Fuzzy-based Intelligent Expert Search for Knowledge Management Systems

Kun-woo Yang [a] and Soon-young Huh [b]

[a] *Graduate School of Management, Korea Advanced Institute of Science and Technology*
207-43 Cheongryangri-Dong, Dongdaemoon-Ku, Seoul 130-012, South Korea
Tel: +82-2-958-3650, Fax: +82-2-958-3604, E-mail: simpact@kgsm.kaist.ac.kr

[b] *Graduate School of Management, Korea Advanced Institute of Science and Technology*
207-43 Cheongryangri-Dong, Dongdaemoon-Ku, Seoul 130-012, South Korea
Tel: +82-2-958-3626, Fax: +82-2-958-3604, E-mail: syhuh @kgsm.kaist.ac.kr

Abstract

In managing organizational tacit knowledge, recent researches have shown that it is more applicable in many ways to provide expert search mechanisms in KMS to pinpoint experts in the organizations with searched expertise. In this paper, we propose an intelligent expert search framework to provide search capabilities for experts in similar or related fields according to the user's information needs. In enabling intelligent expert searches, Fuzzy Abstraction Hierarchy (FAH) framework has been adopted, through which finding experts with similar or related expertise is possible according to the subject field hierarchy defined in the system. To improve FAH, a text categorization approach called Vector Space Model is utilized. To test applicability and practicality of the proposed framework, the prototype system, "Knowledge Portal for Researchers in Science and Technology" sponsored by the Ministry of Science and Technology (MOST) of Korea, was developed.

Keywords:

expert search; knowledge management; text categorization

Introduction

To manage valuable organizational tacit knowledge effectively, which is usually embedded in the operating procedures as routines or standards in the organizations, a lot of time and efforts among knowledge management system (KMS) researchers have been devoted. However, the results of those research efforts to develop an effective and efficient way to store, retrieve, and share tacit knowledge have not been successful enough to be widely accepted in industry due to its limited applicability and inflexibility.

Some of researchers in this field claimed that deliberate separation of tacit knowledge from its holders, meaning codifying it, inevitably degrades its values and tacit knowledge should be handled in tacit ways [1, 6, 10]. One of the tacit ways proposed for managing tacit knowledge is providing a helpful search method for experts possessing needed expertise in the organization. Conventional query processing, which is usually involved with SQL (Structured Query Language), can only provide exact answers to users' queries if and only if all the conditions of the queries are satisfied. Therefore, if there are not any instances, which match all the search conditions, it provides nothing as the query result. In the expert search case, if there are not any experts having searched expertise, the system cannot provide any useful information. To overcome this limitation and increase the level of satisfaction in performing expert searches, we can utilize cooperative query answering mechanisms [3, 4, 9], which were developed to give flexible query results through interactions with users.

In this paper, we adopt *Fuzzy Abstraction hierarchy* (FAH) [12], which incorporates fuzzy relations and operations to calculate similarity measures among data values in a knowledge abstraction hierarchy. By adopting FAH, KMS users can be given the ordered list of approximate answers to their expert search queries based on the similarity measures calculated using fuzzy relations, which enables searching for experts in similar or related fields. FAH-based search operations require the pre-defined subject field hierarchy with initial similarity measures assigned for pairs of fields having the same parent nodes.

We improve the FAH framework through the automation of the initial similarity measure assignment. To derive initial similarity measures among subject fields, we utilize a text categorization method called Vector Space Model (VSM) [2], through which the needed measures are calculated using training documents with pre-assigned subject fields. By combining the improved FAH framework proposed in this paper with an expert search mechanism in KMS, we can (1) eliminate the constant maintenance cost for subject field hierarchy, (2) reduce the calculation complexity of similarity measures compared to using only a text categorization method, and (3) lift the burden of learning high-level query language syntax from a user's point of view.

Automatic Field Classifier Using Text Categorization

To derive initial similarity measures among subject fields and assign them, the automatic field classifier should be trained using training documents. Among the text categorization techniques, VSM has been chosen in this paper to train the classifier and measure initial similarity values among subject fields since it shows the optimal performance among other techniques in many research results. The following is the list of 3 steps to train the automatic field classifier and figure 1 depicts the training process.

Step 1. *Preparing training documents*: Documents to be used for training a classification model should be prepared with pre-assigned subject fields to which each document belongs. Meaningful terms are to be extracted from training documents to build a vector space for each subject field, which later is used to calculate similarities among subject fields.

