International Journal of u-and e-Service, Science and Technology
Vol.7, No.4 (2014), pp.201-214
http://dx.doi.org/10.14257/ijunnesst.2014.7.4.19

An Analysis Model of R&D Project's Similarity

Jong-Bae Kim[1*], Hyo-Jung Sohn[2], Min-Gyu Lee[3], Baek-Min Seong[4] and Yu-Jin Jo[5]

[1,2,3,4,5]Graduate School of Software, Soongsil University, Seoul 156-743, Korea*
E-mail, [1]kjb123@ssu.ac.kr , [2]hyojung.sohn@gmail.com, [3]marse101@naver.com,*
[4]feeling127@naver.com, [5]lovejung81@naver.com

Abstract

For the government, it is very important to analyze the similarity between various R&D projects, as it can contribute to reducing waste in government budgets and enhancing the investment efficacy of R&D investment. Thus far, there have been studies that attempted to analyze the similarity between two different documents by focusing on the keywords that represent the content of the document, or by using the similarities between sentences, with a view to identifying overlapping R&D projects. However, for various reasons these studies failed to achieve a desirable level of accuracy. In this regard, our purpose in this study is to suggest a means to analyze the similarity between R&D projects using the patent analysis DB from the government's patent technology trend survey program, which examines and collects information on patents related to past R&D projects. To realize this goal, we suggested a similarity assessment model based on set theory and probability theory.

Keywords: *Research & Development Project, Project Similarity, Patent Information, Set theory, Probability theory*

1. Introduction

The government's investment in R&D projects has been rising by 10% per annum thanks to a proactive science and technology policy. However, as various government departments are competing with each other to fund new projects, it can be seen that the age-old problem of Korea's government, wasted funds, still persists. In order to prevent overlapping investment of R&D budgets and increase the efficiency of investment, it is imperative that similar projects be filtered out as early as during the proposition process. In this regard, the government made it mandatory to review the similarities of project propositions when planning R&D projects, by means of the National Science and Technology Knowledge Information Service as per the relevant regulations regarding the project management of the government R&D projects. However, since this service relies simply on keyword matching to assess the similarities, there are still limitations to the accuracy of similarity assessment when there are partial changes in the project or simple substitutions of the technical contents [1].

In order to improve this problem, we developed a similarity analysis model using the patent analysis DB created through the government's R&D patent technology trend survey project, in which the information on the patents related to the past R&D projects is examined and collected. In fact, this patent technology trend survey project (http://ipas.rndip.re.kr) is now in operation with a view to analyzing the trend of the patents in the technical fields when planning R&D projects, which will in turn present the direction of study for new projects and

[1*] Corresponding author. Tel. : +82-10-9027-3148.
Email address: kjb123@ssu.ac.kr(Jong-Bae Kim).

ISSN: 2005-4246 IJUNNEST

prevent overlapping supports for R&D projects for which there are pre-existing patents already. In addition, since planning and performance information on the previous R&D projects is available, it is possible to compare and analyze such information with the planning information of new R&D projects to identify similar projects.

In this study, therefore, we will examine the existing similarity analysis methods and identify possible ways of utilizing and improving them. With this, we will develop a similarity analysis model using the patent analysis DB. In addition, we will present the input/output information for analyzing the similarity, and the analysis algorithm.

2. Related studies

The most realistic way to perform a quantitative analysis of the similarities is to use the outputs of the projects; in particular, the patent documents. However, this document-focused identification of similarity has been considered an important issue not only in the field of R&D but also in many other fields as well. The word 'similarity' here can be defined as a 'quantitative measurement of the volume of information which is shared or not shared by two different entities.’ Also, similarities can be divided into different levels, including complete overlap, partial overlap, and similar overlap [2].

