**Product Functional Information Based Automatic Patent Classification: Method and Experimental Studies**

*LI Wen-qiang a, LI Yan \*a,CHEN Jian b ,HOU chao-yi a*

*a School of Manufacturing Science & Engineering, Sichuan University, Chengdu 610065, China*

*b Faculty of Engineering and the Environment, University of Southampton, Southampton SO17 1BJ, UK*

**Abstract:** In order to effectively extract the hidden information from the patent texts and to further provide this information to support the product innovation design process, this paper proposed an automatic patent classification method based on the functional basis and Naive Bayes [theory](javascript:void(0);). The functions of products are regarded as the [innovation](javascript:void(0);) [attribute](javascript:void(0);)s, and the function co-reference relations of the patents in different areas are established. Patent classification methods are proposed based on the functions of products and the general steps of the patent classification process are proposed. In addition, three training methods are studied in the experiments, including multi-classification fully supervised training, multiple dichotomous supervised training and semi-supervised training. Through comparing and analyzing the experimental results, a patent text classifier is developed. In summary, this paper provides a general idea and the [relevant](javascript:void(0);) technologies on how to build a patent knowledge space by automatically extracting and expanding the patent texts. .

**Key words**: Innovation design; functional basis; patent text classification; Naive Bayes; EM algorithm

# 0 Introduction

Product innovation design can be considered as a result of the transition, superposition and reconstruction of the knowledge in different areas. The innovation level of product design is strongly subject to the level of knowledge’s extraction and processing. As the world's largest innovative knowledge carrier, patent is an important knowledge resource to extend designers’ knowledge space and assist their product innovation design. Statistics shows that new technologies reported by patents are always 3-5 years earlier than those from other information sources, such as scientific journals and conference papers. According to the investigation from the World Intellectual Property Organization, 90%-95% of inventions in the world are reported by patents and 80% of them are not recorded by other texts **[1]**. In addition, patents also have different innovation levels. Some patents’ innovation ideas are obtained through changing function targets, some are obtained from new scientific effect principles, and others are obtained by just improving local structure of products or process. Normally patent texts only include the information about inventor, invention contents and right request and etc, but do not contain the obvious information on which invention method and what principle knowledge are used. However, this information is actually hidden in the patent texts. This hidden information needs the readers to understand and to obtain by themselves. Therefore, it significantly restricts the efficiency of the patents’ application and the knowledge’s transition. How to effectively extract the tacit innovation information which are hidden in patent texts in different areas and how to use this information to support designer during the product innovation design process have become important issues of patents’ utilization.

Existing studies on the mining and utilization of the patent innovation information can be generally classified into two categories. The first one is the general information analysis on the patent collection, so the distribution attributes and the development trends of the patent collection can be obtained. It can, in general, help designers to explore the innovation maturity level of products and avoid the patent infringement issues. Hodes et al **[2]** proposed a patent analysis method by applying the ontology technology. This method can compare and analyze the concepts in the patent claims, so it can help the designers/users to avoid the patent infringement issues. Yoon et al **[3]** proposed a keyword-based morphology method for the patent trend analysis. Tseng et al **[4]** carried out the dynamic study of the patent technology by using the test methods and the patent feature vector model. Xiao et al **[5]**developed a Patent Mapmethod that can analyze the patent information and summarize the distribution and trends of technologies based on the rearrangement of different priorities of the patent collection. Atsushi et al **[6]**  established a conceptual vector model which can automatically generate patent maps based on the clustering model of target patents. Liu et al **[7]**  proposed a product maturity prediction model by mining the technology contents from patent texts. All these methods can quickly find the general distribution and trends of patents. However, because all of them use keyword-based searching and clustering technologies, the obtained trends and analysis results are always concentrated in a certain area or focus on a certain category of product, and they cannot break the patent boundaries between different areas. This limitation hurdles the transition and reconstruction of patent knowledge in different areas.

The second type of patent innovation information mining and utilization focusing on the analysis of individual patents and they can extract the innovation information hidden in patents and support designers to quickly carry out the innovative thinking and design. Cascini et al **[8-9]** developed a Subject-Action-Object ternary structure model to describe the technical features of patent text and it can provide the corresponding patent knowledge support for the TRIZ’s implement process. Based on TRIZ theory, Liu et al **[10]** studied the scientific effects in patents. Liang et al **[11]** established a classifier which used 40 [inventive](javascript:void(0);) principles as the classification standard, and the classifier can automatically classify patent texts and extract the [inventive](javascript:void(0);) principles from patent texts,. Soo et al **[12]** used the ontology technology to develop a patent representation model, which described the configuration relationship of mechanical parts. Boyko et al **[13]** analyzed the patent texts and extract structure information from patents by using the ontology and natural language processing techniques. Xue et al **[14]** put forward a patent automatic categorization method to support the innovation design process, but the method has no formalized and standard classification criteria. Wang et al **[15]** used the "demand-Function-Principle-Structure" mapping of product conceptual design to develop a patent innovation information mining method. Liu **[16]** used the functional basis and the TRIZ principle to extract information attributes from patent texts and established a patent knowledge base.

Based on the patent mining and application methods above, some commercial Computer Aided Innovation (CAI) software companies have implemented the patent knowledge base module into their software packages as an important resource to support the innovation design, such as Pro/Innovator **[17]**, Goldfire Innovator **[18]**, CREAX **[19]** and so on. Through delivering the useful knowledge to the designers at the right time and managing the designers thinking in a right way, this patent knowledge base module can provide the technical support for the designers during the innovative design and product development process.