Step 2. *Building aggregate document vectors for each subject field*: The field vector (FV_i) for each subject field is assembled from the aggregated document (AD_i) of each field. Each field vector has the same number of dimensions as the total number of elements in the index term set (TS) extracted from the entire document collection for training. FV_i, which is the vector of a specific aggregated document (AD_i), is composed of terms belonging to the entire index term set and the explanatory level (or comparative importance) of those terms in that specific field. We can gain the comparative importance of term j in AD_i (field i) from tf_{ij} and t_i in the equations, $w_{ij} = tf_{ij} \times idf_j$, $tf_{ij} = freq_{ij} / \mathrm{Max}_l(freq_{il})$, and $idf_j = \log(N/t_i)$.

Step 3. *Calculating the initial similarity measures among subject fields*: Using a text categorization method, it is possible to derive initial similarity measures among subject fields since an individual similarity measure between two subject fields can be derived from similarity measures between groups of documents belonging to each subject field.

Since the aggregate document vectors derived from the training contain the vectors of each term's explanatory level about each subject field, it is possible to calculate the similarity level among specific subject fields themselves if *Cosine Similarity Function* [2] method applies. The following equation (1) shows the mathematical representation of this similarity function.

$$Sim(FV_1, FV_2) = \cos\theta = \frac{FV_1 \cdot FV_2}{\|FV_1\| \times \|FV_2\|} \qquad (1)$$

where $FV_1 \cdot FV_2$ denotes the inner product of two vectors, FV_1 and FV_2, and $\| FV_1 \|$ represents the absolute value of vector FV_1.

Fuzzy Abstraction Hierarchy

The ultimate purpose of automating expert profiling and providing a search mechanism for needed experts is to return satisfactory answers to a KMS user's queries for experts he or she wants. To improve the user satisfaction on the search results, not only finding the exact matches to the search criteria but also providing approximate answers with similar characteristics is needed when the user's query does not return appropriate answers using the exact match mechanism. In more detail, when a specific KMS system does not have expert profiles with specific expertise that a user wants to find or the query result is not satisfactory enough, more search satisfaction can be achieved if the system is able to provide an additional search capability to find experts with similar or related expertise by improving the query processor of the system.

To facilitate similar expert searches, Fuzzy Abstraction Hierarchy (FAH) is utilized, which is a knowledge representation framework equipped with an intelligent query processing capability to provide approximate answers to user queries. FAH has been developed to remedy the shortcomings of conventional query processing that does not possess any intelligence to cooperate with users in providing flexible query results according to the user's

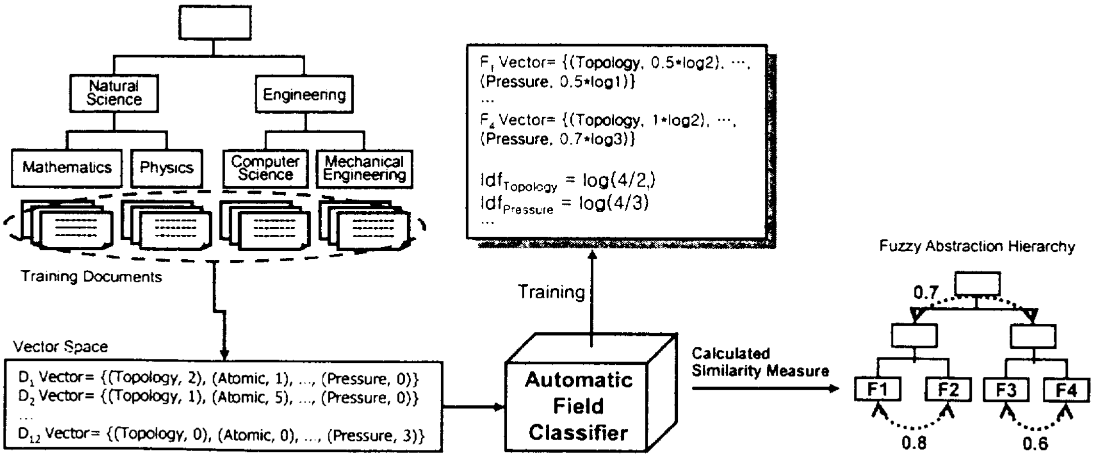
Natural Science
Engineering
Mathematics
Physics
Computer Science
Mechanical Engineering
Training Documents
Vector Space
D₁ Vector= {(Topology, 2), (Atomic, 1), ..., (Pressure, 0)}
D₂ Vector= {(Topology, 1), (Atomic, 5), ..., (Pressure, 0)}
...
D₁₂ Vector= {(Topology, 0), (Atomic, 0), ..., (Pressure, 3)}
F₁ Vector= {(Topology, 0.5*log2), ···, (Pressure, 0.5*log1)}
...
F₄ Vector= {(Topology, 1*log2), ···, (Pressure, 0.7*log3)}
Idf_Topology = log(4/2,)
Idf_Pressure = log(4/3)
...
Training
Automatic Field Classifier
Calculated Similarity Measure
Fuzzy Abstraction Hierarchy
0.7
F1
F2
F3
F4
0.8
0.6