In order to analyze the level of similarity between two different documents, one of the two documents must have a property that can distinguish it from the other, and it is only possible to measure the similarity between these two documents when such identified properties show a higher level of similarity. Reference [6] shows one of the most widely known similarity analysis models by means of dimensionality reduction of the document, which is at a higher dimension, to fingerprints, which is a lower dimension, to use this as the distinguishing properties of the document. However, this model has its own shortcomings, as it can neither take the ambiguity of the natural language nor compare sentences that do not abide by the standard rules of the grammar. Also, both the quality and the performance are important in the analysis of similar documents. If, for example, we compare 'n' sets of different documents by means of the fingerprints, a total inspection is needed, and to calculate the level of similarities among these documents requires the level of complexity to be O (n2.)

Another option to address this problem is the multi-level indexing structure [3]. In this method, the number of lines in a document, as well as the k-bit figures is controlled to suggest a subject of comparison. In this case, the number of the fingerprint groups decreases, allowing faster comparison. Another similar example is Reference [4], where the words that are used in the documents are listed up and some of them are extracted as the characteristics features of the document. This particular model is the most widely used of all the similarity analysis models today. However, it has its own shortcoming, which is that heuristic decision making by an expert is necessary in order to recognize the keywords [5].

An alternative approach of finding similar documents to a certain document is, on the other hand, to find the nearest neighbors during the initial analysis, rather than analyzing the entirety of all the documents at hand. These nearest neighbors are put to a more rigorous analysis. This way, the range of documents to be analyzed is reduced, increasing the efficiency of the similarities analysis [6]. Another example is an algorithm designed to measure the level of similarities in terms of the R&D projects, which is called the comprehensive formation network [7] This model was developed in order to overcome the limitations of the similarities analysis models based on keywords, which works by extracting the science technology standard classification items of the projects to form the unique vector for each of the project classes and designate the original technology pattern of the tasks that are included in a project, showing the similarities between these elements as a comprehensive formation network. However, this approach is not free from its own limitation, which is that the meanings cannot be considered, either.

3. A project similarity assessment model utilizing patent information

3.1. Identification of the patent information to measure the similarities between projects

In this study, the patent information is used as the characteristics to distinguish one document from another. The patent information is based on the data acquired from the government's R&D patent technology trend survey project, the database for which contained various types of information such as project proposal, technical classification, keywords, valid patents, etc. We calculated the similarities between different projects when there is a new project entry, primarily using such data. One of the prerequisites for this concept to work is that each of the existing and new projects has a valid patent of its own. That is, as we mentioned above, it is another prerequisite that the government R&D patent technology trend survey project already checked the existing project for valid patents. Within such valid patent information, those with the same 'national announcement' and 'application number' are recognized as the same patent, and the level of matching between these valid patents is used to measure the level of similarities.

3.2. Scales of measuring similarities

Each project has a set of valid patents, and in order to analyze the level of similarities between these groups, we used Set Theory [6]. The level of similarities based on sets is shown as the ratio of intersections between the two sets among the patents in the sum of the two sets. That is, the level of similarities based on set theory is a scale that reflects the fact that as the ratio of overlapping patents between the two different projects is high, so is the level of similarity. The formula to measure the level of similarities based on set theory is as shown in equation (1).

Set-based similarities = the intersection of the two valid patents / the sum of sets between two valid patents *(1)*

The advantage of this scale is that it can easily be understood by anyone, which could serve to ensure the rationality of the interpretation. On the other hand, since the sum of the two sets is used as the parameter, the level of similarity of the new project with the old project has the same value as the similarity level of the old project with the new one. This exposes one drawback of this method, which is that it is not capable of "showing the level of inclusion of a certain project in others." Also, as the number of sets influences the level of similarity, it should be noted that the number of valid patents may influence the level of similarity measured using this method, which is another drawback of this method.

