text{cardinal}(\text{Samples})$  is the cardinal number of the set of training samples.

Noise and outlier appear in the process of constructing decision tree. Algorithms of pruning decision tree can be used to improve accuracy of classification (Quinlan, 1987).

## 4 APPLICATION OF THE ALGORITHM

### 4.1 Discipline Evaluation Indicator System

This paper uses evaluation indicators from Discipline Evaluation Indicator System in 2012, which is promulgated by China Academic Degrees and Graduate Education Development Center (CDGDC, 2012). The content of Discipline Evaluation Indicator System is shown in Table 1. The Discipline Evaluation Indicator System is composed of 4 primary indicators, including Teaching Staff and Resources, Scientific Research Level, Talent Cultivation Quality, and Academic Reputation. Each primary indicator is composed of

several secondary indexes, 17 secondary indicators in all. Each secondary indicator contains a number of observation points with different weight, which can be considered as tertiary indicators. For example, Teaching Staff and Resources, one of the primary indicators, contains 4 secondary indicators. There are 12 observation points in expert team, one of the secondary indicators, such as the number of academicians in China Academy of Sciences or China Academy of Engineering, the number of distinguished experts in “Thousand Talents Plan”, the number of Yangtze River Scholar and so on.

Table 1: Discipline Evaluation Indicator System promulgated by CDGDC.

Primary	Secondary
A. Teaching Staff and Resources	A1. Expert Team
	A2. The ratio of Students to Teachers
	A3. Full-time Teacher Staff
	A4. Outstanding Disciplines and Laboratory
B. Scientific Research Level	B1. Representative Academic Papers (including domestic and foreign, quality and quantity)
	B2. Scientific Research Award
	B3. Academic Publications and Patent transformations
	B4. Representative Research Projects
	B5. Creativity of Art (only for disciplines of arts)
	B6. Architectural Design (only for disciplines of architecture)
C. Talent Cultivation Quality	C1. Quality of Teaching and Teaching Materials
	C2. Quality of Dissertation
	C3. International Exchange
	C4. Sports Competitions (only for disciplines of sports)
	C5. Outstanding Students and Graduates
	C6. Number of Granting Degrees
D. Academic Reputation	D1. Discipline Reputation (including academic reputation, social contribution, and academic ethics, etc)

The weight of the primary indicator is  $r_i$ . The  $i$  belongs to [A, B, C, D]. The weight of the secondary indicator is  $r_j$ . The  $j$  belongs to [1, 2, ..., m]. The  $m$  is the cardinal number of corresponding secondary indicator. The weight of the tertiary indicator is  $r_k$ . The  $j$  belongs to [1, 2, ..., n].  $n$  is the cardinal number of corresponding tertiary indicator. The records in database map to the tertiary indicator. The weight of the evaluation attribute  $r_{ijk}$  is defined as  $r_{ijk} = r_i * r_j * r_k$ .

This paper analyses the relationship between evaluation attributes by principal component analysis (Yang and Feng, 2012). Non-associated characteristics of evaluation attributes are recorded in database to calculate the mean value of non-associated information entropy, which helps to measure information gain ratio of evaluation attribute

## 4.2 Data Selection

To ensure authenticity, reliability and authority, the related data come from officially released information for public use and the materials submitted by the universities and colleges for evaluation, both of which are subject to vigorous verification. They are all focusing on tertiary indicators. The data are integrated into the discipline basic information table in the database. The table structure is shown in Table 2, which defines 76 evaluation attributes together with some primary key and foreign keys. The table of weight is set to keep the weight of each evaluation attributes. The table of relationship is set to keep the associated characteristics.

Table 2: Table structure of discipline basic information.

N	Meaning	Name	Type	...
1	University or College ID	UCID	char	...
2	University or College Name	UCName	varchar	...
3	Discipline ID	DID	char	...
4	Discipline Name	DName	varchar	...
5	Number of Academicians	Academicians	smallint	...
6	Number of Experts in “Thousand Talents Plan”	ExpertsTTP	smallint	...
7	Number of Distinguished Young Scholars for The National Science Fund	YScholarTNF	smallint	...
8	The ratio of Doctors to Teachers	RatioDT	float	...
9	The ratio of Masters to Teachers	RatioMT	float	...
10	Number of National Outstanding Disciplines	NODisciplines	smallint	...

Table 2: Table structure of discipline basic information (Cont.).