Figure 1 - Automatic Field Classification

needs unless complex and strict query language syntax is followed. It analyzes the intent of a query and transforms the query into a new query of either greater scope by relaxing the original query conditions or smaller scope by strengthening them. FAH has a couple of advantages compared to other data abstraction methods in applying for expert searches. First, since FAH represents the semantic relationships among data values based on data abstraction as a hierarchy, it is the most appropriate and applicable representation framework for expert categorization. This is because it has its own hierarchical structure due to the categorical trait of expertise. Second, based on the mechanism proposed in FAH, we can calculate and represent the exact level of similarity among values, through which other valuable information such as fitness scores or similarity measures between search results can be given.

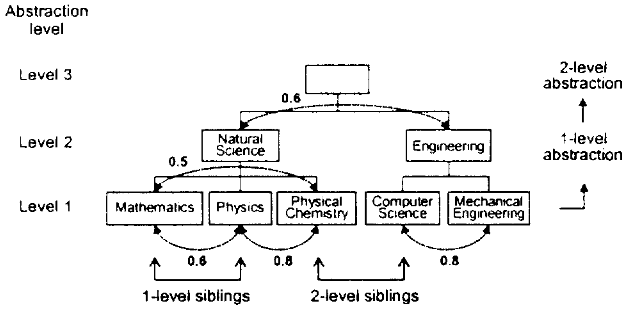

Figure 2 - Example of Fuzzy Abstraction Hierarchy

Providing a wider range of approximate answers by relaxing search conditions requires a human expert's knowledge of the underlying database semantics, e.g., similarity measures among data values. Thus, a system administrator or an experienced expert should evaluate, determine, and finally assign the values for the elemental similarity measure among sibling nodes sharing the first-level parents as shown in figure 2. Also, constant maintenance effort is required whenever needs for updating those similarity measures arise. This type of manual assignment of similarity measures is the major limitation of FAH and has the following problems. First, the similarity values assigned for pairs of fields in the hierarchy are fully dependent on the person who performs evaluation and value assignment and, therefore, the values are rather subjective. Second, it is not possible for one individual to assign all the similarity measures correctly if the hierarchy is large resulting in the large number of data pairs requiring initial value assignment because, most of time, one person cannot be knowledgeable about all the defined fields in the hierarchy. Last but not least is that subject fields themselves keep changing. In other words, as time goes by, the similarity measure between two fields can be different from what it was before due to many reasons. Whenever this kind of update requirement arises, human intervention, which costs time and money, should be involved if the manual assignment process is adopted.

Considering these limitations, there is enough room for developing a more efficient as well as more cost-effective similarity value evaluation and assignment methodology. Thus, an automatic method of assessing the similarity measures is needed for effective knowledge maintenance. A text categorization approach can be applied to automatically derive the similarity measures among the subject fields by analyzing knowledge contents belonging to each subject field.

Intelligent Expert Search Using FAH

Similarity Calculation among Subject Fields

Similarity measures among 1-level sibling fields are derived through the training process of the automatic field classifier. On the contrary, we calculate the similarity measures for pairs of siblings having 2-level or above relationships with each other using fuzzy relation's Max-Min operation [13] and derived initial similarity values from the training. It is possible to assign all the similarity measures for pairs of siblings regardless of their level of relationships using the same procedure we follow to derive similarity measures for 1-level siblings. However, this approach requires the considerable amount of calculations depending on the size of the subject field hierarchy since it considers all the possible pairs of siblings and calculate the desirable values for them.

