The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0263-5577.htm

An R&D knowledge management method for patent document summarization

Amy J.C. Trappey
Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan, and
Charles V. Trappey
Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan

Received 29 July 2007
Revised 24 September 2007
Accepted 9 October 2007

Abstract

Purpose – In an era of rapidly expanding digital content, the number of e-documents and the amount of knowledge frequently overwhelm the R&D teams and often impede intellectual property management. The purpose of this paper is to develop an automatic patent summarization method for accurate knowledge abstraction and effective R&D knowledge management.

Design/methodology/approach – This paper develops an integrated approach for automatic patent summary generation combining the concepts of key phrase recognition and significant information density. Significant information density is defined based on the domain-specific key concepts/phrases, relevant phrases, title phrases, indicator phrases and topic sentences of a given patent document.

Findings – The document compression ratio and the knowledge retention ratio are used to measure both quantitative and qualitative outcomes of the new summarization methodology. Both measurements indicate the significant benefits and superior results of the method.

Research limitations/implications – In order to implement the methodology with practical success, the accurate and efficient pre-processing of identifying key concepts and relevant phrases of patent documents is required. The approach relies on a powerful text-mining engine as the pre-process module for key phrase extraction.

Practical implications – The methodology helps R&D companies consistently and automatically process, extract and summarize the core knowledge of related patent documents. This enabling technology is critical to R&D companies when they are competing to create new technologies and products for short life cycle marketplaces.

Originality/value – This research addresses a new perspective in R&D knowledge management, particularly in solving the knowledge-overloading issue. The methodology helps R&D collaborative teams consistently to summarize the core knowledge of patent documents with efficiency. Efficient R&D knowledge management helps the firm to take advantage of IP positioning while avoiding patent conflict and infringement.

Keywords Document handling, Knowledge management, Information management, Intellectual property

Paper type Research paper

This research was partially supported by Taiwan National Science Council and Ministry of Economic Affairs research grants. The authors also wish to acknowledge Mr Burges H.S. Kao's technical support in prototyping and testing the summarization algorithm.

Industrial Management & Data Systems
Vol. 108 No. 2, 2008
pp. 245-257

0263-5577
DOI 10.1108/02635570810847608

1. Introduction

Organizations are utilizing an ever expanding number of electronic knowledge documents. Engineers and researchers, in particular, are retrieving and processing large numbers of technical reports and patent documents to safeguard existing intellectual properties and to avoid infringing upon the rights of others. The internet has greatly facilitated the search and retrieval of knowledge documents. However, R&D teams rarely have the time or budget to read and understand all of the documents retrieved and, as a result, are increasingly overloaded with too much information. In order to develop new and innovative technologies and minimize infringing upon the rights of existing patents, we propose using a combination of text mining (TM) and summarization methodologies to automatically organize and abstract patent documents for collaborative teams to review. This research uses TM to abstract and summarize the embedded knowledge of patent documents. The methodologies applied in the research include key phrase recognition, derivation of term frequency (TF) relations, paragraph concept clustering, computation of significant information density, and an integrated procedure for automatically generating patent summaries. Several researchers have studied knowledge management for the collaborative environment (Gu *et al.*, 2005; El-Korany, 2007). Different from the previous research emphasizing the management of collaborative processes and workflows, a novel concept and its enabling methodology for automatic patent document summarization is developed to enhance collaborative design knowledge management. Thus, the patent documents collected from patent corpuses can be summarized concisely and, thereafter, shared efficiently and accurately among the collaborating team members.

The motivation behind the research was derived from an industrial case. Taiwan original equipment manufacturers have struggled with repositioning their small and medium sized industries as brand and design companies when production shifted to developing economies across Asia. The hand tool industry in particular, saw many of its manufacturing sites relocated overseas. As a result, manufacturing engineers were often given new assignments to design branded tools for the global market. Thus, the Taiwan Ministry of Economic Affairs funded research to help these engineers work collaboratively to create new designs that did not infringe on the intellectual property of other global brands. The greatest problem for the engineers was to sort through the hundreds of patent documents in their specific hand tool area and:

- understand the English;
- categorize the documents into areas of research and design interest; and
- discover areas of design opportunity.

If the patents could be quickly and accurately summarized, then the information became immediately more valuable. Instead of requiring extensive translation of complete patent documents, the design team could better manage English summaries and pursue full text details from patents directly related to their design effort. The design teams also noted that it was often difficult to predict future technology trends given the large amount of text represented by all of the patents in their field of interest. Patents, by their nature, are large text technical documents that are difficult to understand. The amount of time taken to understand a single document, and the analysis effort required to derive trends and opportunities across the global marketplace for hand tools, led to the conception of this research. Accurate patent summaries, properly sorted and clustered, provide fundamental knowledge of:

- the history of technology;
- leading competitors;
- technology trends in design; and
- gaps in design and new opportunities for development.

