Expert Systems with Applications 39 (2012) 12618–12625

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Technology trends analysis and forecasting application based on decision tree and statistical feature analysis

Jinhyung Kim, Myunggwon Hwang, Do-Heon Jeong *, Hanmin Jung

Software Research Laboratory, Korea Institute of Science and Technology Information, Republic of Korea

ARTICLE INFO

Keywords:
Technology trends analysis
Technology trends forecasting
Feature analysis
Decision making
Strategic intelligence

ABSTRACT

Analyzing mass information and supporting foresight are very important task but they are extremely time-consuming work. In addition, information analysis and forecasting about the science and technology are also very critical tasks for researchers, government officers, businessman, etc. Some related studies recently have been executed and semi-automatic tools have been developed actively. Many researchers, annalists, and businessmen also generally use those tools for strategic decision making. However, existing projects and tools are based on subjective opinions from several experts and most of tools simply explain current situations, not forecasting near future trends. Therefore, in this paper, we propose a technology trends analysis and forecasting model based on quantitative analysis and several text mining technologies for effective, systematic, and objective information analysis and forecasting technology trends. Additionally, we execute a comparative evaluation between the suggested model and Gartner's forecasting model for validating the suggested model because the Gartner's model is widely and generally used for information analysis and forecasting.

1. Introduction

The precise information analysis and new opportunity discovery are very important for future forecasting, future countermeasures decision, and future plan establishment. However, as the amount of information in science and IT field increases exponentially every year, data analysis about that information or extraction of new opportunity from documents, papers, patents, etc. becomes more difficult and complicate. Until now, there have been researches regarding information analysis of mass data and new opportunity discovery (Dereli & Durmusoglu, 2009; John, 1995; Kim, Suh, & Park, 2008; Kim, Lee, Lee, Lee, & Jung, 2011; Rann, 1998; Richard, 1983). Traditional studies had focused on information analysis and conclusion deduction based on the scenario method or the Delphi method or AHP method. These methods are based on non-systematic process and depends on subjective opinions of experts. The scenario method (Hetmanska & Nguyen, 2011; Wright & Goodwin, 2009) is a strategic planning method that a group of experts analyze base information for decision making. However, because there are several irregular and arbitrary cases in the scenario processing, its reliability is low. The Delphi method (Hanafizadeh & Mirzazadeh, 2011; Okoli & Pawlowski, 2004) is a structured communication technique, originally developed as an interactive forecasting method which relies on a panel of experts. In the Delphi method, the experts answer questionnaires in two or more rounds. After each round, a facilitator provides an anonymous summary of the experts' forecasts from the previous round as well as the reasons they provided for their judgments. However, because the Delphi method depends on subjective opinions from experts, that cannot guarantee credibility of forecasting results. The AHP method (Liu, Jin, & Li, 2011; Duran, 2011) is a structured technique for organizing and analyzing complex decision based on mathematics and psychology. Like the Delphi method, the AHP method depends on subject view of evaluator or group of experts. As a result, the AHP method also cannot assure objectivity of forecasting results.

For overcoming limitations mentioned above, many systematic and objective methods are suggested such as Foresight and Understanding form Scientific Exposition (FUSE) (FUSE, 2010), Combining and Uniting Business Intelligence with Semantic Technology (CUBIST) (CUBISTP, 2008), Text Mining Software for Technology Management (VantagePoint) (VantagePoint, 2009), and so on. These projects aim to support decision making by analysis, pattern recognition of scientific documents. However, many researches and projects focus on information analysis and are insufficient to support new opportunity discovery or future forecasting.

Korea Institute of Science and Technology Information (KISTI) have researched regarding information analysis about science and technology field, and technology opportunity discovery since 2010. The research is named InSciTe and, information analysis and

* Corresponding author.
E-mail addresses: jinhyung@kisti.re.kr (J. Kim), mgh@kisti.re.kr (M. Hwang), heon@kisti.re.kr (D.-H. Jeong), jhm@kisti.re.kr (H. Jung).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.021

technology opportunity forecasting are based on suggested Model. We will describe suggested Model at the next section in detail. The suggested model consists of 3 sub-models; Technology Life Cycle Discovery (TLCD), Technology Maturity Forecast (TMF), Emerging Technology Discovery (ETD) models. In the suggested model, we analyze research trends based on papers and patents and execute feature extraction and selection. Based on several features extracted by feature selection, we can recognize technology trends and predict future prospective technologies by decision tree, machine learning, and several kinds of data mining technologies

2. Related works

VantagePoint (2009) developed by Search Technology in 2002 is a powerful text-mining tool for discovering knowledge in search results from patent and literature databases. VantagePoint helps users rapidly understand and navigate through large search results, giving users a better perspective on their information. The perspective provided by VantagePoint enables users to quickly find WHO, WHAT, WHEN and WHERE, enabling users to clarify relationships and find critical patterns among information. VantagePoint's capabilities can be broadly classified into five categories: importing, cleaning, analyzing, reporting, and automating.

The CUBIST project (CUBISTP, 2008) develops methodologies and a platform that combines essential features of Semantic Technologies and Business Intelligence. The CUBIST project is led by Sheffield Hallam Univ., Ontotext, and SAP. This project aims to develop new ways to interrogate not only the massive volume of data on the Internet, but also analyze the different formats it exists in – such as blogs, wikis, and video. With CUBIST, we envision a system with the following core features:

- Support for the federation of data from a variety of unstructured sources.
- A data persistency layer in the form of a semantic Data Warehouse; a hybrid approach based on a BI enabled triple store.
- Semantic information used to improve BI best practices in, for example, data reduction and preprocessing.
- A semantic data warehouse that realizes the advanced mining techniques of FCA.
- FCA guides the user in performing BI and helps the user discover facts not expressed explicitly by the warehouse model.

The FUSE Program (FUSE, 2010) will explore theories and models for the detection of significant technical capability emergence that can be observed from the worldwide scientific, technical, and patent literatures. FUSE will develop and test quantitative techniques that scan the full-length technical text across a large number of documents for time-dependent, pattern-based signals within a wide range of technical areas and multiple human languages. The Program will include empirical testing against examples of real-world capability emergence.

