Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 76, No. 3 (2008) 503–525
DOI: 10.1007/s11192-007-1898-2

How to identify emerging research fields using scientometrics: An example in the field of Information Security

WOO HYOUNG LEE

Institute for Information Technology Advancement, Daejeon (Korea)

In the highly competitive world, there has been a concomitant increase in the need for the research and planning methodology, which can perform an advanced assessment of technological opportunities and an early perception of threats and possibilities of the emerging technology according to the nation's economic and social status.

This research is aiming to provide indicators and visualization methods to measure the latest research trend and aspect underlying scientific and technological documents to researchers and policy planners using "co-word analysis". Information Security field is a highly prospective market value. In this paper, we presented an analysis Information Security.

Co-word analysis was employed to reveal patterns and trends in the Information Security fields by measuring the association strength of terms representatives of relevant publications or other texts produced in the Information Security field. Data were collected from SCI and the critical keywords could be extracted from the author keywords. These extracted keywords were further standardized. In order to trace the dynamic changes in the Information Security field, we presented a variety of technology mapping. The results showed that the Information Security field has some established research theme and also rapidly transforms to embrace new themes.

Introduction

The world has demanded a shift to a new paradigm due to the limitations of the growth strategies centered around labor and capital, and as a result, a shift to the

Received August 30, 2007

Address for correspondence:
WOO HYOUNG LEE
Institute for Information Technology Advancement, 58-4 Hwaam-Dong, Yuseong-Gu, Daejeon, Korea
E-mail: leewh@iita.re.kr or leewh1149@hanmail.net

0138–9130/US $ 20.00

knowledge-based economy where technology and knowledge form the core elements has been successful. The catalyst that stimulates a shift to the knowledge-based economy is information on science and technology.

For instance, approximately 60% of the important discoveries and achievements in science and technology in the 2,000-year-long history of mankind have taken place for the past 100 years. Moreover, the achievements that have been made for the past 20 years are known to take 80% of those that took place in the 20th century. As shown by the statistics, science and technology are now developing exponentially [MOON, 2004].

To cope with such global trends, the government in Korea has made its utmost effort by establishing a basic plan for science and technology and investing approximately 13 trillion won for new fields of technology (6 T's including Information Technology, Bio Technology, and Nano Technology) from 2002 to 2006. As the amount of investment for new technology is increasing, the orientation of investment must be adjusted in a consistent manner without any discrepancy with the strategic goals of the nation and the demands in the market, and especially strategies expected to enhance efficiency must be established. In reality, however, numerous possibilities inherent in new science and technology coexist with uncertainty, and the ways to find, foster, and develop promising science and technology must be seriously taken into account. In accordance with such trends, related research efforts have been actively made on domestic and global scales. One of the representative methodologies is the analysis of science and technology information resources, which will be the main focus of this research.

It is very important to schematize the association among concepts, ideas, and problems in the fields of pure science and social science, and several methods have been attempted to meet the requirement. The traditional method that has been mainly used for science research and policies is to ask a relatively few specialists for advice [LAW & WHITTAKER, 1992]. Another method used to carry out such research from a quantitative perspective is the analysis of science and technology information resources or the analysis of bibliographies.

A large number of reference materials in the science of bibliography helped transform several perspectives in the analysis of bibliographies into the process of knowledge search and data-mining. This is a field derived from the traditional way of analysis where the entire domain of knowledge is regarded as one unit of analysis. Domain visualization places an heavy emphases on the role of information visualization when the structure and driving force of the domain of knowledge are examined, scrutinized, and developed. A promising trend has emerged with many expectations through the synergy of a number of academic specialties including philosophy, sociology, bibliography, information visualization, and the analysis of domains.

Faced with an innovation-led economy and a paradigm shift in science and technology policies, Korea is now going through a phase where the source technology must be acquired by selection and integration while the traditional catch-up strategy is

being maintained. Rather than the mere pursuit of science and technology, the establishment of strategic knowledge and innovation system in accordance with the national agenda is necessary [OH & LEE, 2004]. Moreover, the research and development effort, in and of itself, requires a substantial amount of national budget and can cause a huge ripple effect on every corner of science and technology as well as the economic society on the whole, so a careful, objective approach to deciding which type of business is to be cultivated or whether a certain type of business is to be launched must be taken. A number of scholars and researchers have made countless attempts through objective methods of research to find the most suitable type of business that meets such demands. Designed for such purposes, this research requires an effort to find a business to be raised at an early stage, and as part of such an effort, research on the deduction of new technology through the analysis of science and technology information resources has been conducted.

In the course of this research related theories have been studied to present the methodology of deducing new technology through the analysis of information together with a positive analysis to ensure the validity of the deduced methodology. The first step in conducting research on theories is to scrutinize advanced principles most suitable for this type of research through discussion and study with specialists in the field. Next, the information security field was selected to conduct positive research, and new technology in that field was deduced. The results of the analysis were then discussed with specialists in the field, and the most ideal conclusions were attained.

Theoretical consideration

The co-word analysis is the analysis of contents using a pattern where a pair of items simultaneously appears in the linguistic part of texts in order to identify the relationship between ideas in a particular division suggested by the texts. The co-word analysis was first attempted in a system called "LEXIMAPPE" developed by Ecole Nationale Superieure des Mines Centre de Sociologie de l'Innovation in a concerted effort with CNRS (Centre National de la Recherche Scientifiqie) in France in the 1980's. The co-word analysis was consolidated as a methodology in a book titled "Mapping of the Dynamics of Science and Technology" [CALLON & AL., 1986], proliferating into other nations (the U.K., Holland, and the U.S. to name a few) and enhancing the process, measurement, and interpretation of data [HE, 1999].

The preceding examples of research that conceptualized the tendencies and periodical changes of a theme by using the co-word analysis are as follows:

PETERS & VAN RAAN [1993] used the direct-MDS method that directly applies the multi-dimensional scaling to analyze a domain in Chemical Engineering as well as the cluster-MDS method that combines clustering with the multi-dimensional scaling.

Moreover, PETERS & VAN RAAN [1993] used 3 different sample groups, the measuring index of a correlation between three words, and two types of words in order to compare domain maps in Chemical Engineering.

NOYON & VAN RAAN [1998] conducted research on the neural network from 1989 to 1990 and from 1992 to 1993 in order to inspect the past occasions within the framework of the present and to investigate the present cases based on the past makeup.