In conclusion, a patent summary system acts as a knowledge filter that makes the information contained in a patent more concise and brings the knowledge contained in the set of patents down to a more manageable level for collaborative design teams.

Section 1 has described the research background, motivation and objectives of the paper. In Section 2, literature covering TM and summarization methods are reviewed with a focus on applications in the intellectual property domain. Section 3 depicts the proposed automatic summarization methodology which integrates heuristic methods for TM, data mining, and summarization. In Section 4, the empirical evaluation of the methodology is conducted using a sample of power hand tool patents. The conclusion and future works are described in Section 5.

2. Literature review

This section reviews the relevant issues and current research literature in the areas of key phrase recognition, phrase relevance determination, document summarization and patent document TM, respectively.

2.1 Keyword recognition

Key phrase recognition is typically the first step in text document content analysis for English language information extraction. The stopping, stemming and splitting processes are used to segment sentences (Selamat and Omatu, 2004). Stopping is the process of removing repetitive and low-meaning words (e.g. to, and, it). Stemming is the procedure of reducing words to their original roots (Lovins, 1968) and segmenting is the process of splitting a sentence into segments or individual words separated by blanks. After stopping, stemming, and segmenting, the weight of each phrase in the document is calculated. The simplest way to measure the weight of a phrase within text is to use the TF weighted with the inverse document frequency (TF-IDF) (Aizawa, 2003; Salton and Buckley, 1988), expressed as:

$$w_{jk} = tf_{jk} \times idf_j, \tag{1}$$

where w_{jk} is the weight of term j in document k, tf_{jk} is the number of term j that occurs in document k, and idf_j is the inverse document frequency of term j as derived in equation (2):

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

where n is the total number of documents in the target set, and df_j is the number of documents among the set containing term j. When the idf_j value increases, the term j representing specific documents becomes more significant as proposed by Horng and Yeh (2000). Finally, the top ranked terms with high w_{jk} values are identified as the key phrases for the given document k.

2.2 Phrase relevance

Phrase relevance is often measured using co-occurrence and location-based approaches. Co-occurrence determines how many times two phrases occur in the same documents or sentences. A location-based approach uses the positions of phrases in a given document to measure the relevance between pairs of phrases. Phrase relevance is usually used to synthesize two or more phrases into one key concept to simplify and enhance the accuracy of keyword extraction for query and retrieval. Statistical correlations are commonly used to extract and merge keywords and key phrases based on the keyword frequencies and keyword locations. One method extracts the most frequently appearing phrases and calculates a χ^2 value of the co-occurrence distribution to identify important phrases within a document (Matsuo and Ishizuka, 2004).

2.3 Document summarization

ISO 215 (1986) defines a summary as a:

> ... brief restatement within the document (usually at the end) of its salient findings and conclusions, and is intended to complete the orientation of a reader who has studied the preceding text.

In addition, Hovy and Lin (1999) define text summarization as "the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)." Owing to development of internet and digital content technologies, interest has been growing for automated summarization techniques to assist fast knowledge processing.

Some approaches score sentences of a given document using term frequencies (Kupiec *et al.*, 1995) or TF-IDF (Teufel and Moens, 1997) of manually selected keywords. The n top-scored sentences are selected to construct a summary. The drawback of the method is the requirement for the manual selection of keywords. Similar to keyword-based methods, some researchers consider the importance of headings (Lorch *et al.*, 2001). For heading-based methods, phrases within the titles or the subtitle headings are assigned greater weights in the evaluation scheme and the n top scoring sentences are extracted and rewritten as the summary. In addition to keywords and headings, researchers also combine many other document features or methods to improve the quality of summary.

Kupiec *et al.* (1995) use a training corpus of documents with matching abstracts to statistically calculate scores and rank the importance of sentences. In the pre-process, they use heuristic rules to define sentence features and then calculate the Bayesian probability of each sentence within a document assuming statistical independence of the features. Afterward, a set of sentences with the highest Bayesian probabilities are identified to form the summary. Nonetheless, this method requires manually created abstracts (often called golden summaries) by professional abstractors in order to calculate the probability values. Lin and Hovy (1997) describe the optimal position policy which uses a well-known set of documents following a rigorous writing style such as the Ziff-Davis corpus and the Wall Street Journal. Using these document sets, the data-analysis process identifies patterns in the text. The research approach shows that for these style documents, important sentences are usually located at certain positions. Although genre dependent, the positions can be determined automatically using training algorithms. The weakness of the method is the requirement of large-scale corpora in specific domain to train the algorithms for accurate summarization results.

