Journal of Universal Computer Science, vol. 17, no. 14 (2011), 1991-2008
submitted: 6/12/10, accepted: 28/9/11, appeared: 1/10/11 © *J.UCS*

An Intelligent System for Automated Binary Knowledge Document Classification and Content Analysis

Tzu-An Chiang
(National Taipei College of Business, Taipei City, Taiwan
phdchiang@gmail.com)

Chun-Yi Wu
(National Tsing Hua University, Hsinchu, Taiwan
d9534524@oz.nthu.edu.tw)

Charles V. Trappey
(National Chiao Tung University, Hsinchu, Taiwan
trappey@faculty.nctu.edu.tw)

Amy J. C. Trappey[*]
(National Tsing Hua University, Hsinchu, Taiwan
trappey@ie.nthu.edu.tw)

Abstract: Many companies rely on patent engineers to search patent documents and offer recommendations and advice to R&D engineers. Given the increasing number of patent documents filed each year, new means to effectively and efficiently identify and manage technology specific patent documents are required. This research applies a back-propagation artificial neural network (BPANN), a hierarchical ontology technique, and a normalized term frequency (NTF) method to develop an intelligent system for binary knowledge document classification and content analysis. The intelligent system minimizes inappropriate patent document classification and reduces the effort required to search and screen patents for analysis. Finally, this paper uses the design of light emitting diode (LED) lamps as a case study to illustrate and verify the efficiency of automated binary knowledge document classification and content analysis.

Keywords: BPANN, document classification, hierarchical ontology, normalized term frequency
Categories: H.3.1, H.3.3

1 Introduction

Companies are driven by consumers and competitors to introduce new products in a timely matter. However, introducing complex product designs that involve different technical domains of R&D noticeably increases the risk of failure for new products. If a firm searches knowledge documents to find the most advanced technology available

[*]Corresponding author: Professor Amy Trappey, Department of Industrial Engineering, National Tsing Hua University, Hsinchu (300), Taiwan; trappey@ie.nthu.edu.tw; Tel: +886-3-5742651; Fax: +886-3-5722204.

to develop new products, then they avoid repeating research development and shorten the delivery time of new products. According to the report of the European Patent Office [EPO, 07], up to 80% of the newest and most advanced technical knowledge can be accessed through patent databases. Patent documents reveal considerable details of the underlying knowledge. Patent databases are the most prolific and up-to-date source of R&D technological knowledge. Hence, R&D engineers can learn and monitor state-of-the-art technology simply by obtaining and studying the right patents. In addition, when a company wants to apply for a new patent, a patent search better enables the company to understand and evaluate the novelty of their new patent and how to strategically position their R&D efforts.

Intellectual property (IP) rights protection is a fundamental aspect of business. If a company detects patents that infringe upon their IP rights, then appropriate actions are taken to reduce potential business losses. Moreover, when someone files an infringement suit against a company, the company being sued can search for patent documents that support prior arts in order to invalidate the lawsuit. In addition, a firm avoids infringing upon other companies' patents by thoroughly reviewing existing patents. Hence, a company must collect and understand related patents when conducting R&D. The attack and defence of IP rights among companies is a critical issue since it affects strategic business operations and investments. Consequently, patent analysis plays a critical role. Due to the rapid development of technology and the needs for IP protection, the total patent applications surpassed 500,000 in 2010 [USPTO, 10]. Traditional patent analysis, that relies on well-trained patent engineers, patent attorneys and R&D engineers, is hindered by the overload of technology claims. If patent engineers use the international patent classification (IPC) or the US patent classification (UPC) to search patent documents in a specific domain, they often acquire too many documents and cost excessive time to manually screen documents. In addition, patent engineers use keywords to search patent documents and there are no ontological standards for keywords. If patent engineers cannot provide proper keywords, the search result will find too few or too many patents. In order to solve the above mentioned problems, this research employs a back-propagation artificial neural network (BPANN) and a hierarchical ontology to develop an intelligent system for automated binary knowledge document classification and content analysis. The approach assists patent engineers to better search and screen domain-specific and technology-specific patents. In addition, this paper applies a hierarchical ontology and normalized term frequency (NTF) values to create an analytical method for computing the content ratio of each critical technology disclosed in a patent document. The analytical results avoid retrieving irrelevant patent documents. Finally, the intelligent system allows patent engineers to input multiple search conditions to find relevant patents with higher accuracy comparing to the previous approaches.

2 Literature Review

In order to enhance global competitiveness, technology companies have increased their efforts in patent analysis and management. This section organizes research related to patent management, document management, and knowledge management techniques. Almonayyes [2006] incorporates a Naive Bayes classifier with case-based

reasoning to classify Arabic documents into categories appropriate for analysis. The results show that the classification accuracy is improved by integrating the explanation patterns with the Naive Bayes classifier. Brank et al. [2008] used a classifier training and feature selection method for large documents. The feature selection uses the weights obtained from the linear classifier and incorporates the relative importance of features for classification. The results show that feature selection using weights from the linear support vectors classifies documents better than the odds ration and information gain method. Yoon and Park [2004] proposed a network based patent analysis method to show the relationship between domain patents in a virtual network. This approach enables the quantitative evaluation of a patents' degree of importance, degree of newness, and degrees of similarity. Lee et al. [2009] proposed a method for creating and utilizing keywords derived from patent maps and applied these words to facilitate the new technology creation processes. Trappey et al. [2009b] used an ontology combined with TF-IDF (Term Frequency - Inverse Document Frequency) concept clustering to automatically summarize patents for efficient IP sharing and exchanges. Li et al. [2009] proposed a snowball rolling procedure to retrieve significant keywords for patent document searches.

