Development of a patent document classification and search platform using a back-propagation network

Amy J.C. Trappey a,*, Fu-Chiang Hsu a, Charles V. Trappey b, Chia-I. Lin c

a Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu 300, Taiwan
b Faculty of Business, Queensland University of Technology, Australia
c Center for Aerospace and Systems Technology, Industrial Technology Research Institute, Hsinchu 300, Taiwan

Abstract

In order to process large numbers of explicit knowledge documents such as patents in an organized manner, automatic document categorization and search are required. In this paper, we develop a document classification and search methodology based on neural network technology that helps companies manage patent documents more effectively. The classification process begins by extracting key phrases from the document set by means of automatic text processing and determining the significance of key phrases according to their frequency in text. In order to maintain a manageable number of independent key phrases, correlation analysis is applied to compute the similarities between key phrases. Phrases with higher correlations are synthesized into a smaller set of phrases. Finally, the back-propagation network model is adopted as a classifier. The target output identifies a patent document’s category based on a hierarchical classification scheme, in this case, the international patent classification (IPC) standard. The methodology is tested using patents related to the design of power hand-tools. Related patents are automatically classified using pre-trained neural network models. In the prototype system, two modules are used for patent document management. The automatic classification module helps the user classify patent documents and the search module helps users find relevant and related patent documents. The result shows an improvement in document classification and identification over previously published methods of patent document management.

Keywords: Knowledge document management; Document classification; Patent search; Neural networks

1. Introduction

With the rapid development of information technology, the number of electronic documents and the digital content of documents exceed the capacity of manual control and management. People are increasingly required to handle wide ranges of information from multiple sources. As a result, knowledge management systems are implemented by enterprises and organizations to manage their information and knowledge more effectively. Knowledge management includes sorting useful knowledge from information, storing knowledge in good order, and finding knowledge in an existing knowledge base (Turban & Aronson, 2001). In this research, we focus on explicit knowledge management, i.e., management of well-structured documents such as patent documents (called “patents” in brief). Patents provide exclusive rights and legal protection for patent inventors. In addition, patents play an important role in the advancement and diffusion of technology. The objective of this research is to develop an effective methodology to automatically classify and identify patent documents. Furthermore, a prototype system is implemented and tested using hand tools patents sourced from the World Intellectual Property Organization (WIPO) database for scenario demonstration.

There have been many research efforts devoted to automatic document classification. Some of the classification
methodologies are difficult to implement, and others are neither efficient nor effective, requiring developers of knowledge management systems to expend considerable resources testing and evaluating algorithms. The purpose of this research is to develop a document classification method based on neural networks and benchmark the performance against published standards. Through the implementation of a document classification module and a document search module, a prototype patent document management system is created. The patent document management system automates the classification of patent documents and improves the search for documents.

The automatic document classification methodology is described in the following steps: First, significant terms are abstracted from patent documents and are used to build a key phrase database. Second, the similarities between phrases are computed and depicted in a correlation matrix in order to synthesize phrases into a smaller set representing key concepts within the patent domain. After the steps of key phrase extraction and synthesis, a consolidated set of key phrases are treated as inputs of the back-propagation network model. The neural network model is trained using key phrases and the frequency of key phrases from the sample documents. The trained model is assessed until it reaches a satisfactory level of accuracy. After the network model is trained, the final step is to use the model for automated patent document classification and search.

2. Literature survey

This section reviews the relevant topics including knowledge and e-document management, document categorization, clustering methodologies and related patent analysis research.

2.1. Knowledge and e-document management

According to Turban and Aronson (2001), knowledge management is a process that helps organizations identify, select, organize, disseminate, and transfer knowledge. The previous research indicates expertise is a part of the organizational memory that typically resides within the organization in an unstructured manner. Document management is the management of various documents of an organization. Effective document management is a great help to the implementation of knowledge management (Hou & Chan, 2003).

Traditional organization documents vary in form and are referred to by a variety of common business terms such as contracts, reports, manuals, memos, patents, drawing, invoices, ledgers, spreadsheets, and newsletters. With the development of information technology, the number of electronic documents can match or exceed the numbers of paper documents unless the business processes are redefined. Thus, electronic document management is an important issue for managers. Electronic documents are more beneficial than paper-based documents in three ways (Meier & Sprague, 1996):

1. Electronic documents with increased media richness helps users quickly understand document contents. Communication is also enhanced by the speed and timing of electronic transmission.
2. Electronic document management can be integrated with workflow processes.
3. Electronic document management leads to an expanded organizational memory with better access.