Recent research focuses on generating multi-document summaries and abstracts using the concept of sentence fusion. The approach is to first identify common information or themes across a group of articles. Then, the key phrases and sentences of a given theme are derived using TM techniques. Finally, the representing sentences of key themes are synthesized as the abstract or summary of the given group of documents. This approach is particularly useful for summarizing news articles reporting common events (Barzilay and McKeown, 2005).

2.4 TM patent documents

The purpose of applying TM techniques to organizational information systems is to improve knowledge management using keyword extraction, automatic text categorization, clustering and the other derivative applications. Trappey *et al.* (2004) and Hsu *et al.* (2004) implement a multi-channel legal knowledge service center, called the legal knowledge management (LKM) platform, which integrates collaborative knowledge management functions using data-mining techniques. The system uploads legal documents, e.g. patent documents, for key phrase extraction, document categorization, metadata management, version control, and authority management. For enhancing the LKM's strategic analysis, Hsu *et al.* (2006) further address the use of knowledge document cluster analysis for enterprise R&D strategic planning.

An electronic document management system applying neural network technology was developed by Trappey *et al.* (2006) to automate patent document categorization. Finally, an improved document classification method using an ontology-based neural network approach was proposed and implemented in (Kantrowitz *et al.*, 2000). Using the above described approaches, R&D teams can better manage and utilize patents and other knowledge documents for innovative product designs.

3. System methodology

The automatic summarization method for this research uses key phrase recognition, paragraph concept clustering, and significant information density measures. The goal of the methodology development is to capture key concepts for patent summaries, which engineers can read and organize with greater efficiency. The approach follows two processing steps consisting of key phrase recognition and automatic summary generation. For the first step, the primary task is to identify a representative set of key phrases within a given patent domain. The output of this step provides information to measure information density which in turn is used to identify the key paragraphs to be used for summary generation.

3.1 Key phrase recognition

The system outputs key phrases automatically (without the document's author or a domain expert's assistance) following the sequence described below:

(1) *Word segmentation.* The patent content is retrieved and the text is transformed into a set of words.

(2) *Stopping.* Words with insignificant meaning (based on a stop-word list) are removed.

(3) *Morphology diagnostics.* Using a dictionary, nouns and verbs are extracted to represent the content of the document (Matsuo and Ishizuka, 2004).

(4) *Stemming.* The porter stemming algorithm (Selamat and Omatu, 2004) is used to deconstruct the tense and plurality of words into their root words. Stemming is used

to derive term weights while considering words with the same root word (Kantrowitz *et al.*, 2000). This approach reduces system memory and computing time.

(5) *Phrase weight calculation.* The easy-to-implement and effective TF-IDF weight scheme is used to measure phrase information weights (Horng and Yeh, 2000).

(6) *Phrase synthesis.* Phrases are checked and those with similar weights and meanings are merged.

(7) *Key phrase output.* The final key phrase output is used to construct a frequency matrix for analysis.

(8) *Phrase relevance measures.* The phrase relevance values are computed using phrase co-occurrence and are fed back into the phrase relevance database which provides input for the summary generation module.

The phrase relevance measure uses the following algorithm:

(1) *Step 1.* Build a phrase frequency matrix based on the number of key phrases occurring in each paragraph (Table I). The cosine similarity vector space model is used to measure the relevance value for two key phrases (Grossman *et al.*, 1997; Hou and Chan, 2003). The relevance value of key phrase i and key phrase j are calculated as:

$$\mathrm{Re}\,v(\mathrm{KP}_i, \mathrm{KP}_j) = \mathrm{Sim}(\mathrm{KP}_i, \mathrm{KP}_j) \quad = \frac{\mathrm{KP}_i \cdot \mathrm{KP}_j}{|\mathrm{KP}_i||\mathrm{KP}_j|} = \frac{\sum_{k=1}^{m} \mathrm{tf}_{ik}\mathrm{tf}_{jk}}{\sqrt{\sum_{k=1}^{m} \mathrm{tf}_{ik}^2 \sum_{k=1}^{m} \mathrm{tf}_{jk}^2}}.$$

If the value is close to 1, the two phrases co-occur with high relevancy and if the value is close to 0, then the two phrases are not relevant.