Figure 3 shows the process of similarity measure calculation among 2-level siblings using Max-Min operation with simple FAH as an example. In the figure, rectangles represent subject fields and solid lines depict the hierarchical structure among subjects such as p_1 being the parent for a_1 and a_2. Also, a thick line shows the feasible path, which represents the possible way to connect two subject fields along with the hierarchical structure on FAH. Numbers along with dotted arrows represent the similarity measures among two fields connected by the arrow.

Let us calculate the similarity measure between a_1 and b_3 from the figure. Here, a_1 and b_3 are subject fields belonging to higher level fields, p_1 and p_2, respectively and their parent nodes, p_1 and p_2, share the same parent node though not shown in the figure, making a_1 and b_3 2-level siblings. The result of fuzzy relation composition using Max-Min operation for a_1 and b_3 is as follows.

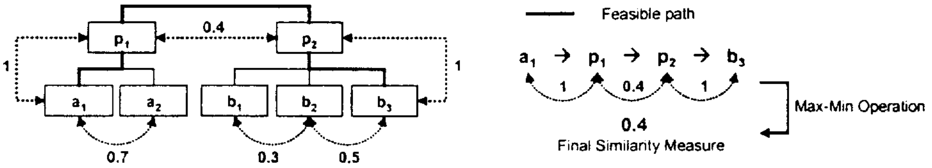

Figure 3 - Similarity Measure Calculation Using Max-Min Operation

$(Derived)\ sim(\mathrm{a}_1, \mathrm{b}_3)$

$$= \mathrm{Max}\ [\mathrm{Min}\ [sim(\mathrm{a}_1, \mathrm{p}_1), sim(\mathrm{p}_1, \mathrm{p}_2), sim(\mathrm{p}_2, \mathrm{b}_3)]] \qquad (2)$$

As illustrated in figure 3, FAH has the hierarchical structure in which one child has only one parent. Consequently, from a_1 to b_3, there exists only one feasible path, which is $\mathrm{a}_1 \rightarrow \mathrm{p}_1 \rightarrow \mathrm{p}_2 \rightarrow \mathrm{b}_3$. If we assume that the similarity measure between a parent and a child is always 1 uniformly, equation (2) can be transformed into the following.

$(Derived)\ sim(\mathrm{a}_1, \mathrm{b}_3)$

$= \mathrm{Min}\ [sim(\mathrm{a}_1, \mathrm{p}_1), sim(\mathrm{p}_1, \mathrm{p}_2), sim(\mathrm{p}_2, \mathrm{b}_3)]$

$= \mathrm{Min}\ [1, sim(\mathrm{p}_1, \mathrm{p}_2), 1]$

$= sim(\mathrm{p}_1, \mathrm{p}_2)$

$= 0.4$ □□□□□□□ □□□□□□□□□□ (3)

We can see that by fixing the similarity measure between the parent and the child as 1, the derived similarity measure between a_1 and b_3 is the same as the similarity measure among their parents, p_1 and p_2. Meanwhile, it should be noticed that *(Derived)* *sim*(a_1, b_3), 0.4, resulted from equation (3) is greater than *sim*(b_1, b_2), 0.3, which does not satisfy the decreasing similarity proposition below. To remedy this anomaly, we define the extended similarity as in equation (4) including the concept of sibling level formulated from the abstraction level in the hierarchy.

$$(Extended)\ sim(a_1, b_3) = \frac{\{Max(level\ difference) - level\ difference\} + (Derived)\ sim(a_1, b_3)}{Max(level\ difference)} \qquad (4)$$

where level difference among *n*-level siblings is *n* and *Max(level difference)* is the largest possible sibling level between two subject fields defined in the hierarchy. Both of them are 2 in this example.

Proposition 1. *Monotonously Decreasing Similarity*

The similarity measure between (*n*+1)-level sibling fields are smaller than those between *n*-level sibling fields, where *n*>=1.

Proof.

Let *a* and *b* be *(Derived) sim* between *n*-level siblings and (*n*+1)-level siblings respectively.

