International Journal of Production Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tprs20

Ontology-based neural network for patent knowledge management in design collaboration

Amy J. C. Trappey [a], Charles V. Trappey [b], Tzu-An Chiang [c] & Yi-Hsuan Huang [a]
[a] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan
[b] Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan
[c] National Taipei College of Business, Taipei, Taiwan
Published online: 01 Aug 2012.

To cite this article: Amy J. C. Trappey, Charles V. Trappey, Tzu-An Chiang & Yi-Hsuan Huang (2013) Ontology-based neural network for patent knowledge management in design collaboration, International Journal of Production Research, 51:7, 1992-2005, DOI: 10.1080/00207543.2012.701775

To link to this article: http://dx.doi.org/10.1080/00207543.2012.701775

International Journal of Production Research, 2013
Vol. 51, No. 7, 1992–2005, http://dx.doi.org/10.1080/00207543.2012.701775

Taylor & Francis
Taylor & Francis Group

Ontology-based neural network for patent knowledge management in design collaboration

Amy J. C. Trappey[a]*, Charles V. Trappey[b], Tzu-An Chiang[c] and Yi-Hsuan Huang[a]

[a]*Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan;* [b]*Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan;* [c]*National Taipei College of Business, Taipei, Taiwan*

(*Received 28 April 2012; final version received 4 June 2012*)

In order to stimulate innovation during the collaborative process of new product and production development, especially to avoid duplicating existing techniques or infringing upon others' patents and intellectual property rights, the collaborative team of research and development, and patent engineers must accurately identify relevant patent knowledge in a timely manner. This research develops a novel knowledge management approach using ontology-based artificial neural network (ANN) algorithm to automatically classify and search knowledge documents stored in huge online patent corpuses. This research focuses on developing a smart and semantic oriented classification and search from the sources of the most critical and well-structured knowledge publications, i.e. patents, to gain valuable and practical references for the collaborative networks of technology-centric product and production development teams. The research uses the domain ontology schema created using Protégé and derives the semantic concept probabilities of key phrases that frequently occur in domain relevant patent documents. Then, by combining the term frequencies and the concept probabilities of key phrases as the ANN inputs, the method shows significant improvement in classification accuracy. In addition, this research provides an advanced semantic-oriented search algorithm to accurately identify related patent documents in the patent knowledge base. The case demonstration analyses 343 chemical mechanical polishing and 150 radio-frequency identification patents sample sets to verify and measure the performance of the proposed approach. The results are compared with the previous automatic classification methods demonstrating much improved outcomes.

Keywords: knowledge-document categorisation; artificial neural network; ontology

1. Introduction

The competitive nature of business and the fast-changing consumer demand forces enterprises to accelerate the speed of new product development and market entry. Companies gain competitive advantages when their research and development (R&D) is innovative and timely. According to the reports from World Intellectual Property Organization (WIPO) (Gurry 2011) and other sources (EPO 2007, Leeds City Council 2011), 80% or more of the technology regularly disclosed in patent systems is not commonly revealed in other publications. The United States Patent and Trade Office (USPTO) indicates that the counts of US patent applications and granted patents have exceeded 490,000 and 210,000 respectively in 2010 (USPTO 2011). The intellectual property rights (IPR) of patented technologies are protected by the intellectual property (IP) law. Hence, a patent not only provides technical knowledge and indicates trends, but also supports potential litigation defensively or offensively for the patentee (Kim *et al*. 1999). In order to develop state-of-the-art technology, patent document analysis plays a crucial role during R&D. By applying patent knowledge effectively, an R&D collaborative team can significantly decrease the time and the expense of new product development. With the growing notion of rapid technology development and vigorous IPR protection, there is significant increase in the numbers of applied and granted patents. IPR and R&D managers and engineers in any company have difficulty finding critical information, let alone knowledge, effectively within the large numbers of related patent documents and their inter-linking claims. Frequently, hundreds of patents are published every week waiting to be studied. Hence, it is crucial to apply smart text and data-mining techniques to automatically categorise patent documents with efficiency and accuracy as references for IPR analysts and R&D engineers. This research proposes an ontology-based semantic ANN approach which systematically and automatically classifies large number of patent documents within patent knowledge bases and assists users to effectively search and retrieve related patents for collaborative product development. In order to demonstrate the

*Corresponding author. Email: trappey@ie.nthu.edu.tw

methodology for practical applications, a prototype system is developed and tested using a total of 493 chemical mechanical polishing (CMP) and radio frequency identification (RFID) patents sourced from the patent corpuses of the World Intellectual Property Office (WIPO) and the United States Patent and Trademark Office (USPTO).

The paper is organised into several sections. In Section 2, we review the related research literature and their proposed methodologies. Section 3 describes the systematic methodology of the ontology-based semantic ANN for intelligent patent categorisation and search. Section 4 depicts the case results for categorising and searching CMP and RFID patents using the methodology developed in Section 3. The comparative evaluation of the proposed method is demonstrated using the case examples with significantly improved results. Finally, the conclusion, contributions and limitations of the research are discussed in Section 5.

2. Literature review

Many organisations realise that collaborative networks have a clear and positive impact on business performance by sharing enterprise resources (Romero *et al*. 2009). Owing to the rapid growth of available online knowledge documents in a collaborative environment, it is very difficult to read, understand, classify, and find knowledge documents using traditional man-power operations (Crowder *et al*. 2010). Therefore, Zhen *et al*. (2011) proposed a distributed knowledge sharing model to recommend potentially useful knowledge from personal repositories to those members who may need them in a collaborative network. Our research approach proposes to screen and target colleagues with similar knowledge interests. The filter identifies the top-k correlated knowledge resources based on scores given by collaborating colleagues.

With the development of data-mining techniques, some academics have focused their attention on content analysis of knowledge documents (Liu and Harding 2009) to perform document categorisation and clustering. For example, based on the set of keywords in a document, Svingen (1998) used a genetic programming algorithm to calculate a fitness value to decide whether a document of interest can be linked to a specific user. Ko and Seo (2000) constructed an unsupervised text classifier by using the Naive Bayes algorithm. They divided the documents into sentences and categorise each sentence using keyword lists of each category using a sentence similarity measure. The proposed method shows a similar degree of performance, compared with the traditional supervised learning methods. Lam and Han (2003) unified the strengths of k-Nearest-Neighbour and linear classifiers to develop a generalised instance set (GIS) algorithm for automatic document categorisation. To further enhance the performance of document classification, they proposed a meta-model framework, which uses the category feature characteristics derived from the training document set to capture inherent properties of a particular category.

