International Journal of Fuzzy Logic and Intelligent Systems, vol. 5, no. 4, December 2005 pp. 353-359

Prediction of User's Preference by using Fuzzy Rule & RDB Inference: A Cosmetic Brand Selection

Jin Sung Kim

School of Business Administration, Jeonju University
Hyoja-Dong 3-1200, Wansan-Ku, Jeonju, Jeonbuk 560-759, South Korea
Tel: +82-63-220-2932, Fax: +82-63-220-2787, E-mail: kimjs@jj.ac.kr

Abstract

In this research, we propose a Unified Fuzzy rule-based knowledge Inference Systems (UFIS) to help the expert in cosmetic brand detection. Users' preferred cosmetic product detection is very important in the level of CRM. To this purpose, many corporations trying to develop an efficient data mining tool. In this study, we develop a prototype fuzzy rule detection and inference system. The framework used in this development is mainly based on two different mechanisms such as fuzzy rule extraction and RDB (Relational DB)-based fuzzy rule inference. First, fuzzy clustering and fuzzy rule extraction deal with the presence of the knowledge in data base and its value is presented with a value between 0 ~1. Second, RDB and SQL (Structured Query Language)-based fuzzy rule inference mechanism provide more flexibility in knowledge management than conventional non-fuzzy value-based KMS (Knowledge Management Systems).

Key words : Cosmetic, Data mining, Expert systems, Fuzzy clustering, Fuzzy rule, Knowledge management, RDB, SQL.

1. Introduction

Internet firms offer products, services, message boards, reference tools, search engines, and many other specialized customer values and can be an entry point to other sites in the Internet [2][21]. Recently, many of the firms are interested in using the *web mining* techniques which refer to the use of *data mining (DM)* techniques to improve the customer value. It helps the forms to automatically retrieve, extract and evaluate (generalize/analyze) information for knowledge discovery from web documents, services and their customers [5][23]. However, early research in DM field concentrated on *Boolean association rules*, which are concerned only with whether an item is present in a transaction or not, without considering its quantity [3][4]. In addition, traditional DM technologies originally have some type of uncertainty, for instance, when the boundaries of a class of objects are not sharply defined [10][16][31][29].

The most common, useful and widely accepted solution for this problem is the introduction of fuzzy sets [7][9][11][13]. Because of the fuzzy sets provide mathematical meanings to the natural language statements and become an effective solution for dealing with uncertainty [30].

Fuzzy rule-based knowledge management and/or inference models are often used to model systems in an input/output sense by means of IF-THEN rules. It is desirable that the rule base covers all the situations of the system that are of importance for appropriate decision making. In that case, the number of rules should be kept low to increase the generalizing ability of the system, and to ensure a compact and transparent model [28]. In some cases, to gain a compact and transparent model and to overcome these limitations, fuzzy classification mechanism is used [15].

Nevertheless, the limitation comes from the size of knowledge base (or rule base) is still remained as a tackling point of the development of knowledge management systems (KMS) and ES. To resolve this problem, Veryha [29] suggested a framework for implementing fuzzy classification in information systems using conventional SQL querying. The first main contribution of Veryha [29] is that the fuzzy classification and use of conventional DB-based SQL queries which provide easy-to-use functionality for data extraction. Second, the approach proposed a new mechanism can be used as and effective DM tool in large information systems and easily integrated with conventional relational databases (RDB). Third, the approach has several benefits including RDB-based flexible data combination/analysis and improvement of information presentation at the report generation phase because it is based on the RDB who presenting the relationships among knowledge sets.

To improve the effectiveness of DM, with these advantages, we suppose an UFIS (Unified Fuzzy rule-based knowledge Inference Systems) based on fuzzy rule extraction and RDB-based fuzzy rule inference. Figure 1 shows the structure and components of prototype UFIS.