In order to improve such shortcomings of the set-based similarities method, I suggest the similarities level scale based on probabilities. Probability theory [8] is capable of showing the probabilities of results that are yet to occur. Based on the fact that the similarity between the projects cannot have a fixed, confirmed value at the moment, the application of probabilities theory is likely to be a good idea in terms of measuring and showing the similarities. The formula for calculating the similarities level using probability theory is as shown in equation (2) below:

o The chance of A and B being different from each other under the conditions of A *(2)*

=> p (A not B| A) = P (A-B) / P(A) => 'The chance of A being different from B"

*o The chance of A and B being different from each other under the conditions of B
=> p(A not B |A) = P (B-A) / P(B) => The chance of B being different from A*

o 1.0 – the chance of A being different from B => The chance of A being similar to B

o 1.0 – the chance of B being different from A => The chance of B being similar to A

o The chance of A being similar to B x the chance of B being similar to A =>The chance of A and B being similar to each other =>Similarity based on probability theory

In the analysis of the similarities level based on probability theory, each of the steps in the process can be calculated as a quantitative probability, without being influenced by the number of valid patents, which is an advantage of this method. On the other hand, the conditional probability theory cannot ensure the rationality of the interpretation of the results in the comparison. And, since the chances of similarities between the old and new projects are combined, it is not capable of showing the inclusion relationship, which is a drawback of this method.

In addition, it is also true that both the probabilities-based and the set-based similarities analysis results cannot explain the inclusion relationship between the new and old projects. In order to address this problem, it is necessary to develop a new method to combine the chance of the new project being similar to the old one and the chance of the old and new projects being similar to each other. In this study, for the sake of clear expressions of the inclusion relationship, we adopted the method of using the ones with the higher values. Here, however, it is necessary that the probability-based similarities in the equation (2) be taken into account as well. Therefore, the final formula resulting from this may look like equation (3) below:

*o The chance of A and B being different from one another under the conditions of A
=> p(A not B | A) = P (A-B) / P(A) => The chance of A being different from B*

o 1.0 – the chance of A being different from B => the chance of A and B being similar to one another

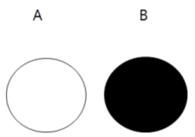
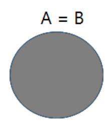
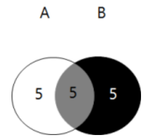
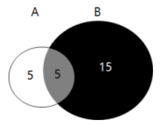
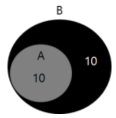
o the chance of A being similar to B x the chance of B being similar to A => the chance of A and B being similar to one another => the similarities level based on the probabilities theory (whole) *(3)*

o Max (the chance of A being similar to B, the chance of A and B being similar to one another) => the similarities level based on the probabilities theory (part)

3.3. Comparison between the overlapping projects based on valid patents / similarities scales

The similarity scales mentioned above, which are based on valid patents, can be compared to one another in terms of the relationships between the valid patents in the new and old projects, as shown in Table 1. The relationships between the valid patents can be divided into three categories: valid patents not matching, valid patents matching with one another completely, or a partial match between these valid patents. Of these, the partial match can again be sub-divided into three specific relationships between the projects: the number of valid patents is the same between the old and new projects, the number of valid patents is different, and the case in which there is an inclusive relationship between them.

Table 1. Comparison of the Metric for Similarity

Item		Total Mismatch	Perfect Match	Partial Match		
				Numbers identical	Numbers not identical	Included
						
Set (total)	A -> B	0	1	0.333	0.200	0.500
	B -> A	0	1	0.333	0.200	0.500
Probability (Total)	A -> B	0	1	0.250	0.125	0.500
	B -> A	0	1	0.250	0.125	0.500
Probability (Part)	A -> B	0	1	0.500	0.500	1.000
	B -> A	0	1	0.500	0.250	0.500

However, for these similarities scales, it should be noted that they are not selective scales but trade-off scales. Normally, it is from the perspectives of the new projects that the old projects are viewed (that is A → B.) Therefore, if there is a mistake in the interpretation of the probabilities, it is possible that A and B may be interpreted as being a perfect match. In order to see if the relationship between A and B is either a perfect match or an inclusive one, it should be compared along with the probabilities in equation (2) (whole.) Here, if it is a perfect match, both the probability in the equation (2) (whole) and the probability in the equation (3) (part) should be 1, while only the probability of the equation (3) (part) is 1 when the relationship is inclusive.