The term 'frequency method' is widely applied for document categorisation and clustering (Lam *et al*. 1999, Liu and Zhang 2001). Farkas (1993) published a method that uses term frequencies to represent documents as numeric concept vectors in a semantically meaningful way and then combines the results with a back-propagation artificial neural network (BPANN) and self-organising maps (SOM) to build an automatic document classification system. Karras and Mertzios (2002) utilised the learning and generalisation capabilities of neural networks to develop a document categorisation system for non-domain specific full-text documents. This system applies word semantic category co-occurrence analysis. Their method resolves the problem of dimensionality reduction for document categorisation feature extraction. When processing large numbers of patent documents, Trappey *et al*. (2006a) applied a BPANN to develop a patent document classification and search platform. The document's categories are pre-defined based on the international patent classification (IPC) standard. The classification process begins by extracting key phrases from the patent document set and then determines the significance of key phrases according to their frequencies in text. After extracting all high frequency terms, a correlation matrix of terms is created by calculating their frequency of occurrence within the document set. Then, highly correlated phrases are merged. Since there are fewer variables, training the BPANN model is simplified. In order to improve the effectiveness of document classification and clustering, later studies used ontology to express knowledge concepts that are readable by computer programs. For example, Trappey *et al*. (2006b) presented an ontology-based neural network approach to classify patent documents. The system extracts the features of a document. These features, matched with concept classes and their relationships in the domain ontology, are then transferred as inputs into the artificial neural network (ANN) models. Based on the assumption that the higher the relative frequency of a word, the stronger the relationship between this word and its associated class, Hung and Wermter (2004) proposed an extended significance vector model (ESVM) to represent a preference for a specific semantic class and then used the hypernym relationship from the WordNet ontology to build the hypernym significance vector model (HSVM). Finally, the

self-organising map (SOM) based on ESVM, HSVM, and their hybrid vector space model (HyM) are developed. According to their experiments, the proposed models improve the classification performance of SOM. In addition, Trappey *et al*. (2009) proposed a fuzzy ontological document clustering (FODC) methodology which uses a fuzzy logic control to create suitable document clusters. The benchmarking results show that the FODC approach outperforms the simple K-means clustering approach.

Although there are several academic papers focusing on patent document categorisation, little research combines the concept probabilities of key phrases with an ontology to develop an ontology-based semantic concept ANN to improve the effectiveness of knowledge document self-categorisation and search. In order to aid IPR and R&D engineers in locating relevant patents, this research provides a semantic-oriented search function to improve patent search time and accuracy.

3. System framework and detailed methodology

The system framework for knowledge document categorisation and search using an ontology-based semantic ANN is depicted in Figure 1. For the pre-processing part, an ontology schema is constructed using Protégé (http://protege.stanford.edu/) before it is imported into the system. There still remain many research challenges related to the building of domain-specific ontologies that are not covered by the scope of this research paper. The system is

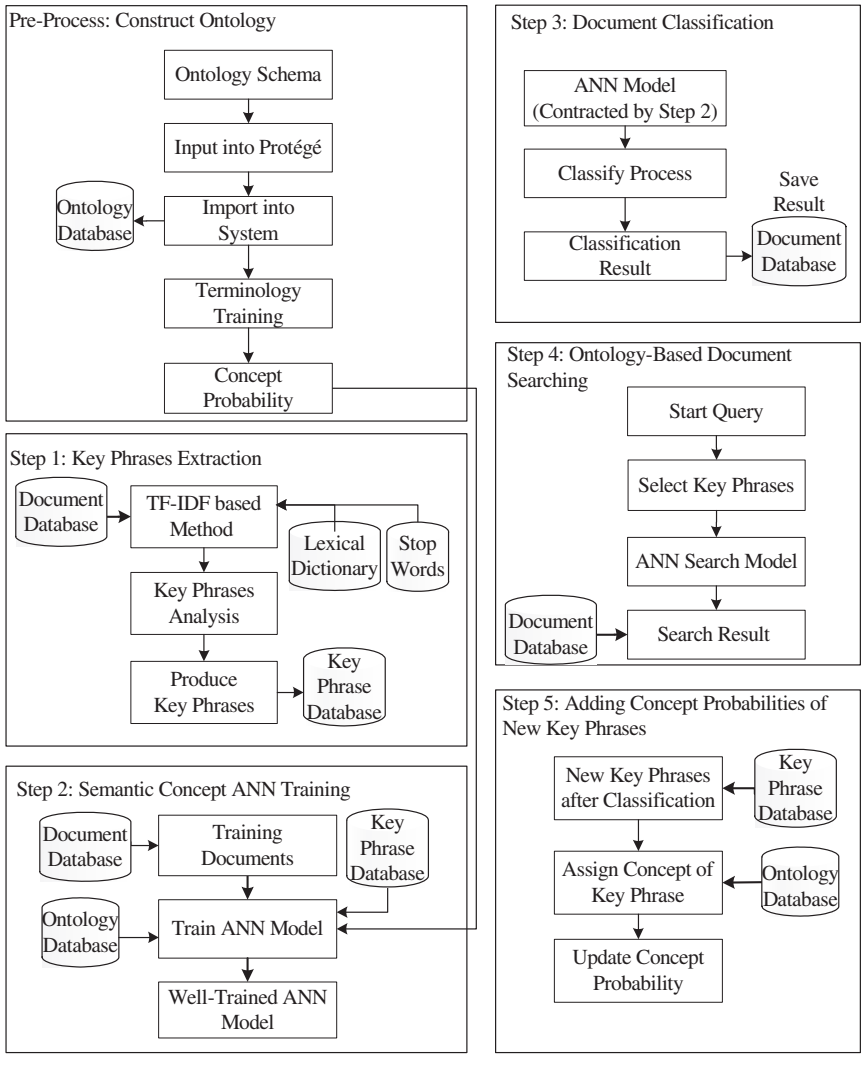

Figure 1. Ontology-based semantic knowledge document categorisation and search framework.

trained to derive the probabilities of a specific key phrase representing given concepts. After completing the pre-process, the first step of the analytical procedure is to extract high frequency key phrases appearing in the knowledge documents based on the combined value of Term Frequency and Inverse Document Frequency (TF-IDF). Further, the semantic concept ANN model is trained using the combined frequencies and probabilities of key phrases as ANN input parameter values. The trained model is then used to classify patent documents. Finally, a user can specify the key phrases and assign their weights for retrieving related knowledge documents from the knowledge base. If new key words are found in given documents, the system will immediately update the concept probabilities of key phrases. In the following subsections, we describe the system methodology in detail.