DING & AL. [2000] suggested BIRS (Bibliometric Information Retrieval System), a system that searches information on the web and the online database by applying Computation & Information Science to information search. BIRS is a system with a visual interface realized to expand the inquiries of a user through an analytical method used in Computation & Information Science. The user can search information through the author map, the journal map, and the keyword map, among which the keyword map is designed such that the user can access the desired information from a variety of dimensions to see results. A domain map with a wide range, a domain map with greater details, and top 20 words with great similarities are suggested at phase 1, 2, and 3 respectively.

DING & AL. [2001] analyzed changes according to a thematic domain and a particular period on the basis of the documentation found by SCI and SSCI in the field of information search. They collected 3,227 keywords excerpted from 2,012 documents and selected 240 keywords in total through a manual method that adds keywords excerpted from a title and an abstract as well as a standardization method that applies the LISA thesaurus. By using the frequency rate of simultaneous appearance of such keywords, they created 5 clusters while generating a domain map by using the multi-dimensional scaling. Moreover, they generated a detailed domain map categorized into each cluster.

MORRIS [2005] had developed what is known as the Timeline Map. This methodology asserted that if the frequency distribution of the simultaneous citation of references consulted and applicable references are shown on a graph, this kind of distribution will have the form of a power law.

LIN & CHIANG [2007] is a study that revealed whether Chinese herbal medicine and its use have enjoyed a rapid increase in recent times. For this purpose, it analyzed the *American Journal of Chinese Medicine* for the years 2002~2004. The AHP was utilized as an analytical method, and the study explained how the AHP method was employed.

GALVEZ & MOYA-ANEGÓN [2007] delineated the method of raising the quality of data when the database is utilized for bibliometric purposes. The study clarified the Enhanced-State Transducers (E-FST), which is an improved form of existing Natural Language Processing (NLP).

TIJSSEN [2007] is a study that showed different characteristics of science research in Africa, especially in terms of the quantitative 'scientometric.' In this endeavor, the study has analyzed the research results of Africa-based authors for the extended period

between 1980~2004 by using the SCI DB. The results of the analysis revealed that research productivity in Africa is progressively declining and that international co-publication is on the rise.

LEYDESDORFF [2007] discussed the combining methods for a journal-map analysis of the relationship between SCI2004 and SSCI2004. For this purpose, the study used the Vector-space model and, as part of its conclusion, proposed an indicator that can specify the interdisciplinary journal.

The methodology of deducing new technology through S&T information resources

The research methodology

Synopsis. The first thing to be considered in analyzing science and technology information resources is the characteristics of new technology. The characteristics of new technology imply, in other words, the uncertain nature of new technology. As a preliminary step toward the deduction of new technology through the analysis of science and technology information resources, using varied methods of analysis in accordance with the uncertainty of new technology has been advocated (refer to Table 1). This research is focused on the development of methods capable of analyzing new technology with over 3–4 of the uncertainty level.

Therefore, the method of deducing new technology to be developed in this research is applicable only to technology with the highest level of uncertainty. As a method of analysis to overcome such uncertainty inherent in new technology, analysis through a network has been debated by many.

The purpose of analyzing such a network is to deduce the characteristics of a structure or a link and to explain the characteristics of the system based on the association as well as the behavior of each unit constituting the system. Understanding the interconnectivity of knowledge resources necessary for the research through such a network helps anticipate or restrict the promising fields. The consideration of a network helps anticipate the dissemination path of promising fields and strengthen the nodes, presenting the orientation of research to researchers and showing guidelines to whether the research in progress is headed in the right direction in terms of assignments and goals set on national or corporate scales.

Table 1. The analysis method corresponding to the uncertainty level of new technology

		Establishment of the range	Investigation	Evaluation	Execution	Analysis method
H ↑	Uncertainty Level 4	Establishing the strategic intention and analyzing the national capacity	Investigating into science and technology information resources	Analyzing the development pattern of new technology	Investing for systematic signal detection	KDD, Text Mining, Data Mining, Bibliometrics, TRM, Anticipating Technology Trends, Scenario Techniques
	Uncertainty Level 3	Deducing the range of technical development	Detecting weak signals and anticipating potential markets	Deducing a group of technical candidates and deducing a scenario as to market development	Creating and planning on prospective stories for each scenario	
Uncertainty Level	Uncertainty Level 2	Deducing technical alternatives	Detecting strong signals and understanding competitive technology and market	Establishing evaluative criteria and selecting technology	Creating and executing technical development plans	Real Option
↓ L	Uncertainty Level 1				Creating and executing commercialization plans	Traditional economic analysis techniques (ROI, NPV)

Source: [YOON & AL., 2004]

The aspects of a network that takes place in the activities of science and technology are virtually countless. A policy network, a cooperation network, and an organization network are only some of the examples. Integrating a variety of networks and deducing a universal principle may be literally impossible. However, it is good to know that there indeed exists a universal principle that generalizes various networks at any phase. The question now is whether such a principle can be applied to the science and technology network in an appropriate manner.

Accordingly, the fact that the method of deducing new technology can be applied to the analysis of a network is indeed of importance in that a network can be linked to the phenomena of knowledge resources necessary for other research activities and that a network is now given prominence in research on a complicated system.

The research framework. The method of deducing new technology through the analysis of science and technology information resources, which is the main focus of this research, can be established with a network taken into account.

First, the nature of reference data to be analyzed must be revealed prior to the establishment of an analysis method with a network being considered. The characteristics of data are categorized on the basis of whether they coincide with

random distribution or scale-free power law distribution. Second, a knowledge map at the point of analysis is generated through the analysis of words with simultaneous appearance, by which the characteristics of data to be analyzed can be examined. The data hub is then examined through the analysis of a network. The network analysis coordinates are used at that stage. Coordinates like connectivity, centrality, and centralization were used for the network analysis coordinates. Third, once the data hub has been found, it is necessary to trace back to the period when the hub first appeared and analyze the domain map in that period of time. Now the association between a young hub (i.e. the hub at the moment) and the hub in that period of time is analyzed with the coordinates mentioned above. Once the overall characteristics of this field have been made explicit, it is necessary to come back to the present time and search for a domain that has a possibility of turning into a hub in the future. That is, a domain whose characteristics are most similar to those of the coordinates that have a possibility of turning into a future hub in this field centered around the hub marked on the current domain map is selected.

