

Increasing Search Success Rate by Comparing Query Formation Patterns

Soonjin Kwon¹, Seongjoon Yoo², Hanku Lee³, Changhee Han⁴, Kangseok Kim⁵ and Wonil Kim^{1,*}

¹ Department of Digital Contents, Sejong University, Seoul, Korea

² Department of Computer Engineering, Sejong University, Seoul, Korea

³ Department of Internet and Multimedia Engineering, Konkuk University, Seoul, Korea

⁴ Department of Computer Science, Korea Military Academy, Seoul, Korea

⁵ Department of Knowledge Information Security, Ajou University, Suwon, Korea

Received: 25 Apr. 2013, Revised: 20 Aug. 2013, Accepted: 21 Aug. 2013

Published online: 1 Apr. 2014

Abstract: Retrieving suitable information to a user's information request is a crucial task of search engines. However, most of conventional search engines that are based on pattern matching schemes tend to provide too many improper results, most of which are not corresponding to a user's request. It is due to the uncertainty of queries. To overcome this problem, we distinguished two different patterns after extensive investigation about the process of query formation. One is the general information seeking pattern and the other is the scholastic information retrieving pattern. The general information seeking pattern has a tendency to correspond to cross-relational query formation process pattern, and the scholastic information retrieving pattern corresponds to goal-oriented structured query pattern. This paper proposes a method to improve search results by utilizing these two query formation patterns. The study on the formation of general cross-relational queries, and goal-oriented structured query pattern by which comparing or modeling will improve the efficiency of navigation. These proposals will benefit not only search engine developers and library managers, but also search engine users and researchers. This paper is expected to provide more suitable searching results than those obtained by using current typical information retrieval engines.

Keywords: Information Seeking Pattern, Search Success Rate, Query Formation Pattern, Query Formalizer

1 Introduction

Information retrieval (IR) engines such as Google and Yahoo have developed and brought considerable value creation. Nevertheless, users are dissatisfied and began to wonder what the information they attained online means to themselves. They sometimes do not know what they are looking for. The statistics of unsuccessful search results justify this fact. Users are tired of the ranking search-results that are from retrieving index matched by keywords. Almost half of the searches are unsuccessful (up to 50%) [1]. Among the 35% of users sessions initiated by a query, the success rate of B2B site search was only 58%. Even though the visitors succeeded, 30% of them abandoned the sites because they typed in wrong queries [2]. Slightly over half of respondents (about 54%) said they would be likely to use a search engine that focuses specifically on serving their specific needs [3].

This research claims that IR system has to be complemented because that current IR systems do not have a process determines suitability. It is believed that the retrieval systems should ask users not only to say his/her intention but also to compare the intentions with results. There is too much information in the current IR system, because there is no suitability criterion in the typical retrieval system. One of the methods to avoid users wandering off is letting the system to recognize the users intent and suitability criteria.

Therefore, this paper proposes a methodology for accommodating users intent and suitability criteria. The same query presented by users can be expressed by a different meaning. It is necessary to deal with the uncertainty of the query, and to spare a place which can let the users intent and suitability criteria under the users circumstances. The important improvement point can be on the process before users queries are compromised.

* Corresponding author e-mail: wikim@sejong.ac.kr

After reviewing some previous works about query formation process in the next sections, newly designed concept is presented, and discussed.

2 Discussion on Previous Researches

Google's IR engine is composed of automatic relations extraction, global ranking, and page ranking [4]. This engine compares compromised queries with the data extracted from storage, ranks them, and then shows the list to searchers. The IR engines, however, only line up data lists because the suitability for the users' information needs cannot be known. The structure cannot display the information which meets the users' needs. This is the limitation of current IR engines. The genuine role of search engines has to include the users' information needs and to know the criteria in the suitability of those. After then, the search engines should bring the appropriate information so that the user would be able to satisfy about the useful information. The important starting points are information needs and suitability criteria. These are not considered under the condition of the traditional IR engines, Google and Yahoo, for instance. Their functions are just as information retrievers not information searchers.