Manuscript received Nov. 2, 2005; revised Dec. 14, 2005

2. Related Works

2.1 Fuzzy *c*-means clustering

Due to the relevance of fuzzy *c*-means in OR (Operations Research), many researchers have widely used it in a wide range of applications. Recently, in the field of *Internet business*, it seems that Internet portals can also benefit from this method. Indeed, one of the key decisions that Internet portals need to address is what to offer to their potential users. The type of information provided through an Internet portal depends on the interests of its potential users. However, users/customers differ with respect to their interests. For example, some people can use the same service to find different jobs with their own interest [17][25]. In this case, the probability of mismatching will grow continually. To prevent the unexpected situations, there is need to develop a new mechanism to improve the system flexibility.

One of the mechanisms, we recommend the fuzzy *c*-means clustering mechanism, which can minimize the sum of squared errors [8][25]. The goal function of fuzzy *c*-means is as follows:

$$\text{Min: } \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(\|\underline{x}_k - \underline{v}_i\|^2) \tag{1}$$

Where n=number of individuals to be clustered, c=number of clusters, u_{ik}=degree of membership of individual k in cluster i, $\underline{x}_k$=a vector of h characteristics for individual k, v_i=a vector of the cluster means of the h characteristics for cluster i, and m=the weighing exponent.

Equation (1) represents the sum of squared errors and is a goal function that the fuzzy *c*-means algorithm tries to minimize. The values of c (number of clusters) and m (number of individuals to be clustered) are empirically determined. The constraints for goal function are as follows:

$$0 \le u_{ik} \le 1, \quad \forall i,k \tag{2}$$

$$\sum_{i=1}^{c}(u_{ik}) = 1, \quad \forall k \tag{3}$$

Constraint (2) ensures that the degrees of memberships are between 0 and 1. Constraint (3) means that, for a given individual, the degrees of the membership across the clusters sum to one. The cluster means are given by

$$\underline{v}_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \underline{x}_k}{\sum_{k=1}^{n}(u_{ik})^m}, \quad \forall i \tag{4}$$

and the degrees of membership are given by

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|\underline{x}_k - \underline{v}_i\|}{\|\underline{x}_k - \underline{v}_j\|}\right)^{2/(m-1)}}$$

$$\text{for } x_i \neq v_j;\ \forall i,k;\ \textit{and}\ m > 1 \tag{5}$$

Equations (4) and (5) are the necessary conditions for obtaining the minimum of the sum-of-square criterion function (equation (1)). Solution is obtained by iteration through these conditions. An iterative algorithm, also called "alternating optimization," is used to solve these equations and to identify clusters and associated cluster memberships. It starts with an initial solution for U_0 (eq. (5)) and loops through a cycle of estimates for U_{t-1} (eq. (5)) → V_t (eq. (4)) → U_t (eq. (5)). The iteration stops when the difference between Ut and Ut-1 is very small [8].