The process of deducing new technology

Revealing the nature of source data. The first step toward investigating into new technology through the analysis of science and technology information resources is to reveal the nature of data to be analyzed. In other words, the characteristics of data to be analyzed are categorized on the basis of whether they have the nature of a random network or a scale-free network. If the data to be analyzed have the nature of a scale-free network, new technology can be deduced while the remaining process is being executed. A scale-free network is as follows:

Up until now we have not had any alternative plans to describe the interconnected world, and a random network has dominated our thoughts in modeling a network. That is, the networks in complex reality have been regarded as random entities in essence. However, in the course of research that has lasted for the past several decades, scientists have been able to perceive the fact that nature under certain circumstances generate an amount that coincide with the distribution of the power law rather than the bell-shaped function.

The power law distribution is a distribution characterized by a curve decreasing toward the right. In a scale-free network where the number of lines connecting nodes corresponds to the power law distribution, most nodes have only a small number of links, and they are interconnected by a few nodes with a large number of connecting lines. The principle of the power law is visually very similar to that of an aviation route network where a number of small airports are interconnected through a small number of major hubs.

The power law distribution that decreases with a gentle slope can naturally integrate hubs with a large number of links to a varying degree. This helps us expect that all scale-free networks will have several hubs with huge dimensions that regulate the phase structure of the entire networks. Through a number of positive analysis efforts, most theoretically important networks have turned out to be scale-free networks, and as a result, the validity of a hub has been acknowledged.

The examination of a hub

a. The concept of a hub. By testing the socializing behavior of people, GRADWELL [2000] concluded that regardless of social status there are always a small number of people with superb skills at making friends or acquaintances, and they can be defined as hubs or connectors. In addition, BARABASI [2002] proved through a positive analysis that hubs (i.e. nodes with an exceptional number of links) are found in a variety of complex systems ranging from economy to body cells. He claimed further that hubs are the fundamental attribute of most networks, stimulating the curiosity of scientists in various fields including biology, computer science, ecology, etc.

As GRADWELL [2000] and BARABASI [2002] claim, such hubs form an important part of our social networks, leading fads and trends, making important business transactions possible, spreading fads, and helping open a restaurant. They are like social threads that seamlessly interconnect people with different races, education levels, and family backgrounds. Therefore, also in the fields to be analyzed through an early warning system, which is the theme of this research, such hubs are expected to form a paradigm by interconnecting different academic fields, nations, and research areas and propagating a new type of study.

Hubs are indeed special, deserving our attention without a doubt. Moreover, hubs dominate the structure of all networks, playing a role in making a small world. That is, a substantial number of nodes and links inherent in each hub help make the distance between two nodes in a system shorter. As defined above, a network whose number of lines interconnecting nodes corresponds to the power law distribution is referred to as a scale-free network. Thus, we can expect that all scale-free networks will have several hubs with huge dimensions that regulate the phase structure of the entire networks.

b. The method of measuring hubs. Finding hubs may be one of the biggest issues for scholars interested in networks. However, the results of research that has been conducted up until now do not suggest any methods that can be universally applied to the analysis of a network. The most outstanding findings are as follows: First, CHEN [2004] introduced nodes with 3 different concepts including a turning point, a hub, and a pivot node as part of his attempt to find the turning point of knowledge through the analysis of a network. Unfortunately, a method that can accurately measure each node was not presented.

In this research, on the other hand, the coordinates used in the analysis of a network (in particular, those related to centrality) were employed for a positive analysis.

Centrality. In case a network is examined in a field of study, the following questions may come to your mind. What attributes make a field of research popular and what topics make it unpopular? What attributes does a popular theme of research have?

What is meant by popularity is that the corresponding field of research is located right in the center of a research network, and there are various ways to measure centrality, that is, how closely a field of research is located to the center. The analysis of a network makes it possible to measure the number of other defects connected to a certain defect and the number of steps taken for a certain defect to reach all the other defects.

For instance, the frequency of a research theme being selected becomes the yardstick of centrality, and a theme with a small amount of distance can be defined as having a high level of centrality. Sometimes the level of some themes reaching other themes via my own theme of research becomes my betweenness centrality. In the research conducted by CHEN [2004], for instance, a theme of research that falls under a pivot node can be a theme with a high level of centrality.

Centrality is one of the coordinates used most frequently in that it is closely associated with the concepts of power and influence. In most empirical analyses, a theme of research with a high level of centrality is given a special rank, and in case of a particular field of research, a high level of centrality guarantees a high survival rate and a good result. That is why centrality is often used as a good independent or dependent variable in a statistical analysis.

FREEMAN [1979], who made the biggest contribution to the development of the coordinate of centrality mainly used in the analysis of a network, categorized centrality into local centrality and global centrality.

As the level of direct connection between a certain defect and those surrounding it increases, the local centrality of that particular defect also increases. On the other hand, as one defect takes a strategically important position in the overall connection structure of a network, the overall level of centrality also increases. A defect with a high level of local centrality may be also high in global centrality, yet the two concepts do not necessarily coincide with each other.

Let us think about a network where mathematics and physics, for example, do not have any connectivity with each other. A person with the largest number of connections in the field of mathematics or physics is known to have a high level of local centrality. However, a theme of research that may not be directly connected to other themes in each group yet plays a role as a bridge between fields of mathematics and physics has a high level of global centrality although the level of local centrality may be low.

Local centrality. A degree, which implies the number of defects interconnected, is a good coordinate to measure local centrality. Local centrality in a vector graph is

referred to as in-centrality as to the connections headed "to" a defect and out-centrality as to those headed "from" that defect. As in the case of density, local centrality for those networks with different dimensions cannot be directly compared. It is because the larger a network is, the smaller the level of centrality for the points belonging to the network becomes. Thus, it is advisory that local centrality be compared for those networks with the same dimensions.