(2) *Step 2.* Feed resulting relevance values back into the phrase relevance database to compute the significant density measure and summary query expansion. The equation for the relevance value is:

$$\mathrm{Re}\,v_{\mathrm{DB}}(\mathrm{KP}_i, \mathrm{KP}_j) = \frac{\sum_{k=1}^{n} \mathrm{Re}\,v_k(\mathrm{KP}_i, \mathrm{KP}_j)}{n},$$

where $\mathrm{Rev}_{\mathrm{DB}}(\mathrm{KP}_i, \mathrm{KP}_j)$ is the relevance value of key phrase i and key phrase j in the document database, $\mathrm{Rev}_k(\mathrm{KP}_i, \mathrm{KP}_j)$ is the relevance value of key phrase i and key phrase j in document k, and n is the number of documents.

	P_1	P_2	P_3	P_4	...	P_m
$\mathrm{KP}_1 =$	$[\mathrm{tf}_{11}$	tf_{12}	tf_{13}	tf_{14}	...	$\mathrm{tf}_{1m}]$
$\mathrm{KP}_2 =$	$[\mathrm{tf}_{21}$	tf_{22}	tf_{23}	tf_{24}	...	$\mathrm{tf}_{2m}]$
$\mathrm{KP}_3 =$	$[\mathrm{tf}_{31}$	tf_{32}	tf_{33}	tf_{34}	...	$\mathrm{tf}_{3m}]$
$\mathrm{KP}_4 =$	$[\mathrm{tf}_{41}$	tf_{42}	tf_{43}	tf_{44}	...	$\mathrm{tf}_{4m}]$
⋮	⋮	⋮	⋮	⋮	...	⋮
$\mathrm{KP}_n =$	$[\mathrm{tf}_{n1}$	tf_{n2}	tf_{n3}	tf_{n4}	...	$\mathrm{tf}_{nm}]$

Table I.
Key phrase frequency matrix

Notes: KP_i = key phrase i; P_j = jth paragraph in a document; tf_{ij} = the number of key phrase i occurring in jth paragraph of the document

3.2 Summarization algorithm

The summarization algorithm uses document concept clustering and significant information density to generate a summary which follows the domain specific rules and the format of a summary template.

3.2.1 Document concept clustering. Documents consist of many concepts which are presented in separate paragraphs. For this reason, paragraphs that describe similar concepts are grouped together as shown in Figure 1. The algorithm for paragraph concept clustering is described as follows:

(1) Construct the paragraph and the key phrase frequency matrix (equals the transpose of the frequency matrix shown in Table I) based on key phrases recognized in the previous sub-section.

(2) Eliminate data outliers. The paragraphs that do not include key phrases are deleted to increase computing efficiency.

(3) Calculate the similarity of each paragraph using the cosine similarity formula (Farkas, 1994) in two steps:

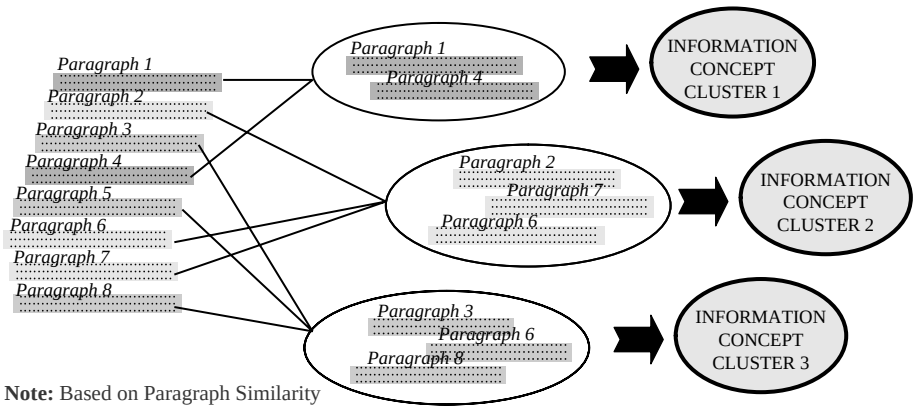
- List phrase frequency vector of two paragraphs with m key phrases:

$$\begin{cases} P_i = [\mathrm{tf}_{i1}, \mathrm{tf}_{i2}, \ldots, \mathrm{tf}_{im}], & i = 1, \ldots, n \\ P_j = [\mathrm{tf}_{j1}, \mathrm{tf}_{j2}, \ldots, \mathrm{tf}_{jm}], & j = 1, \ldots, n \end{cases}$$

- Calculate the similarity of each paragraph pair based on similarity formula with a resulting value between 0 and 1:

$$\mathrm{Sim}(P_i, P_j) = \frac{\sum_{k=1}^{m} \mathrm{tf}_{ik}\mathrm{tf}_{jk}}{\sqrt{\sum_{k=1}^{m} \mathrm{tf}_{ik}^2 \sum_{k=1}^{m} \mathrm{tf}_{jk}^2}} \quad (3)$$

(4) Build a paragraph similarity matrix (Table II) using the values computed in equation (3).