Although these patent mining and application methods could effectively break the boundaries among different areas, they didn’t establish ontology models or the co-reference relations among different areas, and they can only search patent collection and associate the relevant patents through the simple form of upper and lower keyword, so the accuracy and pertinence of the searching results are poor. Additionally, these patent knowledge bases cannot be built until the patents are manually read and understood based on the innovation attributes which was established beforehand. As a consequence, the efficiencies of these patent mining and application methods are quite low. Therefore, the current research focus points on the patent knowledge mining and application are how to break the area boundaries of different professional patents and establish the innovative attribute ontology for the patent knowledge transition, consequently establish the co-reference based on the innovative attribute ontology, accurately extract and classify the patent knowledge from different areas, and finally efficiently help designers to carry out product innovation design. This paper proposed a co-reference model of product functional attributes in different areas, and also developed an automatic mining and classification method of patent information. According to the product functional attribute ontology, the labeled patent knowledge was divided into training set and test set. Then based on Naive Bayes (NB) classification theory, from single-class point of view and multi-class point of view respectively, the patent innovation information of a large number of unlabeled patents were extracted and automatically classified in fully supervised and semi-supervised experiments, and the classifier has been improved as well. It has been demonstrated that the proposed method can automatically extract and classify the innovation information from any new patent text by means of the ontology relationship.

# 1 Patent knowledge based innovation design process

Product design process is the process, through which the designers organize and reconstruct the design knowledge to enter their cognitive access based on a design goal. As the most important knowledge resource in the world, patent can support designers to carry out various design activities. Normally, designers only search the relevant patents in their own professional area in accordance with the analysis of the design objective. Even though the designers do not gain some patent knowledge before, they can obtain this knowledge by technical term based search or keyword based search. This patent knowledge is belonging to the explicit knowledge of the designers. But if only searching the patent knowledge in a certain professional area, the designers’ innovation space will be seriously restricted. Therefore it is necessary for the designers to jump out their own professional area and obtain the patent knowledge from other professional areas. However, knowing and understanding the patent knowledge from various areas is normally quite challenging for the designers. Since this cross-area knowledge cannot be accurately obtained by the designers, it belongs to the tacit knowledge of the designers. After the designers obtain a lot of explicit and tacit patent knowledge through various ways, they will analyze and introduce this knowledge into their cognitive access. The patent knowledge, which just enters the cognitive access is always scattered, isolated and disorderly. Thus the designers need to judge, filter and sort the knowledge based on their own innovative thinking. According to the design tasks, the designer may remove some “noise” patents knowledge; meanwhile they may introduce some new patents into their cognitive access. As results, the quality of the patents in the cognitive channel is further improved. While the design tasks are gradually decomposed and specified during the design process, the patent knowledge entering the cognitive access changes periodically. So the quantity of patents reduces gradually, and the contents of the patents are transformed from the principle level to the behavior knowledge, and further to the structural level. The relationship between the different patents becomes closer and closer, and the designers’ design schemes inspired by the patent knowledge become clearer and more specific. Following the design constraint and optimization process, the designers finally obtain the schemes that satisfy the product design requirements, as shown in Figure 1.



Figure 1 the patent knowledge based product innovation design process

*The creative design process of product is the result of migration, combination and reconstruction of knowledge in different domains. Patent, the most significant innovation information of human society, will enhance the level of design personnel's creative design and innovation efficiency when the patent information is effectively utilized to support the innovation of the design activity.* Due to characteristics of patent text and format, the current patent knowledge base and patent resources platform are mostly organized by individual professional area and consequently they can only be searched by the keywords in the professional area. It tremendously restricts the possibility to find the relevant patent information in cross-area, namely the hidden patent information. Although some organizations have begun to research and  to organize the patents from various areas in accordance with the patents’ innovative attributes, they still rely on the keyword-frequency method for the patent’s analysis and classification, and the effectiveness and accuracy of patent classification are not high. Therefore, which patent attribute to use to express the patent knowledge, and how to extract the relevant innovation knowledge from the patents in cross-area and help designers in knowledge transfer among different areas have become critical points for the patent knowledge mining and application.

# 2 Function attributes and construct co-referential relationship

Product function is not only the goal and starting point of product design, but also the basis of meeting the needs of users. The axiomatic design model defines the design process as a mapping process among the user domain, functional domain, structure domain and process domains. The FBSE (function - behavior - structure - constraints) design model **[20]** defines the design process as a mapping among the function domain, behavior domain, structure domain and constraint domain. Most of the current product innovation design models are based on these two models. In this paper, the functions of products are considered as the [innovation](javascript:void(0);) [attribute](javascript:void(0);)s. The function attribute ontology and the function co-reference relations of multidisciplinary patents were established by extracting function attributes from the patents in different areas. A patent search and classification method was proposed based on the product function and the function co-reference relationship. The product innovation process begins with the analysis of the demands of users and market. After obtaining the design goals and transforming it into the corresponding function target, the total function of product design process can be achieved. Next, the total function is abroken into several functional elements and each of the functional element can be described as a corresponding standard functional group. The corresponding patent texts are searched and conveyed to the designer based on the description of standard functional groups. Based on the corresponding behavior in patents, the specific physical structure can be found to implement various functional elements. After comprehensively analyzing on each design constraint and optimizing the structure carriers, the structure or the structure combination, which can meet the demands of products, can be finally obtained. The whole process is shown in Figure 2.



