Feature Selection for Cloud Computing Patents Classification

Jia-Yen Huang

***Abstract*—Nowadays, many enterprises have considered cloud computing as a seminal technology, and have exploited various types of service models to respond to different customer needs. Patent analysis is an essential ability of survival and development for high technology enterprises. It takes a huge number of patents to support the generation of a business service model of cloud computing. Patent engineers usually fail to collect and analyze patents efficiently due to their large number of professional glossaries and unknown patent classification. This study uses patents in lawsuit as partial important components of pearl patents and proposes a compound retrieval strategy to completely collect the patents of cloud computing. By using text mining as a tool for data processing and keywords extraction, we adopt the technique for order preference by similarity to ideal solution (TOPSIS) to pick out features with high degree of distinguishability for classification. These results establish an important foundation for developing a patent classification system in the future.**

***Index Terms*—Cloud computing, text mining, Feature selection, TOPSIS.**

I. Introduction

Ever since Eric Schmidt, the Executive Chairman for Google, first proposed the concept of cloud computing in the Search Engines Strategies conference held in San Jose 2006, a tremendous wave of interest in this field has been growing. Cloud computing is now regarded as an emerging business opportunity for network technology. Lowendahl *et al*., [1] defined cloud computing as "a style of computing where scalable and elastic IT-related capabilities are provided 'as a service' to external customers using Internet Technologies“. Rosenthal *et al*., [2] examined how the biomedical informatics community can take advantage of cloud computing, and pointed out that cloud computing has four merits: reduced management, extendable, better restore ability and homogeneity. In the past, domestic companies usually have had to transfer local information to overseas subsidiary companies, and then transfer it back after it has been used.

With the development of cloud computing technology, people have gradually changed their patterns of using Internet. For the customer, with the development of cloud computing, more and more computer users have changed their way of using internet. For the enterprises, to integrate numerous computer resources to complete a more extensive operation is believed to be the trend in the future. In view of the importance of cloud computing, the strategy of patent application of this field should be organized with overall consideration so as to lower the risk of R&D investment.

Patent information is an invisible but critical asset in today's information-based and information-dependent businesses. It can not only be used to protect enterprises' R&D results and detect movements of competitors, but also to find important opportunities and to set strategic plans for technological innovation. The assessment of the development trend in cloud computing industry through patent analysis is indispensable for enterprises with intent to have prosperous development in the cloud computing industry. Making good use of patent information allows the tracing of the development of technologies so as to save research fees and shortens research time.

Before conducting patent analysis, the work of patent retrieval has to be accomplished in advance. Although National Institute of Standards and Technology (NIST) has clearly defined three service models, namely the SPI model, on cloud computing, there are still no widely accepted definitions for the cloud computing (Wang *et al*., [3]), in addition, there are few research on ontology in this field (Androcec *et al*., [4]), it is difficult to completely retrieve patents of cloud computing. Most of the studies of cloud computing are focused on the technologies development, security and commercial application (one may refer to the relevant literature: Cartlidge and Clamp, [5]; Brandic and Buyya,[6]; Han and Sim, [7]; Hudic and Weippl, [8].), there is limited literature dealing with the issue of patent analysis. Moreover, there are many difficulties in retrieving the cloud patents with the conventional methods (detailed descriptions are presented in Section II we are therefore motivated to propose a compound patent retrieval strategy so as to correctly determine the classification and the scope of cloud patent.

In general, patent documents required in-depth reading so as to clearly determine its classification of technology. However, due to a great number of professional and technical terms and the increasing volumes of technical data presented in patents, this task is very labor intensive and time consuming. It is not easy for patent engineers to read and analyze patents effectively, especially for industry without widely accepted definitions. Since engineers can no longer completely rely on their own knowledge and skills to analyze the patents, this has consequently facilitated the use of computer aided tools for analyzing the patents (Abbas *et al*., [9]).

The emerging methods of patent analysis which attract many attentions in recent years are through the use of text mining and visualization of the analysis results. In the past,

Manuscript received January 4, 2015; revised March 15, 2015. This work was supported in part by the Ministry of Science and Technology in Taiwan under Grant MOST 103-2410-H-167-008 -MY2.

J. Y. Huang is with the Department of Information Management, National Chin-Yi University of Technology, No.57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung City 41170, Taiwan (e-mail: jygiant@ncut.edu.tw).