2.2. Document categorization methodology

Document categorization is the process of assigning a document into one or more pre-defined document classes. On the other hand, document clustering splits many documents into groups according to the similarity between documents. The similarity is measured by evaluating key representing attributes and features among documents. Both document categorization and document clustering extract and use features of the document for group assignment. The main distinction between categorization and clustering is that document categorization compares document features and pre-defined class features and selects the most suitable document class. Document clustering divides a set of documents into groups without using pre-defined classes (Antonie & Zaiane, 2002; Deng & Wu, 2001).

The traditional means of document categorization is to classify documents by experts within a specific domain. Since experts are costly and vary in capabilities, the result of classification varies and can be unreliable. For these reasons, automatic document categorization has become an important research area. There are several document classifiers proposed by previous researchers including the vector space model, the $k$-nearest-neighbor approach, naive Bayes analysis, neural networks, and genetic algorithms. These approaches are reviewed in the following subsections.

2.2.1. Vector space model (VSM)

The VSM has been widely used in traditional information retrieval (Farkas, 1994; Grossman, Frieder, Holmes, & Roberts, 1997) and for automatic document categorization. There are three key steps where terms are first extracted from the document text, then the weights of the indexed terms are derived to improve the document retrieval accuracy, and then the documents are ranked with respect to a similarity measure (Raghavan & Wong, 1986). VSM is a multi-dimensional vector where each feature of a document is a dimension. For instance, term frequency (TF) and inverted document frequency (IDF) are two features of a text document. After the vector of a text document is derived, a cosine function is applied to measure the similarity between two documents (Tam, Santos, & Setiono, 2002). For example, the set $X = (x_1, x_2, \ldots, x_n)$
represents the VSM of document $X$, where $x_i$ represents the $i$th feature of document $X$. Likewise, $Y = (y_1, y_2, \ldots, y_n)$ represents the VSM of document $Y$ with the similarity between $X$ and $Y$ calculated as

$$
\cos(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}
$$

(1)

2.2.2. k-Nearest-neighbor ($k$NN)

$k$NN is an algorithm that uses pre-trained documents to classify new documents based on a similarity measure. $k$NN uses the distance between two document vectors as a measure for their similarity. The similarity $z(x, c_i)$ is used as the confidence score to indicate $x$ belonging to the particular category $c_i$ (Tam et al., 2002) and is calculated using

$$
z(x, c_i) = \frac{\sum_{d_j \in kNN} \text{sim}(x, d_j) y(d_j, c_i)}{\sum_{d_j \in kNN} \text{sim}(x, d_j)} - b_i
$$

(2)

where $\text{sim}(x, d_j)$ is the similarity between tested document $x$ and trained document $d_j$ which is calculated using Euclidean distance or the cosine value between two document vectors. $y(d_j, c_i)$ is 1 (or 0) when the trained document $d_j$ (not) belongs (or does not belong) to $c_i$. Finally, $b_i$ is the threshold of classifying a document to category $c_i$. Chiang and Chen (2001) apply $k$NN to news document categorization with good results. Further, Lam and Han (2003) develop a modified method of $k$NN called Generalized Instance Sets (GIS) to improve the performance of traditional $k$NN algorithm.

2.2.3. Naïve Bayes

Naïve Bayes classification does not require many observations for all possible combinations of the variables. Rather, the variables are assumed to be independent to each other. In other words, naïve Bayes classifiers assume that the influence of a variable is independent of other variables for a given class, an assumption called class conditional independence. This algorithm uses the joint probability of document features to calculate the probability that a new document belongs to a specific class:

$$
P(c_i | d^\prime) = \frac{P(d^\prime | c_i) \cdot P(c_i)}{\sum_{c_j \in C} P(d^\prime | c_j) \cdot P(c_j)}
$$

(3)

$P(c_i)$ is the probability of a given document belonging to class $c_i$ and $P(d^\prime | c_i)$ is the conditional probability of document $d^\prime$ belonging to a specific class $c_i$. The probability of document $d^\prime$ belonging to class $c_i$ can be derived using the following equation:

$$
P(d^\prime | c_i) = \prod_{j=1}^{|d^\prime|} P(w_j | c_i)
$$

(4)


2.2.4. Genetic algorithm

The genetic algorithm (GA) was developed by Holland (1992) and works well on mixed (continuous and discrete) and combinatorial problems. Three operational components (selection, crossover and mutation) are used to generate various models. The GA can provide an optimal solution but is possible for the algorithm to stick at local optimal solutions. Another shortcoming of GA is the computational requirements of the algorithm, which may not be a concern when there is strong computing power.