Figure 2 Product design process and design elements

Moreover, because of the different description methods on the features of products, the existing function body models are normally established in regard to different areas and design goals **[21-22]**. Based on the above description and expression methods, a functional ontology model, in the form of “operation + object", is put forward in this paper (as shown in Figure 3). Functional group comprises functional operations and functional objects. The function operations consist of 11 sub-classes: shift, regulate, absorb, combine, detect, stabilize, import, accumulate, output, produce, and separate. The functional objects include material, information and energy. So the proposed functional ontology model comprises: 11 primary level functional groups and 50 secondary-level functional groups, as shown in Figure3.



Figure 3 Operation-object based functional group decomposition

Due to the diversity of natural languages, patent applicants in different professional areas normally use different ways to express the functions which are essentially carried by the patents; consequently, it is impossible to accurately search the cross-area patent knowledge by the limited number of functional groups or functional group keyword. Therefore, it is necessary to unify the patent functional expression in different professional areas and build the co-reference relations library which has the ability of semantic migration, and finally support the effective patent knowledge migration in different areas. This paper proposes a semantic crawler method to realize the semantic extension of functional groups, obtain the extended terms (also referred as the weights of the terms) which semantically associate with each functional group and build co-reference relation library and ontology relationship. Using the co-reference relations library and the semantic relevance weights among functional groups, the design task can be preliminary described and then matched to the relevant functional groups, and finally the corresponding patent information can be reached by capturing the designers’ intention. The idea of building the co-reference relations library is, based on the functional ontology, to establish the public relation network about the synonymous relationship words, similar relationship words, up & down relationship words, and structural relationship words in different fields according to different similarity calculation of Sim(O,Sn), as shown in Figure 4. Extended ontology vocabularies can be associated with the semantic correlation among the functional groups and the semantic similarity can be calculated from the semantic correlation between words based on the method from the Wiktionary. The vocabularies that have the relevance with the function ontology but do not have a similarity to the above relationship can be obtained from the existing patents training examples through the machine leaning. The relationship between the functional ontology and extended vocabulary can be described by means of the Web Ontology Language.



Figure 4 Relationship set used for defining the ontology of co-reference relationship

Table 1 ontology [relationship](javascript:void(0);) [type](javascript:void(0);) and [relationship](javascript:void(0);) [express](javascript:void(0);)

|  |  |  |
| --- | --- | --- |
| [**Relationship**](javascript:void(0);)  [**type**](javascript:void(0);) | [**Relationship**](javascript:void(0);)  [**express**](javascript:void(0);) | [**Relationship**](javascript:void(0);) **illustration** |
| [synonymy](javascript:void(0);) [relationship](javascript:void(0);) | same-as | There is synonymous relationship between two concepts |
| accumulation relationship | part-of | There is partial and overall relationship between two concepts |
| [inheritance](javascript:void(0);) [relationship](javascript:void(0);) | kind-of | There is a father-son relationship between two concepts |
| Co-classification | sibling-of | Two concepts belong to the same function |
| [relation](javascript:void(0);) [attributes](javascript:void(0);) | attribute-of | Concept A is an attribute of concept B |
| Instance relationship | instance of | Concept A is an example of concept B |
| existent relation | intersection-of | Two concepts is not on the same tree but have other relationship |

# 3 Function based patent mining and classification

## 3.1 Mining methods of patent knowledge

*Function is the goal of product design and the essential attribute to meet user demands. This paper selects the functions of patent invention subject as patent label and extract and classify the patent from different fields. Patent includes basic information, abstract, claims, description part and so on. The section of patent abstract is a general description of subject invention. Although it cannot fully describe the major content and the concrete structure of subject invention, it includes the function information which the invention subject need to implement. So, it can mine the function information from the patent abstract text. Moreover, the patent abstract text can be obtained directly from the official patent web site in the form of normative documents and it doesn't need to be converted the format like description part, which will greatly improve the efficiency of mining and classification of patent.*

Because patent is a kind of text information, patent knowledge mining belongs to the text information mining. Text mining is a process that discovers knowledge by extracting implicit, previously unknown and potentially information from the text data. The text mining technologies **[23-24]** mainly include:

* Feature extraction: extract nouns, phrases, date, time, currency, and other digital information properties from the texts.
* Subject indexing: use subject terms rather than keyword to index texts.
* Text categorization: automatically classify the natural language texts into the predefined categories and it is a typical machine training method.
* Text clustering: automatically cluster the texts in accordance with the similarity of the given texts, and it is a typical unsupervised machine training method.
* Automatic summary: the process of analyzing the structure of the texts by computer technology, finding the topic statement of the article and finally constituting the abstracts based on the process of sorting, combination and modification.

However, the patent information mining is to provide designers the specific patents that can realize each functional group while the above text mining technologies cannot meet this requirement properly. So a patent text classification method based on the functional group was proposed in this paper and it can achieve the automatic mining of patent texts and provide the patent information to the designers in accordance with different functional needs in the product innovation process.

## 3.2 Function based patent text classification process

According to the general method of text classification, the major steps of patent text classification process proposed in this paper are shown in Figure 5. First, some relevant patents are downloaded from official patent database. After reading them, the functional group of individual patent is defined and the function of each patent is manual labeled, so this labeled text set becomes the text set of the next step classification training experiment. Then, the text set is divided into the training set and test set by a certain separation ratio. The features were extracted from the training text set to build the corresponding feature vector. Then appropriate classification training methods are selected to establish a classifier. The test text set is classified by the classifier and the results are compared to its manual labels, so accuracy of the classifier can be evaluated. When the classification results finally meet the defined target, the classifier can be applied to the classification of any new patent text with any unknown function. The detail of each step will be explained in the latter sections.