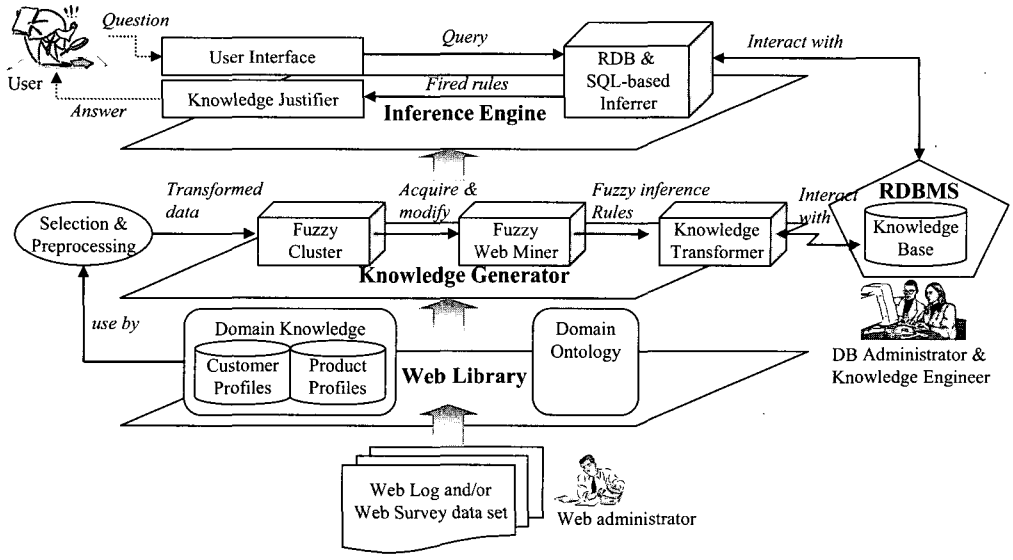


Figure 1 Prototype systems of UFIS

2.2 RDB and SQL-based fuzzy classification

A number of different schemes and tools to implement the fuzzy sets in DBMS (database management systems) have been proposed in recent years, such as fuzzy querying [7][1], fuzzy extension of SQL [15][16][11][31] and fuzzy object oriented database schemes [10][13].

One of them, Veryha [29] proposed a general framework for fuzzy classification in information systems using conventional SQL querying. To confirm usefulness of the framework he developed a prototype based on the stored procedures and DB extensions of MS-SQL Server 2000. In his mechanism, the goal of SQL querying of data with fuzzy classification is to provide DB views and/or reports of fuzzy classified data. To implement the functionality of SQL for fuzzy classified data querying in RDB, Veryha [29] developed an interpreter as stored procedure that will translate conventional SQL commands into native SQL queries of a particular DB (e.g. RDB). The scheme of fuzzy classification framework implementation in RDB is shown in Figure 2.

The framework executes the following steps to classify the data set.

Step-1: *Design of DB tables or views* to query them later using SQL for fuzzy classified data. This step has to be carried out by DB owners.

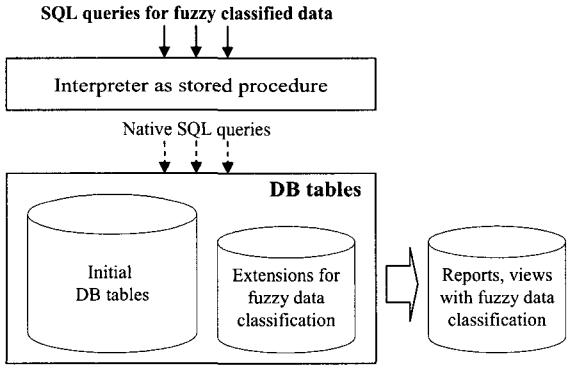


Figure 2 A general scheme of fuzzy classification framework implementation in RDB

Step-2: *Design of DB extensions* (additional tables that contain linguistic variables, membership values and descriptions of atomic values). This step should be carried out by an expert in the given application area. DB extensions can be generated automatically (additional programming may be required in this case).

Step-3: *Design and implementation of interpreter* for SQL transformation into native SQL for the given RDBMS using lexical and syntactical analysis of queries. This step should be carried out by software developer will develop an interpreter in the form of the stored procedure for the given DBMS.

Step-4: *Generation of DB reports and views* using SQL querying of fuzzy classified data formed on Steps 1 and 2 [29].

2.3 The ID3 and the C4.5 algorithms

When we use the ID3 and C4.5 methods, tests are basically based on single attribute selection, so the possible tests are related to the possible attributes.

To extract the relevant patterns/rules, test selection examines the training examples and finds the attribute that separates the examples most perfectly considering their class (conclusion) membership. The ID3 algorithm uses a function from the filed of information theory, the entropy, to measure how separated the elements are in the original training set and in the subsets after partitioning. Formally, the criterion for ID3 algorithm is the following:

$$\frac{\max\{\text{Gain}(A)\}}{A} \tag{6}$$

Where

$$\text{Gain(A)} = \text{I(T)} - \text{E(A,T)} \tag{7}$$

gives the improvement in the entropy, the gain,

$$I(T) = -\sum_{j=1}^{k} \frac{|C_j|}{|T|} \cdot \log_2\left(\frac{|C_j|_T}{|T|}\right) \tag{8}$$

is the entropy function of ID3. $|C_j|_T$ denotes the number of elements of T (training example) belong to class C_j and

$$E(A,T) = \sum_{i=1}^{m} \frac{|T_i|}{|T|} \cdot I(T_i) \tag{9}$$

is the weighted sum of the entropies in the subsets.