3.4 Application of the similarity measurement method.

Based on the similarity measurement method suggested herein, I used a total of 156 projects and 160,218 valid patents as the basis to measure the similarities of projects based on valid patents. The result of the measurement was as shown in Table 2.

Table 2. Similarity Analysis based on Patent Information

Target of analysis	Project name	Set	Probability (whole)	Probability (part)
Crop seed gene functional analysis and industrialization (top 5)	Exploration of useful agricultural genes using the rice mutation group and biological information	0.137	0.427	0.485
	GMO safety assessment technology and industrialization	0.036	0.180	0.186

	Development and commercialization of genetically engineered rice with improved utilization of phosphoric acid in soil	0.002	0.043	0.043
	Development of biological material that produces high value-added lipid	0.003	0.014	0.014
	Industrialization of organic farming techniques for specialty crops	0.003	0.011	0.011

In addition, the result was compared to that of the keyword-based similarity assessment (Table 3), which functioned as the control group. In this example, the case of 'Crop seed gene functional analysis and industrialization' project turned out to be the most similar to the project “Rice mutation group and biological information used to explore useful agricultural genes” from the perspectives of set and probability method. In terms of the sets, it could be interpreted that 13% of the sum of the valid patents were matching. Furthermore, it could also be said that about 48% of the valid patents in “Crop seed gene functional analysis and industrialization” were matching with those of “Rice mutation group and biological information used to explore useful agricultural genes.” In addition, the interpretation using the probability (whole) showed that the probability of these two projects being similar to one another is approximately 42%.

Table 3. Similarity Analysis based on Keywords

Target of analysis	Project name
Crop seed gene functional analysis and industrialization (top 5)	Development of technology to promote a new species of peach and reduce disasters
	Analysis of the characteristics of immune enhancement substances in Korean agricultural products and development of the production process
	Development of the technology to mechanize the entire production process of sweet potatoes
	Value assessment of organic agricultural technology and estimation of new demands
	Field validation study for the new rice drying system utilizing rice husk energy

4. Verification of the Suggested Scale Using Statistical Analysis

The method suggested in this study was verified through two separate analyses. The first verification was the verification of the feasibility of the suggested method. This was performed to find out whether or not the results obtained through the scales for each method could reflect the heuristic analysis results with the subjects of common interests [9]. The second verification related to the reliability and efficiency of the suggested method. Under the assumption that the suggested method is feasible, it involves comparing the results derived with the opinions of experts in order to determine whether or not they were reliable.

For the first verification of the suggested method, the initial verification of reliability was performed as a wide-ranging survey of members of the general public with an ordinary level of knowledge. The questions used in the survey were designed to check the similarity of projects arbitrarily selected among two candidates already determined to be similar. Each of the interviewees was informed of the project name, and given descriptions for each project. The replies were given using a seven-point Likert scale. Each questionnaire consisted of seven questions. The survey was conducted through a website, and the responses obtained were analyzed using SPSS 18.0, a statistical analysis software package. In order to enhance the reliability of the survey outcome, we used Reverse Questions, Reverse Answers, and Interval Request to filter out irrelevant replies. Individual interviewees were identified using the last four digits of their phone numbers and the date of birth. The survey was performed from October 2013 to November of the same year, over a period of 30 days.

Out of the 132 interviewees, 94 were identified as valid using the verification methods of Reverse Question, Reverse Answer, and Interval Requests. As a result, we obtained a total of 4722 replies on the similarities between projects. The hypotheses used for the statistical analysis in this study were as follows:

H0 (Null Hypothesis): There is no correlation between the heuristic decisions and the suggested method in terms of the results of verification of similarity.

H1 (Hypothesis Testing): There is a correlation between the heuristic decisions and the suggested method in terms of the verification of the similarity.