2.2.5. Artificial neural network

Artificial neural network (ANN) is an information processing method inspired by biological nervous systems. An ANN uses interconnected processing nodes computationally linked to solve problems. Neural networks are frequently used for pattern recognition and document classification and learn by using training data to adjust the weights between connecting nodes. Some research has applied artificial neural networks to text classification. Farkas (1994) provides a thesaurus and neural network to generate document clusters. Massey (2003) uses ART1, or Adaptive Resonance Theory, to cluster documents. Selamat and Omatu (2004) propose a Web-page classification system that uses a neural network with inputs gained by principal component analysis and class profile-based features that contain the most regular words in each class. Principal components analysis is a statistical method that transforms a number of correlated variables into a smaller number of uncorrelated variables whereas class profile analysis utilizes the categorical profile built by domain experts. Some researchers propose a term frequency method to select the feature vectors for neural network document categorization (Lam & Han, 2003; Lee, 2003). Karras and Mertzios (2002) propose a feature extraction method based on semantics and co-occurrence analysis for neural network document classification. Mostafa and Lam (2000) propose a multi-layer feed-forward neural network to classify medical documents. Trappey, Lin, and Wang (2005) propose a term frequency method using neural network categorization for electronic document management. Table 1 shows the advantage and shortcomings of the commonly used categorization models.

2.3. Patent content analysis

In the last few years, research concerning the analysis of patents and the automatic classification of patents has grown. Patent analysis technologies include patent bibliometric data analysis, patent citation analysis, patent statistical analysis and automated patent cluster and classification using machine learning techniques. Patent bibliometric analysis was proposed by Narin (1994) using
and search is presented to demonstrate the industrial application of the methodology.

3. System architecture and methodology

This section depicts the detailed methodologies for document classification and document search. First, a document content extraction model is built to represent the document content with a vector consisting of key phrase frequencies. Second, a document classification model based on the back-propagation network (BPN) approach is developed. Finally, a document search model is implementing using a trained back-propagation network.

3.1. Key phrases extraction and frequency

Before document categorization, the key phrases are extracted from the document content. First, the sentences of the document are split into terms and these terms are then used to construct a key phrase base. A phrase correlation matrix is constructed for synthesizing two or more highly co-related phrases into a smaller set of significant phrases. After the document phrase extraction, the document is represented by a vector containing the frequencies of all key phrases.

A domain term that appears in a document with high frequency indicates that the term is a significant keyword or key phrase. The first step to build the key phrase base is to extract high frequency terms from the training documents. After extracting all high frequency terms, a correlation matrix of terms is created by calculating their frequency of occurrence within same documents. The correlation of two key phrases (KP, and KP) appeared in a set of patent documents is determined using the equation proposed by Hou and Chan (2003):

$$ R_{ij} = \frac{\sum_{l=1}^{N_D} X_{i,l} X_{j,l} - N_D \bar{X}_i \bar{X}_j}{\sqrt{\left(\sum_{l=1}^{N_D} X_{i,l}^2 - N_D \bar{X}_i^2\right) \left(\sum_{l=1}^{N_D} X_{j,l}^2 - N_D \bar{X}_j^2\right)}} $$

where $R_{ij}$ is the correlation of KP, and KP, that appear in a set of patent documents; $X_{i,l}$ is the frequency that KP, appeared in document $D_l$; $X_{j,l}$ is the frequency that KP, appeared in document $D_l$; $\bar{X}_i$ is the average frequency that KP, appeared in all documents (all $D$s); $\bar{X}_j$ is the average frequency that KP, appeared in all documents (all $D$s); $N_D$ is the total number of documents.

The highly correlated key phrases are selected and stored as the related phrase list. Finally, the key phrase list is completed when highly correlated phrases are merged, a necessary step since it is easier to train BPN models using fewer variables. When a new document is uploaded into a patent knowledge management system, the key phrases and their frequencies are extracted from the document.