In order to help engineers obtain domain specific patent documents, Lai and Wu [2005] employed patent co-citation analysis to establish a patent classification system. The classification system reveals the relationship of technologies and the evolution of a technology category. This method has been applied to research planning, the creation of patent portfolios, and the formation of licensing strategies. Kim and Choi [2007] use the k-nearest neighbor approach to categorize Japanese patent documents. By considering several factors within a technological field, their research calculates similarity scores between a given set of documents. Hsu et al. [2004] developed a multi-channel legal knowledge service platform. This platform applies key phrase clustering technology to categorize patent documents without displaying the hierarchical structure of the underlying technology. Trappey et al. [2006a] used a back-propagation network to identify a patent document's category based on a hierarchy of international patent classification (IPC) standards. Moench et al. [2003] pointed out that the use of semantic technologies produces high quality search results and decreases the time spent searching for documents. Ontology-based methodologies yield a variety of standardized knowledge representations [Bergmann, 03]. Consequently, ontologies are better suited to express domain-specific knowledge [Li, 03]. Trappey et al. [2006b] developed an ontology-based neural network document categorization system. This system combines the frequency of key phrases and ontology-based neural network to classify patent documents. However, the use of pre-defined categories based on either international patent classification (IPC) or the US patent classification (UPC) is too general to satisfy the need for specific patent analysis. A self-developed classification method is required. Hsu and Trappey [2006] presented a method for technology and knowledge document cluster analysis. Their approach helps companies with patent map analysis, sub-technology clustering, patent document clustering, and technology maturity measurement. Trappey et al. [2009a] proposed an improved methodology using fuzzy ontological document clustering approach using domain ontology as clustering criteria. Moreover, patents can be grouped using a non-exhaustive ontological clustering approach [Trappey, 10] for patents consisting of multiple claims.

3 The NTF-based Binary Document Classification and Content Analysis Methodology

Figure 1 shows the architecture for NTF-based binary document classification and content analysis. The methodology includes four parts, which are (1) creating the key phrase dictionary, (2) extracting domain-specific patent documents, (3) analyzing the content ratio of each critical technology for the domain-specific patents and (4) the semantic search for finding the technology-specific patents. The following subsections elaborate each part of the methodology proposed in the paper.

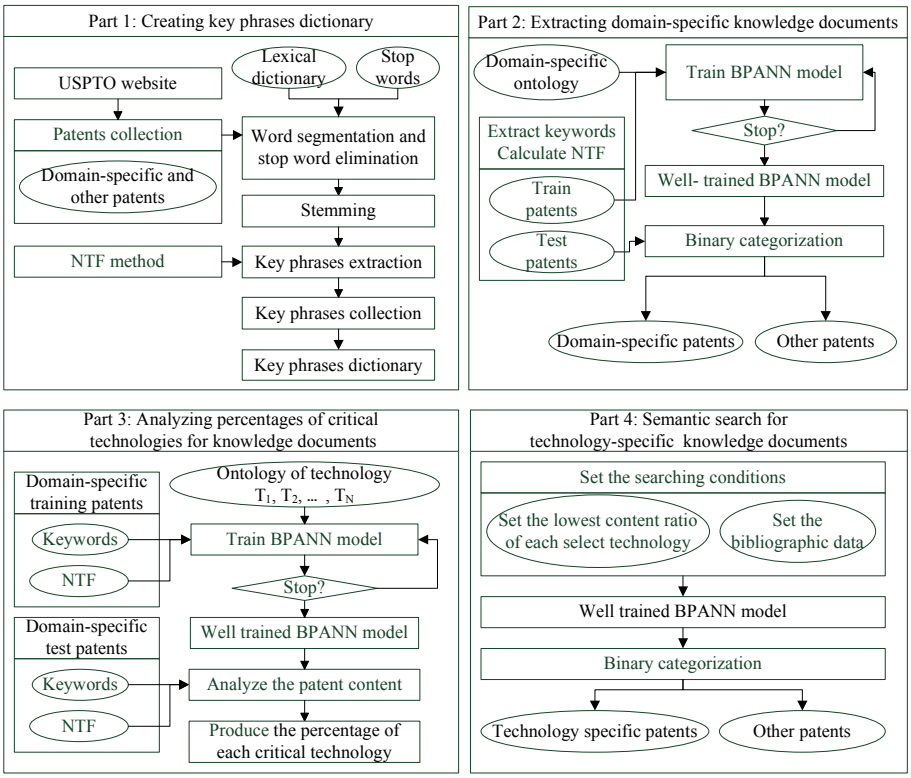

Figure 1: The architecture and methodology for NTF-based binary document classification and content analysis

3.1 Creating key phrase dictionary

This research downloads patent documents from the United States Patent and Trademark Office database. The patent documents include the domain-specific patent documents and other patents. These patent documents are uploaded into the system, which segments the patent documents into single words. Since the stop words contain little information, these words are removed. To recognize the morphology of words, a lexical dictionary [Matsuo, 04] is used to identify the nouns and verbs. Afterward, the

Porter algorithm [Porter, 80] is used to merge words with different tenses and plurality. Consequently, words with the same roots are integrated [Kantrowitz, 00] and reduce the dimension of the word and phrase vector.