The FUSE Program will build upon substantial prior research and development in diverse technical areas, including: information extraction, machine learning, classification, clustering, time series summarization and analysis, network analysis, graph theory, statistical inference, technology forecasting, research and development management, business innovation, diffusion of innovation and market dynamics, bibliometrics and scientometrics, history of science, sociology of science, and psychology of science and emergence.

As mentioned above, there are many projects and researches regarding information analysis and pattern recognition. However, the most part of researches and projects just focusing information analysis and support low intensity forecasting services.

3. Technology trends analysis and forecasting model

Fig. 1 shows architecture of InSciTe and suggested model. The suggested model suggested in this paper is a part of InSciTe service which supports several kinds of an services for technology opportunity discovery based on agents (Company, Nation, and Person), and technologies. InSciTe is based on ontology data and consists of typical seven modules: SS& AE module for managing sub services, OntoPipeliner for resource allocation and constitution of each service, OntoURI for identifying URI of ontology resources, OntoURIResolver for managing duplication of ontology resources, OntoVerifier for ontology inference verification, and OntoRelFinder for tracking relation among ontology information.

Technology discovery model consists of three sub-models: technology life cycle discovery (TLCD) model, technology maturity forecasting (TMF) model, and emerging technology discovery (ETD) model. The TLCD model decides emerging phase of a specific technology through feature selection and analysis extracted from papers and patents. Emerging phase is comprised of 5 steps based on general technology life cycle concept: irruption, frenzy, turning point, synergy, maturity. The TMF model calculates technology development speed and technology maturity of a specific technology. The TMF model uses the exponential moving average (EMA) method (Song, Hao, & Hao, 2011) for calculating technology development speed. At present, because the TMF model is being developed conceptually, we do not discuss on the TMF model in this paper. The ETD model selects emerging technology among lots of technologies in various kinds of fields such as information technology, physics, life science, telematics, environments, and so on.

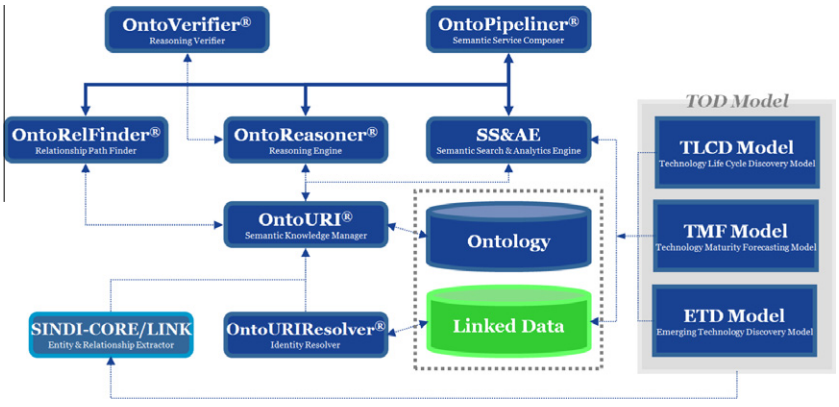

Fig. 1. Architecture of InSciTe and suggested model.

Definition 1. Each feature set is a combination of constitution elements from papers and patents information.

$$S(Pp) = \{Pp_1, Pp_2, \ldots, Pp_n\}, S(Pt) = \{Pt_1, Pt_2, \ldots, Pt_n\}$$

$$FS_{AbsoluteGrowthRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}\} \iff (AN^k_{Pp} - AN^{k-1}_{Pp})/AN^{k-1}_{Pp}$$

$$FS_{RelativeGrowthRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}\} \iff \left(N^k_{Pp} - N^{k-1}_{Pp}\right)/N^{k-1}_{Pp}$$

$$FS_{AuthorRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}, author\} \iff \left(A^k_{Pp}/AA^k_{Pp}\right) * 100(\%)$$

$$FS_{AuthorGrowthRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}, author\} \iff \left(AA^k_{Pp} - AA^{k-1}_{Pp}\right)/AA^{k-1}_{Pp}$$

$$FS_{DomainRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}, domain\} \iff \left(D^k_{Pp}/AD^k_{Pp}\right) * 100(\%)$$

$$FS_{DomainGrowthRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}, domain\} \iff \left(AD^k_{Pp}/AD^{k-1}_{Pp}\right)/AD^{k-1}_{Pp}$$

$$FS_{JournalRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}, Journal\} \iff \left(J^k_{Pp}/AJ^k_{Pp}\right) * 100(\%)$$

$$FS_{JournalGrowthRate}(S(Pp)^k) = \{number_{Pp}, date_{Pp}, Journal\} \iff \left(AJ^k_{Pp}/AJ^{k-1}_{Pp}\right)/AJ^{k-1}_{Pp}$$

$$FS_{AbsoluteGrowthRate}(S(Pt)^k) = \{number_{Pt}, date_{Pt}\} \iff \left(AN^k_{Pt} - AN^{k-1}_{Pt}\right)/AN^{k-1}_{Pt}$$

$$FS_{RelativeGrowthRate}(S(Pt)^k) = \{number_{Pt}, date_{Pp}\} \iff \left(N^k_{Pt} - N^{k-1}_{Pt}\right)/N^{k-1}_{Pt}$$

$$FS_{InventorRate}(S(Pt)^k) = \{number_{Pt}, date_{Pt}, inventor\} \iff \left(I^k_{Pt}/AI^k_{Pt}\right) * 100(\%)$$

$$FS_{InventorGrowthRate}(S(Pt)^k) = \{number_{Pt}, date_{Pt}, inventor\} \iff \left(AI^k_{Pt} - AI^{k-1}_{Pt}\right)/AI^{k-1}_{Pt}$$

$$FS_{ApplicantRate}(S(Pt)^k) = \{number_{Pp}, date_{Pp}, Applicant\} \iff \left(A^k_{Pt}/AA^k_{Pt}\right) * 100(\%)$$

$$FS_{ApplicantGrowthRate}(S(Pt)^k) = \{number_{Pp}, date_{Pp}, Applicant\} \iff \left(AA^k_{Pt}/AA^{k-1}_{Pt}\right)/AA^{k-1}_{Pt}$$

$$FS_{PatentFamilyRate}(S(Pt)^k) = \{number_{Pp}, date_{Pp}, PatentFamily\} \iff \left(P^k_{Pt}/AP^k_{Pt}\right) * 100(\%)$$

$$FS_{PatentFamilyGrowthRate}(S(Pt)^k) = \{number_{Pp}, date_{Pp}, PatentFamily\} \iff \left(AP^k_{Pt}/AP^{k-1}_{Pt}\right)/AP^{k-1}_{Pt}$$

3.1. Technology life cycle discovery model

TLCD model supports deciding the emerging phase of technologies. The emerging phase consists of five steps: irruption, frenzy, turning point, synergy, and maturity. Above five steps are defined as 'Great Surges of Development' by Carolta (2007). The irruption step means emergence of a new technology and the frenzy step represents that financial capital mobilizes to explore the potential, and a range of business models develop. The turning point step illustrates a financial crash and recession and the synergy step means emergence of new institutions and industry structures of the new technology and re-growth. The maturity step represents the final stable steps. The TLCD model consists of typical two parts: Feature extraction and selection, decision making and machine learning.

