J Syst Sci Syst Eng (Mar 2009) 18(1): 071-094 ISSN: 1004-3756 (Paper) 1861-9576 (Online)
DOI: 10.1007/s11518-009-5100-7 CN11-2983/N

AUTOMATIC PATENT DOCUMENT SUMMARIZATION FOR COLLABORATIVE KNOWLEDGE SYSTEMS AND SERVICES

Amy J.C. TRAPPEY[1,2] Charles V. TRAPPEY[3] Chun-Yi WU[2]

[1]*Dept. of Industrial Engineering and Management, National Taipei University of Technology, Taiwan, China*
trappey@ntut.edu.tw

[2]*Dept. of Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan, China*
trappey@ie.nthu.edu.tw

[3]*Dept. of Management Science, National Chiao Tung University, Taiwan, China*
trappey@faculty.nctu.edu.tw (✉)

Abstract

Engineering and research teams often develop new products and technologies by referring to inventions described in patent databases. Efficient patent analysis builds R&D knowledge, reduces new product development time, increases market success, and reduces potential patent infringement. Thus, it is beneficial to automatically and systematically extract information from patent documents in order to improve knowledge sharing and collaboration among R&D team members. In this research, patents are summarized using a combined ontology based and TF-IDF concept clustering approach. The ontology captures the general knowledge and core meaning of patents in a given domain. Then, the proposed methodology extracts, clusters, and integrates the content of a patent to derive a summary and a cluster tree diagram of key terms. Patents from the International Patent Classification (IPC) codes B25C, B25D, B25F (categories for power hand tools) and B24B, C09G and H011 (categories for chemical mechanical polishing) are used as case studies to evaluate the compression ratio, retention ratio, and classification accuracy of the summarization results. The evaluation uses statistics to represent the summary generation and its compression ratio, the ontology based keyword extraction retention ratio, and the summary classification accuracy. The results show that the ontology based approach yields about the same compression ratio as previous non-ontology based research but yields on average an 11% improvement for the retention ratio and a 14% improvement for classification accuracy.

Keywords: Semantic knowledge service, key phrase extraction, document summarization, text mining, patent document analysis

1. Introduction

High technology firms maintain global competitiveness by quickly introducing innovative products that satisfy customers and build brand loyalty. If a conventional product development approach is used, there is a

tendency for research and development efforts to be split among groups with a resulting loss in information and knowledge sharing. A critical task for R&D groups is the analysis of patent documents and the synthesis of product knowledge. The management and analysis of patented technology provides competitive insights and better enables the formulation of legal strategies for the control of new inventions, processes, and products. In order to increase knowledge visibility and rapid sharing among R&D groups, a new approach is developed to facilitate patent document summarization. By summarizing patent documents into more concise text, knowledge can be readily categorized and managed. The improved use and storage of patent document knowledge increases the success of product lifecycle management, particularly during the design and redesign stages.

As societies become knowledge oriented and trade increases across borders, the need to guarantee the protection of intellectual property becomes unavoidable. Governments grant inventors the right to make, use, and sell new technological processes and devices. The grant of rights to protect an invention, i.e., a patent, provides the unique details of the invention and specifies exactly what is protected from copying or duplicating by competitors. Patent documents, as knolwedge documents, are also the key documents used for settling legal disputes (Kim et al. 1999). Technology oriented firms continuously analyze patent databases to protect their technology from infringement and to avoid duplicating the inventions of others. Emerging technologies stimulate the growth of large numbers of global patents. Analysis of these patents provides a means to gain competitive advantage. Firms that can best process and map the emergence of technologies and establish new boundaries ultimately profit from a time limited monopoly on the new invention.

The purpose of this research is to automate patent document summarization as a key step toward efficient patent analysis for knowledge services. Most of the existing patent document management systems use keyword matching as a search method to find related patents. Keyword searches often generate irrelevant results when words represent different meanings in different contexts or when the word indices do not preserve the relationships between words. Therefore, the goal of this research is to develop an ontology based methodology to analyze patents and abstract their content into shorter summaries for design teams to share during the collaborative design process.

2. Related Research

In this paper, we present the design, implementation, and evaluation of an automatic document summarization system using ontology based concept clustering. The related research includes ontology for semantic knowledge services, algorithms for document key phrase extraction, document summarization and abstraction, and text mining applications for patent analysis and synthesis.

2.1 Ontology

In the field of computer science, the definition of ontology has shifted from a focus on formal, explicit specifications of shared concepts (Gruber 1993) to an emphasis on determining a rigorous and exhaustive

organization of a knowledge domain. Wordnet from Princeton University (2006) defines an ontology as a hierarchical representation of a knowledge domain that contains all the relevant entities and their relations. A growing number of research studies refer to ontology as a means to represent knowledge that is characterized by multiple concepts with the relationships between the concepts also represented.

Rodrigues et al. (2008) noted that an ontology can be mapped using computer based tools thus enabling the knowledge domain to be standardized and shared. The Oracle 10g tool for example, uses ontology for metadata description, domain model specification, and data integration. As Bobillo et al. (2008) indicated, ontologies create patterns for representing relations between context descriptions and domain subsets. The ontological pattern can be utilized to enable reasoning, analysis, and reuse of domain knowledge.

Recent publications and empirical studies show that knowledge bases constructed from documents can used to model the meaning of the text. Since the meaning of the text is not immediately obvious from the words or phrases, the semantic patterns extracted from the text are used for writing summaries and analyzing the meaning as constrained by the ontology. According to Buitelaar et al. (2008), an ontology can also be used as a database schema, which formalizes the data relevant for summaries and expresses meaning according to a vocabulary. The ontological schema provides valuable background knowledge and represents the causal relationships in the text.

Zheng et al. (2008) demonstrated how to represent domain knowledge with concepts and relations that are weighted by relevance scores that are in turn used to judge whether or not a summary is relevant to a specific topic. Experts annotate key phrases so that meaningful relationships between words can be inferred. Therefore, the ontology describes the semantics of the information sources and makes the content explicit.

Finally, Joung and Chuang (2009) provide an extensive discussion of the applications for peer to peer (P2P) development of ontologies. These ontologies are derived among peers that identify words and phrases that facilitate the processing and extraction of text for summaries. The ontology based approaches adopted by these researchers are shown in Table 1.

2.2 Keyword Extraction

Within the field of text summarization, a gradual but marked change in the area of keyword extraction research has taken place. Researchers commonly accept keyword extraction as a fundemental process in document context analysis and information retrieval. Keywords model dominant patterns of words within documents and key phrases represent a series of keywords. Therefore, keyword extraction is particularly beneficial for text summarization. Keywords are often extracted from text using a corpus approach, a lexical approach, or a statistical approach.

The corpus approach uses a predefined corpora or dictionary to extract the key words and phrases within subject documents. This approach is easy to implement but requires time and effort to build and maintain the latest corpora. As Chung and Nation (2004) have noted, the technical vocabulary is defined by

Table 1 Comparison with corpus based approaches

Author	Ontology Type	Software Application	Contribution
Rodrigues et al. (2008)	A series of important concepts	Oracle JDeveloper 10g	Systematic data integration
Bobillo et al. (2008)	Models of context knowledge	Document context analysis, discovery of ontological patterns	The creation of a context-domain relevance model
Buitelaar et al. (2008)	Formalization of relevant data	Ontological database schema	Extraction and integration of causal text relations
Zheng et al. (2008)	Computation of text relevance scores	Focused crawling	The classification of web pages
Joung and Chuang (2009)	Peer definition of meaningful relationships	Peer to Peer (P2P) network	Semantic search

experts who have knowledge of the subject domain. The expert derived corpora is used to mark which words are important for retrieval and provides contextual links to related terminology. Tseng et al. (2007) report that the most important feature of the methodology is the rigorous approach used to verify the usefulness of the extracted text segments. A keyword analysis method using a corpora can split text into a series of key phrases with equivalent semantics, and then aggregate these phrases as a summary. The corpus based extraction uses sentences composed of key phrases as the smallest units for summarization. In addition, the corpora does not extract unrelated proper nouns or undefined terms. However, when the scale of the corpus is large, the speed of term mapping decreases.