Note: Based on Paragraph Similarity

Figure 1.
Document information concept clusters

(5) Cluster paragraphs based on the derived paragraph similarity matrix using the K-mean algorithm (Han and Kamber, 2000) with the criteria of maximizing the root mean square standard deviation and minimizing r^2 for inter- and intra-class similarities, respectively, (Sharma, 1996). The K-means approach begins with the selection of K desired number of clusters with K random seed points. The seeds, then, define the initial cluster boundaries of entities assigned to a cluster within the boundaries. The recursive steps will calculate and adjust the centroids for the entities grouped within each cluster and the cluster boundaries until the centroids and the boundaries converged and cease to change.

3.2.2 Significant information density. After the document concept clusters are identified (Figure 1), the most significant paragraph in each cluster must be chosen to form the candidate summary set. The most significant paragraph in a cluster is identified by comparing the significant information densities of all paragraphs in the given cluster (Salton *et al.*, 1994). The significant information density is identified using six layers of information, including key phrases (KPL), title phrases (TPL), phrases that have high relevancy with key phrases (RPL), topic sentences (TSL), domain specific phrases (DPL) and indicator phrases (IPL). The six layers are to be stacked into a consolidated representation for all paragraphs in the concept clusters (Figure 2).

Significant information is quantified using the significant information mass (SIM) for each layer. Therefore, after concept clustering, the SIM for each of the six layers (KPL, TPL, RPL, TSL, DPL, and IPL) for each paragraph is calculated. After the individual layer SIM calculations, the significant information density (ρ) of each paragraph in every concept cluster is calculated as:

	P_1	P_2	P_3	...	P_n
P_1	1	Sim(P_1, P_2)	Sim(P_1, P_3)	...	Sim(P_1, P_n)
P_2	Sim(P_2, P_1)	1	Sim(P_1, P_2)	...	Sim(P_1, P_n)
P_3	Sim(P_3, P_1)	Sim(P_2, P_2)	1	...	Sim(P_2, P_n)
⋮	⋮	⋮	⋮	...	⋮
P_n	Sim(P_n, P_1)	Sim(P_n, P_2)	Sim(P_n, P_3)	...	1

Note: $P_j = j$th paragraph in a document

Table II.
Paragraph similarity matrix

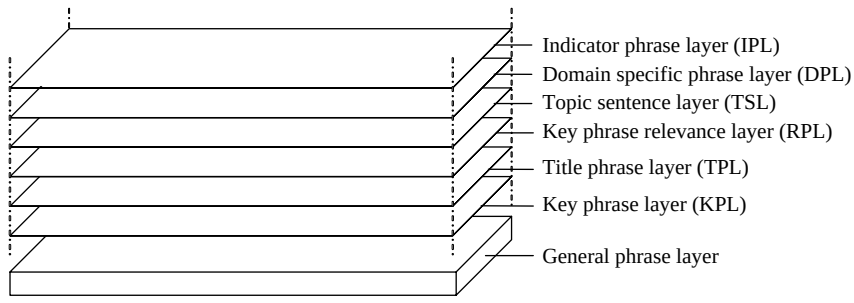


Figure 2.
Significant information layers

$$\text{Significant information density}(\rho) = [(\text{SIM of KPL})_m + (\text{SIM of TPL})_m + (\text{SIM of RPL})_m + (\text{SIM of DPL})_m + (\text{SIM of IPL})_m] \text{ total number of phrases in paragraph } m.$$

The paragraphs with the highest significant information densities are selected from all concept clusters to form the candidate summary set.

3.2.3 Domain specific rule and summary template. If special patterns of domain specific documents are mined and classified, it is possible to create rules to enhance the quality and efficiency of the summarization methodology. For example, USA patents are divided into two sections. One section describes the knowledge content and details of the invention and technology. The second section describes the claims, a declaration of the boundaries and rights owned by the patent holder under the legal claim of the patent. Therefore, when the system summarizes a patent document, there are two parts of the report. In the claim section, the independent claims are often more important than the non-independent sub-claims. Thus, the paragraphs describing independent claims are selected prior to recognizing key phrases and calculating significant information densities for the candidate summary set.

As shown in Figure 3, the summary template consolidates the key phrases (KP_i) and the candidate summary set (i.e. significant information paragraphs) derived from Sections 3.1 to 3.2.2 to form a compressed summary page.