First, frequencies of all terms are derived. Then, KP, is used to present the frequency of key phrase KP, in the document, and RPf, to represent the frequency of bibliographic data such as patent numbers, inventors, assignees, title, applicant country, date of application, and international patent classifications (IPC) to compare and contrast citation contributions. Karki (1997) notes that patent citation analysis provides information such as the most highly cited patents, the current impact index, the technology cycle time, and the technological strength of patents. However, patent citation analysis has several drawbacks. First, Kostoff (1998) points out the links between two patents using citations can be positive or negative yielding an inaccurate relationship. Secondly, as indicated by Karki (1997), if there is no machine readable form of the patent available, citation analysis becomes a very time intensive task. Further, patent documents may fail to cite basic patents or background information.

A number of studies have focused on patent statistical analysis. Andersen (1998) describes patent statistical data as a measure of invention. Breschi, Lissoni, and Malerba (2003) propose that patent application data be used as a measure of a firm’s technological advancement and diversification. Moreover, Cantwell and Giovanna (2004) utilize patent statistics as an indication of a country’s technological specialization. In addition to patent statistical analysis, automated patent cluster and classification techniques provide an efficient way to manage large numbers of patent documents. The following papers demonstrate the application of machine learning techniques to patent content analysis. Hsu, Trappey, Hou, and Trappey, and Liu (2004, in press) use clustering techniques for patent technology clustering, patent document clustering and technology maturity measurement. Chakrabarti, Dom, Agrawal, and Raghavan (1998) propose a Bayesian patent classification system to automate patent classification, Larkey (1999) develops a kNN-base system for U.S. patent search and classification, and Kohonen et al. (2000) implement a self-organization map based patent classification system. In this paper, a back-propagation network categorization and search method for patent classification is developed, validated, and evaluated. Further, a case for power hand tool patent document categorization.
related-phrase \( \text{RP}_{ij} \). The correlation of \( \text{RP}_{ij} \) and KP, are listed as \( R_{ij} \) and the final frequency of \( \text{KP}_{i} \) is

\[
\text{KP}_{i} = \text{KP}_{i} + \sum_{j=1}^{n} \text{RP}_{ij} \cdot R_{ij}
\]  

(6)

After calculating the KPF of all key phrases, a vector of key phrase frequencies is listed as

\[
[\text{KP}_{1}, \text{KP}_{2}, \text{KP}_{3}, \ldots, \text{KP}_{n}]
\]  

(7)

This vector serves as the input to the neural network model.

3.2. Automated document categorization

In this section, we describe the document categorization methodology based on the back-propagation network (BPN) algorithm. The BPN algorithm is a multi-layered network first proposed by Rumelhart, Hinton, and Williams (1986). The algorithm is also called a supervised learning method that can be used to solve non-linear problems. An advantage of BPN is that it does not need to change the network structure or activation functions of nodes and only adjusts the weights between nodes in the network (Massey, 2003). Another advantage of using BPN is its rapid execution when a trained network is applied. The learning stage of BPN involves a forward pass and a backward pass. The following sections describe these two phases.

3.2.1. Forward pass

The forward pass phase starts from the input layer. The input of every node is calculated. Then, the output from the activation functions of the nodes is derived and is passed forward to the next layer where processing continues until reaching the final output layer. The net input from input layer to the hidden layer node \( j \) is calculated using

\[
\text{net}^{h}_{j} = \sum_{i \in \text{previous layer}} w_{ij}^{h} x_{i} + b_{j}
\]  

(8)

where \( w_{ij}^{h} \) is the weight of connection between input layer node \( i \) and hidden layer node \( j \), \( x_{i} \) is the input of node \( i \), and \( b_{j} \) is the bias associated with node \( j \). The output of node \( j \) can be determined using the following equation:

\[
H_{j} = f(\text{net}^{h}_{j})
\]  

(9)

The activation function, \( f(x) = \frac{1}{1 + e^{-x}} + c \), where \( c \) is a constant (e.g., \( c = 0 \) or \( c = 5 \)) and the net input from hidden layer to output layer node \( k \) is computed using

\[
\text{net}^{o}_{k} = \sum_{j} w_{jk}^{o} H_{j}
\]  

(10)

where \( w_{jk}^{o} \) is the weight of connection between hidden layer node \( j \) and output layer node \( k \). Finally, we can determine the output of the neural network.

\[
O_{k} = g(\text{net}^{o}_{k}) = g\left(\sum_{j} w_{jk}^{o} H_{j}\right)
\]  

(11)

where \( g(x) \) is the activation function of node \( k \).