The correlation between the similarity analysis results using heuristic decisions and those using the methods suggested herein was analyzed. Both of the scales were quantitative in nature. Therefore, we conducted Pearson's test, and the results of the analysis are as shown in Table 4.

Table 4. Correlation Analysis between the Proposed Metric and Response

		Survey Answer	Set	Probability _Total	Probability _Part
Survey Answer	Pearson's (coefficient)	1	.564**	.507**	.533**
	Significant Probability (Both sides)		.000	.000	.000
	N	4722	1734	1734	1734
Set	Pearson's (coefficient)	.564**	1	.978**	.983**
	Significant Probability (Both sides)	.000		.000	.000
	N	1734	1734	1734	1734
Probability_Total	Pearson's (coefficient)	.507**	.987**	1	.999**
	Significant Probability (Both sides)	.000	.000		.000
	N	1734	1734	1734	1734

Probability_Part	Pearson's (coefficient)	.533**	.983**	.999**	1
	Significant Probability (Both sides)	.000	.000	.000	
	N	1734	1734	1734	1734

The result of the analysis of the correlation showed that the correlation between the answers obtained from the survey and the analysis based on the valid patents (set, probability_total, probability_part) was statistically significant. It was also confirmed that the value of significant probability (both sides) was 0.01 or less (at a 99% confidence level.) Of these, since the coefficient was 0.5 or higher in the analysis of the survey results and the analysis results based on the valid patents, it could be concluded that there is a correlation. Furthermore, since the value was positive, the correlation was also positive. This implies that if the value in the answers from the survey was higher (that is, because the interviewees answered that the projects were more similar), the value obtained from the calculation based on the suggested method increases (the value in the scale gets bigger.) In order to ensure the reliability and accuracy of the analysis results based on the valid patents, we also conducted a regression analysis. The results of the regression analysis were as shown in Table 5 and Table 6.

The results showed that the regression model that employs the heuristic responds as the dependent variables, and the results of patent-based analysis as the independent variable was statistically significant. The significance probability was less than 0.01 (a confidence level of 99% or higher), and the R square was 0.671 (67.1% accuracy of interpretation ratio). This means that the results of the project similarity analysis based on the valid patents could interpret the heuristic determinations at a 99% confidence interval with 67.1% accuracy. Meanwhile, the remaining 32.9% (1 - accuracy) needs to be interpreted using some other factor.

Table 5. Model Summarization

Model	R	R square	Modified R square	Standard error of the estimated value
1	.819[a]	.671	.670	.024

a. estimated value : (constant), probability_part, set, probability_total

Table 6. Distribution Analysis[b]

Model		Sum of Squares	Freedom	Average square	F	Significant Probability
1	Regression model	2.009	3	.670	1174.377	.000[a]
	Residual	.986	1730	.001		
	Sum	2.995	1733			

a. estimated value (constant), probability_part, set, probability_total

b. dependent variables: survey answers

The results showed that the regression model that employs the heuristic responds as the dependent variables, and the results of patent-based analysis as the independent variable was statistically significant. The significance probability was less than 0.01 (a confidence level of 99% or higher,) and the R square was 0.671 (67.1% accuracy of interpretation ratio). This means that the results of the project similarity analysis based on the valid patents could interpret the heuristic determinations with 67.1% accuracy, at a 99% confidence interval. Meanwhile, the remaining 32.9% (1 - accuracy) needs to be interpreted using some other factor.