The traditional approaches for key phrase extraction include Term Frequency-Inverse Document Frequency (TF-IDF) and Normalized TF (NTF) as shown in Equations 1 and 2 respectively. The TF-IDF approach often eliminates the domain key phrases because when the key phrases appear in each patent document, the value of idf_j is zero. Since the term frequencies are affected by the total word counts (or length) of the document, this paper adopts the NTF method [Salton, 88] to avoid the TF bias due to document length variation. If the values of NTFs exceed the threshold, then the system retrieves these phrases and saves them in the key phrase dictionary.

$$TF - IDF = tf_{jk} \times idf_j$$
$$idf_j = \log_2 \left(\frac{n}{df_j} \right) \tag{1}$$

where

tf_{jk} is the number of the key phrase *j* in the document *k*.

n is the total number of documents in the document set.

df_j is the number of documents containing the key phrase *j* in the document set.

$$NTF_{jk} = tf_{jk} \times \frac{\left(\sum_{s=1}^{n} dn_s \right) / n}{dn_k} \tag{2}$$

where

NTF_{jk} : the NTF of the key phrase *j* in the document *k*,

tf_{jk} : the number of the key phrase *j* that occurs in the document *k*,

dn_s : the total number of key phrases in the document *s*,

n : the total number of documents in a document set.

3.2 Extracting domain specific patent documents

After key phrase extraction, the neural network model for binary document categorization is discussed. An artificial neural network (ANN) is a data modeling tool used to capture and represent complex input/output relationships. An ANN is composed of a large number of highly interconnected processing neurons working in unison to solve problems. Therefore, an ANN can be configured for pattern recognition or document classification. There are two features of an ANN. First, an ANN acquires knowledge through learning. Second, the ANN's knowledge is stored within inter-neuron connection strengths known as synaptic weights.

Chen et al. [2006] proposed a hierarchical neural network to classify documents and discovered that hierarchical neural networks improve the accuracy of categorization. Farkas [1993] proposed a method that represents concept vectors in a semantically meaningful way that combines BPANN with self-organizing maps (SOM) to build efficient and effective automatic document classification systems.

In this research, the BPANN method is used as a multi-layer network to solve non-linear problems. The BPANN learning algorithm is a supervised learning method with the advantage of BPANN is that it adjusts weights between nodes without changing the network structure or the activation functions. In addition, Protégé, an ontology building tool, is used to construct a domain-specific ontology. An ontology is a formally organized knowledge concept, expressed by the relationship between objects, or concepts [Trappey, 09a]. An ontology is a unified structure of concepts which enables the communication or sharing of information [Chen, 05]. In this research, the domain experts define the tree hierarchy of the domain-specific knowledge. The ontology is built by domain experts. After establishing the ontology, the intelligent system employs NTF to identify high frequency phrases in a knowledge document and then import a domain-specific ontology into the system. The intelligent system automatically identifies which high frequency phrases are key phrases. Since a domain specific class includes key phrases there is seldom a 1:1 mapping relationship among high frequent phrases, ontological nodes and domain specific classes create the basic structure. In addition, because the intelligent system is scalable and flexible, users define domains based on specific requirements. Hence, a domain can contain the critical technologies of a product, a module, or merely a component. The highest layer of a hierarchical ontology tree represents the domain concept. The second layer shows the sub-domains of a specific and critical technology. The ontology is then used to assist the intelligent system to extract key phrases and identify phrase sub-domains. The following section describes the means for extracting domain specific knowledge documents where binary document classification is based on the ontology. If the ontology is modified, the binary patent document classification will also change.

The learning algorithm of BPANN is expressed as follows. The equation of hidden layer values is shown in Equation 3. The equation translates the weights from the input layer to the hidden layer using the activation function of the weighted NTF input values. The activation function is a sigmoid function as shown in Equation 4.

$$net_j^h = f\left(\sum_i w_{ij}^h X_i\right) \tag{3}$$

w_{ij}^h are the weight from the input layer to the hidden layer,

X_i is the NTF value of the input *i*.

$$f(x) = \frac{1}{1+e^{-x}} \tag{4}$$

The output values, from the hidden layer to the output layer, are calculated using Equation 5.