N	Meaning	Name	Type	...
11	Number of Highly Cited Papers by ESI	HCPapers ESI	medium int	...
12	Number of papers in SSCI,AHCI&CSSCI,CSCD	PapersSA CC	medium int	...
13	Number of Natural Science Award	NaturalSA	medium int	...
14	Number of Academic Publications	APublications	medium int	...
15	Number of Patent Transformations	PatentTSF	medium int	...
16	Number of National Natural Science Fund	NationalNSF	medium int	...
17	Number of Projects from Provinces and Ministry	ProjectsPM	medium int	...
18	Number of National Teaching Achievement Award	NTeachingAA	medium int	...
19	Number of Projects supported by National Science eand Technology Ministry	NationalSTM	medium int	...
20	Number of Outstanding Graduates	OutstandingGD	medium int	...
...	...	...	...	...
77	Passing Rate of Doctoral Dissertation	PRateDD	float	...
78	Number of Foreign Students	ForeignST	medium int	...
79	Number of Exchange Students	Exchange ST	medium int	...
80	Number of Granting Degrees	GrantingDegrees	medium int	...
81	Year	Year	medium int	...
82	Classification Type	CType	varchar	...

We shall do some preprocessing works, such as data cleaning, data integration, data transformation,

data reduction and so on since data from source databases are incomplete, inconsistent, and redundant (Carlo, 2010).

### 4.3 Automatic Discipline Classification System

#### 4.3.1 System Structure

The system structure of automatic discipline classification system based on weighted gain ratio is shown in Figure 1. The process is as follows.

Create model: Train from sample dataset by Discipline Decision Tree Classification Algorithm. Create decision tree, and deduce classification rules.

Optimize model: Evaluate and analyse existing rules by testing dataset. Optimize classification rules.

Apply model: Apply the optimized classification rules on new data to classify disciplines.

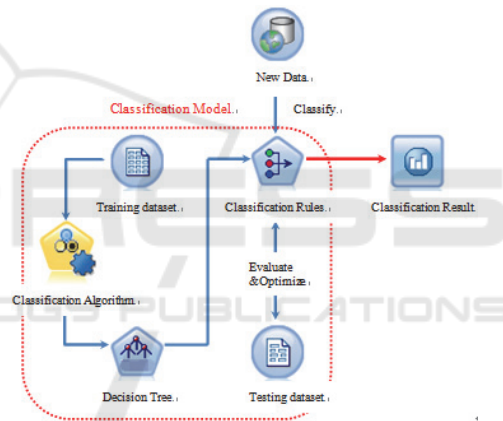


Figure 1: System structure of automatic discipline classification system.

#### 4.3.2 Create Sample Dataset

This paper focuses on discipline evaluation in universities and colleges of Shanghai. It makes discipline classification in Shanghai. According to the result of China Discipline Ranking from CDGDC, disciplines in universities and colleges of Shanghai are defined as 4 categories, outstanding disciplines, advantageous disciplines, newly-emerging disciplines and general disciplines. Classification attributes are created in database. For a few disciplines which were not evaluated in China Discipline Ranking in 2012, they are evaluated comprehensively by experts from universities and from educational administrative department of Shanghai according to Discipline Evaluation Indicator System from CDGDC in 2012. Their

classification attributes are also created in database.

The sample of training dataset is shown in Table 3. 98 disciplines from 26 universities and colleges of Shanghai are selected as samples in training dataset. Attributes  $A_i$  ( $i = 1, 2, \dots, 76$ ) are defined as evaluation attributes in Table 2 of Table Structure of Discipline Basic Information. The Discipline Decision Tree Classification Algorithm based on Weighted Information Gain Ratio is applied to train the sample dataset.

Table 3: Sample of training dataset.

Evaluation Attributes				Classification
...	$A_i$	$A_j$	$A_k$	Outstanding
...	>80	high	>16	Advantageous
...	$\leq 20$	high	10..16	Newly-emerging
...	21..50	medi	>16	Advantageous
...	51..80	high	10..16	Newly-emerging
...	$\leq 20$	medi	<10	General
...	21..50	low	10..16	Newly-emerging
...	51..80	high	10..16	Advantageous
...	>80	high	>16	General
...	$\leq 20$	low	<10	General
...	51..80	high	>16	Advantageous
...	51..80	high	>16	Outstanding
...	21..50	medi	<10	Newly-emerging
...	51..80	medi	>16	Advantageous
...	$\leq 20$	low	10..16	General
...	>80	medi	>16	Advantageous
...	>80	medi	>16	Advantageous
...	21..50	low	10..16	General
...	$\leq 20$	low	<10	General
...	51..80	high	>16	Newly-emerging
...	...	...	...	...
...	21..50	medi	10..16	Advantageous
...	>80	high	<10	Newly-emerging
...	>80	medi	>16	Advantageous
...	21..50	high	<10	General

### 4.3.3 Create Discipline Classification Rules

When running the automatic discipline classification system, the prediction model extracts samples from training dataset, and creates discipline classification decision tree by the algorithm. It prunes the decision tree by the post-pruning algorithm based on

Bayesian theory (Cai, 2011). Figure 2 shows parts of the decision tree.