Table 7. Expert Questionnaire Form

Project Name 1	Functional analysis of the seed genes and industrialization of crops									
Project Name 2	GMO safety assessment technology and industrialization									
AREA	Technical area	Low	1	2	3	4	5	6	7	High
	Support area	Low	1	2	3	4	5	6	7	High
	Application Area	Low	1	2	3	4	5	6	7	High
Description	Title	Low	1	2	3	4	5	6	7	High
	Purpose	Low	1	2	3	4	5	6	7	High
Execution	GOAL	Low	1	2	3	4	5	6	7	High
	Achievements	Low	1	2	3	4	5	6	7	High
	Implementation system	Low	1	2	3	4	5	6	7	High
Entity	Implementation Details	Low	1	2	3	4	5	6	7	High
	Project entity	Low	1	2	3	4	5	6	7	High
	Beneficiary	Low	1	2	3	4	5	6	7	High
Total score (sum of each item)										

The second verification of the suggested method was the verification of the analysis results by experts. Three experts analyzed the project similarity using the Delphi method. 156 projects were analyzed, and the main areas of the projects were agriculture, food, horticulture, foodstuffs, and research. Of these projects, we selected two projects arbitrarily and presented them to experts, who evaluated the similarities between the projects using a seven-point scale. For the analysis, we also provided the project proposals for the research project. The analysis criteria consisted of the area of project, the contents, the performance, and the parties responsible for the execution of the project. The similarities between the project areas were determined in the areas of technology, support, and applications. Similarities in contents were determined in terms of the title, purpose, goals, and achievements. Similarities in performance were determined based on the implementation system and the implementation details. The parties of the project were evaluated based on the similarities with the project entity and the beneficiaries. The result of the evaluation was determined through a unanimous agreement among the evaluators using the Delphi method.

The reliability of the expert meeting results was analyzed using Cronbach's Alpha method. This method can be used to determine whether the experts maintained a consistent set of criteria as they evaluated the similarities. The alpha value obtained from the analysis result was 0.771, meaning that the alpha value of the reliability analysis was 0.7 or higher. This shows that the experts assessed maintained their consistency with a confidence level of 95%. In addition, we performed an analysis to examine the correlation between the analysis results on the similarities between the projects through a survey with the experts and the method suggested in this study. The result of the survey of experts showed a strong correlation with the analysis result based on the valid patents. The coefficient was 0.994, while the confidence level was 99%, indicating a very strong correlation.

Table 8. Correlation Analysis between the Proposed Metric and Expert Survey Result

		Set	Probability_Total	Probability_Part	Experts
Set	Pearson's (coefficient)	1	.943**	.989**	.994**
	Significant Probability (Both sides)		.000	.000	.000
	N	1048	1048	1048	1048
Probability_Total	Pearson's (coefficient)	.943**	1	.981**	.927**
	Significant Probability (Both sides)	.000		.000	.000
	N	1048	1048	1048	1048
Probability_Part	Pearson's (coefficient)	.989**	.981**	1	.980**
	Significant Probability (Both sides)	.000	.000		.000
	N	1048	1048	1048	1048
Experts	Pearson's (coefficient)	.994**	.927**	.980**	1
	Significant Probability (Both sides)	.000	.000	.000	
	N	1048	1048	1048	6527

The analysis result shown in Table 8 can be understood as follows: Firstly, the analysis by experts determined similarities for the 4 areas, 11 items, as mentioned above. The result can be understood to be very similar to the results from the classification methods based on the patent information. However, while the keyword-based analysis also had a correlation, the interpretation rate could be considered as very low. The reason here is that the same technical classification, such as 'improvement of productivity,' may be used in a different manner for different project contents between horticulture and agriculture. The interpretation of 10% may indicate that some of the core contents used in horticulture for the improvement of productivity were also used for the same purpose in agriculture. The result of the statistical analysis for the two methods suggested herein based on the survey with the expert was as follows:

- *The result of the expert survey supported the similarity analysis results based on the patent information.*

- *The expert survey results found that the higher the values of the patent-based similarity analysis, the higher the level of similarity.*

- *The expert survey result was found to show that the patent-based analysis has a higher rate of interpretation (accuracy) compared to that of keyword-based analysis.*

- *All of the results mentioned above have a statistical confidence level of 99% or higher.*

5. Conclusion

The aim of this study was to suggest a method of analyzing the similarities between projects using patent information. To this end, we analyzed the existing literature on similarity analysis. Based on the result of this analysis, we first suggested a project similarity analysis model using patent information. Based on the patent information, we suggested the input and output information to analyze the project similarities. Using such information, we suggested analytical methods for two different aspects. Then, we showed the feasibility (the results of survey of the general public) and the accuracy (the results of the expert survey.)