Z_{ijk} implies the relationship from agent i to agent j in network k.

$$Outdegree_{ik} = \sum_{i=1}^{N} Z_{ijk} = Z_{ik}$$

In case of asymmetry, out-degree is the number of relationships from agent i to all the other agents j.

$$Indegree_{ik} = \sum_{i=1}^{N} Z_{ijk} = Z_{ik}$$

On the other hand, out-degree is the number of relationships from all the other agents j to agent i. Degree centrality is measured as the ratio of in-degree to out-degree for each agent as to the total number of connections.

$$DegreeCentrality\ C_i = \sum_{j=1}^{N} (Z_{ij} + Z_{ji}) / \sum_{i=1}^{N} \sum_{i=1}^{N} (Z_{ij}) \quad (\text{only if } 0 \le C_i \le 1)$$

Global centrality. A representative coordinate that describes the global centrality of a point is close centrality and can be measured by closeness with or distance from other points. Now the distance between two points implies the path distance, that is, the shortest distance between the two points. A point with the smallest amount of path distance is a central point with the highest level of global centrality. As in the case of local centrality, global centrality in a vector graph is measured as in-closeness and out-closeness in accordance with the corresponding direction.

Betweenness centrality. FREEMAN [1979] suggested betweenness centrality as the third method to measure the centrality of a point. This type of centrality measures the level of one point being positioned in-between other points in a network. Betweenness centrality measures the level of a mediating role in-between other points.

Standardized Betweenness Centrality

$$C'_B(P_m) = \frac{\sum_{i}^{N} \sum_{j}^{N} \frac{G_{imj}}{G_{ij}}}{(N^2 - 3N + 2)/2} \quad (\text{only if } i<j,\ i \neq j)$$

In general, there exist several geodesics, the shortest paths connecting a pair of points.

$\sum_{i}^{N}\sum_{j}^{N} g_{ij}$, g_{ij} is the number of geodesics connecting i and j whereas $\sum_{i}^{N}\sum_{j}^{N} g_{imj}$ is the number of cases where point m is positioned on the geodesic between i and j. If all geodesics are identical, the probabilities of paths to be used by i and j are equal, and in case of $1/g_{ij}$, the probability of a path to be used increases as the frequency of appearance on the geodesics increases. The denominator is a number divided by the maximum value of the numerator so that the coordinates can be normalized. The numerator has the largest value in a radiating star network, and the value is expresses as ${}_{n-1}C_2$, which is $\{(N^2-3N+2)/2\}$.

The investigation into promising research areas through a positive analysis

Collecting data for analysis

The first step toward applying the method of selecting a theme of research as to information security technology is to collect data to be analyzed. The representative data commonly used for analysis include theses and patents, and as far as this research is concerned, only theses were considered.

The SCI-CD version, listing theses published from 1994 to 2003, was used for the analysis of theses. The range of keywords associated with information security technology was determined under discussion with professors in the information system department at Boston University. Based on the range of analysis, the following 28 keywords were selected.

Table 2. Primary keywords

Security, network, cryptography, encryption, authentication, secure, Intrusion, antivirus, detection, spam, vulnerability, digital signature , Verification, firewall, traffic, recovery, monitoring, hacking, attack, penetration, incident response, risk analysis, filtering, buffer overflow, Denial Service, DoS, audit, scan

As a result of the analysis, 50,392 keywords for author (294,394 without duplications) were selected from 8,464 documents excerpted. As a result of the cleansing process for these keywords, a substantial number of improper keywords turned out to be undesirably included, and an adjustment was made to the pool of keywords as shown in the Table 3.

Table 3. Secondary keywords

(security or secure or crypto) and (encryption or authentication or intrusion or verification or firewall or recovery or hacking or audit or scan or algorithm)

As a result of the analysis, 2,880 keywords for author (6,857 without duplications) were selected from 976 documents excerpted. Important keywords were directly extracted from titles and abstracts while others were extracted from each of the documents and added to the pool of keywords by the SCI and SSCI indexers. All the keywords added by indexers or those extracted from subjects or abstracts were normalized by the LISA, LCSH, and Information Technology Terms thesauruses (indices of information stored in a computer system) so that the keywords could have consistency (singular and plural), unity (synonyms), and clarity (homonyms).

Table 4. Final keywords for analysis

Rank	Keyword	Number of documents	Rank	Keyword	Number of documents
1	Security	264	31	Discrete Logarithm	22
2	Cryptography	176	32	Detection	21
3	Encryption	135	33	Neural Network	21
4	Authentication	124	34	Security Verification	21
5	Algorithm	98	35	Recognition	21
6	Systems	73	36	Cryptosporidium	20
7	Internet	64	37	Chaos	20
8	Verification	54	38	Experimental	20
9	Communication	51	39	Information	19
10	Network	50	40	Multiplication	19
11	Cryptosystem	48	41	Complexity	19
12	Image	39	42	Access Control	19
13	Protocol	38	43	Biometric	19
14	Cryptanalysis	35	44	Data Encryption	19
15	Design	32	45	Performance	19
16	Synchronization	31	46	Public-Key Cryptosystem	19
17	Model	30	47	Signature	18
18	Monitoring	29	48	Video	18
19	Watermarking	29	49	State	18
20	Optical Security	29	50	Distributed System	18
21	Mage Encryption	28	51	Computation	18
22	Implementation	26	52	Intrusion Detection	18
23	Codes	26	53	Error Correcting Codes	18
24	Applications	26	54	Java	17
25	Protection	25	55	Cryptographic Protocols	17
26	RSA	25	56	Password	17
27	Secure Communication	24	57	Public-Key Cryptosystem	17
28	Public Key	24	58	Privacy	17
29	Identification	24	59	Power System	16
30	Key Distribution	23	60	Optimization	16

The keywords with a low rate of frequency (i.e. once or twice) were incorporated into those with inclusive meanings. Among the keywords appearing only once or twice, those without another words with similar or inclusive meanings were neglected. Finally, 223 keywords appearing more than 5 times were selected as the samples of research on the analysis of words with simultaneous appearance

The investigation into new technology in the field of information security

Understanding the characteristics of data for analysis. The first step toward investigating into new technology is to examine the nature of source data. That is, the nature of data to be analyzed needs to be categorized into a random network and a scale-free network. This process is very important in that whether hubs can be found depends upon the nature of data itself. As a result of analyzing the characteristics of data in the field of information security to be analyzed in this research, there turned out to have the nature of a scale-free network (refer to Figure 2).