Figure 5 the flow chart of text classification

## 3.2.1 Patent text preparation

The first step of patent mining is to determine the patent text data sources. Because the patent text contains a lot of information, to select which information as the basis of the patent mining and classification has been recently studied by some researchers **[25, 26, 27]**.Their principle of the selection is to reduce the workload of information mining as much as possible while ensuring the accuracy of classification. Since the title and abstract contain most of innovation information of the patents, the form of "title + abstract" is selected as the basic information for the patent mining in this paper.

The next step is to choose the class label of the patent classification. Based on the functional ontology model established in the previous section, a number of functional groups were selected as the criteria of patent classification. The principle of functional group selection is starting from the group with high degree of abstraction to the group with low degree of abstraction. The functional groups with higher abstraction can be clearly classified for the manual classification and the co-referential relation library associated with them also have abundant key words so they can provide more criteria for the classifier, and the accuracy of result from the classifier is higher. Moreover, the classification training process of the functional class label with higher degree of abstraction has the transitivity and inheritance effects on that of the functional class label with lower degree of abstraction. After finishing the text classification of the class label with higher degree of abstraction, we only need to further define the classification criteria and the classification accuracy to carry out the text classification of the class label with lower degree of abstraction, and don’t need to re-divide the sample labeled texts, select classification method and train the classification model again. We choose 11 primary-level functional groups as the classification criteria to illustrate this classification process. 2243 invention patents were randomly chosen and downloaded from an official patent website. After reading them manually, and based on the functional group classification criteria, the implementation function of each patent was consequently determined, and the classification label of each patent was marked as well; next, the patent texts with most representative property were selected and saved into the patent database as the standard formats. The class label distribution of text set is shown in Table 1.

Table 1 distribution of the class label of the text set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Class label** | **shift** | **regulate** | **absorb** | **combine** | **detect** | **stabilize** |
| Number of texts | 172 | 233 | 67 | 273 | 242 | 157 |
| **Class label** | **import** | **accumulate** | **output** | **produce** | **separate** |  |
| Number of texts | 269 | 260 | 118 | 216 | 236 |  |

## 3.2.2 Patent text classification and feature vector building

In order to ensure the classification accuracy, this paper divided the patent text set into training set and test set in accordance with a certain separation ratio. If the ratio of training set to the whole text set is larger, the function information contained in feature vector is more comprehensive. When the ratio of test set to the whole text set is larger, the final evaluated result from classifier is more reliable. Because the number of whole labeled text sets is not very big, 20% of labeled texts are randomly selected as the test set and the remaining is used as the training set. The useful features are selected by using the feature selection methods, consequently the vector space model is built for the text segmentation processing of the training set. The segmentation tool is the ICTCLAS system **[28]**.

Following the text segmentation, it needs mining the information from the patent texts, namely obtain the feature keywords or called as feature vectors. Two aspects need to be considered in the feature keyword selection. First, the feature word may be the word which frequently appears in most of the texts of a certain function but rarely appear in the texts of other categories. On the other hand, the feature word may also be the word which does not appear frequently but must appear in all texts of a certain function and rarely appear in the texts of other categories. Therefore, during the feature word mining process, it is necessary to consider not only the frequency in which the word appears in the text, but also the representative characteristics of the word. Currently, there are many methods can be used for the feature selection of the text classification. The Text Frequency (DF), Chi square test (CHI), Information Gain (IG), Term Strength (TS) and Mutual Information (MI) are the most typical methods. The DF method can calculate the number of texts which contain a certain word. The advantage of this method is easy implementation and simple algorithm, and the disadvantage is the risk of overestimating the appearing frequency of some words, such as some adverbs or conjunctions. In this paper, we used a more efficient TF-IDF method to replace the traditionally DF method. Because the CHI method only records whether the word appears in the texts, but doesn't count the frequency of the word appearing, therefore this method is actually in favor of the words which appear in low frequency and is called as "low-frequency words defect" phenomenon. The IG method and the MI method are common feature selection methods and also quite suitable for the feature classification of text information. The algorithm of TS method is quite complicated and the relevant literature about this method is rare **[29]**.

In order to obtain a proper method for the features classification and also get a better classification results, the TF-IDF, IG and MI methods were used to select the features of training set and the results will be tested and compared.

(1) TF-IDF (Term Frequency -Inverse Text Frequency weighting) feature selection

The main idea of TF-IDF is that if a word or phrase appears in a text at high frequency but rarely appears in other texts, it means that this word or phrase has a high distinguish ability among categories and is suitable for the classification.

 (1)

where TF is the Term Frequency and IDF is the Inverse Text Frequency **[30]**.

 (2)

where is the times of the word *i* appears in the text *j*, *k* is the text serial number and is the total times of the words appear in the text *j*.

 (3)

where n is the number of all texts, *n* represents the number of texts that contain the term *t*. TF-IDF tends to filter out the common words and reserve the important words **[31]**.

（2）IG (Information Gain) feature selection

The IG method is widely used in machine training. It calculates the information gain to a function by judging whether a word entry appears in an article. If is the set space of the target categories, its information gain from the entry *t* is

 (4)

（3）MI feature selection

MI (Mutual Information of words and categories) reflects the relevancy between entry and category and it is a standard widely used to establish the relevant statistical model of terms, the more mutual information between the entry and function, the higher co-occurrence probability between them.

 (5)

where is the joint probability between word entry *t* and function *c*.  and  are the prior probabilities of word *t* and function *c* respectively **[32]**.

In order to measure the importance of a word in the whole feature selection, two ways were proposed to determine the characteristic value of this word.