Information needs consist of four levels: compromised needs, formalized needs, conscious needs, and visceral needs. Current IR systems are mostly dealing with compromised needs [5]. According to Taylor, they are also in the last process of the needs [5]. Compromised needs are expressed by compromised queries. Queries in current IR engines are problematic because they are unable to include cognitive, situational and affective suitability criteria. In order to expand the span of query from compromised needs to formalized needs, interactions should dramatically increase [6,7]. In fact, current IR engines such as Google and Yahoo still could not overcome the barriers within the compromised need level. As yet they are not search engines but retrieval engines to extract data or information from computational resources. It is a natural phenomenon that the numbers of studies about users are increasing as the research trends transit from Retrieval engine to Search engine. Current retrieval engines have abandoned the suitability or eligibility criteria to determine whether the retrieved information is suitable for the users' information seeking objects or goals. The search engines leave users astray so that users have to spend more time in the search process. Current retrieval engines ignore users' efforts to get closer to the concept of commitment. It is now necessary to bring suitability criteria of users into query-boxes than ever. How to capture the suitability criteria is required far more than to retrieve the information.

According to our analysis of previous works, experts have different procedures and methods from common people while forming a query [8,9,10,11]. Many researchers studied on modeling the impact of

information seeking behavior. Their studies are mostly theoretical deductive models, but contain useful implications. They researched various models and added empirical data, and also included various professional fields to the information seeking behavior based on empirical models. Wilson, Leckie, Savolainen, Johnson studied the influences on the different models [12,13,14,15], and Dervin, Ellis, Kuhlthau, Lenz researched the process of models [16,17,18,19]. Krikelas, Miller, Palmer investigated style and pattern of models [20,21,22].

Among the various groups on information seeking, the two categories were chosen to be investigated in this paper. They are news reporters group and academic scholars group who show different information seeking behavior patterns (Table 1). News reporters tend to start interactive information navigations, in which the information retrieval engine used as a starting tool.

News reporters have a wide range of patterns in their information search behaviors. The patterns were monitoring, extracting, browsing, and connecting.

Compared to them, academic scholars and researchers tend to show the pattern of information search in depth. They tend to use the retrieval engine after deciding or formulating queries. They have been identified by the importance of the information accuracy, timeliness of information retrieval, and convenience of understanding. This research distinguishes the trends of search engines usage as a explorer (or filter) and retriever. Search engines should have some different features from IR engines depending on in what phase of needs a user is standing. A user's purpose to navigate the information is to find the appropriate information for all intents and purposes. The suitability of navigating information is the basic step and very significant. Suitability has been a major research area of information retrieval over 50 years [23], but it still has not reached an agreement [24].

Early in the development of IR technologies, the system-oriented concepts, such as compliance of the relevance, precision and recall rates were emphasized. However, the system that is focused on relevance alone could not fully explain the concept of fitness [26]. That is why user-centered perspective studies prosper and this proposal is based on this approach [23,25,26].

Table 1: Information seeking behavior and the characteristics of the query

	reporter	scholars
Behavior pattern	Cross-relational	Objectoriented
Purpose of use	Starting tool(explorer)	Filter or Information retriever
Query formation	Progressively completed	Completed structure

3 New Concept Design based on Query Data Analysis

According to the previous studies and lessons learned on current IR engines, it is necessary to have a query formalization process for better search result. Within the current IR technology, the mass retrieving results are poured by simply comparing a query with index terms. Therefore, it is needed for an intelligent search to pursue users' intents, and criteria of determining the suitability is required. The information suitability in the human information behavior is a dynamic concept depending on the users situations in which they are currently in. It is also a complex and multi-dimensional concept. It is affected by cognitive, emotional, social and cultural factors. Information suitability accepted by users is determined with the subjective and dynamic compliance in the actual search processes. Suitability is influenced by multi-dimensional factors. It could not be defined by just one of any suitability factors. It seems to be that suitability properties include three roles in the information behavior. They are firstly multi-dimensional concept in accordance to a users awareness, secondly dynamic concept based on a users determination at a specific time point, and finally compound concept that can be measured by users perspective.