3.1 *Ontology schema construction*

During the pre-processing phase, the domain experts construct the domain ontology and transfer the ontology schema into machine readable web ontology language (OWL) using the Protégé package. In order to organise the knowledge provided in the knowledge documents, with respect to the defined ontology schema, the system is trained using a set of representing knowledge documents. First, a natural language processing and tagging tool, MontyLingua (2011), is used to tag the parts of speech (POS), chunks, and lemmas in the training knowledge documents. MontyLingua is a commonsense-enriched natural language 'understander' for English, developed in the Media Laboratory at MIT. Given a document as raw English text, MontyLingua automatically extracts subject/verb/object tuples, adjectives, verb and noun phrases from the raw text. Afterwards, knowledge engineers map the extracted words to semantic concepts of the ontology as shown in Figure 2. For example, the phrase 'CMP apparatus and method' maps to the concept *CMP_method*, while 'polishing' represents the concept of *polish*, and 'semiconductor wafers' represents the concept of *substrate*. The system records the probabilities of the semantic concepts that a key word or key phrase implies in the patent document. The conditional probabilities represented by P (The patent concept | The word W in the chunk of C in corpora) are computed during the training session.

3.2 *Key phrase extraction*

The TF-IDF method is used to extract key phrases from knowledge documents. The steps of the TF-IDF method include word segmentation, stopping, morphology diagnostics, and stemming. After these steps, we calculate TF-IDF values for each word and select key phrases for TF-IDF values that exceed assigned thresholds. The procedures for key phrase extraction are described by the following steps. The first step is word segmentation. The system separates words according to several segmentation symbols represented by commas semi-colons, periods, and so forth. The second step is to delete stop words (e.g. a, after, almost, be, been, come), which hold little information related to the domain knowledge. The third step is morphology diagnostics. In this step, a lexical dictionary is imported to identify the morphology of meaningful words. The system converts all words to a set of verbs and nouns to consistently express the knowledge content (Matsuo and Ishizuka 2004). In the last step, by means of the

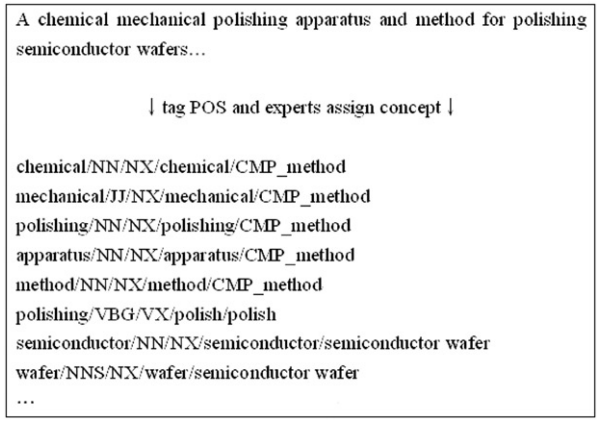

Figure 2. Mapping the extracted words to semantic concepts of the ontology.

Porter Stemming Algorithm (Porter 1980), the system restores the tense and plurality of words to their root words. This process achieves the goal of integrating the term weights (w_{jk} in the following Equation (1)), with words which have the same root words (Kantrowitz, *et al*. 2000). All phrases' TF-IDF values are calculated recursively using Equation (1). The term weight (w_{jk}) is the product of the term frequency (tf_{jk}) and the inverse document frequency (idf_j) of the given phrase (j) in a given patent document (k). If the TF-IDF values of given phrases exceed a defined threshold value, these phrases are considered the key phrases, i.e. well representing the given document, and are saved in the key phrase database. Otherwise, owing to their infrequent appearances, the phrases are interpreted as insignificant and, thus, excluded from the key phrase set. For detailed discussion of TF-IDF and other minor improved algorithms, please refer to further readings in Salton and Buckley (1988), Kantrowitz, *et al*. (2000), Kao (2000).

$$\begin{aligned} w_{jk} &= tf_{jk} \times idf_j, \\ idf_j &= \log_2\left(\frac{n}{df_j}\right), \end{aligned} \tag{1}$$

where w_{jk} is the phrase weight of phrase j in the document k; tf_{jk} is the number of phrases j that occur in document k; n is the total number of documents in a document set; and df_j is the number of documents containing the phrase j in the document set.

3.3 *Ontology-based semantic concept artificial neural network*

A back-propagation artificial neural network (BPANN) is a multi-layer network applied to solve nonlinear problems. A BPANN is one of the most frequently used neural network models applied to document categorisation and word identification. The learning algorithm of a BPANN is a supervised learning method. An advantage of a BPANN is that the network structure and activation functions of nodes are not changed. The algorithm adjusts weights between nodes in the BPANN network. The diagram of a BPANN is shown in Figure 3.