Classification rules are created by the decision tree. Figure 3 shows parts of the classification rules.

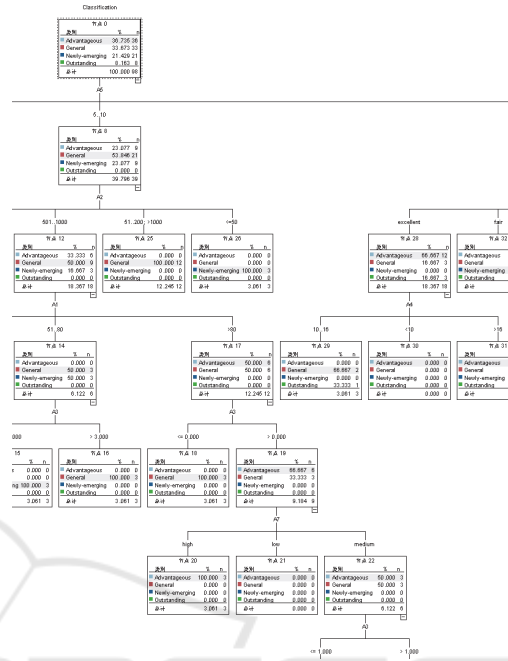


Figure 2: Discipline classification decision tree.

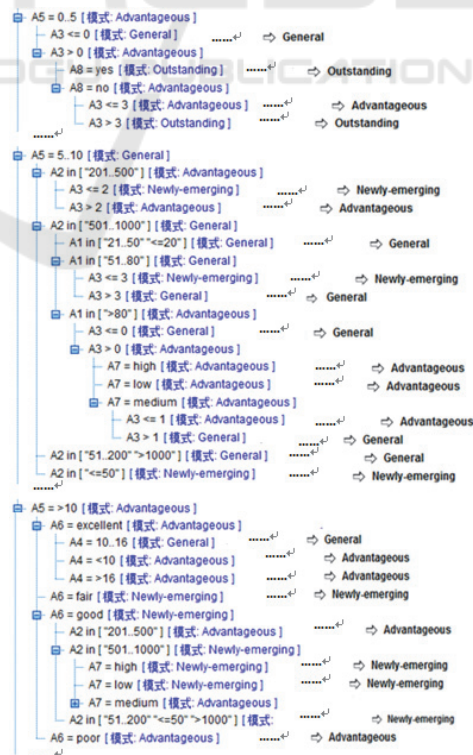


Figure 3: Discipline classification rules.

### 4.3.4 Evaluate and Optimize the Classification Rules

How to evaluate the classification rules is important. Testing dataset is used to evaluate the Discipline Decision Tree Classification Algorithm. We compare the predicting results with the actual results. It is effective if our scheme can achieve about 80% accuracy in forecasts.

The first-class disciplines of universities and colleges in Shanghai are selected as testing dataset, which contains 161 disciplines in 22 universities (Shanghai Municipal Government, 2012). 60 typical items are imported to test the classification rules. Testing result is shown in Figure 4. 49 prediction results are correct. We achieve about 81.67% accuracy in forecasts. Coincidence matrix compares difference between prediction value and actual value. Confidence report is analyzes the accuracy of prediction.

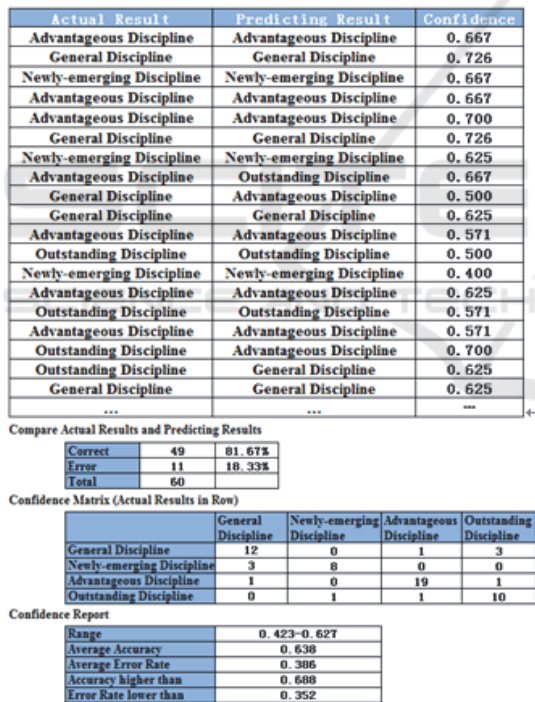


Figure 4: Evaluation result of discipline classification rules.