DOI: 10.7763/IJSSH.2016.V6.707

patent analysis relies on well-trained patent engineers, who must know how to collect information, and understand techniques in specific fields and business intelligence. Various techniques have been developed to assist patent analysis analyzers to discover the patent intelligence. The task of analyzing the patent data using the automated tools to fulfill diverse requirements through techniques such as visualization and text mining is termed as patent informatics (Abbas, *et al.*, [9]). Text mining is a process of editing, organizing, and analyzing a large number of files. Its purpose is to provide analysts or specific users with features and relationship among specific information (Sullivan, [10]). With the above techniques, it helps users more accurately picking up the necessary information in a large amount of text data.

Noted that text mining is not just a matter of extracting frequent words out of text, how to select adequate keywords for subsequent analysis is the key. As for the task of classification, the selection of keywords with high degree of distinguishability for classification is crucial to the success of classification. Commonly used methods include mutual information (Yoon and Lee, [11]) and other traditional indicators such as TF, IDF, and entropy. There are shortcomings no matter using the indicator TFIDF along or using entropy-based mutual information. Therefore, this study proposes a new method, by using TOPSIS, to effectively pick out features with high degree of distinguishability for classification.

II. Literature Review

The SPI model proposed by NIST includes three business models of cloud computing: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). SaaS is a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted. It is the application that consumers use, but it does not control the operation system, hardware or the basic internet structure of operation. Consumers of PaaS can control the environment of operating the application. Consumers of IaaS can control the operating system, storage space, deployed applications and network components (such as firewalls, load balancers, etc.), but cannot control the basic structure of cloud.

To facilitate the classification and description, media often classifies an enterprise into a particular model of service. In fact, specific cloud computing enterprise in the end belongs to which a model is not entirely clear defined; a company may get involved in three service models at the same time. For instance, Amazon is known as an IaaS enterprise for providing IaaS service, such as the web service offered by EC2; but in the past few years, it also became a SaaS enterprise by providing its customers on-demand cloud computing solution.

In the era of competitive knowledge-based economy, intellectual property has become the most concrete and influential intellectual assets, and so is in the development of cloud computing. According to the research by CHI Research, technology companies that own more patents have delivered much better returns to their shareholders. In several industries, such as pharmaceutical industry, the more the patent counts the higher the market value of a company. For companies attempt to enhance their market value, they should raise innovative value of their patens (Chen and Chang, [12]).

Patent analysis is to transform patent data into usable information through sifting, gathering statistics and analysis. It can be applied to evaluate an enterprise's competition position, innovation investment, and patent portfolio (Zhang, [13]). It is almost a standard process to conduct a patent analysis in the early stage of product development in industry.

Proper patent search is the primary work of patent analysis. Patent retrieval refers to a process that reviews past publications and patents to construct the intellectual property landscape surrounding a novel concept. Even though the SPI model is usually adopted as the three major cloud business models, it does not mean that the range of cloud computing patents is already clearly defined. Several reasons lead to the difficulties of retrieving the patents of cloud computing. First, cloud computing involves researchers from various backgrounds, such as distributed computing and grid computing, and they work on cloud computing from different viewpoints. Moreover, the enabling technologies of cloud computing are still evolving, e.g., Web 2.0 (Wang *et al.*, [3]).

Secondly, the development of the ontology of cloud computing is still incomplete. Han and Sim [7] presented a cloud service discovery system (CSDS) which interacts with cloud ontology to determine the similarities between and among services. Unfortunately, they just showed the ontologies (Fig. 3 to Fig.5 in their paper) without giving any explanation about how these ontologies were built or adopted. The authenticity for the correctness of the results remains uncertain.

Thirdly, there are very few analysis reports about the determination of the classification and the scope of cloud patents currently. Xue *et al.*, [14] conducted patent portfolio analysis on cloud computing of six famous cloud computing enterprises. Considering that cloud computing is inherited from distributed computing, they limited the range of the patent retrieval in two IPC: G06 (including Electronic Data-Processing, Data Handling System or Data Processing Means, and Information Storage Memory) and H04 (Telecommunication and Communication). Although relevant patents of cloud computing should be classified under IPC H04 and/or G06, which are vast categories including much that could not be regarded as cloud computing. Since these are very broad classes whose meaning is open to different interpretations, the IPC classes used as an analysis basis for Xue's study [14] are not appropriate for patent retrieval. In view of the difficulties of retrieving the patents of cloud computing, this study is motivated to propose a compound retrieval method to collect patents. Detailed processes are presented in the following section.