Equation (2) shows the formula for computing hidden layer values where w_{ij}^h is the weight from the input layer i to the hidden layer j. The value X_i is the ith input and the activation function is a sigmoid function as shown in Equation (3). The output values from the hidden layer to the output layer are computed as shown in Equation (4). The value w_{jk}^o is the weight from the hidden layer j to the output layer k. Finally, the output values are computed using Equation (5). The error of an output layer's node is defined in Equation (6):

$$net_j^n = f\left(\sum_i w_{ij}^h X_i\right) \tag{2}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

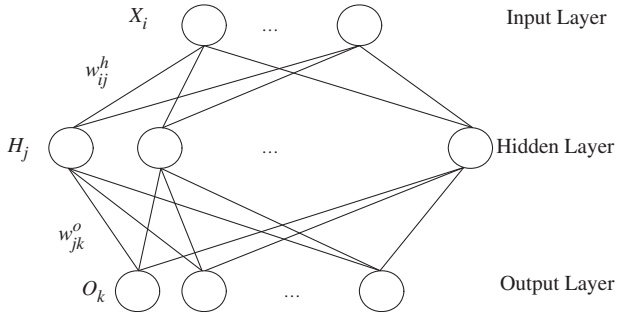

Figure 3. BPANN network diagram.

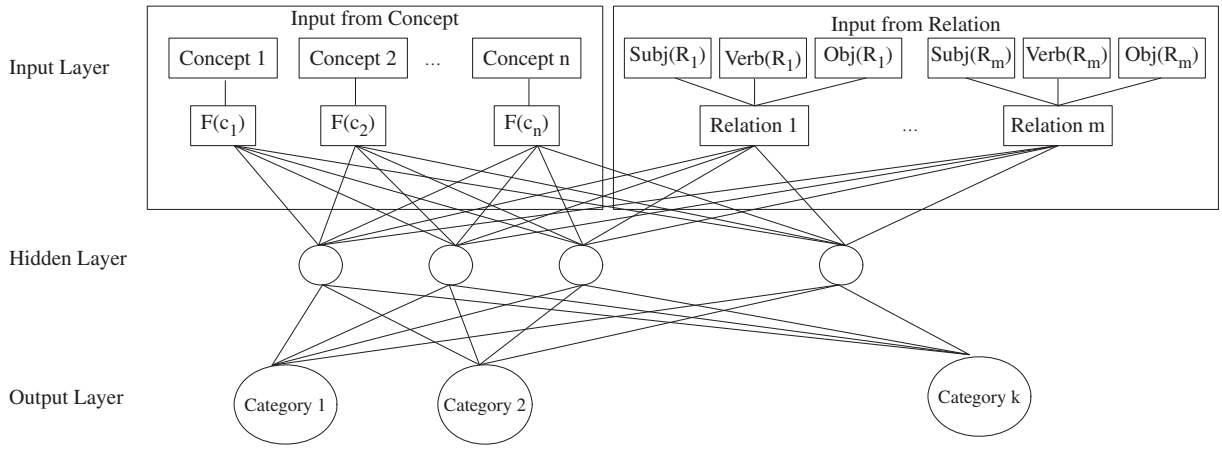

Figure 4. Ontology-based semantic concept artificial neural network.

$$net_k^o = \sum_j w_{jk}^o f\left(\sum_i w_{ij}^h X_i\right) \tag{4}$$

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

$$E = \frac{1}{2}\sum_k \left(T_k - \left(\sum_j w_{jk}^o f\left(\sum_i w_{ij}^h X_i\right)\right)\right)^2. \tag{6}$$

During the backward pass phase, the weights are adjusted by error correction rules to direct the output of the BPANN network toward the desired outcome. The gradient value is computed using Equation (7):

$$\begin{aligned} \Delta w_{ij}^N &= \eta \delta_j^N net_i^{N-1} \\ \delta_j^N &= f\left(net_j^N\right)\left(T_j - O_j\right). \end{aligned} \tag{7}$$

δ_j^N is the error value of node j in the layer N, net_i^{N-1} is the input value of node i in the layer $N-1$, and Δw_{ij}^N is the weight from the node i to the node j in the layer N. When $1 < n < N$, the equation may be re-written as Equation (8):

$$\begin{aligned} \Delta w_{ij}^N &= \eta \delta_j^N net_i^{n-1} \\ \delta_j^N &= f\left(net_j^N\right) \sum_k w_{jk}^{n+1} \delta_k^{n+1}. \end{aligned} \tag{8}$$

In order to achieve the goal of knowledge document self-categorisation using a semantic-oriented search, this paper modifies the neural network model proposed by Trappey *et al.* (2006b). The modified ontology-based semantic concept neural network is shown in Figure 4.

An equation is used to restrict the input values to a range within 0 and 1 as shown in Equation (9) (Kao 2000). When the concepts in the ontology appear in a knowledge document, this research considers the probability that a specific key phrase contains a certain semantic concept. The matrix for the key phrase frequencies and the probabilities of semantic concepts are shown in Table 1. KP_j is the key phrase j, $f\left(KP_j\right)$ is the frequency of the key phrase j, and $P\left(C_{i,j}\right)$ is the probability of the key phrase j in the semantic concept i. With this information, the total frequency of each concept is calculated using Equation (10).

$$f = 1 - k\frac{1}{x^{1/2}} \tag{9}$$

$$Total(C_i) = f(KP_1) \times P\left(C_{i,1}\right) + f(KP_2) \times P\left(C_{i,2}\right) + \cdots + f(KP_m) \times P\left(C_{i,m}\right). \tag{10}$$

Table 1. Key-phrase matrix of frequencies and probabilities of concepts.

Key phrase	Frequency	Probability of Concept 1	Probability of Concept 2	⋯	Probability of Concept n
KP_1	$f(KP_1)$	$P(C_{1,1})$	$P(C_{2,1})$	⋯	$P(C_{n,1})$
KP_2	$f(KP_2)$	$P(C_{1,2})$	$P(C_{2,2})$	⋯	$P(C_{n,2})$
⋯	⋯	⋯	⋯	⋯	⋯
KP_m	$f(KP_m)$	$P(C_{1,m})$	$P(C_{2,m})$	⋯	$P(C_{n,m})$