 (6)

 (7)

## 3.3 [Vectorization](javascript:void(0);) of patent text

After feature words were selected, their TF value for each patent text in the training set were used as the eigenvalue. On the other hand, by using these eigenvalues as the weight of each feature word, we can present each text in the training set and test set as a vector of all feature words. The format is

class label j, feature 1: feature weight 1, feature 2: feature weight 2:

...... feature *n*: feature weights *n*

Therefore, all text set were transferred into the computer-readable mathematical models and the patent eigenvectors were built as well. In the experiment, for each feature selection method, the corresponding feature words were selected in accordance with the threshold value (*β*) ranging from 0.1 to 1 and the step length is 0.1, so the final results can be compared.

### 3.4 Patent text classification

Following the vectorization process, a proper method should be selected for the training and classification process of the feature vector. Many methods can be used for the training and classification of data **[33]**. The Naive Bayes (NB) has the advantage of low cost of computation and is the most widely used method in the data classification. So this paper chooses it for the self-training in the following classification experiments.

The principle of this method is based on the Bayes' Theorem and the features conditional independence.

Assume the unknown classification text is *Xi* (*x1*, *x2* ..., *xn*) and the function set *c* is (*c1*, *c2* ... *cn*), the probability that text *di* belongs to the function *cj* is

 (8)

Because Bayesian classification has an important assumption on conditional independence, the maximum probability that text *Xi* belongs to the function *cj* is

 (9)

If the maximum probability of texts *di* to a function is the largest, it means that the text *di* belongs to this function.

## *3.5 Patent mining and classification algorithm based on functional* *attributes*

*Based on the patent mining and classification methods proposed in this paper, the patent text mining and automatic classification process based on the functional attributes as the characteristics mainly consists of four processes.*

*(1) The first step is the patent text segmentation processing. The main objective of this step is to segment the 2243 patents abstract texts which had been marked to one of functional category artificially. The word segmentation tools used in this paper is Chinese lexical analysis system ICTCLAS which is developed by Chinese academy of sciences. The specific algorithm is as follows:*

*Aalgorithm 1 Segmentation*

*Input：Stay participle document D*

*Output：Segmentation D’*

*Function Word Segmentation(D)*

*for sentence in D do*

*D’←NULL*

*sentence=ictclas(sentence)*

*sentence=is Stop Word(sentence)*

*D’=D’ U {sentence}*

*end for*

*returnD’*

*\*\*\*\*\*\*\**

*(2) The second step is to establish the feature vector of patents abstract text. According to the different classification requirements and characteristics, three kinds of feature selection methods which including TF-IDF, IG and MI had been established. The specific algorithm is as follows:*

*Aalgorithm 2 TF-IDF feature selection*

*Input：Word set W, Threshold value β*

*Output：Feature set T*

*Function Feature Selection(W)*

*for word w in W do*

*if TFidf(w) >β then*

*T←TU{w}*

*end if*

*end for*

*return T*

*\*\*\*\*\*\*\*\**

*Aalgorithm 3 IG feature selection*

*Input：Word set W, Threshold value β*

*Output：Feature set T*

*Function Feature Selection (W,C)*

*for word w in W do*

*if G(w) >β then*

*T←TU{w}*

*end if*

*end for*

*return T*

*\*\*\*\*\*\*\*\**

*Aalgorithm 4 MI feature selection*

*Input：Word set W, Threshold value β*

*Output：Feature set T*

*Function Feature Selection (W,C)*

*for word w in W do*

*if I(w,c) >β then*

*T←TU{w}*

*end if*

*end for*

*return T*

*\*\*\*\*\*\*\*\**

*According to Feature set of T, the 2243 patent abstract text can be converted to vector set of T'. When the threshold value β change from 0.1 to 1(step ratio 0.1), the different Feature vector set T' of patent abstract text based on different threshold value β can be established.*

*(3) The third step is to get the classes set of the testing samples belong to based on the classification learning results of training samples. According to the different groups of training set: test set between 5:5, 4:6, 3:7,2:8 and 1:9, the 2243 patent abstract text will be* [*divide*](javascript:void(0);)*d into training samples group and testing samples group. The prior probability  which the feature set vector T1 '(t1, t2..., tm) of training sample set belongs to the function class vector C (c1, c2..., c11) can be obtained based on the calculation of Naive Bayes arithmetic. In addition, based on the prior probability ,the probability calculation which the test sample set X (x1, x2...xi ,..xn) belongs to the function class vector* *C (c1, c2..., c11) can be realized. When all the maximum probability values of are summarized together, it can get the test* *sample category set G which can reflect the* [*situation*](javascript:void(0);) *of test sample set belongs to the feature vector. In order to improve the success rate of the classification of test sample, the feature vector of test sample set X (x1, x2...xi ,..xn), will be extend*

*to the relevant text features base on the ontology relationship of text characteristic library. The specific algorithm is as follows:*

*Algorithm 5 Samples Training and Testing*

*Input：Training Samples T1'，Testing Samples X，Feature Vector C*

*Output：Sample Category Set G*

*Function Class Selection (D,C)*

*for t in T1' do*

**

*for xi in X do*

**

**

*G←TU{X}*

*return G*

*\*\*\*\*\*\*\*\**

*(4) According to the sample category set G, it can realize the accuracy of analysis based on the situation of actual classification among the situations of manual labeling. It can define as classification accuracy, r as the recall rate and f as the balance value.*

****** *(10)*

****** *(11)*

*******F(P,R)=2RP/R+P (12)*

*According to the different groups of training set: test set between 5:5, 4:6, 3:7,2:8 and 1:9 and the threshold value β change from 0.1 to 1(step ratio 0.1) , the 2243 patent abstract text had been finished the classification training and testing experiment as follows.*