The outputs represent the domain-specific classes of patent documents. The errors of all output layer nodes are defined by Equation 6.

$$O_k = g(net_k^o) = g\left(\sum_j w_{jk}^o f\left(\sum_i w_{ij}^h X_i\right)\right) = g\left(\sum_j w_{jk}^o H_j\right) \tag{5}$$

where w_{jk}^o is the weight from the hidden layer to the output layer.

$$E = \frac{1}{2}\sum_k (T_k - O_k)^2 \tag{6}$$

For the backward pass, the methodology adjusts weights using error correction rules to adjust the expected output. The gradient value is inferred using Equation 7.

$$\Delta w_{ij}^o = \eta\, \delta_k^o\, H_j \tag{7}$$

where η is the learning rate and $\delta_k^o = (T_k - O_k)\, g'\left(net_k^o\right)$.

Keywords are extracted from training and test patent documents. The BPANN model uses the NTF values as input items to automatically classify the knowledge documents according to the binary algorithm.

3.3 Analyzing percentages of critical technologies for patent documents

The hierarchical ontology maps the keywords and their sub-ontologies. Hence, the total number of key phrases in a knowledge document and the frequencies of the key phrases for specific sub-domains are calculated. Afterwards, the well trained BPANN model computes the content ratio of each sub-domain. Table 1 displays the analytical result representing the content ratio of each sub-domain within each patent document.

Patents	**Ratio of Sub-Domain 1**	**Ratio of Sub-Domain 2**	**…**	**Ratio of Sub-Domain n**
No. 1	P(No. 1, 1)	P(No. 1, 2)	…	P(No. 1, n)
No. 2	P(No. 2, 1)	P(No. 2, 2)	…	P(No. 2, n)
…	…	…	…	…
No. n	P(No. n, 1)	P(No. n, 2)	…	P(No. n, n)

Table 1: The content ratio of patent documents

3.4 Semantic search for technology specific patent documents

Figure 2 depicts a general process to support users in finding critical knowledge documents. First, users use the NTF-based content analysis module to analyze a knowledge document. When the critical technologies in a knowledge document are identified, users compute the content percentage of each critical technology. According to these analytical results users set the search conditions including the content ratio threshold of each critical technology. The intelligent system allows users to set the search conditions related to the bibliographic data of knowledge documents. Next, the user oriented semantic search function of the intelligent system performs binary categorization to retrieve the documents which best match the engineers' needs. If the search conditions are modified, the results of the binary patent document classification also change. If engineers desire to better understand the current state of technology, the search is modified using the computer interface. The intelligent system finds and prioritizes the related knowledge documents using the content percentages. The content ratio of the critical technologies in a knowledge document reveals key attributes and claims about the main invention.

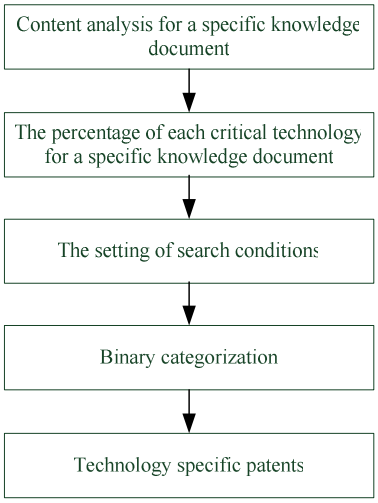

Figure 2: A general process of finding critical knowledge documents

4 Case Study

The case study relates to patent covering claims for light emitting diode (LED) lamps. The data and analysis demonstrate the comprehensive capabilities and practical contributions for automated binary knowledge document classification and content analysis. The case study can be divided into two parts. The first part describes how to create the intelligent system. The second part uses a real case as an example to elaborate the effectiveness and practical contributions of the intelligent system. With regard to the first part, when new product engineers complete the LED lamp development process, the intellectual property (IP) department searches for related

patents to avoid filing patents potentially infringing upon the prior IPs claimed by other inventors. For example, noted that the LED lamp may infringe on the target patent CN 200610066778.8. A prior art patent can invalid the target patent based on the US Patent Laws [35USC103, 07]♦. Thus, if a prior art patent was found, the target patent holder cannot gain business benefits from the patent.

The database of United States Patent and Trademark Office (USPTO) holds over 7,000,000 patents. The purpose of the case study is to discover the prior art of the target patent (CN 200610066778.8). The application date of the target patent is 2006/4 and its priority date is 2005/7. Therefore, the search conditions are set before the priority date (19950701->20050701). The key phrases include light emitting diodes techniques, cooling or heating arrangements of lighting devices, and the arrangement of electric circuit elements in or on lighting devices. Total of 90 LED patents (60 for training and 30 for testing) are collected corresponding to the search conditions given by the domain experts. Further, additional 80 non-LED patents are used (60 for training and 20 for testing) for building the BPANN intelligent system to automatically perform binary knowledge document classification. First, the intelligent system uses the predefined lexical dictionary and stop words to extract the key phrases and calculate the NTF values of all phrases. Then, the intelligent system provides the top 100 NTF values of key phrases used to build the domain ontology. Some of the extracted key phrases are shown in Table 2. Based on the description of claims, the case study analyzes the relationship between key phrases and the LED patent claims. The next step is to build the LED lamp ontology (shown in Figure 3) to train the BPANN model.