4. Empirical evaluation

Two criteria values, the compression and retention ratios (RR), are used to evaluate the patent summarization method and the quality of its outputs (Hovy and Lin, 1999). The compression ratio (CR) is the ratio of the word counts between the summary and its original document. Further, the RR is the average value of the recall ratio and the precision ratio. The CR measures the consistence of the summary and the RR measures the amount of information retained or lost by the summary process. Finally, how well the summary represents the original document is also an indicator of validity. Using summaries for TM to automatically extract key phrases and classify patents is tested and compared to the original documents.

[*Title of source document*].↵
[This | The] [technique | technical | scientific] [report | document | archive | paper | text file] [describe | represent | express | address] [mainly | principally] information about [*KP_1, KP_2, ..., and KP_n*].↵
There [is | are] [*the number of clusters*] concepts(s) of the [report | document | archive | paper | text file], and [it | they] are [described | represented | expressed | addressed] as [follow(s) | following(s)]:↵
(1) [*Significant information paragraphs/sentences extracted from cluster 1*].↵
(2) [*Significant information paragraphs/sentences extracted from cluster 2*].↵
…….↵
(n) [*Significant information paragraphs/sentences extracted from cluster n*].↵
[*Extra information of source document using domain-specific rule*].↵
[*Metadata of source document*].↵

Figure 3.
Summary template

For case study, the digital content of 111 patents from the IPC classification B25 (category of hand tools, portable power-driven tools, and handles for hand implement) was collected from WIPO (2006) for summary generation. The example patents under investigation belonged to four sub-classes as listed in Table III.

After summarization, the average CR of the patent summaries in B25C, B25D, B25F, and B25G are 22.39, 27.48, 22.56, and 20.56 percent with the overall average CR of 23.22 percent. The results of Figure 4 compares favorably with the research results provided in (Mani *et al.*, 1998).

The RR measures whether the summary still holds knowledge expressed in the original document. The key phrase extraction is used to evaluate the RR of the summaries (Witten *et al.*, 1999). The RR of summary-based key phrase extraction is calculated as the combination of recall ratio and precision ratio shown in Figure 5.

From the case study, the key phrase retention rates are about 79 percent with documents compressed to less than one quarter of their original size. Automatic neural network (ANN) classification (Trappey *et al.*, 2006) is used to evaluate the knowledge representation of the summaries. The average accuracy of the classification results is 94 percent as shown in Figure 6. The ANN classification of the compressed summaries yields a higher accuracy result (94 percent) compared to the same ANN classification of the original patent documents (90 percent).

IPC classification	Description	Number of patents
B25C	Hand held nailing or stapling tools; manually operated portable stapling tools	37
B25D	Percussive tools	29
B25F	Combination or multi purpose tools not otherwise provided for; details or components	17
B25G	Handles for hand implements	28

Table III.
Test patent document sub-classes under the IPC B25 category

Note: B25 – hand tools; portable power-driven tools; handles for hand implements; workshop equipment; manipulators