Suitability is defined as logical, situational, and psychological. It is also described and analyzed within various perspectives. But still, it has not reached to an agreement for a concept as well as its properties [27]. People just intuitively understand it [28,29]. Related researches are classified as the types of relationships on some web sites authentic data in which IR technologies had been applied. For example, the query data of two institutes such as Korea National Assembly Library (KNAL) and Korea Institute of Science and Technology Information (KISTI), each total counts appeared occasions of 3,185,187/1month and 411,254/1month respectively. The KISTIs Web site had a fixed user intention to locate the paper (37%), trends analysis (26%), and dissertation (18%). The percentage is not an intention representation, but a media type. Also for the engines employed by these two institutions, mass search results had been poured by simply comparing a query with the index term. The current IR engine by traditional method gives users too many results but they do not fit to users goals or objectives to find any information. The need for an intelligent search fitting users intent is required. Therefore search engines should be composed of a query formalizer, and a retriever that is in the traditional retrieval engines.

A retrieval engine with traditional methods shows a list ranked up after users input the queries. It is not a search, but just a retrieve. Therefore, it is needed to invent a new model to perform the real search process. In the proposed model, both news reporters and scholars should have different search engines. It is necessary to find a place for capturing users intent. The place was designed

for to absorb and to hold the intention. It is a useful concept to absorb the users suitability in searching as query formalizer.

A method to catch the users intent is required. Figure 1 shows a newly designed method. It shows new concept of the query formalizer. User information needs and the information retrieved are compared with the suitability criteria. Candidate of information media and the hint for helping users to make appropriate query are formed and selected. Query is formulated, after which the media or computational / informational sources can perform the process of getting information. Thus, it could be possible to retrieve the information with the conformance enhanced by the query formalizer. The core difference between traditional methods and enhanced methods is to separate functions from processes during the query process. This method has an advantage that can avoid retrieving too many results which are not suitable for users information needs. The query formalizer would be a new concept to guarantee the suitability for users of retrieving information.

4 Proposed Approach on Search Engine

In order to improve query process, it is necessary to accept the users suitability criteria because that is the first step to extract user intention for information retrieval. Problems are defined as follows. First, how to get user intention, second, how to assist the user to formalize the query, third, how to implement them. The following three sections will explain these steps in detail.

4.1 User Intention

In the traditional way that retrieves the information, the users intention was not only unknown in advance but also not the part of the system. In order to know the users intent, the swamp is created. The swamp preserves the absorbed intent-words or pre-stores in the pre-defined storage. The method recommends several words after extracting the intent-word from the queries relations.

In order to capture the intention of the user, a novel idea that are not employed in the traditional search methods are considered. In this new model of search process, the query is determined to construct a bridge to connect user intention and search engine. That is the query formalizer, which assists the search engine to retrieve information according to user intention. By doing so, the intention of the news reporters and scholars is captured and stored differently. It would be able to satisfy all of the patterns.

4.2 Queries Formalization Methods

To confirm the query, news reporters are wandering all over the place to find information and sources. On the

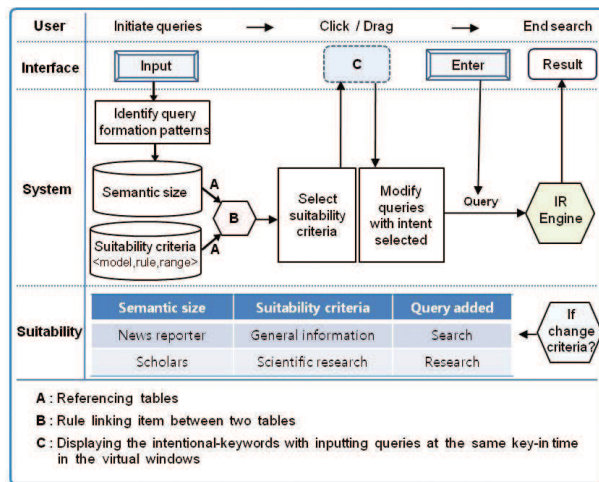


Fig. 1: Process flow to reflect the users needs, in the place to compare whether the purposes to search is fit or not

other hand, the professional scholars or experts usually confirm the query by themselves first.