Lamp Assembly	Wire Winding Box	Light Emitting Diodes	Reflector
pressure discharge lamp	lampshade	conductive	light source
lamp device	light bulb	heat sink	circuit
substrate	temperature	lamp seat	lamp base

Table 2: Partial key phrases derived from patents

♦ [35USC103, 07] … *(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made…*

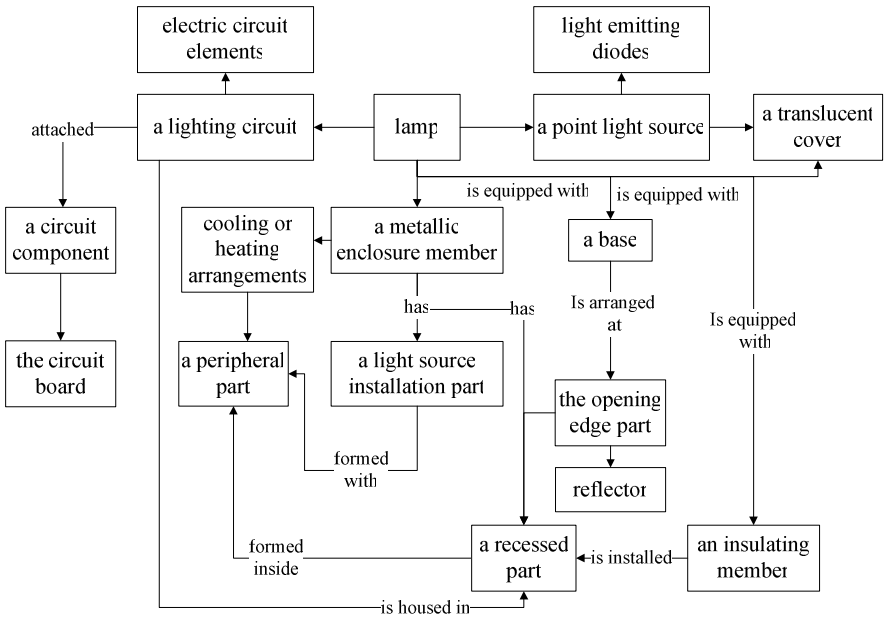

Figure 3: The LED lamp ontology

The intelligent system inputs the NTF values of key phrases to train the BPANN model-based intelligent system. The case study prepares 120 (60 LED and 60 non-LED) training patents and 50 (30 LED and 20 non-LED) test patents. With regards to the size of a training set, this research requires that the accuracy is over 90% for the classification result. The binary classification network of the intelligent system includes the input layer, the hidden layer and the output layer. The number of the input nodes is 50 and the number of the hidden nodes is 26. The transfer function uses a sigmoid function as shown in Equation 4. The training parameters include 1000 iterations, a learning rate of 0.2 and a momentum of 0.8. The performance of the binary categorization network is shown in Table 3.

Number of Training Patents	LED Patents: 60 Non-LED Patents: 60
Training - Root Mean Square Error (RMSE) [Trappey, et al., 2006a]	6.8%
Training - Set Accuracy [Trappey, et al., 06b]	93.5%
Number of Test Patents	LED Patents: 30 Non-LED Patents: 20
Test - RMSE	13.5%
Test - Set Accuracy	83%

Table 3: The results of the technology specific model

60 training patents and 30 test patents for the LED domain are used to build the analytical model defining the content percentage of critical LED lamp technologies. The critical technologies include circuit element configuration, cooling apparatus, and light emitting diodes. The key phrases in the patent document are extracted using a predefined ontology. Then, the extracted key phrases are used to train the content percentage of each critical technology. The numbers of hidden nodes and output nodes of the model are 27 and 3 respectively. The transfer function is the sigmoid formula and the training parameters include 8,000 iterations, a learning rate of 0.15, and a momentum of 0.85. The performance of the analytical model is shown in Table 4. The well trained model enables the IP department to analyze the content ratio of prior art patents effectively and efficiently. The IP engineers use the results of the binary knowledge document classification and content analysis to evaluate the prior art patents to avoid patent infringement lawsuits. Figure 4 shows the target patent invention used in the case study.

Number of Patents	Subclass: Cooling Apparatus: 36 Subclass: Light Emitting Diodes: 15 Subclass: Circuit Element Configuration: 39
Training - RMSE	3.1%
Training - Set Accuracy	96.5%
Test - RMSE	19.6%
Test - Set Accuracy	87.4%

Table 4: The analytical results derived using content percentage of each critical technology in the LED domain