The lexical approach analyzes keywords using natural language processing grammar programs. The programs are also based on corpora or dictionaries. According to Ercan and Cicekli (2007), the lexical metadata hold a series of related words that represent the semantic context of the documents. This approach assumes that keywords of text should be semantically related with the concepts of the text. The number of keywords and semantic relations among the keywords can be different for each metadata set. Using semantic word features based on the lexical approach determines important keywords in the document. The coverage and size of the lexical metadata file can indicate how well the keywords represent the concepts of the text.

Statistical approaches such as Term Frequency (TF) measure the relative frequency of words appearing in a document. If a word appears with a higher frequency, then the system considers the word more important than other words in the document. Therefore, this approach is considered to be a simple way to compute the scores and to weigh the importance of sentence content (Luhn 1957). In general, the importance of a keyword in a document depends on two factors. TF is the relative frequency of the appearance of the keyword in the document while Document Frequency (DF) measures the number of documents containing the word. DF represents the proportion of documents containing the word among all documents. If the frequency of a keyword is lower in all other

documents rather than the current document, then this keyword represents the current document better than other documents. Greiff (1998) demonstrated that Inverse Document Frequency (IDF) can be interpreted as representing a reversely normalized frequency of a given keyword across all documents in a set. As a weighting factor, IDF is useful for determining whether a keyword can describe and represent a specific document. Salton and Buckley (1988) proposed a framework for combining TF with IDF, and the Term Frequency - Inverse Document Frequency (TF-IDF) approach is one of the most commonly used term weighting schemes. TF-IDF represents the word frequency in the document normalized by the domain frequency as defined below:

$$w_{jk} = tf_{jk} \times idf_j \quad (1)$$

w_{jk} = phrase weight of phrase j in document k

tf_{jk} = the number of phrases j that occur in document k

$$idf_j = \log_2\left(\frac{n}{df_j}\right) \quad (2)$$

n = the total number of documents in the document set

df_j = the number of documents containing the phrase j in the document set

Aizawa (2003) identified two computational problems with the TF-IDF approach. One problem is determining the specificity of a term within a given document set and another is reducing the features of the documents. However, if the target is to automatically construct the frequency of terminology within a series of documents using a specific domain corpora, then the approach can be applied effectively. In addition, Hu (2006) discusses a machine readable approach for the extraction of titles from general documents in order to construct a more meaningful representation of a specific knowledge category. Finally, Zhang et al. (2008) noted that using a sequence of two or more words as multi-words with regular collocations also effectively improve the representation of text.

2.3 Text Summarization

A series of brief and succinct statements that represent the main concepts of text in a concise form are described as a summarization. Text summarization has been widely investigated since Luhn (1958) indicated that the most frequently used words represent the most important concepts of the text. With the abundance of available information, people are over whelmed with the problem of extracting useful segments of text and producing summaries. Mani & Maybury (1999) formally defined automatic text summarization as the process of concentrating the most important information from subject sources to generate a retrenched version for a specific task and demand.

As Ye et al. (2007) demonstrated, the automatic summarization process can be simplified by choosing a representative subset of sentences from the original documents. However, summarization requires semantic representation, lexical inference, and natural language processing, areas of research which have yet to reach a mature stage of development. The processes of text summarization was

decomposed into three phases by Mani & Maybury (1999). The first phase analyzes the input subject sources and chooses a set of salient features. Next, the results are transformed into a concise form by ranking the importance of features and selecting those features with higher scores and minimum overlap. Finally, the synthesis phase takes the brief representations and produces a summary. The quality of the summary is evaluated using a compression statistic to evaluate the ratio between the length of the summary and the length of the original text.

There have been many advances in automatic text summarization over the last five decades (Table 2). Pioneering work from the 1950's through the 1960's focused on the heuristic study of term frequency, sentence location, and lexical cues. Luhn (1958) was the first to use term frequencies to extract text for summarization. Edmundson (1968) noted that the first paragraph or first sentences of each paragraph often contain important topic information which further improves summaries. During 1970s to 1980s, the use of frames, templates and complex text processes based on artificial intelligence identified the main concepts from a text and were used to extract relationships between concepts. However, the drawback of artificial intelligence is that the frames or templates used may lead to the incomplete analysis of concepts from the text sources.

The latest research concentrates on information retrieval using symbolic level analysis. Characteristic text units are identified with hybrid approaches and semantic representation using synonyms, polysemy and antonyms. According to Morris & Hirst (1991), lexical cohesion provides semantic information which helps determine the meaning of concepts. Goldstein et al. (1999) provided a query based summarization approach to extract criteria based on the number of words co-occurring in the query and sentence. Ko et al. (2003) constructed lexical clusters which have different semantic categories. The topic keywords from each cluster are then used to create a text summary based on semantics.

Table 2 Automatic text summarization research

Phases	Model	Description
1950s ~1960s	Surface Level Heuristics	Focuses on the study of text genres using term frequency, sentence location and lexical cue phrases. (Luhn 1958; Edmundson 1969)
1970s ~1980s	Discourse Based Entity Level Artificial Intelligence	Uses frames, templates and complex text processes to identify main text concepts and extract relationships between concepts or entities using logic, production rules, and scripts. (Young and Hayes, 1985; Fum, Guida, & Tasso, 1985)
1990s ~2000s	Knowledge Based Information Retrieval	Symbolic level analysis which finds characteristic text units using hybrid approaches and semantic representations. (Morris & Hirst 1991; Kupiec et al. 1995; Aone et al. 1997; Salton et al. 1997; Hovy and Lin 1999; Goldstein et al. 1999; Mani and Bloedorn 1999; Gong and Liu 2001; Yeh et al. 2002; Ko et al. 2003)

The use of hybrid approaches in text summarization has increased in recent years. Yeh et al. (2005) provide a modified corpus based approach and latent semantic analysis based on text relationship maps for automatic text summarization. This approach ranks sentence positions and uses these positions and a scoring function trained by a genetic algorithm to generate summaries. Yeh et al. (2008) extended this research using a sentence ranking method to model a set of topic related documents as a sentence based network. The importance of a node in a lexical network (with a node representing a sentence in the text) is determined by the number of nodes to which it connects and the importance of the connecting nodes. Spreading activation theory is then applied to recursively weight the importance of sentences by spreading their sentence specific feature scores through the network to adjust the importance of other sentences. Consequently, a ranking of sentences indicates the relative importance of the sentences.

Reeve et al. (2007) introduced an extractive summarization method which identifies domain specific concepts and synonymous phrases associated with the concepts. Their approach better enables users to quickly find relevant sources and extract essential information with reduced effort. Lin and Liang (2008) offer a general theoretical basis for four types of text summarization methods. The surface level approach relies on shallow features to extract information including thematic features, location, background, and cues words. The discourse based approach models the format of the document, rhetorical structure of the text, and the relation of patterns using communicative rules. The entity level approach builds an internal representation of the text and analyzes the relationship according to similarity, proximity, and co-occurrence of entities. The knowledge based approach depends on domain knowledge of the concepts and the relationship between concepts.

2.4 Text Mining Applications in Patent Analysis

Patents are an official representation of intellectual property ownership and consist of over 50 million documents contained in a diverse range of databases. The largest databases are maintained by the World Intellectual Property Organization, the State Intellectual Property Office of the P.R.C., the United States Patent and Trademark Office, the European Patent Office, the Korean Intellectual Property Office, and the Japanese Patent Office. Patent analysis or mapping requires significant effort and expertise because these documents are lengthy and include numerous technical and legal terms. Research and development engineers use keywords (e.g., specific terms, technology classifications, and inventors) to retrieve patent documents and then manually review the documents to extract information and write summaries. Automated technologies for assisting analysts in patent processing and analysis are emerging. A multi-channel legal knowledge service center named the Legal Knowledge Management platform (LKM) was implemented by Hsu et al. (2006) to integrate knowledge management and data mining techniques. After patent documents are uploaded to the LKM platform, keywords are automatically extracted and categorized for

analysis.