Some limitations were identified in the course of this study. The domains used in verifying the cases were based on projects for the Rural Development Administration that were carried out from 2010 to 2013. The reason why the verification efforts of this study were limited to certain fields was that the verification process required consultations with experts. However, it is believed that further studies covering other domains as well could contribute to enhancing the generality of the project similarity analysis model suggested in the current study.

The suggested model proposes a number of different scales. Each of these scales has its own unique meaning, and they are needed to ensure the diversity of the determination criteria. While the suggested scales could be used to prioritize similar projects, it is not possible to show whether the priority is high or low through the interpretation of the scales. For such an interpretation, it is necessary for a sufficient amount of patent information on the projects to be gathered to achieve a statistical analysis.

References

[1] J.-B. Kim, H.-J. Sohn, M.-G. Lee, B.-M. Seong and Y.-J. Jo, “A Model for Measuring the Similarity of R&D Projects”, 2014 International Conference on Future Information & Communication Engineering (ICFICE), **(2014)**, pp. 331-334.

[2] M. Bendersky and W. B. Croft, “Finding text reuse on the web”, Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM, **(2009)**.

[3] H. Miihleisen, T. Walther and R. Tolksdorf, “Multi-level indexing in a distributed self-organized storage system”, Evolutionary Computation (CEC), 2011 IEEE Congress on. IEEE, **(2011)**.

[4] G. Chowdhury and S. S. Chowdhury, “Introduction to digital libraries”, Facet publishing, **(2002)**.

[5] J.-H. Kim, Y.-J. Kim and J.-B. Kim, "A study on Similarity analysis of National R&D Programs using R&D Project's technical classification", Journal of Digital Contents Society, vol. 13, no. 3, **(2012)** September, pp. 317-324.

[6] Domâinguez and J. Ferreirâos, "Labyrinth of thought: A history of set theory and its role in modern mathematics", Springer, **(2007)**.

[7] K. J. Seok, L. H. Jai and M. Y. Ho, "Apparatus and method for configuring a comprehensive intellectual property rights star network by detecting patent similarity", Korea Institute Of Science & Technology Information, G06F 17/30, 1020070071793, **(2006)**.

[8] Kolmogorov and A. Nikolaevich, "Foundations of the Theory of Probability", **(1950)**.

[9] D. Freedman, "Statistical models: theory and practice", Cambridge University Press, **(2009)**.

Authors

Jong-Bae Kim, he received his bachelor's degree of Business Administration in University of Seoul, Seoul(1995) and master's degree(2002), doctor's degree of Computer Science in Soongsil University, Seoul(2006). Now he is a professor in the Graduate School of Software, Soongsil University, Seoul, Korea. His research interests focus on Software Engineering, and Open Source Software.

Hyo-Jung Sohn, she received her bachelor's degree of Business Administration in Soongsil University, Seoul(2006). And she is studying her master's degree of software engineering in Graduated Soongsil University, Seoul. Her current research interests include open source development and management information system.

Min-Gyu Lee, he received her bachelor's degree of Information and Telecommunication in Dongguk University (2013). And he is studying her master's degree of software engineering in Graduated Soongsil University, Seoul. His current research interests include open source development and Security.

Baek-Min Seong, he received his bachelor's degree of Business Administration in Soongsil University, Seoul(2014). And he is studying her master's degree of software engineering in Graduated Soongsil University, Seoul. Her current research interests include database.

Yu-Jin Jo, she received her bachelor's degree of Business Administration in Atlanta University, (2013). And she is studying her master's degree of software engineering in Graduated Soongsil University, Seoul. Her current research interests include open source development and management information system.