3.1.1. Feature extraction and selection

TLCD Model extracts 20 features from papers and patents and uses them for deciding emerging phase of technologies. Each feature is a combination of elements from paper and patent such as author, journal, domain, and etc. For example, 'Author Rate' feature is a combination of number, publication date, author of papers. There can be much more features in the papers and patents information, but we define 16 key features as a final feature sets in TLCD model. Definition 1 represents key features, element constitution, and calculation expression of each feature.

Growth rate is a feature that represents increase or decrease rate of number of research by year. We consider two growth rates; an absolute growth rate and a relative growth rate in order to reflect diverse aspects of growth rate. The absolute growth rate is based on accumulated number, the relative one is based on number of Y and Y-1 years of paper and patent. Agent rate means relative weight among totally accumulated number of authors in a specific year. Agent rate represents author rate from papers and inventor rate from patent. Domain growth rate and journal growth rate represent ratio which means how many papers and patents are related to a specific domain or journal in a specific year. Patent family rate is proportion of number of patent family group in applied year to accumulated number of all patent. Patent family is a group patent with same application number, IPC number and inventor.

According to the above formulas, the TLCD model calculates all values of every feature. Fig. 2 shows values of feature sets about 'augmented reality' technology.

After calculating every feature, we decide features of each emerging phase using average and standard deviation. Fig. 3 shows represent trends of 'Journal Growth Rate' feature in the 'Irruption' phase and the 'Frenzy' phase. As shown in Fig. 3(a), the majority

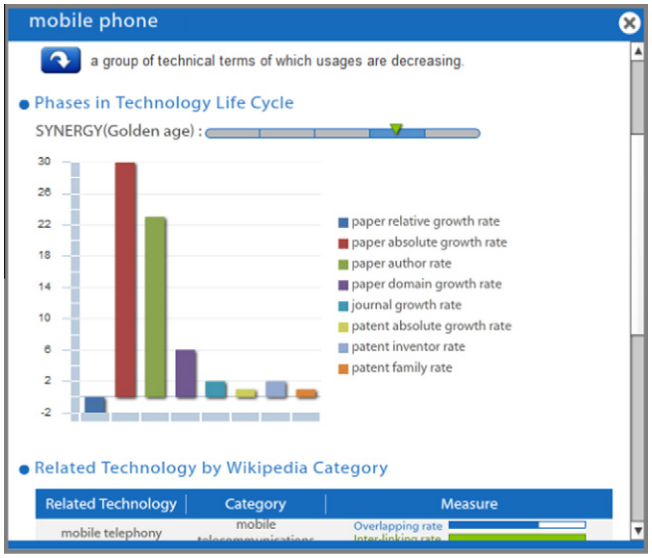

Fig. 2. Values of feature sets.

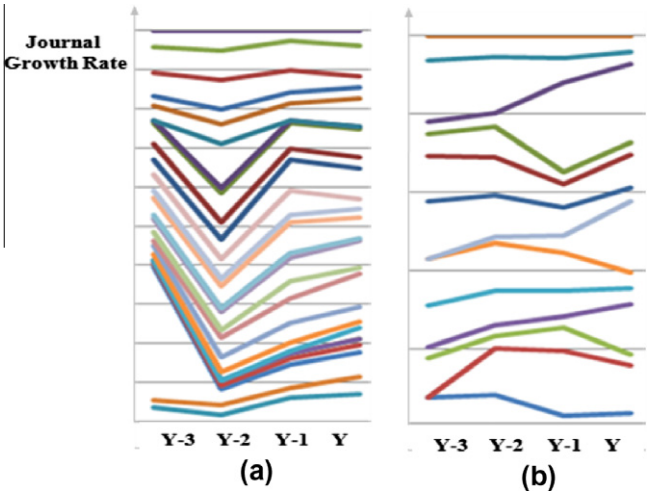

Fig. 3. Journal growth rate in (a) irruption phase and (b) frenzy phase.

Table 1
Decision tree induction algorithm.

Top-down decision tree induction
function GROW_TREE (T: set of examples)
returns decision tree:
t^*:=optimal_test (T)
p:=partition induced on T by t^*
if stop_criterion (p)
then return leaf (into (T))
else
for all P_j **in** P:
tr_j:=GROW_TREE (P_j)
Return node (t^*, ∪j{j,tr_j})
Single node refinement
for all candidate tests t associated with the node:
for all examples e in the training set T:
update_statistics (S[t], t (e), target (e))
Q[t] :=compute_quality (S[t])
$t^* := \arg_{\max(t)} Q[t]$
partition T according to t^*

number of technologies in the 'Irruption' phase have similar pattern such as decrease (Y-3 ∼ Y-2), rapid increase (Y-2 ∼ Y-1), and slight increase (Y-1 ∼ Y). However, technologies in the 'Frenzy' phase have irregular and various patterns as shown in Fig. 3(b). As a result, we can conclude the 'Journal Growth Rate' feature as a representative feature for deciding 'Irruption' emerging phase but the 'Journal Growth Rate' feature does not have big impact on deciding 'Frenzy' emerging phase. By performing feature selection repeatedly, we can decide typical features for each emerging phase.