# 4 Classification experiments and result analysis

### 4.1 Experiment settings

Depending on that the training data set has been accurately categorized, the classification training process can be divided into fully supervised training and semi-supervised training. The fully supervised training is the training from the labeled text set and the semi-supervised training is the training from the both labeled and unlabeled text set. According to the number of classification categories, the classification training process can be divided into multiple classification training and dichotomous classification training. As the name implies, the dichotomous training can only separate two classification categories, namely, yes or no. Because in this experiment, the whole experimental text set has many categories to be classified, but there are not too many labeled texts, if we use fully supervised training for the classification, there are small amount of the corresponding labeled texts for each function and the classification relationship is not clear because of the limited feature vector for each classification function. On the other hand, if we want to increase the number of labeled text sets, it needs to spend a lot of power and time cost. Therefore, the experiment process has been improved from two aspects:

First, based on the labeled text set, the multi-classification training process was transferred into a number of dichotomous training processes. It can improve the efficiency of classification training by increasing the information characteristics of each function. This improved method can save the manpower cost by just increasing the computing cost.

Second, by combining the labeled text set and a large volume of unlabeled text set, we use the semi-supervised training method in the training process so as to keep increasing the information and also to check the information characteristics of each function.

Three different kinds of training methods were chosen in the experiments: multi-classification fully supervised training, multiple dichotomous supervised training and semi-supervised training. In the experiments, the classification accuracies from different feature vector selection methods and different classification training methods were analyzed one by one, so as to gradually improve the classifier and finally reach the best solution.

## 4.2 Fully supervised classification experiment and result analysis

The NB algorithm was used to carry out the fully supervised training experiment and a classifier was established as well. The labeled texts in the test set were automatically classified and the classification results were compared with manually labeled results, so that the performance of the classifier can be evaluated. The corresponding averaged classification accuracies of using different feature selection methods and the classification accuracies for the functions class label are shown in Figure 6.

|  |  |
| --- | --- |
| **Class label** | **accuracy** |
| **shift** | 0.6443 |
| **regulate** | 0.6287 |
| **absorb** | 0.2123 |
| **combine** | 0.6231 |
| **detect** | 0.5645 |
| **stabilize** | 0.5774 |
| **import** | 0.6298 |
| **accumulate** | 0.6756 |
| **output** | 0.6053 |
| **produce** | 0.5941 |
| **separate** | 0.6425 |

Figure 6 The classification accuracies and classification accuracy of functions class label

in fully supervised experiment

The analysis on the experimental results is below.

1) Even though the accuracy of MI [feature](javascript:void(0);) [selection](javascript:void(0);) [method](javascript:void(0);) increased with the increase of values *β*, the overall classification accuracy of MI method is the poorest. The average accuracy rate of MI is less than 45%, and it can be considered that MI is not an effective feature selection method in the supervised experiment. Note that when the values *β* is 1, it means that all the words are feature words and the accuracy of three feature selection methods coincide at this point.

2) For the TF–IDF method and IG method, when *β* is less than 1, the classification results are for the obviously better than that with *β* equal 1. The two methods are effective in both groups and the classification average accuracy up to 55%.

3) *The classification accuracy rate for the function "absorb" is 21.23%. The reason is, for this function, both the number of the labeled texts and the number of the texts in the test set are less than those for other categories. So there is a strong random characteristic on the evaluation of classifier’s performance and a very high possibility of the occurrence on extremely low or high classification accuracy. So, in order to improve the classification accuracy of the classifier, the test and training samples should averagely come from the different functional attribute sets and ensure the diversity of patent as far as possible.*

## 4.3 Multiple dichotomous classification experiment and result analysis

As mentioned above, due to the limited labeled texts, in order to further improve the accuracy of the classification, we replaced the multi-classifier training process by a number of dichotomous training processes in this experiment. For the labeled text set, the labeled texts belonging to a function was called as the positive example set of this function and the remaining texts were called as the negative example set of this function. The classification method and implementation process of NB method in this experiment are similar to that in the previous full supervised classification experiment, and we just need to establish eleven classifiers to correspond to the eleven categories. The corresponding average classification accuracies of using different feature selection methods and the classification accuracies for the functions class label are shown in Figure 7.

|  |  |
| --- | --- |
| **Class label** | **accuracy** |
| **shift** | 0.6510 |
| **regulate** | 0.5243 |
| **absorb** | 0.3212 |
| **combine** | 0.7254 |
| **detect** | 0.7121 |
| **stabilize** | 0.6414 |
| **import** | 0.7512 |
| **accumulate** | 0.6695 |
| **output** | 0.7156 |
| **produce** | 0.6912 |
| **separate** | 0.7410 |

Figure 7 The classification accuracies and classification accuracy of functions class label

in multiple dichotomous experiment

The analysis on the experimental results is below.

1) By comparing the result of each function, it can be found that the classification accuracy rate of each function in the multiple dichotomous experiment is higher than that in the multi-classification experiment. The reason of this improvement is that the original 11 classes have been divided into two categories, and it increased the same function characteristics information and co-reference ontology relationship by increasing the number of labeled texts in each function.

2) The classification accuracy rate of "absorb" function is still only 32.12%. The reason is same as that in the previous experiment. Even though this function has been separated from another 10 categories, the internal characteristics information of this function is still not clear and the internal ontology relationship is not perfect. Therefore, establishing a good ontology relation is the premise to improve the classification accuracy.