$Total(C_i)$ is the total frequency of the semantic concept i. Finally, the output of the semantic concept i is defined by function $F_{(C)}$ as shown by Equation (11):

$$F_{(C)} = \left(1 - k\frac{1}{Total(C_i)^{1/2}}\right). \tag{11}$$

The ontology-based semantic concept neural network also considers the semantic relationship. A relation is represented by a sentence, and there are many types of sentence patterns in a knowledge document. Thus, the most common sentence patterns include a document's common sentences and clauses that are identified in the three steps. First, there is sentence extraction where the proposed system reads the content and transforms the content into a long string. Then, individual sentences in the knowledge documents are parsed using the punctuation marks to build the set $\{S_1, S_2, S_3 \ldots S_n\}$. For the third step of sentence analysis, the morphology analysis and pattern analysis are combined. The extracted sentences derived through phrase 1 are used to analyse the sentence morphologies and sentence patterns are classified into common sentences and clauses. The common sentences (CS_p) are represented by Equation (12):

$$(CS_p) = \{[Art_p][Adj_p]\ Noun_p\} + Verb_p + \left\{\left[Art_p^{\cdot}\right]\left[Adj_p^{\cdot}\right]\ Noun_p^{\cdot}\right\}[Adv_p]. \tag{12}$$

'*Art*' represents the article in the sentence CS_p, '*Adj*' represents the adjective in the sentence CS_p, '*Noun*' represents the noun in the sentence CS_p, and '*Adv*' represents the adverb in the sentence CS_p. After morphology analysis, pattern analysis is conducted. All common sentence patterns are denoted as 'Common sentence = Subject + Verb + Object'. Finally, the system extracts the subject, verb and object of a sentence via the above steps, and the variables are defined as follows. S_p is the sentence p in the document, Subj(S_p) is a subject of the sentence S_p, Verb(S_p) is a verb of the sentence S_p, Obj(S_p) is an object of the sentence S and CS is the set of common sentences.

All relationships represented in a sentence are called the 'relation description'. A relation description includes the extracted subject, verb and object with the following variables defined. Subj(R_v) is the subject, Verbj(R_v) is the verb, and Obj(R_v) is the object of the description of the relation R_v. If $Subj(S_p)$ and $Subj(R_v)$ are synonyms, or if $Verb(S_p)$ and $Verb(R_v)$ are synonyms, or if $Obj(S_p)$ and $Obj(R_v)$ are synonyms, then the probability for the sentence (S_p) is computed using Equation (13):

$$P(S_p) = \frac{(P(Subj(R_v)) + P(Verb(R_v)) + P(Obj(R_v)))}{3}. \tag{13}$$

3.4 *Ontology-based document searching*

The trained semantic concept BPANN model is used to provide a domain-specific document searching service. First, a list of concepts and relations are described by the ontology, and a user can select and set the weight of each key word. Second, the selected concepts and relations are transferred as BPANN input values. Third, the BPANN model is used to search knowledge documents. Fourth, the output displays the classification of a knowledge document. A user can select a classification and read the knowledge documents within this classification.

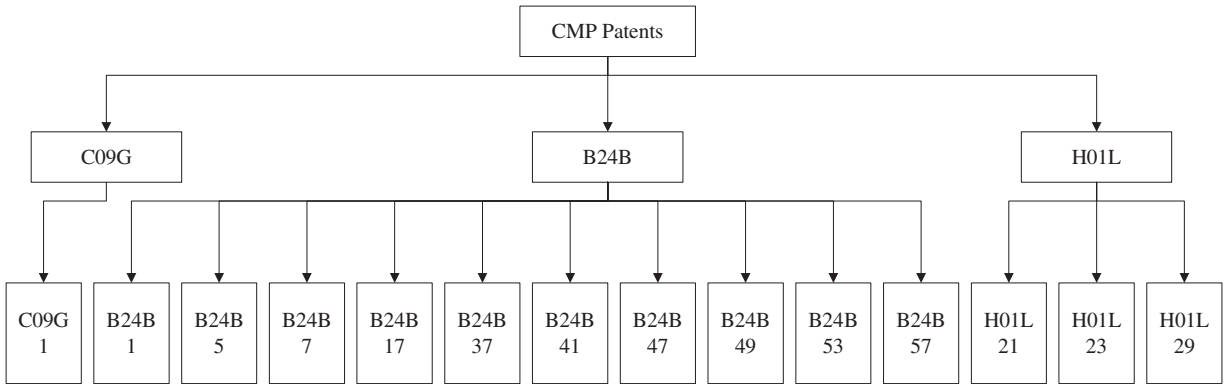

Figure 5. IPC hierarchical structure of the CMP patent domain.

3.5 *Adding concept probabilities of new key phrases*

Upon completion of document categorisation, new key phrases are produced. Since these new key phrases are not yet included in the concept probability database, the system updates the concept probability database to increase the identification accuracy. There are several steps for the modifying module. The system lists all new key phrases and patent engineers assign key phrases to concepts in the ontology. Then, the system calculates the probability of each new key phrase using the methods described in Section 3.3. Finally, the system saves the concept probabilities of the new key phrases.

4. Case analyses of CMP and RFID related patents

This research uses a set of Chemical Mechanical Polishing (CMP) and Radio Frequency Identification (RFID) patent documents as a case study to elaborate and demonstrate the system methodology and evaluate the search performance by comparing the results to previously tested approaches.

4.1 *Hierarchical document classification for CMP and RFID*

Following the International Patent Classification (IPC) standard (WIPO 2012), there are five layers in the classification hierarchy. The first layer is 'Section', the second layer is 'Class', the third is 'Subclass', the fourth is 'Group', and the final layer is 'Subgroup'. In this section, the classification hierarchies for CMP and RFID are constructed. A two-layer hierarchy is used to explain how we use an ontology-based semantic concept artificial neural network to classify and search for patent documents.

CMP is a technique used in semiconductor fabrication for polishing the top surface of in-process semiconductor wafers or other substrates. There are three standard IPC codes related to the CMP technical domain, including C09G, B24B, and H01L (Lin 2005). Figure 5 shows the IPC hierarchical structure of the CMP patent domain.

Radio-frequency identification (RFID) is an automatic identification method, relying on storing and remotely retrieving data using devices called RFID tags or transponders (Stefan 2005). An RFID tag is an object that is attached to or incorporated into a product, animal, or person for the purpose of identification and is activated using radio waves. Chip-based RFID tags contain silicon chips and antennas. Passive tags require no internal power source, whereas active tags require a power source. Figure 6 shows the IPC code structure for the RFID patent domain.