III. Methodology

A. Compound Patent Retrieval Strategy

As mentioned above, there are many difficulties in retrieving the cloud patents with the conventional methods. Hence, we propose compound patent retrieval strategy by bring several constraints into the retrieval, so to collect data In line with the characteristics of cloud computing industry. Four major steps are depicted as follows.

Step 1: We use "cloud computing" as a query string and limit the search domain to title, abstract, and claim parts of the patents. Patents in lawsuit, as part of important components of pearl patents, are also added into the pearl set. To make sure that the pearl are patents of cloud computing, we perused the patents manually and categorized them into three commercial service models of cloud.

Step 2: This study extracts keywords from the pearl patents and uses the values of TF-IDF (Term frequency-inverse document frequency) to rank the keywords.

Step 3: Obtain more patents via block building search.

Step 4: Restrain database by targeted enterprises.

If one or more starting patents (the pearls) that have strong relevance to technology of interest can be located during patent retrieval, then one can control the retrieval techniques out of those pearls so as to promote the achievements and efficiency of retrieval. Generally speaking, there are some possible ways to decide the resources of pearl's information: 1) Master one known patent, take that as a pearl, and utilize its classification or key words to find out other related patents. 2) If searcher only knows one theme, then s/he has to confirm the classification and keywords first; find out one or more patents of interest, and take them as pearls. According to the study of Allison *et al*., [15], patents in lawsuit are considered as valuable patents. Therefore, we use patents in lawsuit, shown in Table I, as part of the important components of pearl patents.

TABLE I: PATENTS LAWSUITS OF CLOUD COMPUTING

Description of patent lawsuits	Patent in lawsuit	Business model
Quantum Corporation vs. Overland Storage, Inc.	7263596	SaaS
Oracle Files Suit Against Oasis Research LLC Vs. Cloud Computing Patents	5771354, 5901228, 6014651, 6327579, 6411943, 7080051	
British Telecommunications (BT) vs. Google Inc.	6151309, 6169515, 6397040, 6826598	
Microsoft vs. Salesforce.com	7251653, 5742768, 5644737, 6263352, 6122558, 6542164, 6281879, 5845077, 5941947	
Patent Infringements Alleged Against Apple and Google	6317831, 6317594, 6532446, 6292657, 6654786	PaaS
British Telecommunications (BT) vs. Google Inc.	6578079, 6650284	
Active video Networks vs. Trans Video Electronics	5991801, 5594936	
Elia Data of Texas, LLC vs. Amazon	7113996	IaaS
Personal Web Technologies vs. Amazon, Caringo, Dropbox, EMC, Google, NEC, NetApp, Vmware and YouTube	7949662, 8001096, 7945544	

The second step in our retrieval strategy is to extract high frequency key phrases out of pearls by TF-IDF. TF-IDF is a product of two indexes: the term frequency (TF), one of the most commonly used term weightings; and the inverse document frequency weighting (Trivedi *et al*., [16]). TF-IDF is used to evaluate the importance of a word by the score of the frequency of the key phrase in one document and the inverse of the frequency of the key phrase in all documents, and it is formulated as

$$w_{ij} = tf_{ij} \times \log\left(N / df_i\right) \tag{1}$$

where w_{ij} is the phrase weight of phrase i in document j, tf_{ij} is the number of phrase i occurring in document j, N is the total number of documents in the document set, and df_i is the number of documents containing the phrase i in the document set.

To be able to accurately grasp the various cloud computing service model patent, this study compares the top-50 rated keywords between three business models, and excludes those in common. Meaningless words are also pruned off. In the end, we have top-5 rated keywords of each business model, which are different to each other. The keywords for SaaS are: services, application, processing, memory and identifier; For PaaS are: wireless, store, nodes, virtual machine and machine; For IaaS are: virtual, physical, computing nodes, virtual computer networks, management.

The third step of patent retrieval uses a block building search technique to formulate a search statement. The block building approach starts with single concept searches which usually results in a very large number of hits. After all of the single concept searches are completed, they are combined using appropriate Boolean operators. This strategy can decompose a complex search task into simplex one, and narrows the topic and reduces the number of hits.

This strategy divides a retrieval problem into numerous concepts, and then identifies major concepts and their logical relationships with one another. Searching out all of the strings in each concept and combing the strings into a single set representing that concept using Boolean operator OR. Eventually, combine the facets sets with Boolean operator AND.