The power function distribution is a distribution characterized by a curve decreasing to the right as shown in Figure 2. In a scale-free network where the number of connecting lines for each node coincides with the power function distribution, it can be anticipated that most nodes have only a small number of links and they are interconnected by a few hubs with a large number of connecting lines. The principle of the power function is visually very similar to that of an aviation route network where a number of small airports are interconnected through a small number of major hubs.

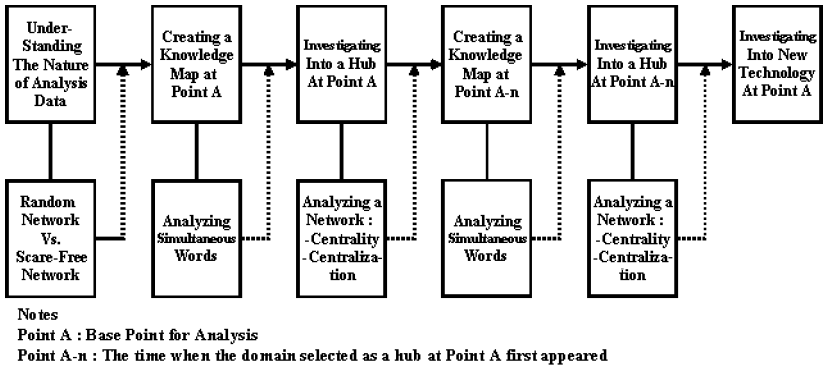

Figure 1. The Framework for the Deduction of New Technology

The power function distribution that decreases with a gentle slope can naturally integrate hubs with a large number of links to a varying degree. This helps us expect

that all scale-free networks will have several hubs with huge dimensions that regulate the phase structure of the entire networks. Through a number of positive analysis efforts, most theoretically important networks have turned out to be scale-free networks, and as a result, the validity of a hub has been acknowledged.

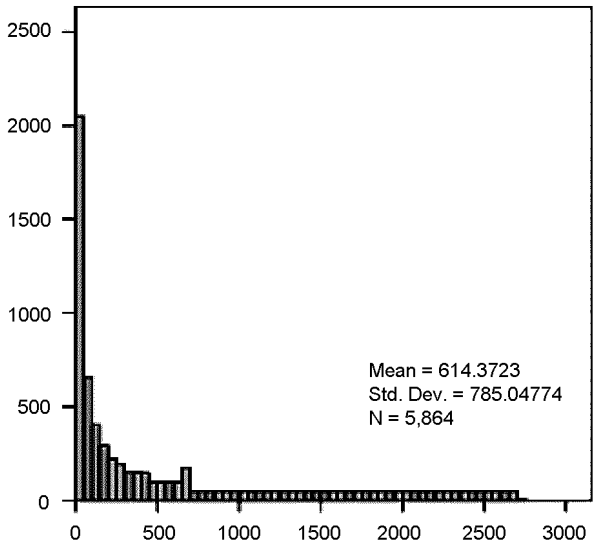

Figure 2. The characteristics of data for analysis

Generating a knowledge map at the current point

(A) Generating a matrix with simultaneous appearance. The specially established MS-SQL program was used to calculate the number of occurrences in the same publication. In doing so the matrix of 223 x 223 keywords with simultaneous appearance was generated. The frequencies of X and Y appearing simultaneously were entered in the cells of keywords X and Y respectively. The value of a cell with a diagonal line through it was treated as the value of loss.

The matrix of simultaneous appearance generated was then converted into a relative matrix that uses the Pearson correlation coefficient, which shows the similarity and dissimilarity between each pair of keywords. That is, the matrix of simultaneous appearance was normalized.

Table 5. The matrix of simultaneous appearance

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
1	19	0	0	2	0	0	1	0	9	0	0	0	0	4	0	0	0	0	0	0	.
2	0	11	6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	.
3	0	6	6	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	.
4	2	0	0	98	1	3	4	0	3	0	0	0	1	1	0	1	0	1	0	0	.
5	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.
6	0	0	0	3	0	10	0	0	2	0	0	0	0	0	0	0	0	0	0	0	.
7	1	0	0	4	0	0	26	0	1	0	0	1	0	0	0	0	0	0	0	1	.
8	0	0	0	0	0	0	0	10	4	0	1	1	0	0	0	0	0	0	0	0	.
9	9	0	0	3	0	2	1	4	124	7	13	0	1	5	1	2	5	0	0	1	.
10	0	0	0	0	0	0	0	0	7	15	0	0	0	0	0	0	0	0	0	1	.
11	0	0	0	0	0	0	0	1	13	0	14	0	0	0	0	0	0	0	0	0	.
12	0	0	0	0	0	0	1	1	0	0	0	5	0	0	0	0	0	0	0	0	.
13	0	0	0	1	0	0	0	0	1	0	0	0	9	0	0	0	0	0	0	0	.
14	4	0	0	1	0	0	0	0	5	0	0	0	0	29	0	0	0	0	0	0	.
15	0	1	2	0	0	0	0	0	1	0	0	0	0	0	7	0	0	0	0	2	.
16	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	5	0	0	0	0	.
17	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	6	0	0	0	.
18	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	20	1	0	.
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0	.
20	0	0	1	0	0	0	1	0	1	1	0	0	0	0	2	0	0	0	0	6	.
…	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

(B) Clustering. For a more accurate description of the information security field, 223 keywords were divided into 13 clusters by using hierarchical clustering based on Ward's methodology.

Each cluster is then to be given a name that can describe the unique characteristics. There are roughly two methods to do this. First, the name that can represent the keywords constituting each cluster is to be newly created. Second, the name of a keyword with the highest rate of frequency among those constituting each cluster is to be used as the name of a cluster. In this research, the second method was used to name each cluster. The names of the 13 clusters are as follows: Security Assessment, Detection, Monitoring, Systems, Optical Security, Encryption, Verification, Cryptography, Computation, Cryptosystem, model, Authentication, and Privacy. On the basis of the names of the 13 clusters, the second matrix of simultaneous appearance was generated.