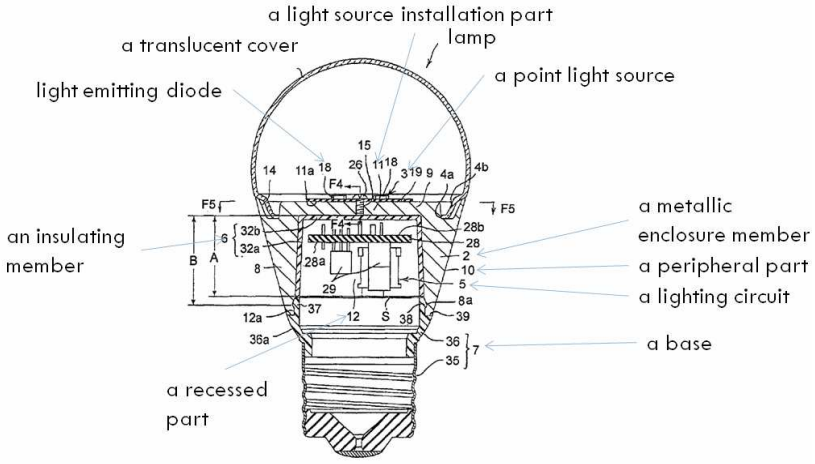

Figure 4: A drawing of the invention related to the target patent

This research uses an intelligent system to acquire the ratio of critical technology test patents. Table 5 shows the content percentage of each critical technology of 3 LED domain patents. Patent No. US5857767, among the test patents, best describes thermal management systems for LED arrays. The content levels for cooling apparatus, LED, and circuit element configurations are 51.7%, 33.5%, and 14.8% respectively. The subject invention relates to LEDs mounted in an array on a circuit to avoid damaging temperatures as shown in Figure 5. Patent No. US6948829 focuses on the design of LED bulbs. The present invention relates to light bulbs that use LED's as light emission elements (Figure 6). The claim of Patent No. US6942365 describes the housing of a high intensity LED lamp that also provides electrical connectivity, heat dissipation, and a reflector device in a compact and integrated package (Figure 7). Using the above explanations, the analytical model of content percentages showcase the benefit of extracting and expressing important information derived from patent documents.

Patent No.	Critical Technologies		
	Cooling Apparatus	Light Emitting Diodes	Circuit Element Configuration
US5857767	51.7%	33.5%	14.8%
US6948829	25.1%	62.4%	12.5%
US6942365	23.6%	12.7%	63.7%

Table 5: Percentages of critical technologies for three LED domain patents

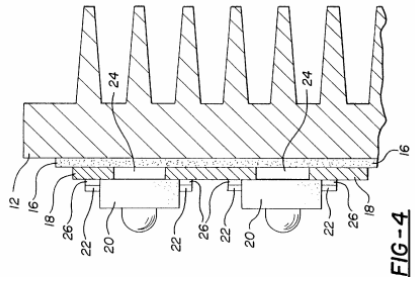

Figure 5: The LED cooling apparatus for Patent No. US5857767

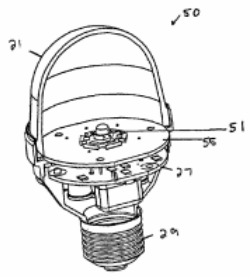

Figure 6: The light emitting diodes for Patent No. US6948829

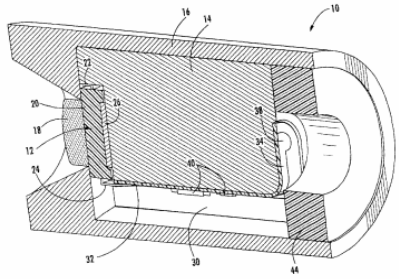

Figure 7: The circuit element configuration for Patent No. US6942365

After training the intelligent system, IP engineers search for patent documents with overlapping prior art claims. First, the IP engineer uses the application date of a target patent and the key phrases of the device to search for relevant patents. Afterwards, the binary knowledge document categorization model classifies 170 relevant patents and identifies 90 patents related to the LED domain. Then, the analytical model computes the content percentage for the 90 LED patents. Finally, the IP engineer selects the critical technologies and sets the lowest content ratio for each critical technology in a patent document. In this case, the IP engineer set the content levels for the cooling apparatus and LED at 40% and 20% respectively. The intelligent system identifies US patent US7165866, which relates to a light with an enhanced and heat dissipating bulb (Figure 8). The content analysis of patent US7165866, shown in Table 6, identifies this patent having similar technical content and can potentially being a prior art of the target patent (CN 200610066778.8). Thus, further investigation of US7165866 is needed when the infringement legal matter of the target patent occurred.

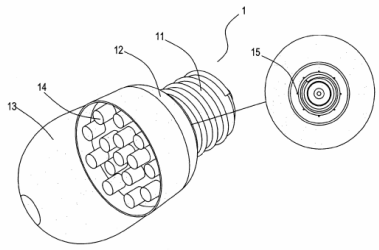

Figure 8: The potential prior art patent US7165866

Technology-Specific	Cooling Apparatus	Light Emitting Diodes
Lowest Constraint	40%	20%
US7165866	**54%**	**25.6%**

Table 6: The result of other technology specific patents

The proposed NTF-based binary document classification is evaluated by comparing results with similar tests conducted using the legal knowledge management (LKM) system [Hsu, 2004], the TF-based ANN classification system [Trappey, et al., 2006a] and the fuzzy ART patent analysis system [Trappey, et al., 2010]. This paper proposes three evaluation indexes, including the degree of accuracy for the domain-specific patent categorization, the degree of accuracy for the technology-specific patent categorization satisfying the searching conditions, and the total average deviation between the threshold conditions and the content ratios of patents selected by the system as shown in Equation 8, 9, 10 and 11. Because other systems cannot provide the analytical function of the content percentage of each critical technology, we use the results of domain-specific patent categorization to calculate the degree of accuracy and the total average deviation.