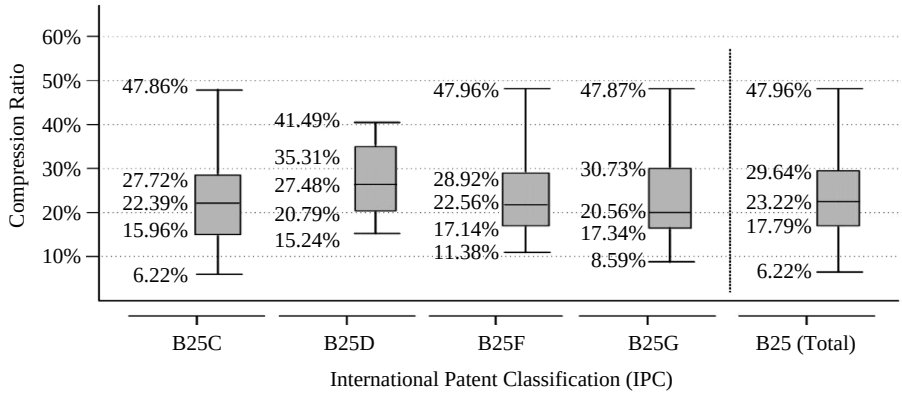


Figure 4.
CR evaluation results

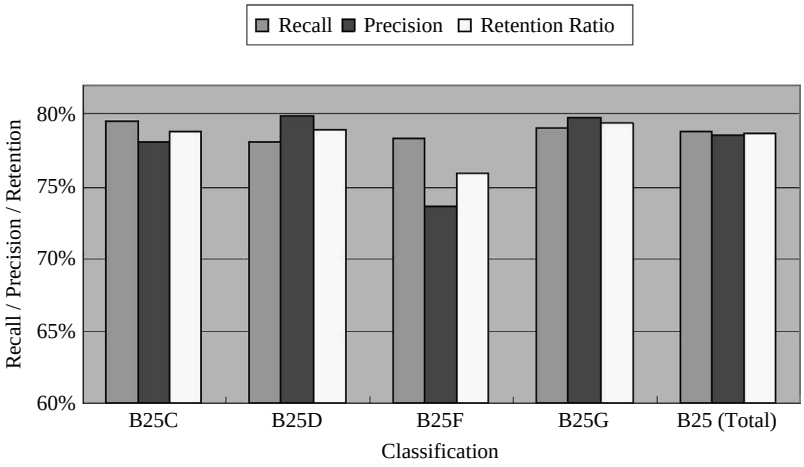


Figure 5.
The B25 patent summary evaluation results

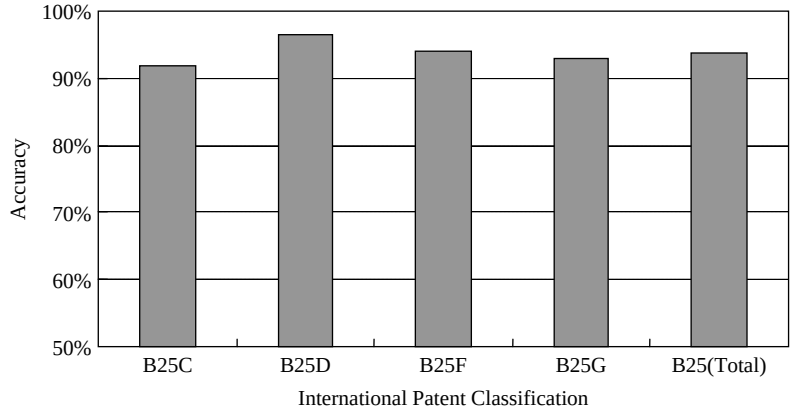


Figure 6.
The accuracy of classifying the B25 patent summaries

5. Conclusion

Engineers must analyze and categorize increasing numbers of patents as part of the research, development, and design process. The large amount of information is not only difficult to organize but also restricts efficient collaborative design efforts. Therefore, better ways of acquiring, organizing, and presenting patent documents are critical issues for collaborative R&D knowledge management. In this research, a novel patent document summarization methodology using key phrase recognition and significant information density is proposed. The approach helps design engineers save time processing large numbers of patent documents by automatically generating summaries without significant loss of knowledge content. The contributions of this research are summarized as follows. First, key phrase recognition technology is incorporated into the system. Second, the significant information density approach which includes information concept clustering based on paragraph similarities is derived. Third, domain specific patent rules and a summary template are proposed to

improve the summarization result. In short, with an automatic patent summarization solution, collaborative design engineers can efficiently review and process patent content. Finally, the summarization module is a critical part of the engineering knowledge management platform which integrates intelligent patent search, automatic patent classification, patent analysis and patent valuation to form a complete collaborative design knowledge environment.

Accurate patent summaries, properly sorted and clustered, provide fundamental knowledge of the history of technology, the leading competitors, trends in design, and new opportunities for development. Future research will focus on creating summarizations across a set of patents in a common domain and clustering patents. The research of Barzilay and McKeown (2005) shows that sentence fusion can be used to generate a multi-document summary. If a set of patent summarizations can be clustered into homogeneous groups, then each group of patents may be summarized using sentence fusion to generate nodes of knowledge. These nodes will then be used to better organize and link technology development within large databases of patents.

References

Aizawa, A. (2003), "An information-theoretic perspective of tf-idf measures", *Information Processing & Management*, Vol. 39, pp. 45-65.

Barzilay, R. and McKeown, K.R. (2005), "Sentence fusion for multi-document news summarization", *Computational Linguistics*, Vol. 31 No. 3, pp. 297-327.

El-Korany, A. (2007), "A knowledge management application in enterprises", *International Journal of Management and Enterprise Development (IJMED)*, Vol. 4 No. 6, pp. 693-702.

Farkas, J. (1994), "Generating document clusters using thesauri and neural networks", *Proceedings of the 1994 Canadian Conference on Electrical and Computer Engineering*, Vol. 2, Halifax, Nova Scotia, pp. 710-3.

Grossman, D., Frieder, O., Holmes, D. and Roberts, D. (1997), "Integrating structured data and text: a relational approach", *Journal of the American Society for Information Science*, Vol. 48 No. 2, pp. 122-32.