## 4.4 Semi-supervised classification experiment and result analysis

We proposed a Semi-supervised training method to further solve the problem of low classification accuracy, which is caused by the limited number of the labeled texts and imperfect generic feature information of the experiment setting. Since there are a large number of un-labeled texts, we classified these un-labeled texts, next based on the classification results, built the text feature vector and co-reference ontology relationship for the following classification. The Semi-supervised classification experimental analysis method is based on EM-NB algorithm.

EM-NB algorithm is an iterative algorithm of maximum likelihood estimation and it is widely used in the condition of incomplete data. EM-NB algorithm has two steps: expectation step (E-step) and maximization step (M-step).The implementation process of the EM-NB algorithm based classification experiment is shown in Figure 8.



Figure 8 Process of classification experiment based on EM-NB algorithm

First the classifier *F* was built based on the labeled text set of *L*. The built classifier was used to classify the un-labeled data set *U* , namely calculate the probability *Pr* (*cj* | *di*) that each non-labeled text *cj* belongs to each function *di*. A new NB classifier was built based on the probability *Pr* (*cj* | *di*) of the set *U*. Because all the texts in the set *U* have the probability identifications *Pr* (*cj* | *di*) of the set *L*, the classifier can use the labeled data set *L* and the non-labeled data set *U* at the same time. This process repeated until the classifier parameters *Pr* (*wt* | *cj*) and *Pr* (*cj*) reached stable conditions or got convergences. *Pr* (*wt* | *cj*) represents the probability that the feature *wt* belongs to the function *cj* and *Pr* (*cj*) represents the probability that each text belongs to the function *cj*.

### 4.5 The experiment process and result analysis

Semi-supervised training process needs both the labeled text set *L* and the non-labeled text set *U*. The *L* text set was built by manually reading and marking. For the creation of the *U* text set, a web crawler program was designed to automatically download 65,000 un-labeled patent texts (in the form of “title + abstract”) from a professional patent database. The separating proportion (training and test) of the text set *L*, the feature selection methods and the thresholds setting are same as those in the previous experiments. The experimental results of corresponding average classification accuracy by using different feature selection methods and the classification accuracy for each function are shown in Figure 9.

|  |  |
| --- | --- |
| **Class label** | **accuracy** |
| **shift** | 0.8542 |
| **regulate** | 0.8201 |
| **absorb** | 0.4215 |
| **combine** | 0.8365 |
| **detect** | 0.8254 |
| **stabilize** | 0.8541 |
| **import** | 0.8145 |
| **accumulate** | 0.8892 |
| **output** | 0.7252 |
| **produce** | 0.7056 |
| **separate** | 0.8275 |

Figure 9 The classification accuracies and classification accuracy of functions class label

in EM-NB algorithm experiment

The analysis on the results from this experimental is below.

1) Compared to the results from previous two experiments, this experiment obtained 85% of the highest average classification accuracy rate and around 81% of the average accuracy. It verified that the semi-supervised classification method is better than the fully supervised classification methods. The reason is that the semi-supervised method can greatly increase the information characteristics and the ontology relationship of the same function.

2) The classification accuracy rates of most categories have been significantly improved except those of the categories "produce" and "output". Two reasons can cause this behavior: first, the categories "produce" and "output" has lower differentiation level in the information characteristics compared with those of other categories; secondly, these two functions have strong semantic similarity and there is a big overlap on the information characteristic individually owned by them and the co-reference relationship between them. Therefore during the function definition, it is necessary to increase the discrimination among the different functions as far as possible, so as to ensure the semantic independence.

3) The classification accuracy rate of function "absorb" is 42.15%. It means that the classification algorithm of semi-supervised can get a good overall accuracy.

**4.6 Further comparative analysis of three experiments**

In overall, all three experiments obtained decent accurate classification results so it proved that the whole experiment design is an efficient attempt to extract the hidden patent information and classify them automatically. We can further study the experiment results and put forward some suggestions on the automatic patent text classification process. (1) For the classification of a large number of texts containing hidden information, the semi-supervised classification method is better than the completely supervised classification method (2) For the function definition of text classification, it should make the functional function independent of each other as far as possible, so as to increase the differentiation among the categories and consequently avoid the poor classification caused by the high similarity of the categories’ definition (3) For the preparation of function experimental text set, it is necessary to evenly separate the manual labeled texts for each function, so that the accuracies of the classifications can reach the stable conditions with more equal convergence ratio .

# 5 Conclusion

This paper studied the automatic mining and classification of the functional information from the patent texts to support the product innovation design. The classification on the primary level functional groups has been carried out and the classifier with the best classification accuracy has been built through a couple of experiments. In general, all three experiments obtained decent accurate classification results so it proved that the whole experiment design is an efficient attempt to extract the hidden patent information and classify them automatically. We can further study the experiment results and put forward some suggestions on the automatic patent text classification process. (1) For the classification of a large number of texts containing hidden information, the semi-supervised classification method is better than the completely supervised classification method (2) For the function definition of text classification, it should make the functional function independent of each other as far as possible, so as to increase the differentiation among the categories and consequently avoid the poor classification caused by the high similarity of the categories’ definition (3) For the preparation of function experimental text set, it is necessary to evenly separate the manual labeled texts for each function, so that the accuracies of the classifications can reach the stable conditions with more equal convergence ratio.

Some works can be further carried out in the future:

* Expand the labeled text set, and improve the performance of the existing classifier.
* Implement the patent text automatic classification on the secondary level and third-level functional groups, so as to establish a comprehensive patent knowledge base.
* Realize the automatically downloading of patent texts from various patent databases and the regular updating of the patent library.
* *In order to further improve the classification accuracy and to optimize the classification algorithm, the research of false rejection rate with this technique plays an important role.*

# Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 51435011).