The error of the network is expressed as

\[
E = \frac{1}{2} \sum_{k} (T_{k} - O_{k})^{2}
\]  

(12)

where \( T_{k} \) is the real output of the training data.

3.2.2. Backward pass

The backward pass passes data back from the output layer to the previous hidden layer. The backward pass is used for determining errors and adjusting weights. Since \( E \) is defined as the function of \( O_{k} \) and \( O_{k} \) is the function of \( w_{jk}^{o} \), the weight adjustment between output layer and hidden layer \( \Delta w_{jk}^{o} \) can be expressed as

\[
\Delta w_{jk}^{o} = -\eta \frac{\partial E}{\partial w_{jk}^{o}} = -\eta \frac{\partial E}{\partial O_{k}} \frac{\partial O_{k}}{\partial H_{j}} \frac{\partial H_{j}}{\partial w_{jk}^{o}}
\]  

\[
= \eta \sum_{k} (T_{k} - O_{k})g'(\text{net}^{o}_{k}) w_{jk}^{0} f'(\text{net}^{h}_{j}) H_{j}
\]  

(13)

where \( \eta \) is the learning rate and \( \delta_{k} = (T_{k} - O_{k})g'(\text{net}^{o}_{k}) \).

Similarly, since \( O_{k} \) is a function of \( H_{j} \), the function of \( w_{ij}^{h} \) is the weight adjustment between hidden layer and input layer:

\[
\Delta w_{ij}^{h} = -\eta \frac{\partial E}{\partial w_{ij}^{h}} = -\eta \frac{\partial E}{\partial H_{j}} \frac{\partial H_{j}}{\partial w_{ij}^{h}} = -\eta \frac{\partial E}{\partial O_{k}} \frac{\partial O_{k}}{\partial H_{j}} \frac{\partial H_{j}}{\partial w_{ij}^{h}}
\]  

\[
= \eta \sum_{k} \delta_{k}^{h} w_{jk}^{0} f'(\text{net}^{h}_{j}) H_{j}
\]  

(14)

where \( \delta_{j} = f'(\text{net}^{h}_{j}) \sum_{k} \delta_{k}^{h} w_{jk}^{0} \).

Therefore, the weight adjustment is depicted as

\[
\Delta w_{ij}^{h} = \eta \delta_{j} f(\text{net}^{h}_{j})
\]  

(15)

where \( \delta_{j} \) is the output error of layer \( j \), \( \text{net}^{h}_{j} \) is the input of layer \( i \).

As described in Section 3.1, the key phrase frequency (KPF) vector is used to represent a patent document. Before importing the vector into the neural net, all key phrase frequencies in the vector are normalized between 0 and 1 using a transformation function:

\[
\text{KPF}' = \frac{\text{KPF}}{\max(\text{KPF}_{1}, \text{KPF}_{2}, \ldots, \text{KPF}_{n})}
\]  

(16)

The output values of the network represent the goodness-of-fit between a test document and all potential classes. In our study, IPC provides the target classes for the patent documents.

3.3. Document searching methodology

In this section, the trained neural network model constructed in Section 3.2 is used to develop a document search model. This model enhances document search since
analysts only need to key in a few phrases. First, a user selects a set of key phrases from the key phrase database. The set of key phrases describes the target document that the user wants to search. In the key phrase vector, the selected key phrase values are set to 1, and the non-selected ones are set to 0. Following the calculation of the KPF vector described in Section 3.2, the output identifies the best fit document class(es) given the input key phrases.

4. System analysis and design

4.1. Function modules of the system

The prototype system includes the System Parameters Management Module, the Automatic Categorization Module, and the Document Search Module. The system parameters management module provides the interface to adjust keyword correlation values and neural network weights. The automatic categorization module contains the functions of document upload, content extraction and document categorization. Finally, the document search module provides users an interface to search and download documents.

4.1.1. Functional flow of document categorization

The process of document categorization begins with uploading documents. After finishing the uploading of documents, the system extracts the vector of key phrase frequency (KPF) and imports the vector into the neural network classifier. The output is calculated automatically, and then the system shows the results of the automatic categorization. Users can accept the results or select a new document class (Fig. 1).

4.1.2. Document search functional flow

When searching for a document, a user selects a set of related key phrases that best represents the target. The system imports the 0/1 vector into the trained neural network classifier. Following the procedure for document categorization, the output is calculated and the system identifies the document class (Fig. 2).