4.2 *CMP and RFID domain ontology schema*

Protégé is used to construct an ontological domain, which is then transferred into OWL language. The OWL file in the system recognises the knowledge that falls within the CMP and RFID domains. An ontology schema for the CMP domain is defined by domain experts (Lin 2005) and the ontology schema for RFID is defined referring to

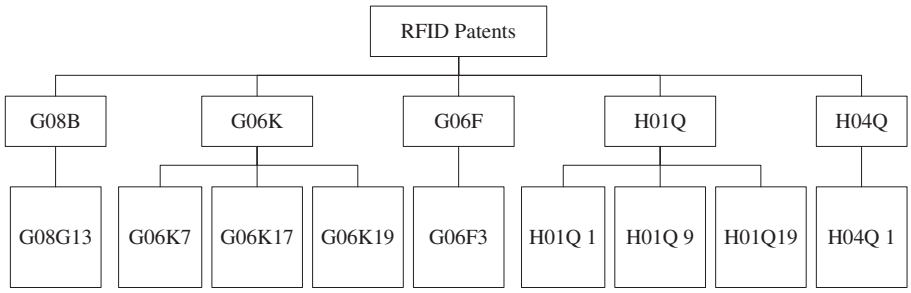

Figure 6. Classification architecture for the RFID patent domain.

Table 2. Hierarchical concepts of CMP and RFID ontology schemas.

Domain	Part	Subclasses
CMP ontology	Application	Substrate, semiconductor, storage
	Apparatus	Cleaning, polishing, detecting, transporting
	Consumable	Film, slurry, storage, and particle (subclass of slurry)
	Parameter	Adhesion, path, pressure, velocity, warp
	Material	Teflon, rubber, plastic, ceramic, metal, and subclasses of metal (copper, silicon, etc.)
	CMP_method	Wafer_transfer_mechanism, CMP_equipment, electromagnetic_radiation, transport_apparatus
RFID ontology	RFID_general	Memory, communication, processor, encoding, security, range, protocol, circuit, tolerance
	Antena	Wave, direction, frequency, band
	RFID_application	Portable, identification, interaction, tracking
	Value	Gain, directivity, impedance
	Tag	Active, passive
	Device	Wireless or wired connector
	Item	Creature or part (non-creature)

(Stefan 2005). The hierarchical concept classes (i.e. parts and subclasses) of CMP and RFID ontology schemas are outlined in Table 2.

There are six parts of the CMP ontology schema and each part is divided into subclasses of concepts. Concepts relations are further defined using OWL language. Examples of the formal representation using the defined CMP ontology parts and subclasses are described as follows:

- Application – One can select one subclass to describe 'application → substrate → copper'. The subclass indicates that 'The substrate of CMP application is made by copper'.
- Apparatus – We can describe 'apparatus → polishing → pad → surface feature → roughness'. The representation means that 'The polishing pad is an apparatus and roughness is one feature of surface'.
- Consumable – The description 'consumable → slurry' means that 'The slurry is a consumable'.
- Parameter – The statement 'parameter → path' defines path as a parameter of the CMP process.
- CMP_method – The foral representation 'CMP_method → wafer_transfer_mechanism → polish_pad' states that the 'polish pad is one method of wafer transferring mechanism'.

The RFID ontology schema is classified into seven parts. Examples of the formal representation using the defined RFID ontology parts and subclasses are described as follows:

- Antenna – Select one subclass of antenna to describe 'antenna → wave'. The statement means that 'The antenna wave'.
- RFID_application – The representation of 'RFID_application → tracking → person_tracking' states that 'personal tracking is one of RFID applications'.
- Tag – The statement 'tag → active' claims 'the tag is active'.

- Device – One uses 'device → wireless → reader' to depict that 'the reader is a wireless device' in the document.

4.3 *Terminology analysis*

After building the ontology, the terminology is analysed in order to calculate the probability that the key phrase implies the semantic concept. For the CMP domain, this research inputs 15 patent documents as training data. For the RFID domain, 10 patent documents are used as training data. Since the CMP ontology is more complex than the RFID ontology, more documents are used as training data. Tables 3 and 4 provide the partial training results for the two knowledge domains.

4.4 *System evaluation*

The retrieval effectiveness of the system is measured by recall, precision, accuracy, and the F statistic. Recall and precision are well-known evaluators of information retrieval effectiveness proposed by (Salton and Buckley 1988). For system measurement, the A, B, and C definitions are described below. In order to measure and evaluate the system, this research defines the following variables:

- A: A document is relevant to the query, and the system infers the document is relevant to the query.
- B: A document is not relevant to the query, but the system infers the document is relevant to the query.
- C: A document is relevant to the query, but the system infers the document is not relevant to the query.
- D: The total number of documents.

Table 3. Partial terminology for the CMP patents.

Lemma	Chunk	Concept	Probability	Lemma	Chunk	Concept	Probability
apparatus	NX	CMP_method	1.00	slurry	NX	wafer_transfer_mechanism	0.25
chemical	NX	CMP_method	1.00	slurry	NX	slurry_delivery_line	0.50
cleaning	NX	clean	1.00	transfer	NX	efficient	0.25
compound	NX	compound	0.33	transfer	NX	transfer	0.25
compound	NX	comprise	0.33	transfer	NX	wafer_transfer_mechanism	0.50
compound	NX	compound	0.33	allow	VX	allow	1.00
control	NX	control	0.50	clean	VX	clean	1.00
control	NX	position	0.50	comprise	VX	comprise	1.00
mechanical	NX	CMP_method	1.00	condition	VX	position	1.00
method	NX	CMP_method	1.00	configure	VX	configure	1.00
polishing	NX	CMP_method	0.50	control	VX	control	0.50
polishing	NX	polish_pad	0.50	control	VX	position	0.50
slurry	NX	slurry	0.25	polish	VX	polish	1.00

Table 4. Partial terminology for the RFID patents.