In order to satisfy both news reporters and researchers, it recommends intended words on the search window as soon as a user enters queries. After this process, the intent is reflected in the query. A large amount of unnecessary information, which is the defects of the current IR engines, could be overcome by confirming an uncertain query with suitability criteria.

In these processes as in Figure 1, the most critical component is creating the set of intent-words and suitability criteria. The processes are as follows; first, it establishes the types of relations between two queries; second, it infers intention; and finally, it reconstructs queries with recommendation.

Seizing on users intention is a key factor to the process because that may limit the scope of the semantic. It may also reduce the range to search requested information. Figure 2 shows the comparison of process and the process of reducing the query size.

The function of B-module in Figure 1 processes as follows. First, it established the types of intention by calculating the value in the table of query-semantic-size from the relations of identified queries. The table includes the pattern and rules of assigning the query pattern to intentional-words. It also includes the lists of the intentional-words. Second, the table of suitability-criteria is established after determining semantic-size from assigning intentional-word. The table includes suitability and suitability criteria model for each user. It also includes patterns, rules and ranges of assigning the model to the semantic size on the queries pattern. Third, after linking two tables by the combination of rules-sets between query-pattern and suitability-criteria, it displays the intentional-keywords by inputting query in the virtual windows at the same time. Afterwards, the user selects

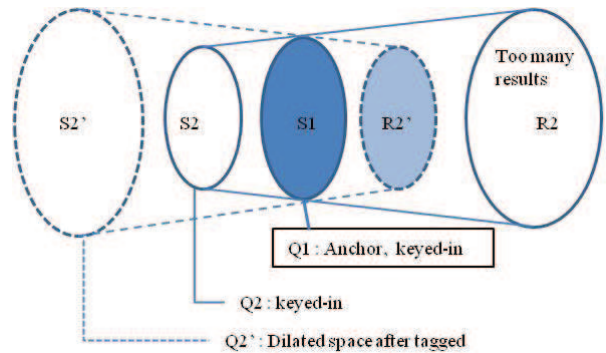


Fig. 2: How to reduce the size of too many results list retrieved

the queries by clicking/dragging the words in the virtual window. Finally, the system modifies the queries clicked/dragged by user. Modified queries would be transferred to the IR engine.

The most important key during this task is to decide the semantic-size of the queries because it is possible to induce the rules that decide the combination sets between query and intention. By doing so, the number of search results could be reduced significantly as much as users want.

Reducing the semantic-size of query can be explained as in Figure 2. Let $S1$ be the space size of the semantic of the anchor keyword to input query, and $S2$ be the space size of the semantic of the second keyword to input query, which is complementary the anchor query ($S1$). And suppose $S2$ be the bigger size than $S1$ and $S2$.

If $S2$ increases larger than $S1$ size, then it becomes $S2'$. If $S2$ expands bigger than $S1$, then *reduced result* ($R2$) becomes much smaller than *too many results* ($R2$). It shows that the way to increase the $S2$ is to reduce the size of $R2$. Therefore by increasing one of the semantic spaces between two queries to input, the result can be equal to a result of reduction, and that means increasing the compliance to users goal or objective.

By limiting the semantic space to find the information, the search results would be reduced. Thus, appropriate response would be obtained with suitability criteria of users intent to query.

4.3 Prototype Configuration

The traditional search box to input search keywords consists of one small box, but the proposed prototype has two boxes in which they have query-formulator box (upper box on the left side of Figure 3) and query-input box (lower box on the right side of Figure 3). When the user input the key-word in the first box, the intent-words are displayed on the virtual window. Recommended words are generated before this task and displayed automatically at the same time with input.