4.2. Case demonstration

In order to demonstrate the proposed platform and methodologies for patent document classification and search, a case for power hand tool patents is presented. In this paper, 300 related hand tools patent documents are extracted from the World Intellectual Property Organization (WIPO) database\(^1\) in electronic format as the test documents.

---

The following sections discuss the case details using the BPN approach for automatic patent classification and search.

4.2.1. Define hierarchical patent document classification

IPC is a hierarchical classification scheme developed by the WIPO and consists of hierarchical class-layers, such as sections, classes, subclasses and groups respectively. Fig. 3 shows the IPC patent/invention classes related to patents of power hand tools downloaded from the WIPO Website. A two-level hierarchical classification structure is used for a test case analysis. The two-layer hierarchical classification structure is described as follows (WIPO):

First layer

- B25C: Hand-held nailing or stapling tools; manually-operated portable stapling tools.
- B25D: Percussive tools.
- B25F: Combination or multi-purpose tools not otherwise provided for; details or components of portable power-driven tools not particularly related to the operations performed and not otherwise provided for.
- B25G: Handles for hand implements.

Second layer

- B25C 01: Hand-held nailing tools.
- B25C 05: Manually-operated portable stapling tools; Hand-held power-operated stapling tools.
- B25D 09: Portable percussive tools with fluid-pressure drive, e.g., having several percussive tool bits operated simultaneously.
- B25D 11: Portable percussive tools with electromotor drive.
- B25D 17: Details of, or accessories for, portable power-driven percussive tools.
- B25F 01: Combination or multi-purpose hand tools.
- B25F 05: Details or components of portable power-driven tools not particularly related to the operations performed and not otherwise provided for.
- B25G 01: Handle constructions.
- B25G 03: Attaching handles to the implements.

4.2.2. Build key phrase base and related key phrase list

The keyword extraction process consists of four steps:

1. Filter non-keywords: In our system, we define nearly 400 stop words such as “to”, “and”, “it”, etc. The stop words in the patent documents are removed before key phrase extraction. A partial list of stop words is presented in Table 2.

2. Calculate the frequency of words: Select the high frequency words from each document. In this step, 67 words with high frequency are identified.

3. Build correlation matrix: Calculate the frequency of each word in each document and, then, build the correlation matrix of words (Table 3), as well as subsequently a related word list.

4. Synthesize phrases: According to the related word list (Table 4), the phrases are merged and act as the input vector. After synthesizing, 59 key phrases remain.

<table>
<thead>
<tr>
<th>Stop-list</th>
</tr>
</thead>
<tbody>
<tr>
<td>are</td>
</tr>
<tr>
<td>all</td>
</tr>
<tr>
<td>always</td>
</tr>
<tr>
<td>both</td>
</tr>
<tr>
<td>beyond</td>
</tr>
<tr>
<td>during</td>
</tr>
<tr>
<td>for</td>
</tr>
<tr>
<td>here</td>
</tr>
<tr>
<td>it</td>
</tr>
</tbody>
</table>

Table 2
A partial listing of stop words

Fig. 3. IPC B25 hierarchical patent classes related to power hand tools.
4.2.3. Construct and train BPN model for classification

The neural network toolbox in Matlab software package\(^2\) is used to construct the neural network model. There are four parameters for the network:

1. **Number of layers:** The IPC network has 2 layers, one hidden layer and one output layer.
2. **Number of output nodes:** The output nodes correspond to the document classes. For example, there are four classes in level 1 (B25C, B25D, B25F, and B25G), so the number of output nodes is four.
3. **Number of neurons:** This number influences the error of network, the training speed, and is adjusted dynamically in the training process.
4. **Activation functions:** The activation function of the hidden layer is set as TANSIG and the activation function of output layer is LOGSIG.

As shown in Fig. 4, after 14 epochs, the error of network drops to less than \(10^{-2.5}\) so we conclude that the network is well trained. Referring to (a)-(d) in Fig. 5, the training results for the classes B25C, B25D, B25F, and B25G are provided. All network models take approximately 30–40 epochs to converge and the network weights are stored in the system database.

4.2.4. Constructing the search model

As described in Section 3.3, the trained network serves as the basis for the search function. The only difference between document classification and search is the determination of the KPF vectors. Users select the target documents’ key phrases from the phrase base manually as the input of the network and the search model generates the vector showing the relationship between the document class and the target documents.