3.1.2. Decision making and machine learning

After feature extraction and selection, we can create a decision tree based on calculated values of feature set as Fig. 4. To determine the phase of technologies, firstly we set two-level decision tree. The reason why we create decision tree is for higher decision accuracy. Mainly, 'Irruption' and 'Synergy' phases are decided at the early part of the decision tree but 'Frenzy' and 'Turning Point' phases are concluded at the end part of the tree. Therefore, decision accuracy for 'Irruption' and 'Synergy' phase is high but that for 'Frenzy' and 'Turning Point' is not. As a result, we use two separate decision tree for guaranteeing much higher accuracy.

Creation of decision tree is based on C4.5 algorithm. The C4.5 algorithm (Du, Wnag, & Gong, 2011; Yi, Lu, & Liu, 2011) is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm.

The constructed decision tree is optimized machine learning method. For machine learning of the decision tree, we use WEKA tool, C4.5 decision tree algorithm, and decision tree induction. WEKA tool (WEKA, 2010) is machine learning and data mining tool coded in Java. The tool was developed by University of Waikato in New Zealand and freeware and open source software. It supports classification, clustering, association, and visualization. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

We also use top-down induction method of decision tree for tree optimization as Table 1. Basically, given a data set, a node is created and a test t^* is selected for that node. A test is a function from the example space to some finite domain (e.g., the value of a discrete attribute, or the Boolean result of a comparison between an attribute and some constant). Each test induces a partition to the data set, with each subset of the partition corresponding to a single test result and containing those data elements for which the test yields that result. Typically the test for which the subsets of the partition are maximally homogeneous with respect to some target attribute (the "class", for classification trees) is selected. For

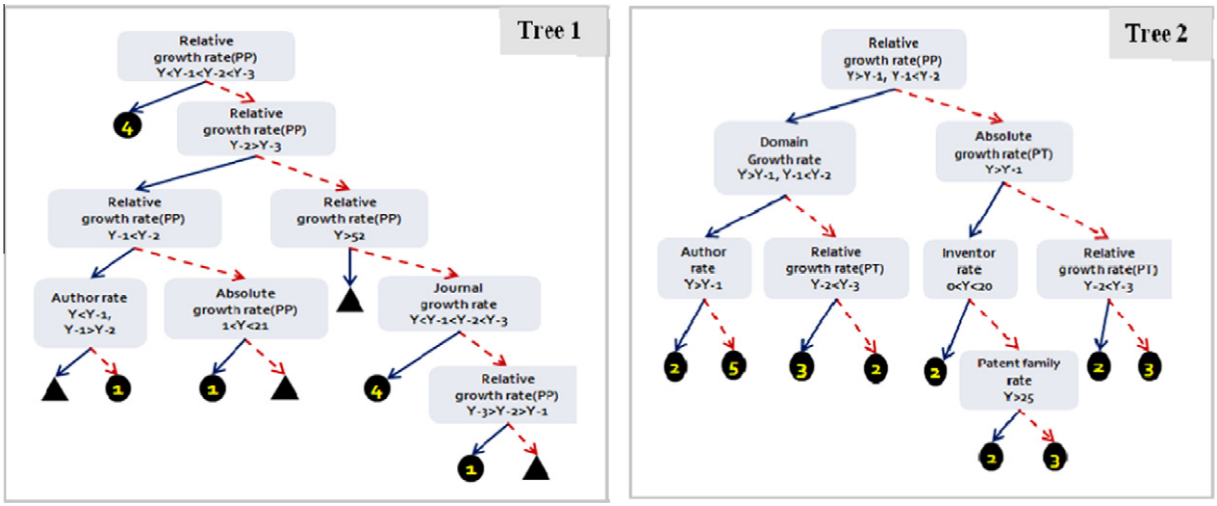

Fig. 4. Decision tree.

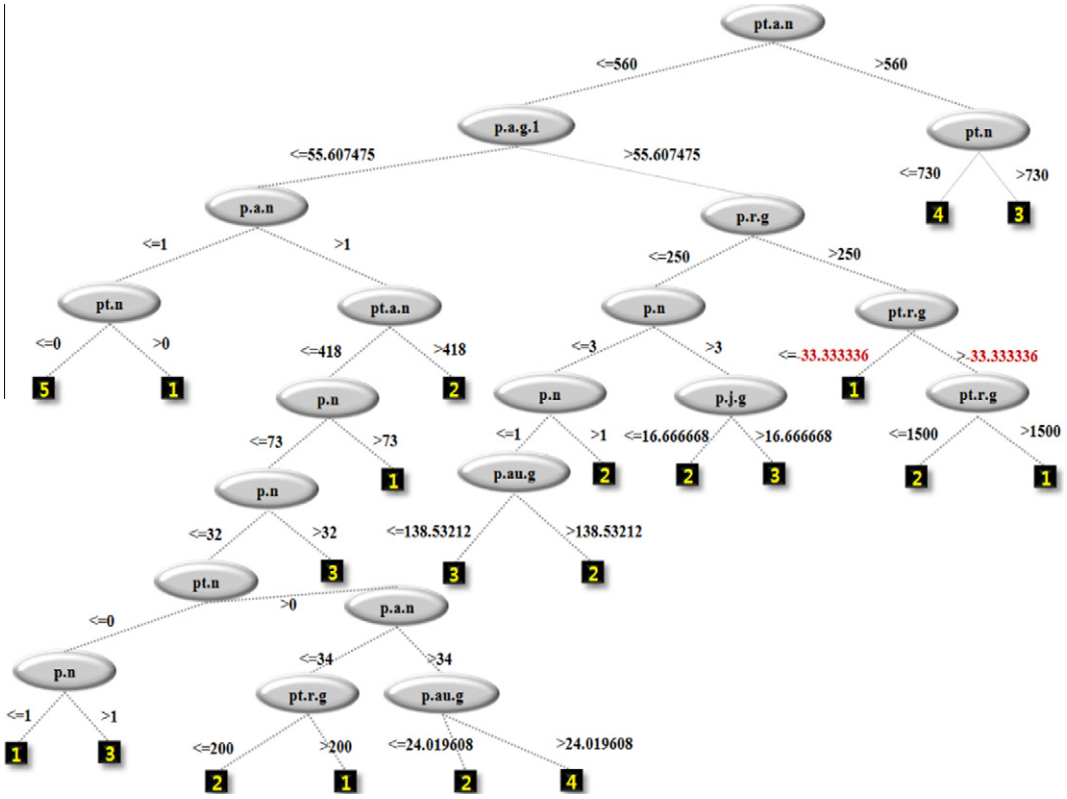

Fig. 5. Decision tree with machine learning.