In the 1st block building retrieval, the study uses "cloud computing" as the searching concept, and the searching scope covers the entire corpus. Next, in the 2nd block building retrieval, we use the top-5 rated keywords of each cloud computing business models for further search. Then, by connecting 1st block building and the 2nd block building with Boolean "and", we can retrieve patents of each cloud computing business models.

The technology trend of cloud computing is so far being actively driven by some huge enterprises who control most of the market share. The movement of these companies is the focus of our attention, this study restraint the scope of retrieval to twenty cloud computing delivery enterprises (called the targets) nominated from famous marketing reports, such as magazine CIO, Network World and Computer Reseller News. The targets include Amazon, Apple, Cisco, EMC, Google, HITACHI, IBM, Microsoft, NTT, Oracle, Salesforce, SAP, Symantec and Vmware. By removing patents irrelevant to the targets, we can truly obtain the patent of cloud computing. Finally, we acquired 393 patents for

PaaS, 167 patents for IaaS, and 650 patents for SaaS.

B. Feature Extraction Based on the TOPSIS

Based on the results of patent retrieval, next, the study conducts features (keywords) extraction, and develops the patent classification system. The features are actually keywords. To make distinction with the keywords extracted from the pearls in section 3.1; here, we call the keywords extracted from the entire corpus the features. The purpose for keywords extraction from the pearls is to extend the retrieval range so as to collect cloud patents as many as possible. Since we don't want to miss important keywords that appear only in few documents, i.e., the key minority, the application of IDF is adequate at this stage. While the features extraction from the entire corpus aim for the usage of classification. Since the TF-IDF scheme is unable to consider the distribution of keywords in each category, it is obviously not suitable for features extraction. If it is used incautiously, those words appear in many documents but in the same class may be ignored; contrarily, those words appear in fewer documents but in different class may be selected, and which actually have little contribution to classification. Therefore, in evaluating the importance of words, their distribution in each class should be taken into account.

Considering words frequently appearing in most patents are redundant and carry less information to users, and keywords appearing in fewer patents are more informative, some studies apply the index of entropy to pick out informative features. The concept of entropy is originally used to measure the level of disorder in thermodynamic system, and it is now applied to many areas, such as mining web informative structures and contents (Kao *et al*., [17]). The entropy of keywords Ti is expressed as follows:

$$E(T_i) = -\sum_{j=1}^{m} \omega_{ij} \log \omega_{ij} \tag{2}$$

where m is the number of events, and ω_{ij} is the value of normalized keywords frequency in the patent set. While a keyword is spread averagely in all of the patent documents, its entropy can be higher; however, the uncertainty can be lower. Accordingly, this study brought the keywords with lower entropy into the keyword thesaurus.

If only TF is adopted for feature extraction, words frequently appear but uniformly distributed across all documents may be selected. Contrarily, If only IDF is adopted, words appear intensively in a few articles may be selected. As for entropy, the value of a word is zero if it only appears in a specific document. In this case, its uncertainty of information might be the lowest; nevertheless, it might be meaningless and therefore should be eliminated. Clearly, it is inappropriate to extract feature simply based on single method.

Some studies adopt mutual information (Yoon and Lee, [11]) or improved TF-IDF (Zhou, *et al*., [18]) to reduce the dimension of word feature space. The mutual information is a measure of the amount of information one random variable contains about another, i.e., a measure of the variables' mutual dependence. Intuitively, mutual information measures the information that two discrete random variables *X* and *Y* share. If *X* and *Y* are independent, then knowing one of these variables does not give any information about the other. Therefore, zero mutual information between two random variables means the variables are independent. At the other extreme, if *X* is a deterministic function of *Y* then all information conveyed by *X* is shared with Y. In this case the mutual information is the same as the entropy of *Y*. From this perspective, mutual information potentially has the same problem with entropy in feature extraction. Considering the deficiency of traditional TF-IDF without considering the distribution of feature words among classes, Zhou, *et al*., [18] proposed a new TF-IDF feature selection method by dividing TF-IDF with entropy. It is inadequate and too subjective to freely combine two numbers into a new indicator whatever by multiplication or division, because it doesn't make much sense. Since the value of entropy could be zero, it is unreasonable to have entropy in denominator. As the value of entropy approaches zero, the value of the new indicator will rise sharply, and it is actually entropy dominated. Although they have tried to add certain positive item to the denominator, however, it gives prominence to the absurdity of this method.