Table 6. The second matrix of simultaneous appearance

	CO1	CO2	CO3	CO4	CO5	CO6	CO7	CO8	CO9	CO10	CO11	CO12	CO13
CO1	52	16	3	1	2	7	6	5	0	1	2	2	2
CO2	16	120	14	14	3	9	12	12	2	5	8	16	23
CO3	3	14	57	4	1	0	6	5	0	0	9	4	7
CO4	1	14	4	114	5	35	8	41	6	4	8	20	18
CO5	2	3	1	5	102	60	38	19	7	8	12	10	12
CO6	7	9	0	35	60	217	30	82	10	35	38	39	24
CO7	6	12	6	8	38	30	154	31	5	19	24	32	47
CO8	5	12	5	41	16	82	31	313	22	54	32	68	35
CO9	0	2	0	6	7	10	5	22	41	2	1	1	5
CO10	1	5	0	4	3	35	19	54	2	119	7	14	12
CO11	2	6	9	8	12	38	24	32	1	7	124	38	16
CO12	2	16	4	20	10	39	32	68	1	14	38	210	66
CO13	2	23	7	18	12	24	47	35	5	12	16	66	155

The second matrix of simultaneous appearance generated was then converted into a relative matrix that uses the Pearson correlation coefficient, which shows the similarity and dissimilarity between each pair of clusters. That is, the second matrix of simultaneous appearance was normalized.

(C) The MDS map. Keywords with a high rate of frequency were selected to represent the corresponding cluster, and the overall diagram was completed by using the MDS (Multi-Dimensional Scaling) technique so that the overall positions of the 13 clusters in the information security field could be made explicit.

All the keywords in each cluster were selected as a group of variables for the MDS maps to be generated and arranged in the order of clusters, and MDS was applied for the second map to be acquired. The 13 MDS maps arranged in the order of the 13 clusters show the unique relationship among keywords in each cluster. This technique is also known as multi-level mapping [PETER & VAN RAAN, 1993]. Multi-level mapping is mainly used to schematize a field with more than one dimension. The overall structure of each field is first generated (a general view). The diagrams arranged in the order of the clusters are used to show the detailed structure of a particular cluster. For more detailed information on a particular field related to each cluster, the multi-level maps can be used to zoom in that specific field.

The data in a matrix with a dimension of 13 x 13 are measured again so that a close diagram can be found on the basis of the 13-vector system (the Pearson correlation coefficient). In other words, the similarity between two words is calculated again on the basis of the frequency of the two words as well as all the other items in the same matrix. Accordingly, words with a high Pearson correlation coefficient are located in the same position on the map, and they have a high similarity in the analysis of simultaneous appearance within the entire matrix.

A cluster can be defined by two different ways. First, a cluster can be viewed as a point of a general network characterized by the position of the cluster, which can be represented by a bundle of points or clusters found in a general network. Second, a cluster can be viewed as a cluster that consists of connections between words. This is defined as the density level of a network, that is, the level of cohesion and intensity [BHATTACHARYA & BASU, 1998]. The inner relationship between the 13 clusters can be proved by the connection between keywords in other clusters.

As shown in Figure 3, a solid line between two words shows a high level of association with over 0.5 of the proximity index whereas a dotted line between two words shows a low level of association with over 0.3 and below 0.5 of the proximity index. Such connections are indeed very interesting in that information on association is visually presented by the corresponding position.

The investigation into hubs at the current point

The knowledge map at the current point can help find a hub, that is, a field that is developing most actively at the moment. In this research the network analysis coordinates, including degree, betweenness, and closeness, were used to investigate into hubs at the current point.

(A) Measuring degrees. The level of degree was measured on the basis of the data analyzed at the current point. As a result of the measurement, cryptography turned out to have the best level of betweenness. Encryption and authentication were next in order (refer to Table 7).

(B) The level of betweenness. Next, the level of betweenness was measured on the basis of the data analyzed at the current point. As a result of the measurement, cryptography turned out to have the best level of betweenness. Encryption and authentication were next in order (refer to Table 7). The results obtained were similar to those of degree. However, it seemed that only the fields belonging to the top group showed consistent results whereas those belonging to the bottom group showed varying results.

(C) The level of closeness. Next, the level of closeness was measured on the basis of the data analyzed at the current point. As a result of the measurement, detection, systems, verification, and model turned out to have the same level (refer to Table 7). Thus, the results of closeness were quite different from those of the two other coordinates stated above.

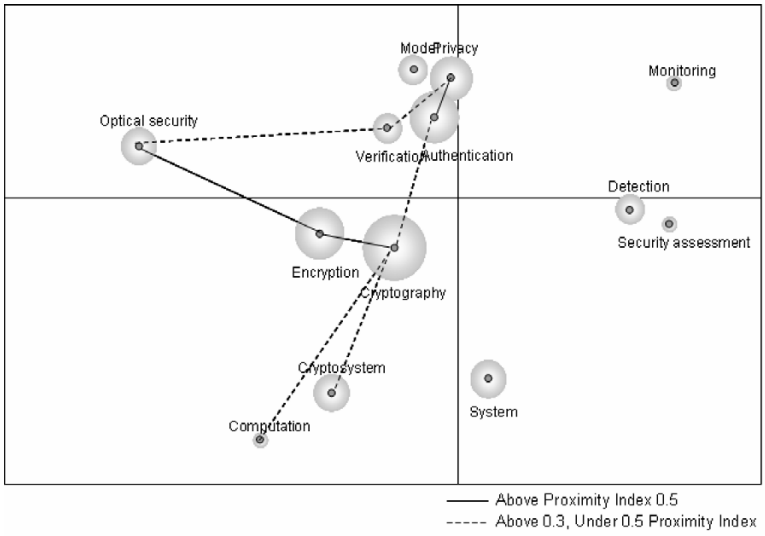

Figure 3. The domain map generated through the analysis of words with simultaneous appearance (1994 to 2003)