Lemma	Chunk	Concept	Probability	Lemma	Chunk	Concept	Probability
accessory	NX	access	1.00	component	NX	device	0.15
activate	VX	active	1.00	computer	NX	device	0.50
actuator	NX	access	0.33	computer	NX	RFID_application	0.50
actuator	NX	encoding	0.33	concave	NX	antenna	1.00
actuator	NX	reader	0.33	conduct	VX	antenna	1.00
adapt	VX	antenna	1.00	conductor	NX	RFID_device	0.50
antenna	NX	antenna	0.90	conductor	NX	unit	0.50
antenna	NX	RFID_device	0.05	configure	VX	RFID_application	1.00
antenna	NX	tag	0.05	connect	VX	reader	1.00

The recall of an information system is defined as the ratio of the number of relevant documents returned to the total number of relevant documents in the collection. The recall is computed using Equation (14). The precision is the ratio of the number of relevant documents returned to the total numbers of retrieved documents in the collection. The precision is calculated using Equation (15). Accuracy is a well-known measure to evaluate classification performance. In our paper, we define accuracy as the ratio of the number of documents classified correctly selected from the total number of documents (Equation (16)). The *F*-measure is a combination of precision and recall used as a single measure of overall performance (van Rijsbergen 1979). The *F*-measure value is calculated using Equation (17).

$$\text{Recall} = \frac{A}{A + C} \tag{14}$$

$$\text{Precision} = \frac{A}{A + B} \tag{15}$$

$$\text{Accuracy} = \frac{A}{D} \tag{16}$$

$$F = \frac{1 + b^2}{\frac{b^2}{\text{Recall}} + \frac{1}{\text{Precision}}}. \tag{17}$$

Patent documents were retrieved from the WIPO Web site and the USPTO Web site. A total of 493 documents, including 343 CMP and 150 RFID patents, were downloaded for the case study, 233 CMP patent documents were used to train the semantic concept BPANN model, and 110 CMP patent documents were used to test the system. The results of CMP classification are shown in Table 5. Similarly, 100 RFID patent documents were used to train the semantic concept BPANN model and 50 RFID patent documents were used to test the system. The results of RFID classification are shown in Table 6.

In order to demonstrate accuracy, the proposed classification methodology is compared with different methodologies including the Legal Knowledge Management (LKM) System (Hsu *et al*. 2006), the Bayes System (Lee 2003), and the TF+Ontology ANN system (Trappey *et al*. 2006b). Hsu's system is a general-purpose system that calculates the term frequencies and co-variances of terms to classify patents. The Bayes-based system uses probability theory to classify patent documents. The TF+Ontology ANN method combines the counts of frequently appearing key phrases and additional semantic terms as ANN inputs to classify patent documents. The new system

Table 5. CMP patent document evaluation results.

IPC classification (sample sizes for testing)	*A*	*B*	*C*	Recall $A/(A+C)$ (%)	Precision $A/(A+B)$ (%)	*F*-measure
B24B(44)	40	4	3	93.02	90.91	0.9259
C09G(20)	19	1	1	95.00	95.00	0.9500
H01L(46)	44	2	2	95.65	95.65	0.9565

Table 6. RFID patent document evaluation results.

IPC classification (sample sizes for testing)	*A*	*B*	*C*	Recall $A/(A+C)$ (%)	Precision $A/(A+B)$ (%)	*F*-measure
G06K(6)	5	1	1	83.33	83.33	0.8333
G06F(24)	21	3	2	91.30	87.50	0.9052
G08B(13)	12	1	1	100	92.31	0.9836
H01Q(3)	3	0	1	75	100	0.7895
H04Q(4)	3	1	0	100	75	0.9375

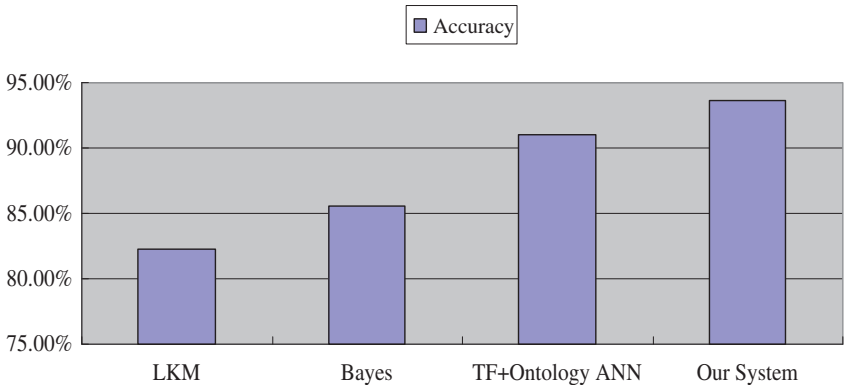

Figure 7. Accuracy comparison across different systems.

proposes an improved TF+Ontology ANN by taking the semantic concept probabilities of key phrases into consideration. The detailed algorithms of the previous approaches can be found in the previously referenced literature. The empirical comparison results of CMP and RFID patent classifications are shown in Figure 7. The experiments of our system and the comparison with previous approaches are conducted following the steps depicted in Figure 1. The steps are systematic, interactive and recursive, which usually run in background (offline) constantly or routinely to update the domain key phrases database and ANN classification model. Thus, the computer process speed is not the major concern in our algorithm development. Nonetheless, the methodology, after proving its superior accuracy in automatic classification comparing to other methods, can be further studied in shorten computer runtime during the coding and implementation stage. Although this paper does not emphasise the runtime aspect, it will be a good future research in regards to the next stage of system implementation efforts.

5. Conclusion

This research proposes an ontology-based semantic concept ANN approach for knowledge document self-categorisation and semantic-oriented search to help R&D and patent engineers derive valuable relevant technical and IPR knowledge quickly and accurately while working in a collaborative network. In order to demonstrate the methodology and evaluate its performance, the ontology schemas for CMP and RFID patents are used as case examples. The ontology schemas and the domain patents are uploaded into the system for BPANN model training and testing for classification and search. The TF-IDF method is used to extract key phrases in documents and to calculate the probability that a specific key phrase contains a certain semantic concept. Afterwards, frequencies and probabilities of key phrases are combined and used as input values to train the semantic concept BPANN. Finally, the test results show that the semantic concept BPANN increases the accuracy of knowledge document self-categorisation when compared with previous approaches. This study provides an improved solution for patent knowledge management, which is particularly valuable for knowledge self-organisation and semantic search which occurs during the complex networking of collaborative design processes. There are important future research directions which are not fully addressed in this paper. One is to develop an effective methodology to capture all ontological features from patents and other knowledge documents. Another research effort is to fully integrate all patent knowledge management functions, such as the smart patent classifier developed in this research, into the collaborative network solution suite for a comprehensive collaborative product development platform.