Figure 3 shows adding the virtual box to store querying users intention. Finally, the purpose of the query is granted, for which the query is suitable for user

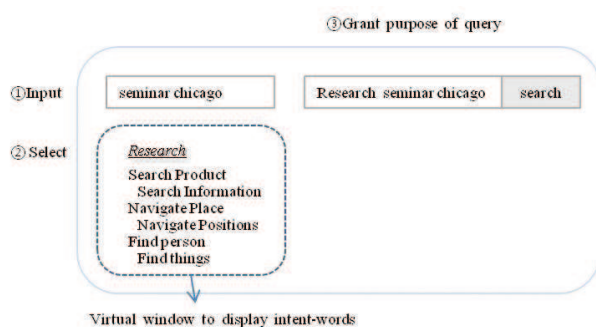


Fig. 3: Explained of the prototype configuration

intention, and the concept of swamp is applied. The swamp absorbs the intent-words by so called Osmotic pressure method. If you input some keywords in the query-formulator box an imaginary box formed in the query is displayed with the users intent. The combined query will be appeared in the search box after users clicks/drags of the intent. As a result, the amount of the search results is expected to be significantly reduced.

According to the mentioned studies above, query formalization is a problem that exists outside the retrieval area [30]. The reason is that the suitability depending on the situation is the concept on the upper hierarchy. Search engine users in South Korea tend not to enter detailed key-words [31]. If the content is not what they wanted, they just change the key-words and re-enter the changed key-words. This seems to be due to the increased documents with the various types of good quality and change of the information fitness criterion to users [27].

5 Discussion

The current IR engines do not reflect the users suitability criteria. Even if it is said that the current IR engines have accurate results (high precision) and high recall-rate, current technologies on the IR have low level of fitness. Dramatically performing improvement of the search engine is needed. Relevance is not a conformance, but a relationship. Suitability to the users request for information means the highest satisfaction, eligibility or fitness. Users would be satisfied with the information in which has suitability to ones needs. Information needs have four-levels such as visceral need, conscious need, formalized need, and compromised need. Current IR engines deal with only the compromised need. The problems are due to the limitation on the query. Improvement or evolution for solving search engines problem is necessary. It is believed to be the right direction shifting from retrieval technologies to search technologies, or reaching question / answering technologies. Suitability to the users request for information, and the maturity of the complex technical infrastructures would be the key directions to satisfy the users.

This paper presents an improved method to search users requested information. It also suggests the effect of the improvement density for the queries presentation by modelling the types of two queries relationships. Besides, its effectiveness will be proven. Although there would be considerable contributions and benefits, the search engines such as Google and Yahoo are problematic and ineffective, in which it shows many irrelevant search results against the users intention. Improving effectiveness of the compliances is possible by accessing to users intention with the constructing model for the relationship between queries.

This model has the advantage that can be separated from the search engine. The model can also be used for giving recommended keywords when the user entered to the search engines. It is expected that the simulation results reduced unnecessary search results rate by considerably depending on the employed features.

6 Conclusions

This paper proposed a novel approach to increase search success rate by comparing query formation patterns. It will reduce unnecessary search for users and developers. The application of this approach would be enable reducing unnecessary effort and unnecessary interactive searching for fine-grained queries relation patterns.

After separating the search process and search function, and complementary functions of IR engine, we were able to discover the clue that IR Engine could be evolved once again. Search success model is very important to all searchers. Formalization of different types of success for informational search has very great significance. The future direction of query formation process requires an in-depth research, and the in-depth research on influencing factors to determine the query has to be made. By doing so, the effect of this study would be confirmed.

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2013-H0301-13-3006) supervised by the NIPA(National IT Industry Promotion Agency).