Tseng et al. (2007) notes that the general directions of text mining for patent analysis requires document preprocessing, indexing, topic clustering, and topic mapping. The steps for document preprocessing include document parsing, segmentation, and text summarization. Moreover, the text mining methodologies focus on keyword and phrase extraction, morphological analysis, stop word filtering, term association, and clustering for indexing. The topic clustering process includes term selection, document clustering and categorization, cluster title generation, and category mapping. Patent clustering techniques provide valuable insight into the existing relationships between different categories of patent documents. Thus, a clustering approach applied to patent text mining is considered beneficial in a business environment where patent information is used to resolve legal issues and generate competitive intelligence (Fattori et al. 2003). Trappey et al. (2006) developed an electronic document classification and search method applying text mining and neural network technologies to classify patent documents. Blanchard (2007) provided algorithms for automatically creating customized stop word lists which are combined with lexical analysis for automatic indexing and information retrieval.

There has recently been an interest in applying text mining techniques to the task of patent analysis as a means to identify significant information for text summarization. For example, keywords can be used for patent information retrieval but this approach has several limitations (Li et al. 2008). Trappey and Trappey (2008) propose a non-ontology based document summarization method which uses keyword recognition plus significant information density (keywords, titles, semantic phrases with high relevancy to keywords, topic sentences, domain-specific phrases, and indicator phrases) to extract paragraphs which are judged to best represent the patent. The goal of text mining patents is to identify content that is most relevant to the research task. Patents use complex and detailed explanations to legally protect the boundaries of the intellectual property claimed by the owner. Our research uses an automatic ontology based patent document summarization approach to help engineers process patent knowledge and share this knowledge with others so as to enhance collaboration. Engineers that better understand the boundaries of the competitors intellectual property will be less likely to waste time re-inventing technologies that are protected by legal claims and will be more likely to discover opportunities for research and development.

3. The System Methodology

The system platform proposed in this paper has two parts. One part extracts key words and phrases. The other part provides the document summary. The patent document summarization algorithm is shown as a procedural flow in Figure 1 where PHT represents Power Hand Tools and CMP represents Chemical Mechanical Polishing. In addition to the text summary, the system uses a comparison mechanism based on a specific ontology defined by domain experts. The system automatically marks, annotates, and highlights the nodes of the ontology tree which correspond to words in the patent document and provides a visual diagram to depict relationships

between words.

3.1 Key-Phrase Extraction

After preliminary processing to divide paragraphs into sentences and divide sentences into words using stop words (blank spaces, punctuation marks, and symbols), the system outputs key phrases automatically using two extraction algorithms. The first extraction algorithm uses the TF-IDF method (Salton and Buckley, 1988). The algorithm measures term information weight and extracts the top key words

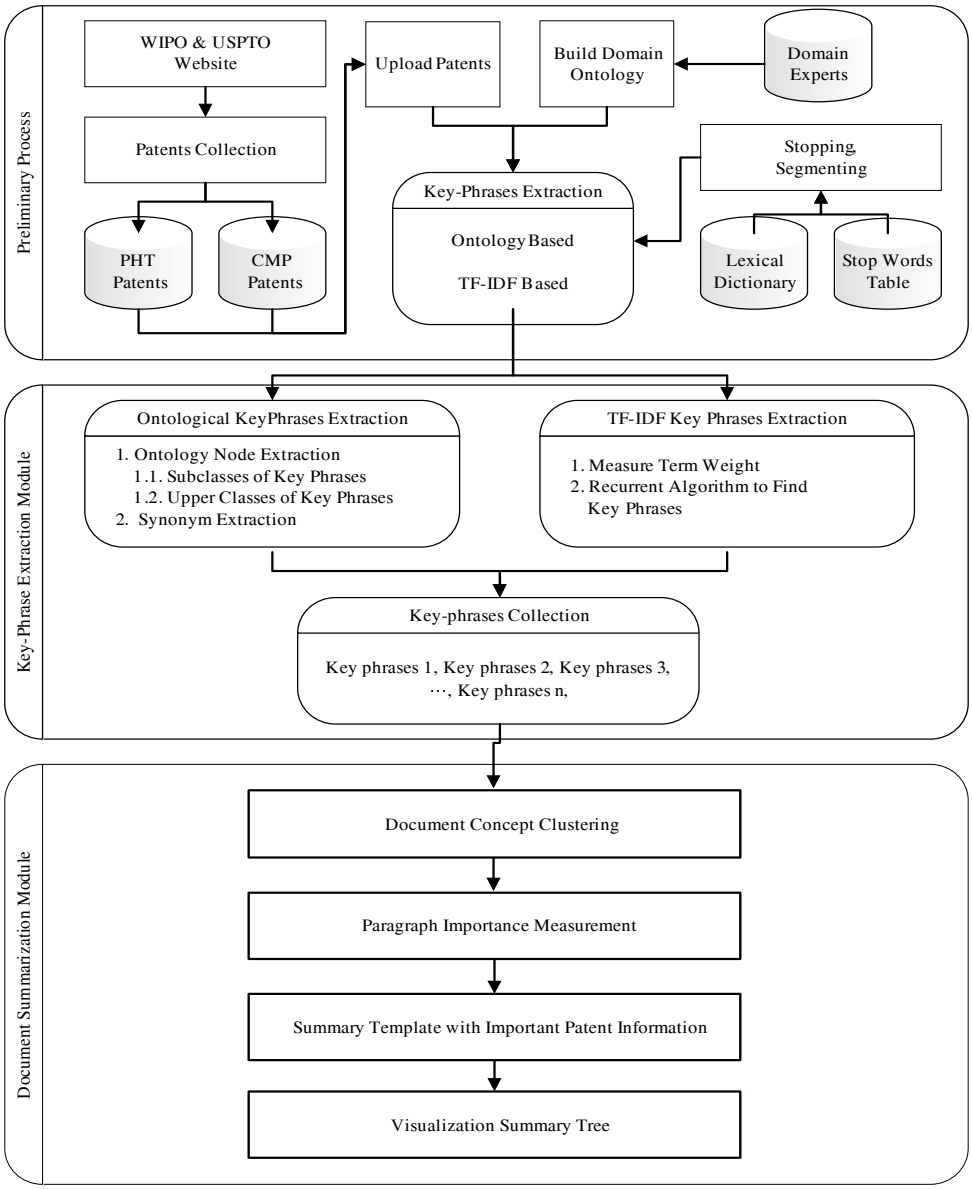

Figure 1 The patent document summarization procedure

or phrases in accordance with their TF-IDF weight. The second extraction algorithm uses a specific ontology embedded in the system to calculate the frequency of mapping words. MontyLingua is used to identify the verbs and nouns while eliminating prepositions and articles. After deriving the vocabulary, the system calculates the occurrence of verbs and nouns and ranks them by frequency. The ontology is built by the domain experts to describe the concepts in vocabularies and determine the linking relationships between vocabularies. The ontology contains descriptions of classes, properties and their instances. This set of terms includes the vocabulary, the semantic interconnections, and suitable rules of inference and logic for the particular domain. The ontology differs from TF-IDF methods in that it represents the relation of classes in patent documents. Through the relationship, the system derives the key phrases including the subclasses of key phrases. For example, the word *implementation* has a subclass phrase *motion unit* which in turn has a subclass *power source* and *electricity*. The upper classes of the key phrases are defined (e.g., *implementation* is the upper class of *mechanism*) as shown in Figure 2. Thus, the semantic based extraction reduces the number of related words and facilitates the construction of ontologies by refining and assembling components. The system acquires key words and phrases from both the ontology based and TF-IDF based methods and removes key words or phrases in common as shown in Table 3. Finally, a list of key words and phrases is derived and is used as the input and foundation for the summary.

3.2 Summary Representation

The summary representation procedure is show in Figure 3. The procedures include key phrase extraction, the creation of the paragraph and key word frequency matrix, the computation of paragraph importance, document concept clustering, and the building of the visualization summary tree.