If *Max(level difference)* is *M* and *(Extended) sim* of *a* and *b* are E_a and E_b,

$$E_a - E_b = \frac{M-n+a}{M} - \frac{M-(n+1)+b}{M}$$

$$= \frac{M-n+a-M+n+1-b}{M}$$

$$= \frac{a-b+1}{M}$$

because $0 < a < 1$, $0 < b < 1$, and $M > 0$,

$-1 < a - b < 1$ and $0 < a - b + 1 < 2$.

Therefore, $\frac{a-b+1}{M} > 0$ and $E_a > E_b$.

Proposition 1 describes the monotonousness property of a decreasing similarity measure with respect to the abstraction level. The similarity measure between *n*-level siblings monotonously decreases as the abstraction level (sibling level) *n* increases. For example, similarity measures among 2-level sibling fields are less than those among 1-level siblings and greater than those among 3-level siblings. Consequently, to reduce the number of sibling pairs requiring similarity measures explicitly to be assigned among them, the similarity values among *n*-level sibling fields (*n*>=2), for which similarity measures are not assigned explicitly, can be calculated using the Max-Min composition operator of fuzzy relations while the similarity measures among 1-level sibling fields are derived from the procedure explained in the previous section using Vector Space Model.

Equation (4) shows that the final similarity measure, *extended similarity*, is calculated from the derived similarity incorporating the level difference to satisfy the monotonously decreasing similarity constraint. By adding '*Max(level difference) – level difference*' to the derived similarity, or just the similarity measure derived from the knowledge classifier in case of 1-level siblings, we can guarantee that the final similarity measure between two fields gets smaller accordingly as the level difference gets larger. In addition, we normalize the calculated similarity measure to have a value between 0 and 1 by dividing it by the maximum level difference. Therefore, the final similarity measure between a_1 and b_3 in the example is *sim*(a_1, b_3) = {(2-2)+0.4}/2 = 0.2. The final similarity measure derived from the above-mentioned calculating process is used to get the list of experts, who are appropriate for a user's search criterion, using FAH operations. We elaborate on the detailed expert search process in the following section.

Process of Intelligent Expert Search

When there are no experts in a specific field, which is requested by a user's query, similar field expert search capability can provide useful search results in a certain level. Addition to such functionality, it is ideal to sort searched experts on their level of similarity to the user's requested expertise. This additional intelligence is enabled by manipulation of given similarity values among subject fields based on FAH.

Figure 4 shows the general process of expert search proposed in this paper using UML's activity diagram [8]. Expert search starts with a user's keyword input, which the system uses to find matches in the field names from the subject field hierarchy. Field name matching starts with the lowest level in the hierarchy. If no matches are found, searching is targeting the next higher level for keyword