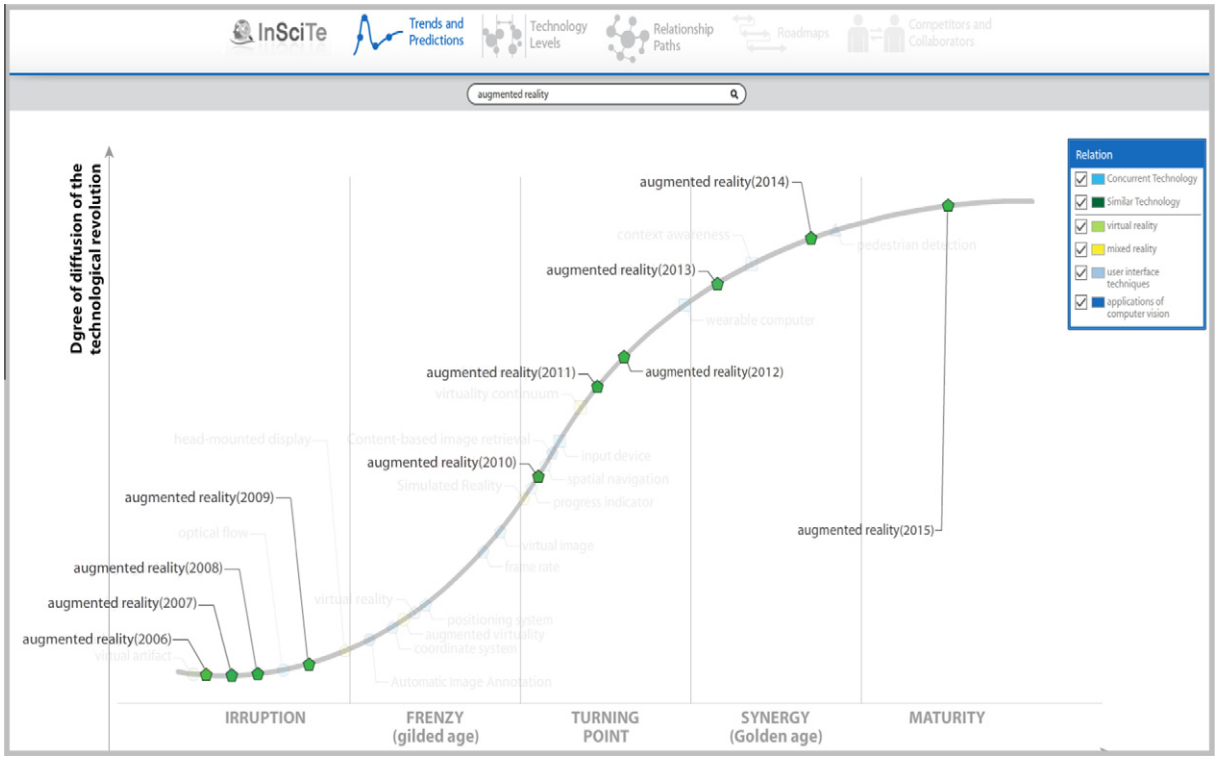

Fig. 6. Service by the TLCD model.

each subset P_j of the partition P induced by t^*, the procedure is repeated and the created nodes become children of the current node. The procedure stops when stop criterion succeeds: this is typically the case when no good test can be found or when the data set is

sufficiently homogeneous already. In that case the subset becomes a leaf of the tree and in this leaf information about the subset is stored (such as the majority class). The result of the initial call of the algorithm is the full decision tree. The computation of the quality of a test t is split into two phases there: one phase where the statistics of t are computed and stored into an array $S[t]$, and a second phase where the quality of t is computed from the statistics. For instance, for classification trees, phase one could compute the class distribution for each outcome of the test. Quality criteria such as information gain or gain ratio can easily be computed from this in phase two. For regression, where variance is typically used as a quality criterion, a similar two-phase process can be defined: the variance can be computed from (y2i,yi;1) where the yi's are the target values.

Through continuous machine learning and tree optimization process, we acquire optimized tree as Fig. 5 which can guarantee higher decision accuracy. By the optimized tree, we can predict future trends and conclude emerging phase of technologies.

As a result, the TLCD model can analyze technology trends and conclude technology emerging phase as shown in Fig. 6.

3.2. Emerging technology discovery model

The ETD model supports discovering emerging technology in various kinds of fields such as IT, medical, physics, mathematics, life science, energy and resources, and so on. In our system, we are managing more than about 70,000 technical terms. Technical terms include a name of technology and technology category. The ETD model decides which technology will be more promising for future than other technologies and selects emerging technologies among technical terms in various kinds of fields. The ETD model defines 5-sub definition of emerging technology as follows.

Definition 2 (*Triggered emerging technology*). technologies appeared within last 2 years and having higher growth rate than average growth rate.

$$S(Pp) = \{Pp_1, Pp_2, \ldots, Pp_n\}, S(Pt) = \{Pt_1, Pt_2, \ldots, Pt_m\}$$

$$S(Pp) \supseteq S^{t_1}(Pp), S(Pt) \supseteq S^{t_1}(Pt)$$

$$S(Pp) \Longleftrightarrow S^{t_1}(Pp) \cup S^{t2}(Pp) \cup \cdots \cup S^{t_o}(Pp)$$

$$S(Pt) \Longleftrightarrow S^{t_1}(Pt) \cup S^{t2}(Pt) \cup \cdots \cup S^{t_p}(Pt)$$

$$N^k_{Pp}(t1) + N^{k-1}_{Pp}(t1) > \sum_{l=k-k_{ini}}^{k-2} N_{Pp}(t1) = AN^{k-2}_{Pp}(t1)$$

$$N^k_{Pt}(t1) + N^{k-1}_{Pt}(t1) > \sum_{l=k-k_{ini}}^{k-2} N_{Pt}(t1) = AN^{k-2}_{Pt}(t1)$$

$$FS^k_{RelativeGrowthRate}(S^{t_1}(Pp)) > \sum_{r=2}^{o} FS^k_{RelativeGrowthRate}(S^r(Pp))/(o-1)$$

$$FS^k_{RelativeGrowthRate}(S^{t_1}(Pt)) > \sum_{r=2}^{p} FS^k_{RelativeGrowthRate}(S^r(Pt))/(p-1)$$

$$FS^{k-1}_{RelativeGrowthRate}(S^{t_1}(Pp)) > \sum_{r=2}^{o} FS^{k-1}_{RelativeGrowthRate}(S^r(Pp))/(o-1)$$

$$FS^{k-1}_{RelativeGrowthRate}(S^{t_1}(Pt)) > \sum_{r=2}^{p} FS^{k-1}_{RelativeGrowthRate}(S^r(Pt))/(p-1)$$