First, similar concepts in different paragraphs are placed in the same cluster. The concept clustering process is depicted in Figure 4. Documents consist of several important concepts

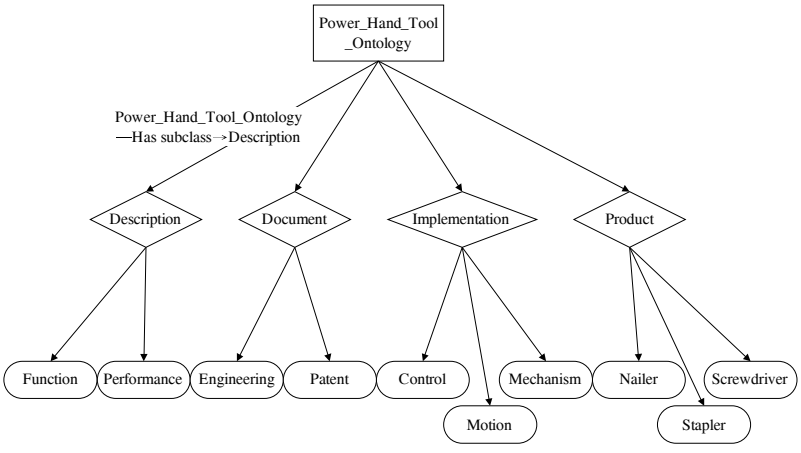

Figure 2 Graphical tree for power hand tool ontology

Table 3 A list of extracted key phrases from a hand tool patent (Brown 2006)

No.	Phrases	TF	Technique
1	striker	43	Ontology
2	shaft	35	Ontology
3	block	32	TF-IDF
4	stapling device	32	TF-IDF
5	cap plate	25	TF-IDF
6	channel	20	TF-IDF & Ontology
7	slide	19	TF-IDF
8	spring	19	Ontology
9	hammer	10	Ontology
10	indentation	9	TF-IDF

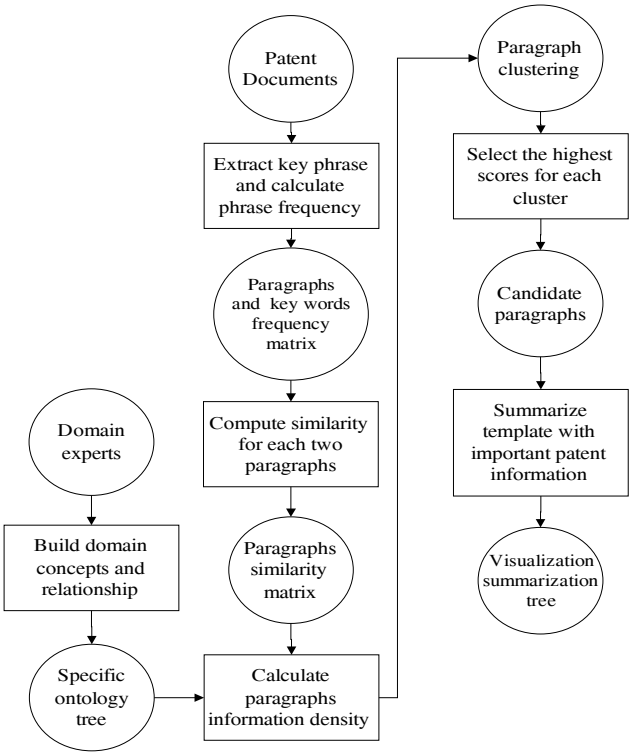

Figure 3 The summary representation procedure

or main topics that exist in different paragraphs. For this reason, a clustering algorithm is utilized to gather similar concepts or main ideas.

On the basis of the key words extracted, the frequency of each key word appearing in each paragraph are tallied using formula 3 and 4, and the similarity between each paragraph pair is computed using formula 5 and 6. The paragraphs and keywords term frequency matrix is shown below:

$$A = matrix([a[1,1],\cdots,a[1,n], a[2,1],\cdots,a[m,1],\cdots,a[m,n]]) \quad (3)$$

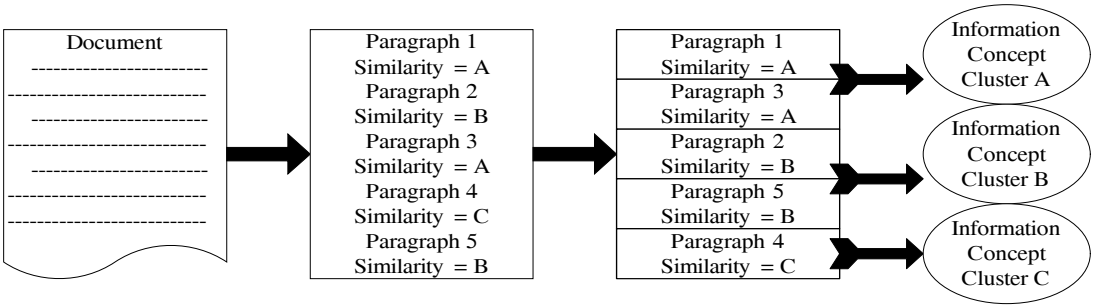

Figure 4 Document concept clustering process

Set

$$\left[tf_{ij}\right]_{m\times n} = \begin{bmatrix} tf_{11} & tf_{12} & tf_{13} & tf_{14} & \cdots & tf_{1n} \\ tf_{21} & tf_{22} & tf_{23} & tf_{24} & \cdots & tf_{2n} \\ tf_{31} & tf_{32} & tf_{33} & tf_{34} & \cdots & tf_{3n} \\ tf_{41} & tf_{42} & tf_{43} & tf_{44} & \cdots & tf_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ tf_{m1} & tf_{m2} & tf_{m3} & tf_{m4} & \cdots & tf_{mn} \end{bmatrix} \quad (4)$$

The concept tf_{ij} = the number of key words j that occur in ith paragraph of the document.

Row vector, $Paragraph_i = \left[tf_{i1}\ tf_{i2}\cdots tf_{in}\right]$ = the number of key words that occur in ith paragraph of the document, key words from 1 to n.

Column vector, $Keyword_j = \begin{bmatrix} tf_{1j} \\ tf_{2j} \\ \vdots \\ tf_{mj} \end{bmatrix}$ = the number of key words j that occur in the paragraphs of the document, paragraphs from 1 to m.

The paragraph similarity matrix is shown below:

Normalizing the paragraph vector,

$$NParagraph_i = \left[tf_{i1}/tf_{\max}, tf_{i2}/tf_{\max}, \cdots, tf_{in}/tf_{\max}\right],$$
$$tf_{in}/tf_{\max} \in [0, 1]$$

$tf_{\max}$ = the highest value of term frequency of keywords in the paragraphs of the document.

Set

$$a_{ij} = Sim(NParagraph_i, NParagraph_j) \quad (5)$$

$$Sim\left(NParagraph_i, NParagraph_j\right) = \frac{NParagraph_i \cdot NParagraph_j}{\left|NParagraph_i\right| \ \left|NParagraph_j\right|} = \frac{\sum_{k=1}^{n} tf_{ik} tf_{jk} \Big/ tf_{\max}^2}{\sqrt{\sum_{k=1}^{n} tf_{ik}^2 \Big/ tf_{\max}^2 \sum_{k=1}^{n} tf_{jk}^2 \Big/ tf_{\max}^2}}$$

$$\begin{cases} NParagraph_i \\ = \left[tf_{i1}/tf_{\max},\ tf_{i2}/tf_{\max},\ \ldots,\ tf_{in}/tf_{\max}\right] \\ = i\text{th paragraph in the document}, i = 1 \sim m \\ NParagraph_j \\ = \left[tf_{j1}/tf_{\max},\ tf_{j2}/tf_{\max},\ \ldots,\ tf_{jn}/tf_{\max}\right] \\ = j\text{th paragraph in the document},\ j = 1 \sim m \end{cases}$$

$Sim(Paragraph_i, Paragraph_j)$ is between 0 and 1. The highest paragraph similarity value, $a_{ij} = 1$, represents that the paragraphs are the same, hence the diagonal a_{ij} of paragraph similarity matrix are

$$\underset{i=j}{Sim}(Paragraph_i, Paragraph_j) = 1.$$

$$A=\left[a_{ij}\right]_{m\times n}=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & a_{34} & \cdots & a_{3n} \\ a_{41} & a_{42} & a_{43} & a_{44} & \cdots & a_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & a_{m4} & \cdots & a_{mn} \end{bmatrix} \tag{6}$$