In view of the weaknesses of the above methods, this study argues that the determination of feature extraction should be treated as a multi-attribute decision problem, which concerns "objects" that can be described through several attributes, including TF, IDF and entropy. In this study, TF is average frequency of a term in a document. We propose to integrate the attributes by TOPSIS, instead of by fundamental operations of arithmetic. Among three attributes, TF has the characteristic of the-larger-the-better, while IDF and entropy are the-smaller-the-better. The reason that we set IDF the-smaller-the-better is because we aim to pick out the representative majority, not the key minority. This is quite different with the concept of using IDF during the process of patent extension by the pearls.

The major difference between TOPSIS and other principle-based decision-making methods is that TOPSIS considers both ideal and non-ideal solutions. A positive ideal solution is the sum of the optimal solutions that maximizes benefit and minimizes cost for each attribute; a negative ideal solution is the sum of the solutions farthest from the ideal solution (Chen *et al*., [19]). The purpose is to find a solution close to the positive ideal solution and far from the negative ideal solution.

Based on the following TOPSIS steps, the order of the preferred solutions can be obtained, and we will use the relative closeness C_i^* as the new indicator.

Step 1: Construct a normalized evaluation matrix.

Step 2: Determine the positive and negative ideal solution.

Step 3: Calculate the separation measures. The distance between solutions i and the positive ideal solution is denoted as the degree of separation S_i^+; and the distance between solutions *i* and the negative ideal solution is denoted as the degree of separation S_i^-.

Step 4: Calculate the relative closeness (C_i^*) of the activity/quality attributes to the ideal solution by $C_i^* = S_i^- / (S_i^+ + S_i^-)$. This index C_i^*, ranging $0 \le C_i^* \le 1$, is now acting as a proxy for patent activity/quality.

By conducting TOPSIS to integrate the values of attributes TF, IDF and entropy into the closeness, we ranked extracted features as shown in Table II. Due to limited space of this paper, only the results of top-10 rated features for IaaS are illustrated.

TABLE II: TOP-10 RATED FEATURES EXTRACTED FOR IAAS

	Features	*TF*	*IDF*	*E*	C_i
1	database	0.0316	1.2283	0.2888	0.6339
2	memory	0.0292	0.9273	0.5407	0.4742
3	virtual machine	0.0271	1.0770	0.4582	0.4394
4	bandwidth	0.0017	1.8304	0.2028	0.4174
5	network	0.0155	1.5293	0.2970	0.3664
6	storage	0.0106	1.3075	0.3879	0.3625
7	application server	0.0043	2.0065	0.1296	0.3515
8	device	0.2693	0.7164	0.6283	0.3428
9	software applications	0.0025	2.0065	0.1198	0.3381
10	infrastructure	0.0062	1.4044	0.3839	0.3302

The top-10 rated features of three classes extracted by TOPSIS and TF-IDF are listed in Table III and Table IV, respectively. The extracted features are basically similar to the ontology presented by Han and Sim [7].

TABLE III: TOP-10 RATED FEATURES OBTAINED BY TOPSIS

IaaS	PaaS	SaaS
database	interface	security
memory	security	proxy
virtual machine	http	access control
bandwidth	web application	database
network	platform	performance metrics
storage	worker node	software
application server	backend	DNS
device	scalable	communication
software applications	hosting	graphical user interface
infrastructure	language	network

TABLE IV: TOP-10 RATED FEATURES OBTAINED BY TFIDF

IaaS	PaaS	SaaS
device	device	device
virtual	application	virtual
servers	virtual	interface
memory	interface	servers
interface	database	database
software	host	web
security	security	security
storage	memory	memory
host	software	workload
workload	worker node	software

Only small part of the features extracted by TOPSIS method is overlapped for the three classes, most of them are different. In contrast, the features derived through TF-IDF method are quiet the same for three classes. In general, the bigger the difference of features constitution between three classes, the better the classification results will be. It is obvious that the features extracted by TOPSIS method have higher distinguishability for classification. In sum, there are 27 different features which are to be used to vectorize the entire corpus and to generate classification rules.

IV. CONCLUSION

The main contribution of this research is two-fold. First, we propose a compound patent retrieval strategy, which make use patents in lawsuit as partial important components of pearl patents, to completely collect the patents of cloud computing. Secondly, this article argues that it is neither appropriate to conduct feature extraction simply based on single index (such as TF, IDF or entropy) nor combining two indexes into a new indicator whatever by multiplication or division. The features extracted by the proposed TOPSIS method are proved having high degree of distinguishability for establishing a classification system.