Gu, N., Xu, J., Wu, X., Yang, J. and Ye, W. (2005), "Ontology based semantic conflicts resolution in collaborative editing of design documents", *Advanced Engineering Informatics*, Vol. 19 No. 2, pp. 103-11.

Han, J. and Kamber, M. (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA.

Horng, J.T. and Yeh, C.C. (2000), "Applying genetic algorithms to query optimization in document retrieval", *Information Processing & Management*, Vol. 36, pp. 737-59.

Hou, J.L. and Chan, C.A. (2003), "A document content extraction model using keyword correlation analysis", *International Journal of Electronic Business Management*, Vol. 1 No. 1, pp. 54-62.

Hovy, E. and Lin, C-Y. (1999), "Automated text summarization in summarist", in Mani, I. and Maubury, M. (Eds), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, MA, pp. 1-14.

Hsu, F.C., Trappey, A.J.C., Hou, J.L., Trappey, C.V. and Liu, S.J. (2004), "Develop a multi-channel legal knowledge service center with knowledge mining capability", *International Journal of Electronic Business Management*, Vol. 2 No. 2, pp. 92-9.

Hsu, F.C., Trappey, A.J.C., Hou, J.L., Trappey, C.V. and Liu, S.J. (2006), "Technology and knowledge document clustering analysis for enterprise R&D strategic planning", *International Journal Technology Management*, Vol. 36 No. 4, pp. 336-53.

ISO 215 (1986), *Documentation – Presentation of Contributions to Periodicals and other Serials*, Technical Report, International Organization for Standardization, Geneva, pp. 3-8.

Kantrowitz, M., Mohit, B. and Mittal, V.O. (2000), "Stemming and its effects on TFIDF ranking", *Proceedings of the 23rd Annual International ACM SIGIR'00 Conference on Research and Development in Information Retrieval, Athens, Greece*, pp. 357-69.

Kupiec, J., Pedersen, J. and Chen, F. (1995), "A trainable document summarizer", *Proceedings of the 18th Annual International ACM SIGIR'95 Conference on Research and Development in Information Retrieval, Seattle, Washington, DC*, Vol. 95, pp. 68-73.

Lin, C.Y. and Hovy, E.H. (1997), "Identifying topics by position", *Proceedings of the Applied Natural Language Processing Conference, Washington, DC*, pp. 283-90.

Lorch, R.F., Lorch, E.P., Ritche, K., McGovern, L. and Coleman, D. (2001), "Effects of headings on text summarization", *Contemporary Educational Psychology*, Vol. 26, pp. 171-91.

Lovins, J.B. (1968), "Development of a stemming algorithm", *Mechanical Translation and Computational Linguistics*, Vol. 11, pp. 22-31.

Mani, I. and Firmin, T. *et al.* (1998), *The TIPSTER SUMMAC Text Summarization Evaluation*, MITRE Technical Report, Washington, DC, pp. 1-47.

Matsuo, Y. and Ishizuka, M. (2004), "Keyword extraction from a single document using word co-occurrence statistical information", *International Journal on Artificial Intelligence Tools*, Vol. 13 No. 1, pp. 157-69.

Salton, G. and Buckley, C. (1988), "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, Vol. 24 No. 5, pp. 513-23.

Salton, G., Allan, J., Buckley, C. and Singhal, A. (1994), "Automatic analysis, theme generation, and summarization of machine-readable texts", *Science*, Vol. 264 No. 3, pp. 1421-6.

Selamat, A. and Omatu, S. (2004), "Web page feature selection and classification using neural networks", *Information Sciences*, Vol. 158, pp. 69-88.

Sharma, S.C. (1996), *Applied Multivariate Techniques*, Wiley, New York, NY.

Teufel, S. and Moens, M. (1997), "Sentence extraction as a classification task", *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Summarization, Madrid, Spain*, pp. 58-65.

Trappey, A.J.C., Hsu, F.C., Trappey, C.V. and Liu, C.I. (2006), "Development of a patent document classification and search platform using a back-propagation network", *Expert Systems with Applications*, Vol. 31, pp. 755-65.

Trappey, A.J.C., Hsu, F.C., Hou, J.L., Trappey, C.V. and Liu, S.J. (2004), "Designing a multi-channel legal knowledge service center using data analysis and contact center technology", *Proceedings of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics, Orlando, FL*, pp. 132-6.

WIPO (2006), World Intellectual Property Organization, available at: www.wipo.int/

Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C. and Nevill-Manning, C.G. (1999), "KEA: practical automatic key phrase extraction", *Proceedings of the Digital Libraries 99, Berkeley, CA*, pp. 254-6.

Corresponding author
Amy J.C. Trappey can be contacted at: trappey@ie.nthu.edu.tw