After evaluating and analysing classification rules, the correlation coefficient, shall be adjusted slightly to get the optimized decision tree and classification rules. We can raise the accuracy to 83.33%. The optimized result is shown in Figure 5.

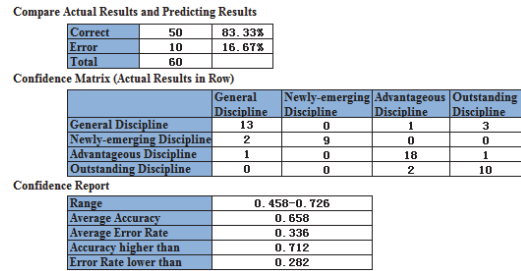


Figure 5: Evaluation result of optimized classification rules.

### 4.4 Use Automatic Discipline Classification System to Do Dynamic Evaluation

The automatic discipline classification system based on weighted information gain ratio classifies discipline according to objective data instead of subjective assumption. It establishes foundation for the objective discipline quality system. It initializes and monitors the disciplines dynamically. The educational administrative department can easily understand the layout structure and status.

## 5 CONCLUSIONS

This paper proposes an algorithm of discipline decision tree classification based on weighted information gain ratio on the basis of extensive investigation and careful analysis of the existing evaluation methods. An automatic discipline classification system is established to analyze the discipline samples in universities and colleges of Shanghai. The study and application of the algorithm is helpful to reveal the development tendency of disciplines. It predicts the progress and breakthrough of disciplines in the future. Meanwhile it provides basis for the educational administrative department to develop a new round discipline strategy.

In the future, we will take further research on fuzzy decision tree according to the discipline characteristics. Fuzzy classification rules will be deduced to evaluate disciplines more scientifically.

## REFERENCES

Han, Wenyu, Mei, Shiwei, 2011. Master discipline law, Cultivate discipline culture, Promote development, *China Higher Education*, vol. 7.

- Hood, W.W, Wilson, C.S., 2001. The literature of bibliometrics, scientometrics, and informetrics, *Scientometrics*, vol. 52.
- Marijk, van der Wender, 2008. Ranking and Classification in Higher Education: A European Perspective, *Higher Education*, vol. 23.
- Jamil, Salmi, Alenoush, Saroyan, 2007. League Tables as Policy Instrument: Uses and Misuses, *Higher Education Management and Policy(OECD)*, vol. 19.
- CDGDC, 2013. Brief introduction of discipline evaluation, <http://www.chinadegrees.cn/xwyyjsjyxx/xxsbdxz/276985.shtml>.
- Moed, H. F., 2006. Bibliometric Rankings of World Universities, *The Netherlands: Centre for Science and Technology Studies (CWTS), Leiden University*.
- Aghion P, Dewatripont M, Hoxby C, 2010. The governance and performance of universities: evidence from Europe and the US The governance and performance of universities: evidence from Europe and the US, *Economic Policy*, vol. 25.
- Han, Jiawei, Kamber, Micheling, Pei, Jian, 2011. The book, *Data Mining: Concepts and Techniques*, 3<sup>rd</sup> edition.
- Quinlan, J. R., 1987. Simplifying decision trees, *International Journal of Man-Machine Studies*, vol. 27.
- CDGDC, 2012. Discipline Evaluation Indicator System in 2012, <http://www.chinadegrees.cn/xwyyjsjyxx/xxsbdxz/277134.shtml>.
- Yang, Xue, Feng, Hui, 2012. An Evaluation on the Input-Output Performance of Universities Based on Principal Component Analysis, *Shanghai Management Science*, vol. 34.
- Carlo, Batin, 2010. The book, *Data Quality: Concepts, Methodologies and Techniques*, 1<sup>st</sup> edition.
- Cai, Z, 2011. Identifying product failure rate based on a conditional Bayesian network classifier, *Expert Systems with Applications*, vol. 38.
- Shanghai Municipal Government, 2012. First-class Disciplines in Shanghai universities and colleges, <http://www.shanghai.gov.cn/shanghai/node2314/node2319/node12344/u26ai33230.html>.