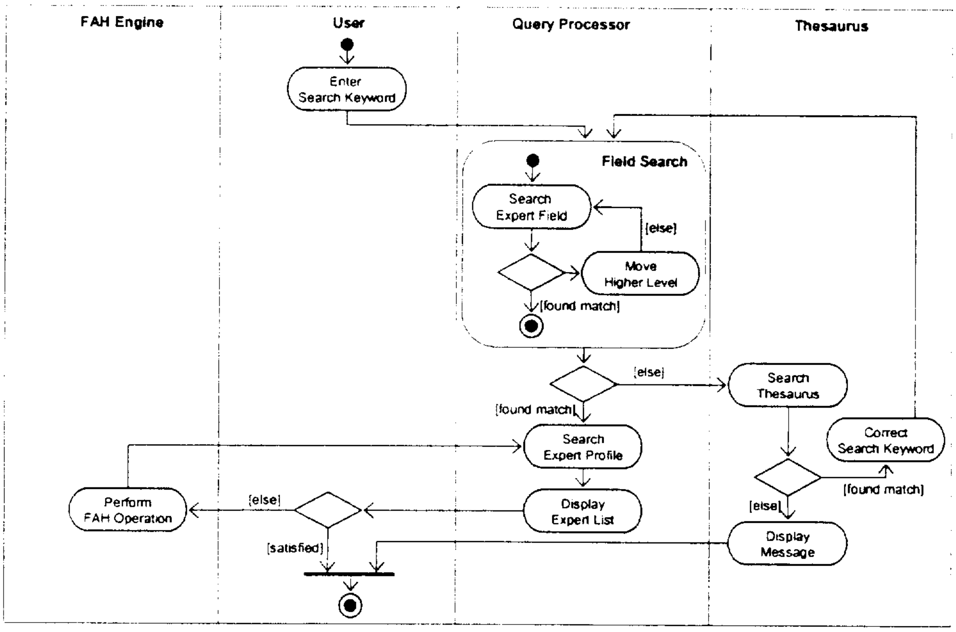

Figure 4 - Activity Diagram for Intelligent Expert Search

matching. From the lowest level to the top, if a field name existed in the hierarchy is found, the expert list for that found field is compiled from the expert profile database. If there are not any matches even though searching is completed all the way to the top level of the field hierarchy, a search thesaurus participates in the process. We adopt this thesaurus approach because it can remedy possible problems that result from typos or synonyms, which are very likely to happen while entering a keyword. For example, let us say there is a subject field called 'Electronic Commerce' in the hierarchy. In this case, if a user types 'EC' instead of its non-abbreviated counterpart, the system would not return any expert lists unless it uses other mechanisms in addition to word-for-word matching. The search thesaurus plays an important role in changing the input keyword, 'EC', to 'Electronic Commerce' so that the system can perform the rest of the search function. By the same token, the search thesaurus can be used to correct a wrong input keyword in case of typos.

Application of Intelligent Expert Search in Knowledge Portal

To test the applicability of the proposed expert search framework, we developed a prototype system, KMS with the intelligent expert search functionality on 'Knowledge Portal' architecture. *Information portal* or *Enterprise Information Portal* (EIP) is recently introduced as the ideal system interface for corporations requiring a proper way to integrate many different types of information systems and provide an effective and efficient as well as highly customized access point to those heterogeneous systems for each employee [5, 7]. Using these portal systems, users can access all the needed information regardless of its source through a friendly system interface, a web browser, and the system provides well-customized information for each individual to help him or her make important business decisions in a timely manner. We adopt this system architecture for the proposed framework since expert search functionality provides part of deliverable knowledge from KMS and a portal architecture is ideal for building a web-based knowledge management system.

The intelligent expert search concept in this paper has been applied to "*Knowledge Portal for Researchers in Science and Technology*" (http://www.z4you.net) sponsored by *the Ministry of Science and Technology* (MOST) of Korea and it is under public service at this time. The automatic field classifier was developed using JAVA, and RAINBOW [11] developed by Carnegie Mellon University was chosen for the text categorization engine in Knowledge Portal. Figure 5 is a sample screen shot of Knowledge Portal providing results for the expert search functionality explained in this section.

To build the subject field hierarchy for FAH in Knowledge Portal, we use the classification system in scientific and technological fields, which was established by '*Korea Science Foundation* (KSF).' KSF classifies all the scientific and technological research fields, which are very huge and complex, using 3 different levels of classes: upper, middle, and lower. The KSF classification system has 4 upper classes such as 'Natural Science,' 'Life Science,' 'Engineering' and 'Multi-disciplinary' and these upper classes include a total of 69 different middle class level subject fields. Each middle class field has 7 to 8 lower classes on the average, making the total number of lower classes in the hierarchy 523. In Knowledge Portal, this subject field hierarchy is used for (1) registering interested subject fields by users, (2) classifying the knowledge artifacts registered on the knowledge base also by users, (3) assigning the expert fields for each expert through the automatic profiling process, and (4) applying FAH operations in the intelligent expert search process while interacting with system users who perform the expert searches.

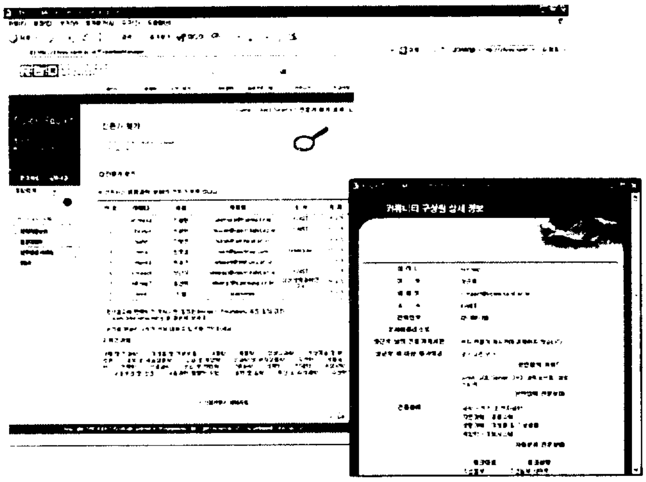

Figure 5 - Sample Screenshot for Expert Search in Knowledge Portal

For the initial training of the automatic field classifier, we used 1400 research project proposal documents, which were also provided by KSF. Since each of these documents had a pre-defined subject field, which was the field the research proposal dealt with and one of the subject field defined in KSF field hierarchy at the same time, we decided that those proposal documents were great sources for training and testing the performance of the automatic expert classifier we designed and implemented.