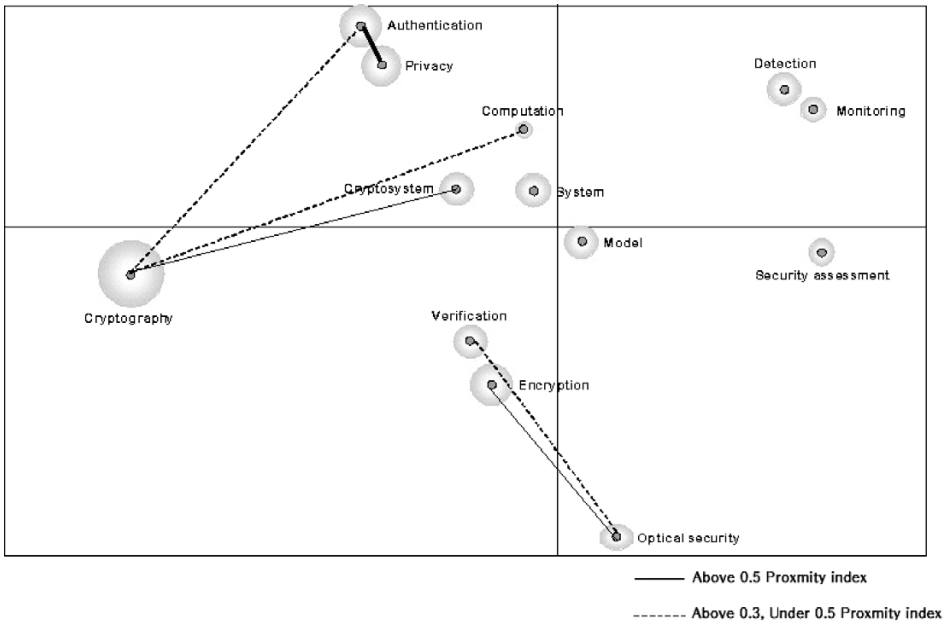

Figure 4. The domain map generated through the analysis of words with simultaneous appearance (1994 to 1999)

Table 7. The level of degree, betweenness and closeness at the current point

	The level of degree	The level of betweenness		The level of closeness	
Measures	Value	In-degree	Out-degree	In-closeness	Out-closeness
Mean	67.745	16.641	16.641	0.956	0.956
Std. Dev.	33.571	9.301	9.301	0.064	0.064
Min.	16.622	3.917	3.917	0.8	0.8
Max.	128.601	33.583	33.583	1	1

Generating a knowledge map at a particular point

The overall aspects of the information security field were analyzed by using the knowledge map and the network analysis coordinates. On the basis of the results and under discussion with professors specializing in this field, encryption was selected as the hub at the current point. In addition, the time when the encryption field first appeared was examined. As a result of the examination, encryption turned out to have first appeared in the year of 1994. Thus, the years from 1994 to 1999 were selected as one point in time and analyzed accordingly. The procedure for creating a knowledge map has been omitted in that it is exactly the same as the procedure explained above, and the results are shown in Figure 4.

Investigating into hubs at a particular point (1994 to 1999)

(A) Measuring degrees. The level of degree was measured on the basis of the data analyzed at a particular point in time. As a result of the measurement, cryptography turned out to have the best level of betweenness. Authentication and privacy were next in order (refer to Table 8).

(B) The level of betweenness. Next, the level of betweenness was measured on the basis of the data analyzed at a particular point in time. As a result of the measurement, cryptography turned out to have the best level of betweenness. Encryption and privacy were next in order (refer to Table 8). The results obtained were similar to those of degree.

Table 8. The level of degree, betweenness and closeness at the particular point

	The level of degree	The level of betweenness		The level of closeness	
Measures	Value	In-degree	Out-degree	In-closeness	Out-closeness
Mean	62.693	5.179	5.179	0.863	0.863
Std. Dev.	31.603	2.968	2.968	0.113	0.113
Min.	17.949	1.25	1.25	0.706	0.706
Max.	131.51	12.417	12.417	1	1

(C) The level of closeness. Next, the level of closeness was measured on the basis of the data analyzed at a particular point. As a result of the measurement, cryptography, systems, verification, and model turned out to have the same level (refer to Table 8). At a particular point in time as well, the results of closeness were quite different from those of the two other coordinates stated above.

Investigating into new technology at the current point

So far, we have examined the knowledge map and hub at the current point (point A: the year 2003) as well as the knowledge map and hub at a particular point in time when the hub at the current point first appeared (point A-n: the year 1994). The relationship between the hub at a particular point in time (the year 1994) and a new hub is as follows:

Table 9. The characteristics of a hub in the information security field

Analysis Index	Result	Implication
Level of degree	Low	– Means the number of connected points and respresents centrality. – A new hub still lacks in the number of connected points, and independent research cannot be conducted yet.
The level of betweenness	High	– Measures the level of one point being positioned in-between other points in a network. – Betweenness centrality measures the level of a mediating role in-between other points. – Independent research still cannot be conducted for a new hub, yet other areas of research can be activated.
Level of closeness	Low	– A representative coordinate that describes the global centrality of a point. – The distance between two points is a geodesic, that is, the shortest distance between the two points. – A point with the smallest amount of path distance is a central point with the highest level of global centrality. – A new hub is a field that is not located in the center of the information security field.

On the basis of such results of analysis, research areas with a high probability of appearing as hubs in the information security field within several years were selected. Research areas were summarized into security assessment, computation, and cryptosystem. All in all, through debate with professors in the related fields, cryptosystem turned out to have a high possibility of becoming the most popular subject in the next few years. Therefore, concentrating research efforts on the present hubs and at the same time conducting research on the cryptosystem field will realize the two goals of the science and technology policies of the nation most efficiently: selection and concentration.

Conclusions and future assignments

Conclusions of the research

This research is rooted in the analysis of a network that has acquired a strong persuasive power in modern physics. The analysis of a network helps prove the fact that there exists a series of particular patterns even in the most complicated phenomena, and such claims have been proved to be true by a number of scholars. Accordingly, this methodology can be a starting point for our nation to solve a number of problems on a national scale and take off as an advanced country through the development of science and technology.

That is, despite the current situation where the traditional catch-up strategy cannot be given up and the source technology must be attained through selection and integration while at the same time the professional analysis of information necessary to effectively support the development of core technology for the growth-driven industry in the next generation must be carried out and the adequate results must be efficiently provided on a governmental scale, there are virtually no systems to produce outcomes.