$$DA = \frac{A}{B} \tag{8}$$

where DA: the degree of accuracy for the domain-specific patent categorization;
A: the number of domain-specific patents selected by the system;
B: the correct number of domain-specific patents.

$$TA = \frac{C}{D} \tag{9}$$

where TA: the degree of accuracy for the technology-specific patent categorization satisfying the searching conditions;
C: the number of technology-specific patents selected by the system that satisfies the searching conditions;

D: the correct number of technology-specific patents categorization that satisfies the searching conditions.

$$TP_i = 1 - \frac{\sum_{j=1}^{n} DP_j^i}{n} \tag{10}$$

where TP_i : the average deviation between the threshold condition and the content ratio of each patent selected by the system for the specific technology *i*;
DP_j^i : the absolute deviation between the threshold condition and the content ratio of the *j*th patent selected by the system for the specific technology *i*;
n: the number of patents selected by the system.

$$AP = \frac{\sum_{i=1}^{I} TP_i}{I} \tag{11}$$

where AP : the total average deviation between the threshold conditions and the content ratios of patents selected by the system.

The results of performance evaluation of the four systems are shown in Table 7. From the analytical results, we can understand that the proposed intelligent system has the best accuracy for the domain-specific patent categorization. In addition, the intelligent system has the ability of content analysis of patents. Therefore, the degree of accuracy for the technology-specific patent categorization and the total average deviation of the proposed intelligent system obviously outperform other systems as shown in Figure 9.

Systems	Evaluation Indexes		
	DA	*TA*	*AP*
LKM [Hsu, 2004]	86%	70%	81%
TF-based ANN [Trappey, et al., 2006a]	91%	74%	74%
Fuzzy ART [Trappey, et al., 2010]	83%	67%	87%
NTF-BPANN (this research)	**93%**	**100%**	**100%**

Table 7: The results of performance evaluation of the four systems

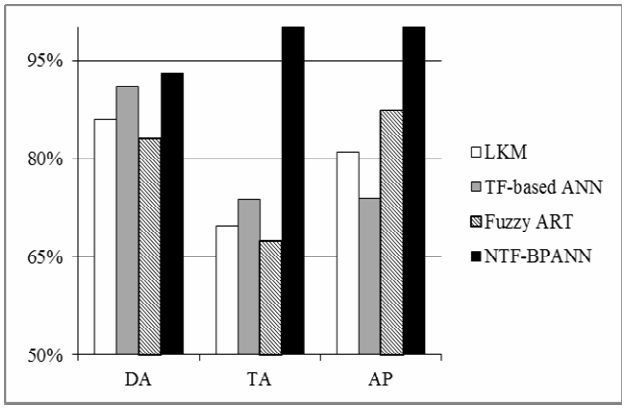

Figure 9: The evaluation results of the indexes for different patent classification approaches

5 Conclusions

While several academic papers focus on patent classification and search, the majority of papers assume that domain specific patent documents are obtained and then automatically categorized according to their IPC codes. The IPC system often fails to satisfy real world requirements for more specific technological classifications. In addition, using an IPC system, patent engineers frequently retrieve too many patent documents unrelated to their search objectives. If IP engineers establish more accurate search conditions, the workload to read, review, and classify emerging technologies is significantly reduced. This paper applies BPANN and an ontology map to develop an intelligent system for automated binary knowledge document classification and content analysis. Patent engineers and R&D engineers identify domain specific patent documents and determine the content percentages of critical technologies in a patent document. By setting the lowest threshold level of each critical technology, engineers are better enabled to efficiently and effectively find patents of interest. Finally, this paper demonstrates the methodology and intelligent system using LED patent analysis case. LED patent screening, sub-technology patent classification, and prior art patent identification are conducted to verify the superior performance of automated binary knowledge document classification and content analysis.

Acknowledgement

This research was partially supported by Taiwan National Science Council research grants. The authors would like to thank the reviewers for their valuable comments and suggestions for the paper revision.

References

[35USC103, 07] Section 103 - Conditions for Patentability; Non-obvious Subject Matter, Chapter 10 – Patentability of Inventions, Part II – Patentability of Inventions and Grant of Patents, Title 35 of the United States Code. Laws in effect as of January 3,: http:// www.gpoaccess.gov, Cite: 35USC103, 31-32 (2007).

[Almonayyes, 06] Almonayyes, A.: "Multiple Explanations Driven Naive Bayes Classifier", Journal of Universal Computer Science, 12, 2, 127-139 (2006).

[Bergmann, 03] Bergmann, R., Schaaf, M.: "Structural Case-Based Reasoning and Ontology-Based Knowledge Management: A Perfect Match?" Journal of Universal Computer Science, 9, 7, 608-626 (2003).