Acknowledgements

This research is partially supported by the National Science Council research grants. The authors thank the editors and reviewers for their valuable comments and suggestions for the manuscript revision.

References

Crowder, R., *et al*., 2010. An information system to support the engineering designer. *Journal of Intelligent Manufacturing*. doi: 10.1007/s10845-010-0458-4.

European Patent Office (EPO), 2007. *European Patent Office (Portal)*. Available from: http://www.epo.org/patents.html [Accessed 20 August 2007].

Farkas, J., 1993. Neural networks and document classification. *In*: *Canadian Conference on Electrical and Computer Engineering*, September 21–27, Vancouver, Canada.

Gurry, F., 2011. *The disclosure of technology in the patent system* (Published: 18 February 2011). Available from: http://www.wipo.int/about-wipo/en/dgo/speeches/dg_who_wipo_wto_med_11.html [Accessed 26 December 2011].

Hung, C. and Wermter, S., 2004. Neural network based document clustering using WordNet ontologies. *International Journal of Hybrid Intelligent Systems*, 1 (3), 127–142.

Hsu, F.-C., *et al*., 2006. Technology and knowledge document cluster analysis for enterprise R&D strategic planning. *International Journal of Technology Management*, 36 (4), 336–353.

Kantrowitz, M., Mohit, B., and Mittal, V.O., 2000. Stemming and its effects on TFIDF ranking. *In*: *Proceedings of the 23th International ACM SIGIR'00 Conference on Research and Development in Information Retrieval*, July 24–28, Athens, Greece (pp. 357–359).

Kao, C.C., 2000. *Personalized information classification system with automatic ontology construction capability*. Thesis (M.S.), National Cheng Kung University.

Karras, D.A., and Mertzios, B.G., 2002. A robust meaning extraction methodology using supervised neural networks. *In*: *Proceedings of Australian Joint Conference on Artificial Intelligence*, December 2–6, Canberra, Australia (pp. 498–510).

Kim, N.H., *et al*., 1999. Patent information retrieval system. *Journal of Korea Information Processing*, 6 (3), 80–85.

Ko, Y., and Seo, J., 2000. Automatic text categorization by unsupervised learning. *In*: *Proceedings of the 18th Conference on Computational Linguistics*, July 31–August 4, Saarbrücken, Germany (pp. 453–459).

Lam, W. and Han, Y., 2003. Automatic textual document categorization based on generalized instance sets and a metamodel. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (5), 628–633.

Lam, W., Ruiz, M.E., and Srinivasan, P., 1999. Automatic text categorization and its applications to text retrieval. *IEEE Transactions on Knowledge Data Engineering*, 11 (6), 865–879.

Lee, T.H., 2003. *XML-based content management platform using web services technique and UNSPSC standard classification schema*. Thesis (M.S.). National Tsing Hua University, Hsinchu, Taiwan.

Leeds City Council, 2011. *Patents and trademarks*. Available from: http://www.leeds.gov.uk/Business/Business_support_and_advice/Business_advice/Patents_and_trademarks.aspx [Accessed 20 December 2011].

Lin, S.H., 2005. *Integrating CMP of innovative knowledge and commercial knowledge in multi-agent prototype model*. Thesis (M.S.). National Taiwan University of Science and Technology, Taipei, Taiwan.

Liu, Y. and Harding, J., 2009. Editorial for the special issue of knowledge discovery and management in engineering design and manufacturing. *Journal of Intelligent Manufacturing*, 20 (5), 499–500.

Liu, Z.Q. and Zhang, Y., 2001. A competitive neural network approach to web-page categorization. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9 (6), 731–741.

Matsuo, Y. and Ishizuka, M., 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13 (1), 157–169.

MontyLingua, 2011. MontyLingua V.2.1 (Python and Java) – a free, commonsense-enriched natural language understander for English. Available from: http://web.media.mit.edu/~hugo/montylingua/) [Accessed 28 December 2011].

Porter, M.F., 1980. An algorithm for suffix stripping. *Program*, 14 (3), 130–137.

Romero, D., Galeano, N., and Molina, A., 2009. Mechanisms for assessing and enhancing organisations' readiness for collaboration in collaborative networks. *International Journal of Production Research*, 47 (17), 4691–4710.

Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Journal of Information Processing and Management*, 24 (5), 513–523.

Stefan, P., 2005. An ontology for the RFID domain. Available from: http://move.ec3.at/Ontology/RFIDOntology0903/RFIDOntology-Report.pdf.

Svingen, B., 1998. Using genetic programming for document classification. *In*: *Proceedings of the Eleventh International Florida Artificial Intelligence Research Society Conference*, May 18–20, Florida (pp. 63–67).

Trappey, A.J.C., *et al*., 2006a. Development of a patent document classification and search platform using a back-propagation network. *Expert Systems with Applications*, 31 (4), 755–765.

Trappey, A.J.C., Trappey, C.V., and Hsieh, E.C.H., 2006b. Automatic categorization of patent documents for R&D knowledge self-organization. *Journal of Management*, 23 (4), 413–424.

Trappey, A.J.C., *et al*., 2009. A fuzzy ontological knowledge document clustering methodology. *IEEE Transactions on Systems, Man and Cybernetics – Part B*, 39 (3), 806–814.

United States Patent and Trademark Office (USPTO), 2011. *U.S. Patent Statistics Report*. Available from: http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm [Accessed 25 December 2011].

van Rijsbergen, C.J., 1979. *Information Retrieval*. London: Butterworth.

World Intellectual Property Organization (WIPO), 2012. *International Patent Classification (IPC)*. Available from: http://www.wipo.int/ipcpub/#refresh=page [Accessed 28 January 2012].

Zhen, L., Jiang, Z., and Song, H.-T., 2011. Distributed knowledge sharing for collaborative product development. *International Journal of Production Research*, 49 (10), 2959–2976.