Definition 3 (*Associated emerging technology*). Top two technologies within the last 2 years and associated and similar technologies to those technologies.

$$N^k_{Pp}(t_1) \geqslant \forall_{m=t} N^m_{Pp}, N^{k-1}_{Pp}(t_1) \geqslant \forall_{n=t} N^n_{Pp}$$

$$N^k_{Pt}(t_1) \geqslant \forall_{m=t} N^m_{Pt}, N^{k-1}_{Pt}(t_1) \geqslant \forall_{n=t} N^n_{Pt}$$

$$WT(AS) * AS_{Pp}\left(N^k_{Pp}(t_1) \geqslant \forall_{m=t} N^m_{Pp}\right) \cap WT(SM) * SM_{Pp}\left(N^k_{Pp}(t_1) \geqslant \forall_{m=t} N^m_{Pp}\right)$$

$$WT(AS) * AS_{Pt}\left(N^k_{Pt}(t_1) \geqslant \forall_{m=t} N^m_{Pt}\right) \cap WT(SM) * SM_{Pt}\left(N^k_{Pt}(t_1) \geqslant \forall_{m=t} N^m_{Pt}\right)$$

Definition 4 (*Matured emerging technology*). Technologies which are already matured and can arrive at the maturity emerging phase within future 2 years.

$$DS^k(t_1) = \sum_{n=k-2}^{k} EP^n(t_1)/3$$

$$FS(t_1) = \sum_{m=k-3}^{k} DS^m(t_1)/4 = \alpha * \{EP^{k-5}(t_1) + (1-\alpha) * EP^{k-4}(t_1) + (1-\alpha)^2 * EP^{K-3}(t_1) + (1-\alpha)^3 * EP^{K-2}(t_1) + (1-\alpha)^4 * EP^{K-1}(t_1) + (1-\alpha)^5 * EP^{K}(t_1)\} + (1-\alpha)^6 * EP^{K+1}(t_1)$$

$$EP^k(t_1) + nFS(t_1) \geqslant 5 \Longleftrightarrow nFS(t_1) \geqslant 5 - EP^k(t_1)$$

$$\Longleftrightarrow 2 \geqslant n \geqslant (5 - EP^k(t_1))/nFS(t_1) \Longleftrightarrow (5 - EP^k(t_1))/nFS(t_1) \leqslant 2$$

Definition 5 (*Referred emerging technology*). Technologies which is commonly referred by several other papers.

$$S^{t_1}(Pp) = \{P^{t_1}_1, P^{t_1}_2, \ldots, P^{t_1}_n\}$$

$$REF_{Pp}(P^{t_1}_1) = \left\{P^{REF(t_1)}_1, P^{REF(t_1)}_3, P^{REF(t_1)}_5, \ldots, P^{REF(t_1)}_m\right\}$$

$$REF_{Pp}(P^{t_1}_2) = \left\{P^{REF(t_1)}_1, P^{REF(t_1)}_2, P^{REF(t_1)}_5, \ldots, P^{REF(t_1)}_o\right\}$$

$$ComREF_{Pp}(t_1) = REF_{Pp}(P^{t_1}_1) \cap REF_{Pp}(P^{t_1}_2) \cap \cdots \cap REF_{Pp}\left(P^{t_1}_p\right) = \left\{P^{REF(t_1)}_1, P^{REF(t_1)}_5\right\}$$

$$TechTerm\left(P^{REF(t_1)}_1\right) = \{t1, t2, t3, t4, t5\}$$

$$TechTerm\left(P^{REF(t_1)}_5\right) = \{t1, t3, t4, t_6\}$$

$$\therefore REFERRED\_EMERGING\_TECHNOLOGY = TechTerm(P^{REF(t_1)}_1) \cap TechTerm(P^{REF(t_1)}_5) \Longleftrightarrow \{t3, t4\}$$

Definition 6 (*Derived emerging technology*). Technologies which is commonly collocated in many papers/patents

$$S^{t_1}(Pp) = \{Pp^{t_1}_1, Pp^{t_1}_2, \ldots, Pp^{t_1}_n\}, S^{t_1}(Pt) = \{Pt^{t_1}_1, Pt^{t_1}_2, \ldots, Pt^{t_1}_m\}$$

$$CoTerm(Pp^{t_1}_1) = \{t_3, t_4, t_6, t_7, t_9, t_{11}\}$$

$$CoTerm(Pp^{t_1}_2) = \{t_2, t_4, t_7, t_9, t_{13}, t_{14}\}$$

$$CoTerm(Pt^{t_1}_1) = \{t_4, t_9, t_{11}, t_{13}\}$$

$$CoTerm(Pt^{t_1}_2) = \{t_2, t_4, t_9, t_{12}, t_{13}, t_{15}, t_{16}, t_{17}, t_{18}\}$$

$$DERIVED\_EMERGING\_TECHNOLOGY = \left\{\bigcap_{k=1}^{n} CoTerm(Pp^{t_1}_k)\right\} \cap \left\{\bigcap_{r=1}^{m} CoTerm(Pt^{t_1}_r)\right\} \Longleftrightarrow \{t_4, t_9\}$$

Table 2 represents description about several kinds of symbols and functions used in definitions of emerging technologies. In this paper, emerging technologies are extracted by quantitative analysis with several criteria defined in Definitions 2–6 finally. Results by

Table 2
Description of symbols and functions.