The greater the a_{ij} value, the greater similarity of these two paragraphs. After computing the similarity value, the paragraph similarity matrix is constructed, and the paragraphs are clustered using the K-means algorithm (Han & Kamber, 2000). The Euclidian distance (d_E) and the total deviation (E) for the K-means algorithm are defined as follows:

$$d_E\left(x_i-x_j\right)=\left(\sum_{d=1}^{k}\left|x_i-x_j\right|^2\right)^{1/2} \tag{7}$$

$$E=\sum_{i=1}^{k}\sum_{x\in S_i}\left\|x-m_i\right\|^2,\quad m_i=\frac{\sum_{x\in S_i}\vec{x}}{\left|S_i\right|} \tag{8}$$

x = input paragraph

m_i = mean of cluster S_i

$\left|S_i\right|$ = the number of data in cluster S_i

A more complete discussion of the proximity functions and the K-means algorithm are provided by Wu et al. (2007). In our research, the paragraphs are fairly balanced across clustering results and the basic K-means algorithm is used for paragraph clustering. Finally, the Root Mean Square Standard Deviation (RMSSTD) and R-Squared (RS) are used to judge the optimal number of clusters for the given data set (Sharma, 1996). RMSSTD represents the minimum variance in the same cluster and RS describes the maximum variance between different clusters. Based on the RMSSTD and RS correlation coefficients, the optimal number of clusters (k) is chosen. Since RMSSTD is the standard deviation of all the variables from the cluster and represents the homogeneity of the clusters, the RMSSTD for each cluster should be as small as possible to derive optimal results. RS equals the sum of squares between groups divided by total sum of squares for the whole data set. In order to obtain a satisfactory cluster result, the cluster with the largest RS value is chosen. RMSSTD and RS are defined as follows:

$$RMSSTD=\left[\frac{\sum_{\substack{i=1,\ldots,nc\\ j=1,\ldots,v}}\sum_{k=1}^{n_{ij}}\left(x_k-\overline{x_k}\right)^2}{\sum_{\substack{i=1,\ldots,nc\\ j=1,\ldots,v}}\left(n_{ij}-1\right)}\right] \tag{9}$$

nc = the number of clusters

v = the number of variables (data dimension)

n_j = the number of data values of j dimension

n_{ij} = the number of data values of j dimension that belong to cluster i

$$RS=\frac{SS_b}{SS_t}=\frac{SS_t-SS_w}{SS_t}$$

$$=\frac{\left\{\sum_{j=1,\ldots,v}\left[\sum_{k=1}^{n_j}\left(x_k-\overline{x_k}\right)^2\right]\right\}-\left\{\sum_{\substack{i=1,\ldots,c\\ j=1,\ldots,v}}\left[\sum_{k=1}^{n_{ij}}\left(x_k-\overline{x_k}\right)^2\right]\right\}}{\sum_{j=1,\ldots,v}\left[\sum_{k=1}^{n_j}\left(x_k-\overline{x_k}\right)^2\right]} \tag{10}$$

SS_w = the sum of squares within group

SS_b = the sum of squares between groups

SS_t = the total sum of squares of whole data set

Table 4 shows the RMSSTD and RS values for the possible numbers of clusters for a US stapling device hand tool patent (Brown, 2006).

Following the formulas and logic above, grouping the patent paragraphs into 7 clusters provides the best solution.

3.3 Paragraph Importance and Scores

This research collects two sets of patent documents as a case study for automatic summarization. The content of each patent includes the summary of the invention, a brief description of drawings, the detailed explanation of the patent, and claims for legal protection. Since patent documents contain specific technology and engineering descriptions, the proposed system requires domain experts in the area of power hand tools and chemical mechanical polishing to construct the domain ontology of specific concepts. The importance measure of each paragraph is related to its information content so the system utilizes the ontology based and TF-IDF approach to derive the summary (Trappey and Trappey 2008) as shown in Figure 5.

The ontology provides the specifications used for information expression, integration, and system development. The vocabulary for the specific domain and the rules for combining terms and relations are also defined. If the extracted key words and phrases correspond to a node of the ontology, it is essential to select the correct position of the node for mapping the importance weights as leaves and branches of the ontology tree (Hsu 2003). The nodes which are furthest from the root nodes provide the greatest information as shown in Figure 6.

There are four levels in the ontology tree. Using (Implementation→Mechanism→Turbine →Blade) as an example, then blade is a component of a turbine, turbine is a mechanism, and mechanism belongs to an implementation of the power hand tool. Thus, the deeper the position

Table 4 The RMSSTD and RS correlation coefficients for clustering the paragraphs of US patent 7,014,088 B2

Number of clusters (k)	2	3	4	5	6	7	8
RMSSTD	4.59	2.29	1.53	1.14	0.91	0.76	0.78
RS	3.98	3.98	3.98	3.98	3.98	3.98	3.98
						Target	