REFERENCES

[1] J. M. Lowendahl, M. Harris, and R. Bonig, *Agenda for Higher Education*, Stamford, CT, USA: Gartner, p. 4, 2012.

[2] A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester, and P. Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing," *J. Biomed Inform.* vol. 43, no. 2, pp. 342-53, 2010.

[3] L. Wang, G. V. Laszewski, A. Younge, X. He, M. Kunze, J. Tao, and C. Fu, "Cloud computing: A perspective study," *New Generation Computing*, vol. 28, issue 2, pp. 137-146, 2010.

[4] D. Androcec, N. Vrcek, and J. Seva, "Cloud computing ontologies: A systematic review," in *Proc. the Third International Conference on Models and Ontology-based Design of Protocols, Architectures and Services*, France, 2012.

[5] J. Cartlidge and P. Clamp. (April 2014). Correcting a financial brokerage model for cloud computing: Closing the window of opportunity for commercialization. *Journal of Cloud Computing: Advances, Systems and Applications*. [Online] 3(2). Available: http://link.springer.com/article/10.1186%2F2192-113X-3-2

[6] I. Brandic and R. Buyya, "Special section: Recent advances in utility and cloud computing," *Future Generation Computer System*, vol. 28, no. 1, pp. 36-38, 2012.

[7] T. Han and K. M. Sim, "An ontology-enhanced cloud service discovery system," in *Proc. International Multi Conference of Engineers and Computer Scientists (IMEC 2010)*, Hong Kong, pp. 644-649, 2010.

[8] A. Hudic and E. Weippl, "Private cloud computing: Consolidation, virtualization, and service-oriented infrastructure," *Computers & Security*, vol. 31, no. 4, pp. 629, 2012.

[9] A. Abbas, L. Zhang, and S. U. Khan, "A literature review on the state-of-the-art in patent analysis," *World Patent Information*, vol. 37, pp. 3-13, 2014.

[10] D. Sullivan, *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*, New York: John Wiley and Sons Inc.., 2001.

[11] Y. Yoon and G. G. Lee, "Efficient implementation of associative classifiers for document classification," *Information Processing and Management*, vol. 43, Issue 2, pp. 393-405, 2007.

[12] Y. S. Chen and K. C. Chang, "The relationship between a firm's patent quality and its market value - The case of US pharmaceutical industry," *Technological Forecasting and Social Change*, vol. 77, issue 1, pp. 20-33, 2010.

[13] Y. Zhang, "Analysis and evaluation of enterprise innovation ability conversions," *International Journal of Innovative Management, Information & Production*, vol. 2, no. 2, pp. 39-46, 2011.

[14] K. Y. Xue, C. L. Liou, and J. Y. Huang, "Patent portfolio of cloud computing industry: A case study of Google," in *Proc. the 1th Cross-Strait Academic and Practice Conference on Trade and Business Management*, Taichung, Taiwan: Feng Chia University, 2011.

[15] J. R. Allison, M. A. Lemley, K. A. Moore, and D. Trunkey, "Valuable Patents," *Georgetown Law Journal*, vol. 92, pp. 435-479, 2004.

[16] A. Trivedi, A. E. Medonca, and B. S. Johnson, "Using Machine Learning for Classifying Documents and Extracting Features," in *Proc. the 11th World Congress of Medical Informatics*, Medinfo 2004.

[17] H. Y. Kao and S. H. Lin, "Mining Web Informative Structures and Contents Based on Entropy Analysis," *IEEE Transaction on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 41-55, 2004.

[18] Y. T. Zhou, J. B. Tang, and J. Q. Wang, "Improved TFIDF feature selection algorithm based on information entropy," *Computer Engineering and Applications*, vol. 43, no. 35, pp. 156-158, 2007.

[19] K. H. Chen, C. N. Liao, and L. C. Wu. (2014). A selection model to logistic centers based on TOPSIS and MCGP Methods: The case of airline industry. *Journal of Applied Mathematics*. [Online]. 10 Available: http://dx.doi.org/10.1155/2014/470128

Jia Yen Huang received the bachelor's degree in mechanical engineering from National Chiao-Tung University in 1982, and the Ph.D. degree in mechanical engineering from National Taiwan University in 1987.

He is currently a professor in the Department of Information Management, National Chin-Yi University of Technology, Taiwan. His present research interests include supply chain management, innovation management, and data mining.