Representation	Description
$N^k_{Pp}(t1)$	Number of paper about tech. term 't_1' in 'k' year
$AN^k_{Pp}(t1)$	Accumulate Number of paper about tech. term 't_1' in 'k' year
$WT(AS)$	Weight for extracting of associated tech.
$WT(SM)$	Weight for extracting of similar tech.
$AS_{Pp}(t_1)$	Associated tech. about tech. term 't_1' based on paper information
$SM_{Pp}(t_1)$	Similar tech. about tech. term 't_1' based on paper information
$DS^k(t_1)$	Draft development speed about tech. term 't_1' in 'k' year
$EP^n(t_1)$	Emerging phase of tech. term 't_1' in 'k' year
$FS(t_1)$	Final development speed about tech. term 't_1' in 'k' year
$REF_{Pp}(P^{t_1}_1)$	Reference lists of 'P_1' paper
$ComREF_{Pp}(t_1)$	Reference lists referred commonly by many papers
$TechTerm(P^{t_1}_1)$	Tech. tem lists of 'P_1' paper
$CoTerm(Pp^{t_1}_1)$	Co-located term lists based on entire contents of 'Pp_1' paper

Table 3
Simulation results of the TLCD model.

Emerging phase	Phase accuracy		
	Gartner		Suggested
Irruption	180		170
		94.4%	
Frenzy	100		80
		80%	
Turning point	90		60
		66.7%	
Synergy	60		60
		100%	
Maturity	60		50
		83.3%	
Total	**490**		**420**
		85.7%	

the emerging technology discovery model are represented in the InSciTe service as Fig. 7.

4. Experiment

In this paper, we use 'Hype Cycle for Emerging Technology' suggested by Gartner (Fenn, 2010) from 2008 to 2011 for evaluating the suggested model. Firstly, we obtain about 500 emerging technologies from Gartner's model and refine duplicated technologies. Additionally, we utilize papers and patents stored in 'National Discovery for Science Leader (NDSL) system (NDSL, 2010) developed by Korea Institute of Science and Technology Information (KISTI). The Gartner's hype cycle model is widely and generally used by several companies, universities, and research agencies for discovery of technology opportunity and emerging technology. In addition, the Gartner's model is utilized for business analysis, future business planning, future strategy establishment, promising research area discovery, emerging research technology discovery, and so on. Because the Gartner's model has better reliability, usability, and credibility than other models, we evaluate our model using Gartner's model as an answer set.

Table 3 shows results of experiment between Gartner's hype cycle model and TLCD model suggested in this paper. Entirely, Suggested TLCD model represents 85.7% decision accuracy compared to Gartner's model. The reason why second and third emerging phase have lower decision accuracy than other phases is second and third emerging phases mainly decided in the leaf nodes in the decision tree.

For validating the ETD model, we perform simulation test with 50 technologies in Gartner's emerging technology 2011 as a comparison target. First of all, we extract 69,000 technologies in several kinds of field from papers and patents information. Then we apply five concepts of emerging technology described in Section 3

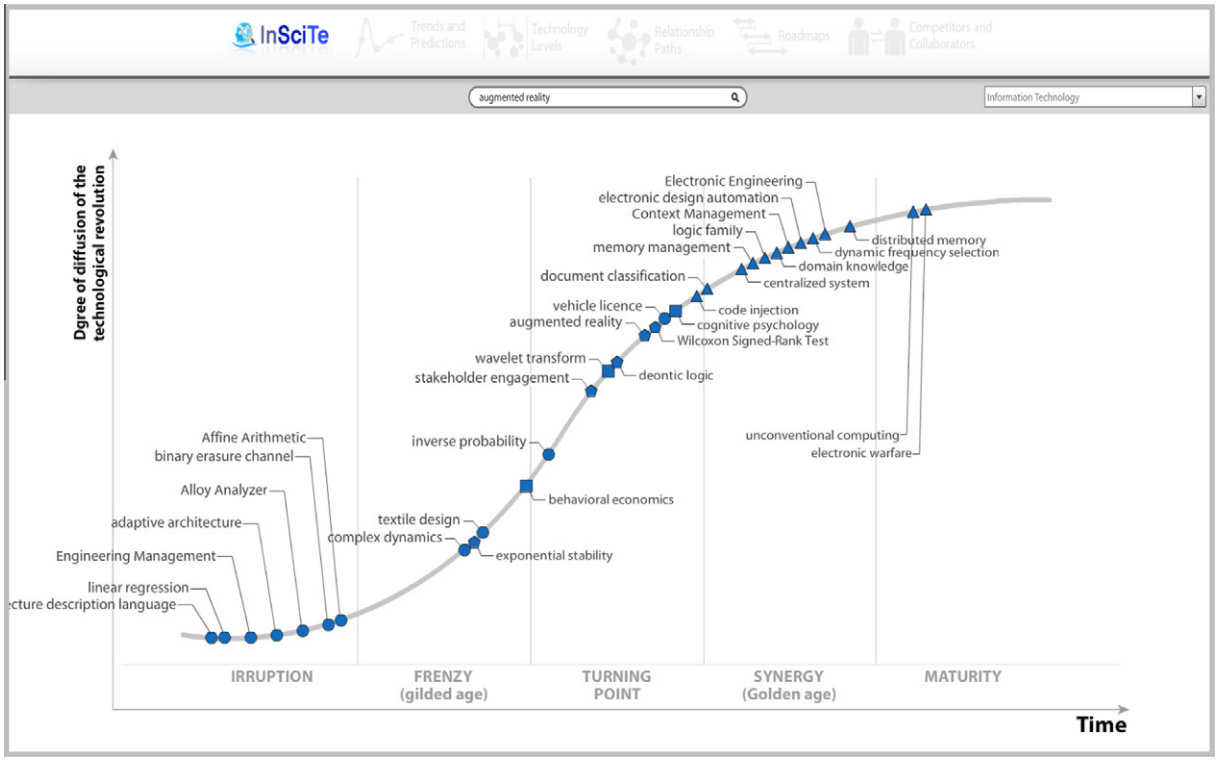

Fig. 7. Service by the ETD model.

Table 4
Simulation results of the ETD model.