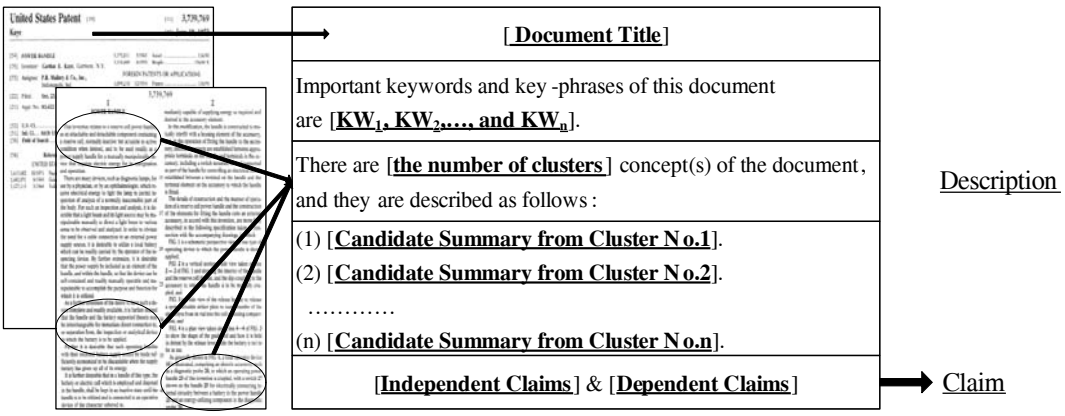

Figure 5 Text summary template

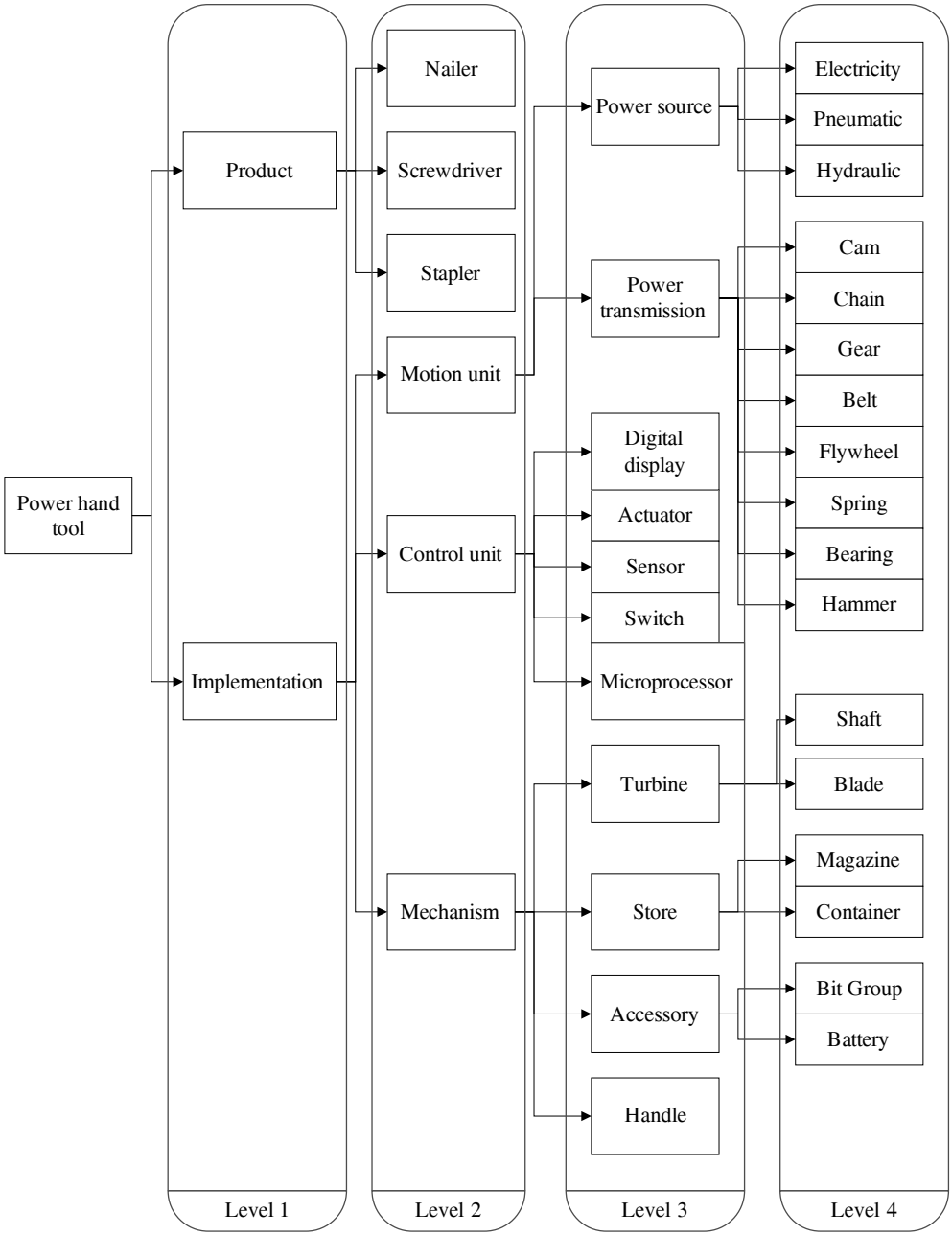

Figure 6 Weight levels for ontological nodes

of the node, then the more detailed the information. When words or phrases in a paragraph match the nodes of the ontology tree, it is necessary to record not only term frequency but also the mapping weights. Formula 11 weighs the scores derived from ontology as shown below:

$$\sum_{i=1}^{N} N_{Ci,CLm} W_{ih} \tag{11}$$

W_{ih} = the weight of node i on h depth.

$N_{Ci,CLm}$ = the frequency of key words and phrases extracted using the ontology in paragraph m.

N = the total number of ontology nodes.

Kupiec et al. (1995) showed that corpus based methods should use different heuristic sentence features and then calculate the Baysïan probability of each sentence assuming statistical independence of the features. The corpus requires a Golden Summary, a summary created by experts and used to calculate the probability values. The Baysïan probability formula is defined as:

$$P(s \in S \mid F_1, F_2, \cdots, F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) P(s \in S)}{\prod_{j=1}^{k} P(F_j)} \tag{12}$$

s = summary created by the summarization system

S = Golden Summary created by the experts

F_j = sentence feature j in the summarization system

Based on the research of Teufel and Moens (1997), subtitles and the opening words of each paragraph fairly approximate the content of documents. The most significant and meaningful sentences usually occur at the beginning and end of a document. Lin & Hovy (1997) argue that important sentences are located at positions which are dependent on the genre and these positions can be determined automatically through an approach called the Optimal Position Policy (OPP). However, the OPP technique reqires a large number of documents from a specific corpora for successful application. As shown by the experiments of Lorch et al. (2001), the phrases appearing in an article title or the paragraph headings (or subtitles) are assigned greater weight for representing key information in the document. Further, Lam-Adesina and Jones (2001) show that the title information contained in a sentence can be scored as follows:

$$\textit{Setence Score} = \frac{\textit{The Total Number of Title Terms Found in a Sentence}}{\textit{The Total Number of Terms in a Document Title}} \tag{13}$$

According to key words and phrases extracted using the TF-IDF method, the system records the frequency of occurrence in a paragraph. Since the length of a paragraph is related to the possibility of key words or phrases appearing in the paragraph, the paragraph information density is used to rank the importance of each paragraph. After S_{CLm} is calculated for each paragraph in each cluster, the system derives the highest scoring paragraphs for each cluster. These paragraphs are regarded as the candidate summaries, but the paragraph scores must be higher than the average scores of the whole paragraph. Therefore, the proposed system considers how the length of a paragraph affects the possibility of key words or phrases appearing in the paragraph, and the paragraphs with the highest scores in each cluster are used to form the candidate summary set as shown in Figure 7. The total score for paragraph m is calculated using equation 14:

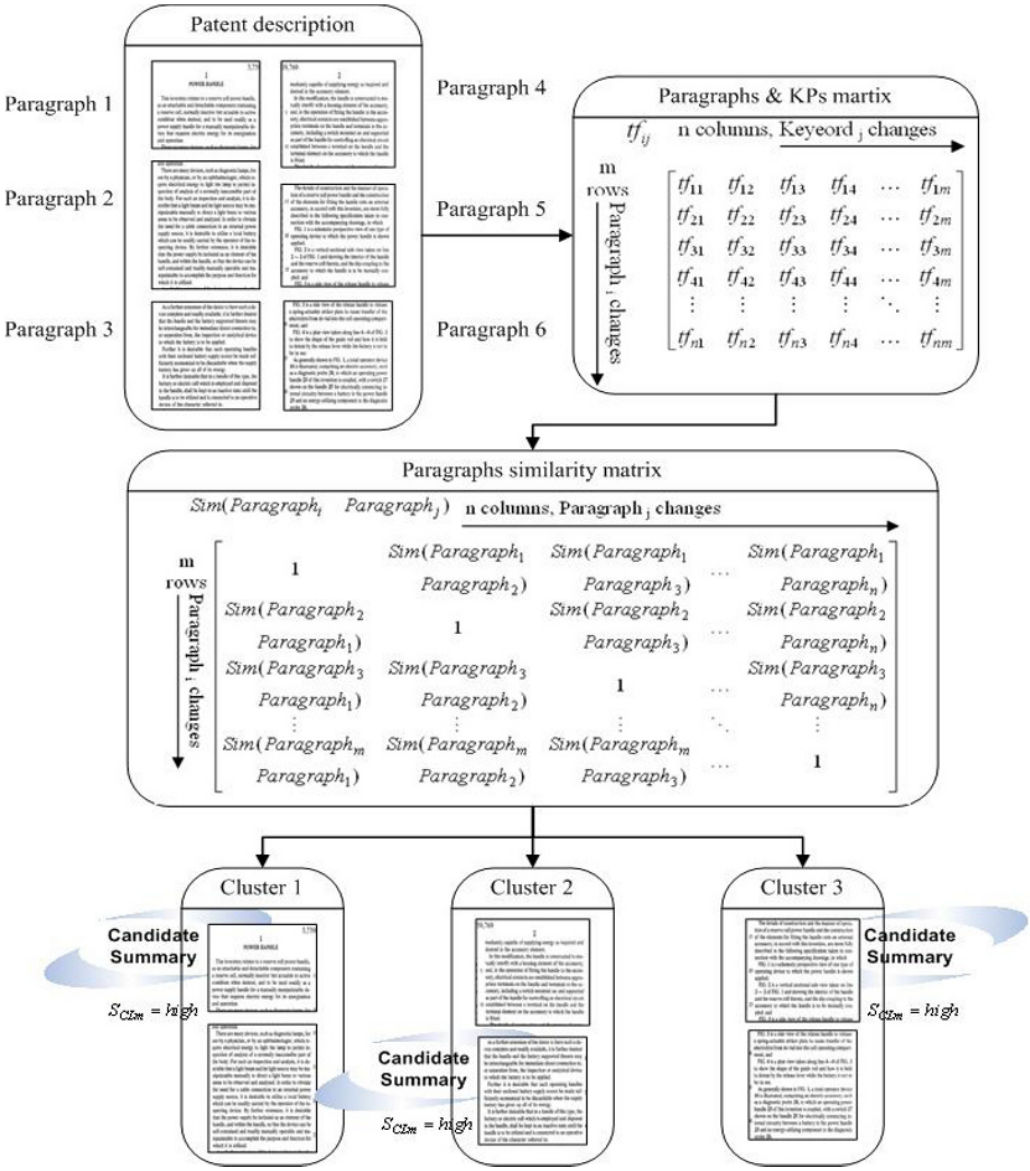

Figure 7 Selecting the highest scores in the concept cluster as the candidate summary

$$S_{CLm} = \frac{\left(\sum_{i=1}^{N} N_{Ci,CLm} W_{ih}\right) + f_{TF}}{TM_m} \qquad (14)$$

TM_m = term number in paragraph m

f_{TF} = the frequency of key words and phrases which are extracted by TF-IDF method in paragraph m

$\sum_{i=1}^{N} N_{Ci,CLm} W_{ih}$ = the scores from key words and phrases which are extracted by ontology-based method in paragraph m

4. Evaluation

In our research, each patent contains paragraphs grouped into clusters using the K-means algorithm (Table 5). The paragraph with the highest information density is selected as the candidate to represent each cluster. Then, all of candidate paragraphs and key information concepts are integrated to form the patent summary.