References

- [1] J. Callan, J. Allan, C. L. A. Clarke, S. Dumais, D.A. Evans, M. Sanderson, and C.X. Zhai, Meeting of the MINDS: An Information Retrieval Research Agenda, ACM SIGIR Forum, 41, 25-34 (2007).
- [2] G. McCracken, Vertical Search: Challenges & Opportunities, 2, 1-22 (2008), [Cited from <http://www.slideshare.net/>]

- [3] L. Gregoriadis, The changing digital environment and vertical search, *12*, 1-21 (2008), [Cited from <http://www.slideshare.net/>]
- [4] S. Brin, R. Motwani, L. Page, and T. Winograd, What can you do with a Web in your Pocket?, *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, **21**, 37-47 (1998).
- [5] R. S. Taylor, Question Negotiation and Information Seeking in Libraries. *College and Libraries*, **28**, 178-194 (1968).
- [6] T. Saracevic, Information Science, *Journal of the American Society for Information Science*, **50**, 1051-1063 (1999).
- [7] T. Saracevic, Relevance reconsidered, In: *Information science : Integration in perspectives*, Proceedings of the Second Conference on Conceptions of Library and Information Science(CoLIS2), 14-17 (1996),
- [8] K. Kim, Review of Research Trend in Models of Information Seeking Behavior. *Journal of the Korean Society for information Management*, **22**, 235-254 (2005).
- [9] J. Y. Lee, An Overview of Developing & Applying Information Use Behavior Models, Proceedings of the 13th Korean Society for Information Management Conference, 5-10 (2006).
- [10] H. Kwon and D. Jeong, Analyzing Patterns in News Reporters Information Seeking Behavior on the Web, *Journal of the Korean Society for information Management*, **22**, 109-130 (2005).
- [11] S. H. Han and J. Y. Lee, Searching Behavior of Scientists and Technologists Based on Their Information Needs, Proceedings of the 13th Korean Society for Information Management Conference, 17-22 (2006).
- [12] T. D. Wilson, Information Behavior: An Interdisciplinary Perspective, *Information Processing & Management*, **33**, 551-572 (1997).
- [13] G. J. Leckie, K. E. Pettigrew and C. Sylvain, Modeling the Information Seeking of Professionals: A General Model Derived from Research on Engineers, Healthcare Professionals, and Lawyers, *Library Quarterly*, **66**, 161-193 (1996).
- [14] R. Savolainen, Everyday life information seeking: Approaching Information Seeking in the Context of way of life, *Library & Information Science Research*, **17**, 259-294 (1995).
- [15] J. D. Johnson, *Cancer-related Information Seeking*. Cresskill, NJ: Hampton Press, (1997).
- [16] B. Dervin, An Overview of Sense-making Research: Concepts, Methods, and Results to Date, Presented at: International Communication Association Annual Meeting, Dallas, Texas, USA. (May 1983).
- [17] D. Ellis, Modeling the information-seeking Patterns of Academic Researchers: A Ground theory approach, *Library Quarterly*, **63**, 469-486 (1993a).
- [18] C. C. Kuhlthau, *Seeking Meaning: A Process Approach to Library and Information Services*, Norwood, NJ: Ablex Publishing, (1993a).
- [19] E. r. Lenz, Information Seeking: A Component of Client Decisions and Health Behavior, *Advances in Nursing Science*, **6**, 59-72 (1985).
- [20] J. Krikelas, Information Seeking Behavior: Patterns and Concepts, *Drexel Library Quarterly*, **19**, 5-20 (1983).
- [21] S. M. Miller, Monitoring and Blunting: Validation of a Questionnaire to Assess styles of Information Seeking under Threat, *Journal of Personality and Social Psychology*, **52**, 345-353 (1987),
- [22] J. Palmer, Scientists and Information: .Personal Factors in Information Behavior, *Journal of Documentation*, **47**, 254-275 (1991).
- [23] P. Borlund, The concept of relevance in IR, *Journal of the American Society for Information Science and Technology*, **54** 913-925 (2003).
- [24] E. G. Toms, H. L. OBrien, R. Kopak, and L. Freund, Searching for relevance in the relevance of search, Proceedings of Fourth International Conference on Conceptions of Library and Information Science(CoLIS2005), 59-78 (2005).
- [25] L. Schamber, M. B. Eisenberg, and M. S. Nilan, A re-examination of relevance: Toward a dynamic, situational definition, *Information Processing & Management*, **26**, 755-775 (1990).
- [26] E. Cosijn and P. Ingwersen, Dimensions of relevance, *Information Processing & Management*, **36**, 533-550 (2000).
- [27] J. Park and Y. Sohn, User-centered relevance judgement model for information retrieval, *Korea Journal of the Science of Emotion and Sensibility*, **12**, 489-500 (2009).
- [28] T. Saracevic, Relevance: A review of the literature and a framework for thinking on the notion in information science, Part II: Nature and manifestations of relevance, *Journal of the American Society for Information Science and Technology*, **58**, 1915-1933 (2007a).
- [29] T. Saracevic, Relevance: A review of the literature and a framework for thinking on the notion in information science, Part III: Behavior and effects of relevance, *Journal of the American Society for Information Science and Technology*, **58**, 2126-2144 (2007b).
- [30] P. Vakkari, Task-based information searching, *Annual Review of Information Science and Technology. Information Today*. Medford, NJ, **37**, 3-48 (2003).
- [31] S. Park and J.-H. Lee, Comparative Evaluation of the Unified Search Services Provided by Major Korean Search Portals, *Journal of Korean Library and Information Science Society*, **39**, 265-278 (2008).