We used about 60% of the documents for training and the remaining 580 documents for validating the classification correctness of the classifier. Due to the limited number of documents, we just measured the classification correctness up to the middle class level. Out of 580, 37.24% of the documents were assigned correct subject fields. We concluded that this somewhat low rate of classification correctness was due to the fact that there were not enough training documents. That is even though we only used up to the middle class level for classification, many fields had less than 5 documents for training, which was statistically not enough to train the classification model correctly. When we considered only the top 10 subject fields, which were sorted decreasingly based on the number of documents belonging to each field, 73.25% of the documents were classified correctly. We expect to incorporate the lower level classes of subject fields in the training process while maintaining the satisfactory level of classification correctness if it is possible to collect enough training documents for each field.

Summary and Conclusion

In this paper, we propose the intelligent expert search framework for KMS. By adopting the proposed framework, KMS can be equipped with intelligent expert search capabilities which provide expert search results in similar subject fields. We improved the cooperative query answering methodology called FAH through adopting a text categorization technique to calculate initial similarity measures among siblings in the hierarchy, which were needed to be assigned by a domain expert before.

The intelligent expert search concept proposed in this paper has been applied to "*Knowledge Portal for Researchers in Science and Technology*," which is under public service. For the further research, we are trying to expand the intelligent expert search framework to incorporate multiple subject hierarchies and algorithmic research efforts are being devoted to improve the efficiency of the field classifier.

Acknowledgments

This work was supported in part by the Ministry of Information & Communications, Korea, under the Information Technology Research Center (ITRC) Support Program.

References

[1] Augier, M. and Vendelo, M. T. (1999). "Networks, Cognition and Management of Tacit Knowledge," *Journal of Knowledge Management*, Vol. 3, No. 4, pp. 252-261.

[2] Baeza-Yates, R., and Riberio-Neto, B. (1999). *Modern Information Retrieval.* New York: ACM Press.

[3] Braga, J. L., Laender, A. H. F., and Ramos, C. V. (2000). "A Knowledge-Based Approach to Cooperative Relational Database Querying," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 14, pp. 73-90.

[4] Cai, Y., Cercone, N., and Han, J. (1993). *Attribute-Oriented Induction in Relational Databases*, in *Knowledge Discovery in Databases*, Cambridge, MA: AAAI/MIT Press.

[5] Deltor, B. (2000). "The Corporate Portal as Information Infrastructure: Toward a Framework for Portal Design," *International Journal of Information Management*, Vol. 20, pp. 91-101.

[6] Desouza, K. C. (2003). "Barriers to Effective Use of Knowledge Management Systems in Software Engineering," *Communications of the ACM*, Vol. 46, No. 1, pp 99 - 101.

[7] Dias, C. (2001). "Corporate Portals: a Literature Review of a New Concept in Information Management," *International Journal of Information Management*, Vol. 21, pp. 269-287.

[8] Fowler, M. and Scott, K. (2000). *UML Distilled 2nd Edition: A Brief Guide to the Standard Object Modeling Language*. Reading: Addison Wesley.

[9] Huh, S. Y., and Moon, K. H. (2000). "A Data Abstraction Approach for Query Relaxation," *Information and Software Technology*, Vol. 42, pp. 407-418.

[10]Kreiner, K. (2002). "Tacit Knowledge Management:

The Role of Artifacts," *Journal of Knowledge Management*, Vol. 6, No. 2, pp. 112-123.

[11]McCallum, A. K. (1996). Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, *http://www.cs.cmu.edu/~mccallum/bow.*

[12]Moon, K. H., and Huh, S. Y. (2002). "An Integrated Query Relaxation Approach Adopting Data Abstraction and Fuzzy Relation," *Submitted for publication to Information Systems Research.*

[13]Zadeh, L. (1973). "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *IEEE Transactions on Systems Management and Cybernetics*, Vol. SMC-3, No. 1, pp28-44.