[Brank, 08] Brank, J., Mladenić, D., Grobelnik, M., Milić-Frayling, N.: "Feature Selection for the Classification of Large Document Collections", Journal of Universal Computer Science, 14, 10, 1562-1596 (2008).

[Chen, 06] Chen, Z., Ni, C., Murphey, Y.L.: "Neural Network Approaches for Text Document Categorization", International Joint Conference on Neural Networks, Vancouver, 1054-1060 (2006).

[Chen, 05] Chen, E., Wu, G.: "An Ontology Learning Method Enhanced by Frame Semantics", The Seventh IEEE International Symposium on Multimedia, California, 374-382 (2005).

[EPO, 07] European Patent Office, http://www.epo.org/patents.html (2007).

[Farkas, 93] Farkas, J.: "Neural Networks and Document Classification", Canadian Conference on Electrical and Computer Engineering, Vancouver, 1-4 (1993).

[Hsu, 04] Hsu, F. -C., Trappey, A. J. C., Hou, J. -L., Trappey, C. V., Liu, S. -J.: "Develop a Multi-Channel Legal Knowledge Service Center with Knowledge Mining Capability", International Journal of Electronic Business Management, 2, 2, 92-99 (2004).

[Hsu, 06] Hsu, F.-C., Trappey, A.J.C.: "Technology and Knowledge Document Cluster Analysis for Enterprise R&D Strategic Planning", International Journal of Technology Management, 36, 4, 336-353 (2006).

[Kantrowitz, 00] Kantrowitz, M., Mohit, B., Mittal, V.: "Stemming and its Effects on TFIDF Ranking", The Annual International ACM SIGIR'00 Conference on Research and Development in Information Retrieval, Athens, 357-359 (2000).

[Kim, 07] Kim, J. -H., Choi, K. -S.: "Patent Document Categorization Based on Semantic Structural Information", Information Processing and Management, 43, 5, 1200-1215 (2007).

[Lee, 09] Lee, S., Yoon, B., Park, Y.: "An Approach to Discovering New Technology Opportunities: Keyword-Based Patent Map Approach", Technovation, 29, 6-7, 481-497 (2009).

[Lai, 05] Lai, K. -K., Wu, S. -J.: "Using the Patent Co-Citation Approach to Establish a New Patent Classification System", Information Processing and Management, 41, 2, 313-330 (2005).

[Li, 03] Li, S. -T., Hsieh, H. -C.: "Managing Operation Knowledge for the Metal Industry", Journal of Universal Computer Science, 9, 6, 472-480 (2003).

[Li, 09] Li, Y. -R., Wang, L. -H., Hong, C. -F.: "Extracting the Significant-Rare Keywords for Patent Analysis", Expert Systems with Applications, 36, 5200-5204 (2009).

[Matsuo, 04] Matsuo, Y., Ishizuka, M.: "Keyword Extraction from a Single Document Using Word Co-Occurrence Statistical Information", International Journal on Artificial Intelligence Tools, 13, 1, 157-169 (2004).

[Moench, 03] Moench, E., Ullrich, M., Schnurr, H.-P., Angele, J.: "SemanticMiner-Ontology-Based Knowledge Retrieval", Journal of Universal Computer Science, 9, 7, 682-696 (2003).

[Porter, 80] Porter, M.F.: "An Algorithm for Suffix Stripping", Program, 14, 3, 130-137 (1980).

[Protégé] Protégé, http://protege.stanford.edu/

[Salton, 88] Salton, G., Buckley, C.: "Term-Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, 24, 5, 513-523 (1988).

[Trappey, 06a] Trappey, A.J.C., Hsu, F. C., Trappey, C.V., Lin, C.I.: "Development of a Patent Document Classification and Search Platform Using a Back-Propagation Network", Expert Systems with Applications, 31, 4, 755– 765 (2006).

[Trappey, 06b] Trappey, A.J.C., Trappey, C.V., Hsieh, E.C.H.: "Automatic Categorization of Patent Documents for R&D Knowledge Self-Organization", Journal of Management, 23, 4, 413-424 (2006).

[Trappey, 09a] Trappey, A.J.C., Trappey, C.V., Hsu, F. -C., Hsiao, D. W.: "A Fuzzy Ontological Knowledge Document Clustering Methodology", IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 39, 3, 806-814 (2009).

[Trappey, 09b] Trappey, A.J.C., Trappey, C.V., Wu, C.Y.: "Automatic Patent Document Summarization for Collaborative Knowledge Systems and Services", Journal of Systems Science and Systems Engineering, 18, 1, 71-94 (2009).

[Trappey, 10] Trappey, C.V., Trappey, A.J.C., Wu, C.Y.: "Clustering Patents Using Non-Exhaustive Overlaps", Journal of Systems Science and Systems Engineering, 19, 2, 162-181 (2010).

[USPTO, 08] USPTO, United States Patent and Trademark Office, http://www.uspto.gov/patents/index.jsp (2008).

[USPTO, 10] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm

[Yoon, 04] Yoon, B., Park, Y.: "A Text-Mining-Based Patent Network: Analysis Tool for High-Technology Trend", Journal of High Technology Management Research, 15, 37-50 (2004).