**Reference**

1. Guo Jieshen and Xiao Guohua. The Study of Patent Information Analysis [J]. Journal of Information, 2008(1):12-15.
2. Hodes A. Patent analysis and formulation using ontology: U.S. Patent Application 10/788,532[P]. 2004-2-27.
3. Yoon B, Park Y. A systematic approach for identifying technology opportunities: Keyword- based morphology analysis [J]. Technological Forecasting and Social Change, 2005, 72(2): 145-160.
4. Tseng Y H. Text mining for patent map analysis [J]. Catalyst, 2005, 5424054(5780101): 6333016.
5. Xiao Guohua. Research and application of patent map [D]. Sichuan University, 2006.
6. Atsushi H U, YUKAWA T. Patent map generation using concept-based vector space model [J]. Working notes of NTCIR-4, Tokyo, 2004: 2-4.
7. Liu Yuqin, Zhu Donghua and Lv Lin. Product technology maturity prediction based on text mining technology [J]. Computer Integrated Manufacturing System, 2008, 14(3): 506-510.
8. Cascini G, Rissone P. PAT-Analyzer: a tool to speed-up patent analyses with a TRIZ perspective[C]Proceedings of the ETRIA world conference–TRIZ future. 2003.
9. Cascini G, Fantechi A, Spinicci E. Natural language processing of patents and technical documentation[M]Document analysis systems VI. Springer Berlin Heidelberg, 2004: 508-520.
10. Liu Renwei. Computer aided product innovation design based on TRIZ system research and development [D]. Shandong University thesis, Jinan, China 2012.
11. Liang yanhong,Tan runhua and Ma jianhong. Patent text classification research of design for product innovation [J]. Computer Integrated Manufacturing System, 2013, 19(2): 382-390
12. Soo V W, Lin S Y, Yang S Y, et al. A cooperative multi-agent platform for invention based on patent document analysis and ontology [J]. Expert Systems with Applications, 2006, 31(4): 766-775.
13. Boyko I. Computer Semantic Search of Inventive Solutions [J]. TRIZ Journal. USA. March. http://www. Triz-journal. Com/archives/2001/03/d, 2001.
14. Xue Chi. Based on the study on the method of patent knowledge of mechanical product innovation design [D]. [Zhejiang](javascript:void(0);) [University](javascript:void(0);) thesis, Hangzhou, China, 2013.
15. Wang Chaoxia, Qiu Qingyin and Feng Peien. The patent knowledge mining for conceptual design method research [J]. Journal of Zhejiang University: Engineering Science, 2008, 42(3): 522-527.
16. Liu Xiang, Li yan and Li Wenqiang. To build support for product innovation design patent knowledge base [J]. Machine Design and Research, 2010, 26(6): 7-11.
17. IWINT Inc. [EB/OL]. (2009-03).http://www.iwint.com.cn/.
18. Invention Machine Corporation. Goldfire innovator [EB/OL]. http://www. Invention -machine. Com/.
19. CREAX. CREAX innovation suite 3.1 software [EB/OL]. [2007-04-02]. http://www.creax. Com/.
20. Qian L, Gero J S. Function-behavior-structure paths and their role in analogy-based design [J]. AI EDAM, 1996, 10(4): 289-312.
21. Tu Jianwei Li yan, and Li Wenqiang. A knowledge retrieval model for product innovation design and implementation [J]. Computer Integrated Manufacturing System 2013, 19(002): 300-308.
22. Wu Chao and Li Yan. For purpose of innovation design based on functional ontology research [J]. [Machinery](javascript:void(0);) [Design](javascript:void(0);) [and](javascript:void(0);) [Manufacture](javascript:void(0);), 2008 (11): 226-228.
23. Xue weimin and Lu yuchang. Text mining technology research [J]. Journal of Beijing Union University: Natural Science Edition, 2006, 19(4): 59-63.
24. Zhang qun. Text mining technology and its application in patent information analysis [J]. Modern Information, 2006, 26(3): 209-210.
25. Fall C J, Törcsvári A, Benzineb K, et al. Automated categorization in the international patent classification[C]//ACM SIGIR Forum. ACM, 2003, 37(1): 10-25.
26. Chen L, Tokuda N, Adachi H. A patent document retrieval system addressing both semantic and syntactic properties[C]//Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20. Association for Computational Linguistics, 2003: 1-6.
27. Cong H, Tong L H. Grouping of TRIZ Inventive Principles to facilitate automatic patent classification [J]. Expert Systems with Applications, 2008, 34(1): 788-795.
28. ICTCLAS [EB/OL]. http://wenku.baidu.com/view/132e30e5097 5f46527d3e1c0.html
29. Liu H, Yu L. toward Integrating Feature Selection Algorithms for Classification and Clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 491-502.
30. Zhang Yufang, Peng Shiming, Lv Jia. Based on TFIDF method improvement and application of text categorization [J]. [Computer](javascript:void(0);) [engineering](javascript:void(0);), 2006, 32(19): 76-78.
31. TF-IDF [EB/OL].http://zh.wikipedia.org/wiki/TF-IDF
32. Yang Y, Ped ersen J O. A comparative study on feature selection in text categorization[C]. ICML. 1997, 97: 412-420.
33. Lewis D D. Naive (Bayes) at forty: The independence assumption in information retrieval [M] Machine training: ECML-98. Springer Berlin Heidelberg, 1998: 4-15.