At the initial stage of this research designed to resolve such conflict, the methodology of deducing new technology from science and technology information resources was closely examined. Driven by such goals of research, the method of analyzing a knowledge map from the angle of a network was suggested. The methodology of deducing new technology suggested by this research can be summarized as follows: First, the nature of data to be analyzed must be investigated. The nature of data is categorized depending on whether it coincides with a random distribution or a scale-free power law distribution. Second, a knowledge map at the point of analysis is generated through the analysis of words with simultaneous appearance. Also, the hub of the analyzed data is examined through the analysis of a network. The network analysis coordinates are used at this stage. Connectivity, centrality, and centralization were used as the network analysis coordinates. Third, once hubs have been found with such methods, it is necessary to track back to the period when the hubs first appeared so that a domain map at that particular point in time can be analyzed. Based on the hubs back then, the association with a new hub (i.e. the hub at the present time) is analyzed with the coordinates introduced above. Once the overall characteristics of this field have been clarified, it is necessary to come back to the present time and search for a domain with a possibility to become a hub in the future. That is, a domain whose characteristics are most similar to those of the coordinates that have a possibility of turning into a future hub in this field centered around the hub marked on the current domain map is selected.

The summary of results as to the deduction of new technology in the information security field based on such methodology is as follows: First, the overall aspects of the

information security field were analyzed with the knowledge map and the network analysis coordinates. On the basis of the results of analysis and discussion with professors specializing in this field, encryption was announced as the hub at the current point in time. Next, the time when the encryption field first appeared was selected and analyzed accordingly. As a result of the analysis, the period when the hub at the current point in time first appeared was characterized by a low level of degree, a high level of betweenness, and a low level of closeness. Finally, as a result of applying such characteristics to the knowledge map at the current point in time, research areas were summarized into security assessment, computation, and cryptosystem. All in all, through debate with professors in the related fields, cryptosystem turned out to have a high possibility of becoming the most popular subject in the next few years.

Future assignments

As this research was intended for suggesting the methodology of investigating into new technology through science and technology information resources, a meaningful approach was, in fact, introduced by this research, and the methodology suggested by this research is believed to be a meaningful first step toward the establishment of the information analysis structure, which is in desperate need.

Furthermore, the methodology of investigating new technology suggested by this research can be applied as follows: First, combined with the existing techniques including text mining, bibliometrics, technometrics, information visualization, and KDD (Knowledge Discovery in Database), the methodology can be applied to the analysis of preliminary validity for a research plan on the national research development project. Second, the results obtained by the methodology in this research are sure to help the Ministry of Planning and Budget as well as the National Science and Technology Council make a reasonable, objective decision as they evaluate the validity of the planning report on a new, large-scale research development project. Third, through consistent research efforts from now on, a variety of quantitative coordinates can be acquired based on science and technology information resources, and they can be systematically integrated into a system used to analyze the validity of a national research development project.

This research, however, has the following limitations. First, the methodology of this research is in lack of validity. That is, the methodology supported by this research is to investigate into hubs through the analysis of a network. The biggest issue in this field at the moment is, without a doubt, a research effort to find hubs. Unfortunately, no one has ever suggested a definite solution, and research in progress is still at a premature stage. Considering such circumstances, I cannot present a good solution as to the validity of the methodology claimed by this research.

Some of the possible solutions to overcome the limitations discussed above and help related research efforts in the future are as follows: First, a research effort can be made to examine whether the conclusions of this research correspond to the anticipated results of theoretical, positive research. Second, an additional positive research effort needs to be made through actual case studies so that the methodology suggested by this research can be verified.

References

BARABASI, A. L. (2002), Statistical mechanics of complex networks, *Reviews of Modern Physics*, 74 : 47–97.

BHATTACHARYA, S., BASU, R. K. (1998), Mapping a research area as the micro level using co-word analysis, *Scientometrics,* 43 (3) : 359–372.

CALLON, M., LAW, J., RIP A. (1986), *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. London: Macmillan.

GALVEZ, C., MOYA-ANEGÓN, F. (2007), Standardizing formats of corporate source data, *Scientometrics*, 70 (1) : 3–26.

CHEN, C. (2004), Searching for intellectual turning points: Progressive knowledge domain visualization, PNAS Early Edition.

DING, Y., CHOWDHURY, G. G., FOO, S. (2000), Journal as markers of intellectual space: Journal co-citation analysis of information retrival area, *Scientometrics*, 47 (1) : 55–73.

DING, Y., CHOWDHURY, G. G., FOO, S.(2001), Bibliography of information retrieval research by using co-word analysis, *Information Processing & Management*, 37 : 817–842.

FREEMAN, L. (1979), Centrality in social networks conceptual clarification, *Social Network*, 1 : 215–239.

GLADWELL, M. (2000), *The Tipping Point*, New York: Little, Brown.

HE, Q. (1999), Knowledge discovery through co-word analysis, *Library Trends*, 48 (1) : 131–159.

LAW, J., WHITTAKER, J. (1992), Mapping acidification research : A test of the co-word method, *Scientometrics*, 23 (3) : 417–461.

LEYDESDORFF, L. (2007), Mapping interdisciplinary at the interfaces between the Science Citation Index and the Social Science Citation Index, *Scientometrics*, 71 (3) : 391–405.

LIN, C. T., CHIANG, C. T. (2007), Evaluating the performance of sponsored Chinese herbal medicine research, *Scientometrics*, 70 (1) : 67–84

MOON, Y. H. (2004), Monitoring and Early Warning of Technological Progress*,* First Workshop of the Research Planning Assessment Study Society, KISTI.

MORRIS, S. A. (2005), Manifestation of emerging specialties in journal literature: a growth model of papers, references, aexemplars, bibliographic coupling, co-citation, and clustering coefficient distribution, *Journal of the American Society for Information Science and Technology*, 56 (12) : 1250–1273.

NOYONS, E. C. M., VAN RAAN, A. F. J.(1998), Advanced mapping of science and technology, *Scientometrics*, 41 (1–2) : 61–76.

OH, S. J., LEE, D. G. (2004), *Reconstruction and Development Direction of the Science & Technology Innovation System,* Public Forum on Reorganization of the National Science & Technology Innovation System.

PETERS, H. P. F., VAN RAAN, A. F. J.(1993), Co-word based science maps of chemical engineering, Part I: Representations by direct multidimensional scaling, *Research Policy*, 22 : 23–45.

TIJSSEN, R. J. W. (2007), Africa's contribution to the worldwide research literature : New analytical perspectives, trends, and performance indicators, *Scientometrics*, 71 (2) : 303–327.

YOON, M. S., LEE, W. H., OH, H. Y., KIM, Y. M. (2004), Methodology of S&T Knowledge Map and Indicators for Ex-Ante Evaluation of National R&D Program, STEPI.