Soonjin Kwon has finished a Doctoral course in Sejong University. He is currently a Researcher with the Department of Digital Contents at Sejong University, South Korea. His research interests are in the areas of intentiondriven information search, queries relations, computational intelligent systems and knowledge discovery.



Kangseok Kim received B.S. in Mathematics from Kyungwon University, Korea, and M.S. in Computer Engineering from Syracuse University, NY, USA and Ph.D. in Computer Science from Indiana University at Bloomington, IN, USA. He is currently research professor of Knowledge Information Security department at Ajou University, Korea. His main research interests include mobile computing, ubiquitous computing, scalable distributed database system, floor control, mobile applications with smart phones, mobile security data mining and bioinformatics.



Seongjoon Yoo received the Ph.D. degree in computer and information science from Syracuse University, Syracuse, New York, in 1996. Since 2002, he has been with the School of Computer Engineering, Sejong University, where he is currently a Professor. Before joining the university, he was a research member of senior technical staff at the ETRI, Korea. His research interests include learning systems, pattern recognition, data mining, and image processing.



Wonil Kim received the B.E in Metal Engineering from Hanyang University, Seoul, Korea in 1982. He worked for Korean Air from 1981 to 1985 as System designer and programmer. He received the B.S., M.S in Computer Science from Southern Illinois University, U.S.A. in 1988, 1990 respectively. He received Ph.D. in Computer and Information Science from Syracuse University, U.S.A. in 2000. From 2000 to 2001, he worked for Bhasha INC, U.S.A. as technical and research staff. He was with Ajou University, Suwon, Korea from 2002 to 2003. Since 2003, he has been with Department of Digital Content, College of Electronics and Information, Sejong University, Seoul, Korea. His research interests include artificial intelligence, multimedia contents, bioinformatics, adaptive systems, and computer security.



Hanku Lee is the director of the Social Media Cloud Computing Research Center and an associate professor of the division of Internet and Multimedia Engineering at Konkuk University, Seoul, Korea. He received his Ph.D. degree in computer science at the Florida State University, USA. His recent research interests are in cloud computing, distributed real-time systems, distributed computing, and compilers.



Changhee Han received M.S. in Computer and Information Science from Syracuse University, USA in 1994, and Ph.D. in Computer Science from University of Southern California, USA in 2004. He is currently a Professor in the Department of Computer Science, Korea Military Academy, Korea. He has published over 20 papers in domestic and international peer reviewed journals. His research interest includes virtual human modeling & simulation with path planning, cognitive mapping, and spatial reasoning; and computational intelligence with genetic algorithm and neural network