4.3. Scenarios discussion

To use the document categorization function, a user uploads a patent document. After the document file is uploaded, the system extracts key phrases from the text and shows the result of the key phrase frequencies. The system calculates the classification results using the neural network model. Consider a patent document abstract without a pre-defined category:

A power tool is provided including a resilient boot having a radial flange mounted in a seat formed by housing

---

halves in order to provide a sealed barrier segregating air of an inlet vent of a motor from air of an exhaust vent and the boot also mounting the motor and providing vibration dampening between the motor and the housing halves. The boot and housing also include wire management slots and the boot absorbs tolerance stack-up of components and provides shock absorption when the power tool is dropped.

In order to identify the category of this patent, the text file is uploaded into the system. After key phrase extraction and model adoption, the classification result shows the document belongs to B25F-05 IPC category code.

For the search function, a user chooses a set of key phrases for target document. The system then determines a document category for the document based on the KPFs. The system shows the category search results and the user reviews all patent documents in the specific IPC category. This function helps users narrow the range of search. For example, an R&D engineer looking for inventions related to handles of power tools selects four key phrases—handle, assembly, shaft, and grip— for search input. The system determines the most suitable category according to the selected key phrases. The automated search narrows the search of patent documents to a specified IPC category. The category is B25G01, which has a collection of 64 patent documents. The titles of these documents are listed in Table 5.

To evaluate the effectiveness of the proposed methodology, 300 training patent documents were imported to derive the neural network model with acceptable error. In this paper, the categorization precision is measured (Salton, 1973) using the following equation:

\[ P_i = \frac{A_i}{A_i + B_i} \]  

where \( P_i \) is the precision of category \( i \), \( A_i \) the number of patent documents that are classified to category \( i \) correctly, and \( B_i \) the number of patent documents that are classified to category \( i \) incorrectly.

Fig. 5. The results of level 2 training: (a) B25C convergence after 35 epochs with two groups, (b) B25D convergence after 30 epochs with three groups, (c) B25F convergence after 30 epochs with two groups, (d) B25G convergence after 40 epochs with two groups.
As shown in Table 6, 124 patent documents are used to test the system performance. The testing precision of the first layer classification is 91.93%. Moreover, the average testing precision of second layer classification is 90.35%. The test results yielded above 90% accuracy for automated patent document classification.

In the end of this section, the proposed classification methodology is compared with other methodologies, such as the legal knowledge management (LKM) system (Hsu et al., 2004) and the Bayes-based system. LKM is a general-purpose system and calculates the term frequencies and co-variances of terms to classify categories. The Bayes-based system uses probability theory to classify patent documents and we reconstruct a system proposed by Lee (2003) to compare with our system. The comparison results are shown in Fig. 6. The result demonstrates the improvement over previously classification methods of patent document management. Furthermore, the consistency of the classification can be guaranteed as compared with the manual-assignment mechanism by various patent document providers.

5. Conclusion

The back-propagation networks (BPN) algorithm provides advantages of non-linear problem solving ability and learning by example. There are limitations to the application of BPN since inadequate training data may yield an unreliable model and the training procedure may require significant computing resources. The first limitation can be solved by compiling a wide range of examples. Since a well-trained model can help companies better manage documents, the cost of computing resources can be easily justified. This research develops a methodology for classifying patent documents automatically using BPN. The patent documents, extracted from WIPO, are originally classified using a hierarchical classification scheme. After key phrases extraction and frequency calculation, the resulting key phrase base adequately represents the characteristics of the patent documents. Furthermore, the approach for the hierarchical document taxonomy is proposed to classify the patent documents with unknown categories using a well-trained neural network. Finally, a patent document search method is also presented. The search method helps
users determine the likely category of the target document and can effectively narrow the searching ranges for R&D engineers under time constraints.

To improve the accuracy of automatic classification, the BPN models are trained continuously and a feedback mechanism deployed to adjust the weights of the BPN model. The extension of this study involves combining the neural network technology with other approaches such as fuzzy decision models and other ontology-based feature extraction methods to improve the flexibility and accuracy of the current method. Besides the analysis of patent documents, the proposed method can be modified for other classifications schemes for general knowledge management applications.

Acknowledgement

This research is funded partially by the Taiwan Ministry of Economic Affairs and National Science Council research grants.

References