The system uploads, automaticly clusters paragraphs, and then selects candidate paragraphs for summarization. Measures for the patent compression ratio, the information density retention ratio, and patent classification accuracy were used for experimental validation. Two hundred full text patent documents were retrieved from the World Intellectual Property Organization (WIPO) and United States Patent and Trademark Office (USPTO) electronic databases. The patents represent technology claims in the areas of power hand tool and chemical mechanical polishing. The classifications for PHT include hand-held nailing or stapling tools, percussive tools, and combination or multi-purpose tools. The classifications for CMP include polishing compositions, machines, devices, processes for grinding or polishing, and semiconductor devices. The test patent classifications are shown in Table 6 and Table 7.

Table 5 Seven paragraph clusters are created for US patent 7,014,088 B2

Clusters	Paragraphs	Information density
Cluster 1	paragraph 16	0.1 (target)
	paragraph 48	0.01
Cluster 2	paragraph 19	0.16 (target)
	paragraph 26	0.14
Cluster 3	paragraph 06	0.14
	paragraph 23	0.16 (target)
	paragraph 50	0.14
	paragraph 58	0.12
Cluster 4	paragraph 39	0.04
	paragraph 40	0.07 (target)
Cluster 5	paragraph 49	0.14 (target)
Cluster 6	paragraph 59	0.07 (target)
Cluster 7	paragraph 60	0.11 (target)

Table 6 Patent classifications for power hand tools

IPC Classification	Description	Quantity Uploaded
B25C	Hand held nailing or stapling tools; manually operated portable stapling tools	33
B25D	Percussive tools	32
B25F	Combination or multi-purpose tools not otherwise provided for; details or components	34

Table 7 Patent classifications for chemical mechanical polishing experiment

IPC Classification	Description	Quantity Uploaded
B24B	Polishing compositions other than French polish; ski waxes.	35
C09G	Machines, devices, or processes for grinding or polishing.	31
H01L	Semiconductor devices; electric solid state devices not otherwise provided for other category.	34

Table 8 Experimental results for compression, retention, and accuracy

Evaluation	Domain	This research (Ontology-Based)	Non-ontology based approach	Difference*
T 1. Compression Ratio	PHT	16%	23%	7%
	CMP	19%	11%	-8%
	Average	18%	17%	-1%
T 2. Retention Ratio	PHT	85%	75%	10%
	CMP	78%	67%	11%
	Average	82%	71%	11%
T 3. Classification Accuracy	PHT	91%	77%	14%
	CMP	89%	76%	13%
	Average	90%	77%	14%

* Positive numbers represent an improvement over the non-ontology based approach.

Using the uploaded PHT and CMP patent document sets, we compare the summarization results between the ontology-based approach (this research) and the non-ontology based approach (Trappey & Trappey 2008). The evaluation uses statistics to represent the summary generation and its compression ratio, the ontology based keyword extraction retention ratio, and the summary classification accuracy. The compression ratio, indicating the text reduction from the original document to the compressed summary, can be stated using the equation (Hassel, 2004),

$$Compression\ \ Ratio\ = \frac{Length\ of\ \ Summary}{Length\ of\ \ Full\ Text} \tag{15}$$

The average retention ratio is about 79% for the CMP case, which surpases 67% achieved by the previous research (Table 8, T2). In order to evaluate automatic document categorization efficiency, a neural network algorothm is used to test the accuracy of summary classifications. After the patents are uploaded, the system extracts the key word frequency vector from the paragraphs and imports the vector into the neural network model. The output is calculated, and then system shows the results of the automatic categorization. Our research uses a parameter neural network model including layers (input layer, hidden layer and output layer), output nodes (the classes of patent documents), neurons, and activation functions. After the network model converges, the weights of the network are stored in the system database for application. The results of the document categorization are shown in Table 8, T3. The accuracy of classification is defined below.

$$Accuracy = \frac{\begin{array}{l}The\ Number\ of\ Summaries\\ Classified\ into\ a\ Correct\ Category\end{array}}{\begin{array}{l}The\ Total\ Number\ of\ Summaries\\ Put\ into\ Classisication\ System\end{array}} \quad (16)$$

Compared with the result of non-ontology based summarization approach (Trappey & Trappey, 2008), the ontology based appoach improves the retention ratios and the classification accuracies for both PHT and CMP patents (Table 8). We find that the ontoloy based approach does not compress the original patents as much as the non-ontology approach, particularly for the CMP case. The reason is that the ontology based approach maintains a larger number of key phrases and keywords which results in longer (but more meaningful) summarys. Nonetheless, for both cases, the compression ratios are maintained at the 20% level and compare favorably with the research results provided by Mani et al. (1998). For the retaintion ratio and the classification accuracy, the ontology based outperforms the non-ontology based approach benchmarked by the previous literature.

5. Conclusion

A methodology for automated ontology based patent document summarization is developed to provide a summarization system which extracts key words and phrases using concept hierarchies and semantic relationships. The system automatically abstracts a summary for a given patent based using the domain concepts and semantic relationships extracted from the document. The system provides a patent text summary and a generates a summary tree to represent key concepts. The ontology based TF-IDF method is used to retrieve domain key phrases and then these key phrases are used to identify high information density paragraphs for the summary. The proposed system for automatic document summarization can be extended to analyze general documents, technology specifications, and engineering specifications. The methodology is not domain specific and can be used to analyze different domains as long as the domain ontology is defined in advance. In this research, patents from the domains of power hand tools and chemical mechanical polishing are used to evaluate the summarization system. Using an automatic patent summarization system, enterprises can efficiently analyze technology trends based on the available patents and IP documents and, therefore, refine R&D strategies for greater competitive advantage.

Acknowledgements

The authors thank the referees for their thorough review and valuable suggestions that helped improve the quality of the paper. This research was partially supported by National Science Council research grants. Please send all enquiries to the corresponding author, Professor Charles Trappey.

References

[1] Aizawa, A. (2003). An information-theoretic perspective of TF-IDF measures. Information Processing & Management, 39 (1): 45-65

[2] Aone, C., Okurowski, M.E., Gorlinsky, J. & Larsen, B. (1997). A scalable summarization system using robust NLP. In: Proceedings of the ACL'97/EACL'97 Workshop on

Intelligent Scalable Text Summarization, 10-17, Madrid, Spain, 1997

[3] Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. World Patent Information, 29 (4): 308-316

[4] Bobillo, F., Delgado, M. & Gómez-Romero, J. (2008). Representation of context-dependant knowledge in ontologies: a model and an application. Expert Systems with Applications, 35 (4): 1899-1908

[5] Brown, C.T. (2006). Stapling Device. United States Patent, No. US 7,014,088 B2

[6] Buitelaar, P., Cimiano, P., Frank, A., Hartung, M. & Racioppa, S. (2008). Ontology-based information extraction and integration from heterogeneous data sources. International Journal of Human-Computer Studies, 66 (11): 759-788

[7] Chung, T.M. & Nation, P. (2004). Identifying technical vocabulary. System, 32 (2): 251-263

[8] Edmundson, H.P. (1969). New methods in automatic extracting. Journal of the ACM (JACM), 16 (2): 264-285

[9] Ercan, G. & Cicekli, I. (2007). Using lexical chains for keyword extraction. Information Processing & Management, 43 (6): 1705-1714

[10] Fattori, M., Pedrazzi, G. & Turra, R. (2003). Text mining applied to patent mapping: a practical business case. World Patent Information, 25: 335-342

[11] Fum, D., Guida, G. & Tasso, C. (1985). Evaluating importance: a step towards text summarization, In: Proceedings of the 9th International Joint Conference on Artificial Intelligence, 840-844, Los Angeles, CA, USA

[12] Goldstein, J., Kantrowitz, M., Mittal, V. & Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. In: Research and Development in Information Retrieval. Available via DIALOG. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.3703

[13] Gong, Y. & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis, In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Available via DIALOG. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.5097

[14] Greiff, W.R. (1998). A Theory of Term Weighting Based on Exploratory Data Analysis. Computer Science Department, University of Massachusetts, Amherst

[15] Gruber, T.R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5 (2): 199-220

[16] Han, J. & Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, California

[17] Hassel, M. (2004). Evaluation of automatic text summarization - a practical implementation. Licentiate Thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden

[18] Hovy, E. & Lin, C.Y. (1999). Automated text summarization in SUM MARIST. In: Advances in Automatic Text Summarization.