Tech.	Results		Tech.	Results	
	Gartner	ETD		Gartner	ETD
3-D flat-panel displays	Y	Y	Speech-to-speech translation	Y	Y
3D printing	Y	Y	Tangible User interfaces	Y	**N**
Augmented reality	Y	Y	Terahertz waves	Y	**N**
Behavioral economics	Y	**N**	Video search	Y	Y
Context delivery architecture	Y	Y	Video Telepresence	Y	Y
Quantum computing	Y	N	Cloud computing	Y	Y
Surface computers	Y	N	E-book readers	Y	Y
Video search	Y	Y	Internet TV	Y	Y
3D printing	Y	Y	Microblogging	Y	Y
Autonomous vehicles	Y	**N**	3D flat-panel TVs and displays	Y	Y
Computer–brain interface	Y	Y	Augmented reality	Y	Y
Context delivery architecture	Y	**N**	Cloud computing	Y	Y
Mesh networks:sensor	Y	Y	Location-aware applications	Y	Y
Public virtual worlds	Y	Y	Pen-centric tablet PCs	Y	Y
Social network analysis	Y	Y	blogs	Y	Y
E-Book readers	Y	Y	Business process analysis	Y	Y
Gesture recognition	Y	Y	traditional EA approach	Y	**N**
Microblogging	Y	Y	Podcasting	Y	**N**
Public virtual worlds	Y	N	Web platforms	Y	Y
……	……	……	……	……	……

to the ETD model. We sort extraction results acquired from five concepts of emerging technology by weight and determine high-ranked 50 technologies as emerging technology. Finally, we compare 50 technologies in Gartner's emerging technology 2011 and emerging technologies extracted by the ETD model. As shown in Table 4, 41 technologies determined as emerging technologies by the ETD model are same as emerging technologies in Gartner's emerging technology lists. As a result, the ETD model shows 82% forecasting accuracy regarding emerging technology determining.

5. Conclusion

In this paper, we designed suggested model for effective information analysis and future forecasting based on papers and patents information. Compared to conventional method, projects, and services, the suggested model supports more systematic process and objective analysis/forecasting results. The suggested model consists of TLCD model, TMF model, and ETD model for much more diverse information analysis and forecasting information provision to interested users. The TLCD model decides emerging phase of technologies in technology life cycle. By the TLCD model, we can calculate technology emerging phases from 2006 to current year. The TMF model predicts technology development speed, maturity, and technology emerging phase in future years. but because the TMF model is in conceptual progress, we did not describe it in this paper. The ETD model selects emerging technology using several definitions such as triggered emerging technology (ET), associated ET, matured ET, referred ET, and derived ET. To evaluate the suggested model, we compared the results of our model to Gartner's emerging technologies and hype cycle. The TLCD model shows 84% accuracy and the ETD model represents 82% forecasting accuracy compared to Gartner's model.

As future works, we will perform simulation test with diverse datasets. Except for Gartner's hype cycle, many other research center and governments such as MIT and Berkley, and etc. predict emerging technologies. Simulation test with several kinds of dataset will improve forecasting accuracy much higher.

Additionally, We have to define more definition about emerging technology except for conventional five definitions. By many kinds of definitions about emerging technology, we can optimize extraction process of emerging technology and acquire much more accurate emerging technology lists.

References

Carolta, P. (2007). Great Surges of Development and Alternative Forms of Globalization, TUT Institute of Public Administration Technical Report.

Combining and Uniting Business Intelligence with Semantic Technology Project. (2008). SAP, OntoText, Sheffield Hallam Univ., Innovantage, Heriot Watt Univ., SpaceApplication, http://www.cubist-project.eu.

Dereli, T., & Durmusoglu, A. (2009). A trend-based patent alert system for technology watch. *Journal of Scientific & Industrial Research, 68*(8), 674–679.

Duran, O. (2011). Computer-aided maintenance management systems selection based on a fuzzy AHP approach. *Advances in Engineering Software, 42*(10), 821–829.

Du, M., Wnag, S., & Gong, G. (2011). Research on Decision tree algorithm based on information entropy. *Advanced Materials Research, 267*(1), 732–737.

Fenn, J. (2010). Gartner's Hype Cycle Special Report for 2010, Gartner Research.

Foresight and Understanding for Scientific Exposition. (2010). DARPA, http://www.iarpa.gov/solicitations_fuse.html.

Hanafizadeh, P., & Mirzazadeh, M. (2011). Visualizing market segmentation using self-organizing maps and fuzzy Delphi method- ADSL market of a telecommunication company. *Expert Systems with Applications, 38*(1), 198–205.

Hetmanska, K., & Nguyen, A. (2011). A method for learning scenario determination and modification in intelligent tutoring systems. *International Journal of Applied Mathematics and Computer Science, 21*(1), 69–82.

John, H. (1995). *Technical Change and the World Economy-Convergence and Divergence in Technology Strategies*. NewYork: Elsevier.

Kim, J., Lee, S., Lee, J., Lee, M., & Jung, H. (2011). Design of TOD Model for Information Analysis and Future Prediction. *Communications in Computer and Information Science, 264*(1), 301–305.

Kim, Y., Suh, H., & Park, P. (2008). Visualization of patent analysis for emerging technology. *Expert System Application with Applications, 34*(1), 1805–1812.

KISTI, Korea Institute of Science and Technology Information, http://www.kisti.re.kr.

Liu, G., Jin, Y., & Li, F. (2011). The application of AHP method in well control risk evaluation by controllable factor analysis. *Journal of Southwest Petroleum University, 33*(2), 137–141.

National Discovery for Science Leaders (NDSL). (2010). http://www.ndsl.kr.

Okoli, C., & Pawlowski, S. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information and Management, 42*(1), 15–29.

Rann, A. (1998). *Handbook of quantitative studies of science and technology*. NewYork: Elsevier.

Richard, S. (1983). Patent Trends as a technological Forecasting Tool. *World Patent Information, 5*(3), 137–143.

Song, F., Hao, S., & Hao, M. (2011). Study on NC instruction interpretations algorithm with high-order differentiability based on moving average. *Mechatronics and Intelligent materials, 211*(1), 900–903.

VantagePoint. (2009). Text Mining Software for Technology Management-Search Technology, Inc., http://www.thevantagepoint.com.

WEKA. (2010). Data Mining with Open Source Machine Learning Software in Java, http://www.cs.waikato.ac.nz/ml/weka.

Wright, G., & Goodwin, P. (2009). Decision making and planning under low levels of predictability: Enhancing the scenario method. *International Journal of Forecasting, 25*(4), 813–825.

Yi, W., Lu, M., & Liu, Z. (2011). Multi-valued attribute and multi-labeled data decision tree algorithm. *International Journal of Machine Learning and Cybernetics, 2*(2), 67–74.