Available via DIALOG. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.2103

[19] Hsu, S.H. (2003). Ontology-based semantic annotation authoring and retrieval (in Chinese). M.S. Thesis, Department of Computer Science, National Dong Hwa University, Hualien, Taiwan, China

[20] Hsu, F.C., Trappey, A.J.C., Hou, J.L., Trappey, C.V. & Liu, S.J. (2006). Technology and knowledge document clustering analysis for enterprise R&D strategic planning. International Journal Technology Management, 36 (4): 336-353

[21] Hu, Y., Li, H., Cao, Y., Teng, L., Meyerzon, D. & Zheng, Q. (2006). Automatic extraction of titles from general documents using machine learning. Information Processing & Management, 42 (5): 1276-1293

[22] Joung, Y.J. & Chuang, F.Y. (2009). OntoZilla: an ontology-based, semi-structured, and evolutionary peer-to-peer network for information systems and services. Future Generation Computer Systems, 25 (1): 53-63

[23] Wu, J., Xiong, H., Chen, J. & Zhang, W. (2007). A generalization of proximity functions for K-means. In: Seventh IEEE International Conference on Data Mining, 361-370

[24] Kim, N.H., Jung, S.Y., Kang, C.S. & Lee, Z.H. (1999). Patent information retrieval system. Journal of Korea Information Processing, 6 (3): 80-85

[25] Ko, Y., Kim, K. & Seo, J. (2003). Topic keyword identification for text summarization using lexical clustering. In: IEICE Trans. Inform. System, 1695-1701. Available via DIALOG. http://sciencelinks.jp/j-east/article/200320/000020032003A0635686.php

[26] Kupiec, J., Pedersen, J. & Chen, F. (1995). A trainable document summarizer. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95), 68-73, Seattle, WA, USA

[27] Lam-Adesina, A.M. & Jones, G.J.F. (2001). Applying summarization techniques for term selection in relevance feedback. In: Proceedings of the 24th Annual International ACM SIGIR'01 Conference on Research and Development in Information Retrieval, 1-9, New Orleans, Louisiana, September 9-13, 2001

[28] Li, Y.R., Wang, L.H. & Hong, C.F. (2008). Extracting the significant-rare keywords for patent analysis. Expert Systems with Applications, In Press, Corrected Proof, Available Online 8 July

[29] Lin, C.Y. & Hovy, E.H. (1997). Identifying topics by position. In: Proceedings of the Applied Natural Language Processing Conference (ANLP-97), 283-290, Washington, D.C., March 31 - April 3, 1997

[30] Lin, F.R. & Liang, C.H. (2008) Storyline-based summarization for news topic retrospection. Decision Support Systems, 45 (3): 473-490

[31] Lorch, R.F., Lorch, E.P., Ritchey, K., McGovern, L. & Coleman, D. (2001). Effects of headings on text summarization. Contemporary Educational Psychology, 26: 171-191

[32] Luhn, H.P. (1957), A statistical approach

to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1 (4): 309-317

[33] Luhn, H.P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2 (2): 159-165

[34] Mani, I. & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. Information Retrieval, 1 (1-2): 35-67

[35] Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T. & Sundheim, B. (1998). The TIPSTER SUMMAC Text Summarization Evaluation. MITRE Technical Report, Washington, D.C., 1-47

[36] Mani, I. & Maybury, M.T. (1999). Advances in Automated Text Summarization. The MIT Press, Cambridge, MA

[37] Morris, J. & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17 (1): 21-43

[38] Princeton University. (2006). WordNet 3.0. Available via DIALOG. http://wordnet.princeton.edu/perl/webwn?s=ontology

[39] Reeve, L.H., Han, H. & Brooks, A.D. (2007). The use of domain-specific concepts in biomedical text summarization. Information Processing & Management, 43 (6): 1765-1776

[40] Rodrigues, T., Rosa, P. & Cardoso, J. (2008). Moving from syntactic to semantic organizations using JXML2OWL. Computers in Industry, 59 (8): 808-819

[41] Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Journal of Information Processing & Management, 24 (5): 513-523

[42] Salton, G., Singhal, A., Mitra, M. & Buckley, C. (1997). Automatic text structuring and summarization. Information Processing & Management, 33 (2): 193-207

[43] Sharma, S.C. (1996). Applied Multivariate Techniques. John Wiley & Sons, Hoboken, New York

[44] Teufel, S. & Moens, M. (1997). Sentence extraction as a classification task. In: Proceedings of the ACL/EACL Workshop on Intelligent Scalable Summarization, 58-65, Madrid, Spain

[45] Trappey, A.J.C., Hsu, F.C., Trappey, C.V. & Liu, C.I. (2006). Development of a patent document classification and search platform using a back-propagation network. Expert Systems with Applications, 31: 755-765

[46] Trappey, A.J.C. & Trappey, C.V. (2008). An R&D knowledge management method for patent document summarization. Industrial Management and Data System, 108 (2): 245-257

[47] Tseng, Y.H., Lin, C.J. & Lin, Y.I. (2007). Text mining techniques for patent analysis. Information Processing & Management, 43 (5): 1216-1247

[48] Wu, J., Xiong, H., Chen, J. & Zhou, W. (2007). A generalization of proximity functions for k-means. In: Seventh IEEE International Conference on Data Mining, 28-31, Omaha, NE, USA

[49] Ye, J.S., Chua, H.T., Kan, W.M. & Qiu, I.L. (2007). Document concept lattice for text understanding and summarization. Information Processing & Management, 43 (6): 1643-1662

[50] Yeh, J.Y., Ke, H.R. & Yang, W.P. (2002).

Chinese text summarization using a trainable summarizer and latent semantic analysis. In: Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology, 76-87, ISBN: 3-540-00261-8. Available via DIALOG. http://portal.acm.org/citation.cfm?id=681381

[51] Yeh, J.Y., Ke, H.R., Yang, W.P. & Meng, I.H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information Processing & Management, 41 (1): 75-95

[52] Yeh, J.Y., Ke, H.R & Yang, W.P. (2008). iSpreadRank: ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. Expert Systems with Applications, 35 (3): 1451-1462

[53] Young, S.R. & Hayes, P.J. (1985). Automatic classification and summarization of banking telexes. In: Proceedings of the 2nd Conference on Artificial Intelligence Application, 402-408

[54] Zhang, W., Yoshida, T. & Tang, X. (2008). Text classification based on multi-word with support vector machine. Knowledge-Based Systems, In Press, Corrected Proof, Available Online 4 April

[55] Zheng, H.T., Kang, B.Y. & Kim, H.G. (2008). An ontology-based approach to learnable focused crawling. Information Sciences, 178 (23): 4512-4522

Amy J.C. Trappey is chair professor in the Department of Industrial Engineering and Management and Dean, College of Management at the National Taipei University of Technology. She is also a faculty member of the Department of Industrial Engineering and Engineering Management, the National Tsing Hua University. Dr. Trappey is an ASME Fellow.

Charles Trappey is a professor of marketing in the Department of Management Science at the National Chiao Tung University.

Chun-Yi Wu is a doctoral student in the Department of Industrial Engineering and Engineering Management at National Tsing Hua University and a system analyst and engineer at Avectec, Inc. His research interests include the development of computerized intelligent systems and the knowledge management of patents.