We should note that the application of *entropy function* and the *Gain criterion*, which is just the simple difference of *I(T)* and *E(A,T)*, has no strong theoretical background, and they are chosen subjectively as one of many feasible solutions. However, ID3 method has a critical limitation. The *Gain criterion* is biased towards discrete attributes with more outcomes. Therefore, the ID3 method is applicable only in problems described by a set of discrete attributes. To overcome the limitation, the *Gain ration* test selection criterion was proposed in C4.5 method, which decreases the Gain in case of many-valued discrete attributes [26][27]. To handle continuous attributes, in C4.5 method, the attribute values are discretized, treated as discrete ones with two outcomes in the following way: different values of the candidate continuous attributes are ordered: $v_1 \le v_2 \le \ldots \le v_n$ and all

$$m_i = \frac{v_i + v_{i+1}}{2}, \quad i = 1,\ldots,n-1 \tag{10}$$

midpoints (m_i) are checked as possible thresholds to divide the training set into two partitions. This test selection criterion in C4.5, however, is also biased towards continuous attributes

with numerous distinct values. Its modified criterion are presented in Quinlan's [27] research. Then, the C5.0 was presented as s commercial version of C4.5.

3. Research Methodology

Our proposed UFIS (*Unified Fuzzy rule-based knowledge Inference Systems*) mainly consists of three main modules *Web Library, Knowledge Generator,* and *Inference Engine.*

Web Library:

Web library contains reusable domain knowledge including domain ontology, product profiles and customer profiles. Web administrators will transfer the data which are summarized and transformed raw data into web library.

Knowledge Generator:

Main functions of knowledge generator are selection & preprocessing of data, fuzzy clustering, fuzzy web mining, knowledge transformation, and interaction with RDBMS to manage the knowledge base efficiently. Especially, as a fuzzy web miner, we will use the C5.0 machine learning (rule extraction) algorithm based on artificial intelligence. It can extract an executable knowledge set, which corresponds to the transformed data generated above it. After the extraction of knowledge it interacts with RDMBS to restore and revise her knowledge bases.

Inference Engine:

Inference engine contains UI (User Interface), SQL-based inferrer, and Knowledge Justifier. Most of conventional ES have text-oriented inference algorithm. However, in this study, UFIS use the RDB-based SQL inference engine. The main benefit of this inference engine is that there is no need to retransformation of text knowledge into a form of executable or inferable knowledge base.

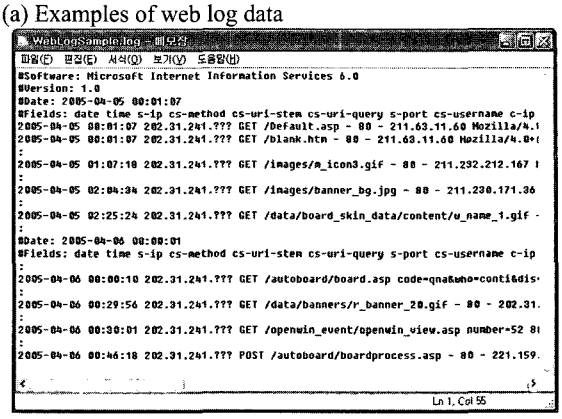
4. Implementation

4.1 Experimental data

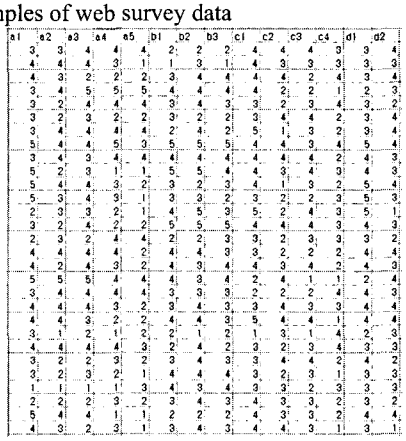
Table 1 show the raw data used in this experiment which contains web log information, customers' profile, and web survey data expressing customers' purchasing behavior on cosmetic-related web site. Using the web resources, UFIS start to clustering and extracting the fuzzy rules.

Table 1 Examples of web log & web survey data

(a) Examples of web log data

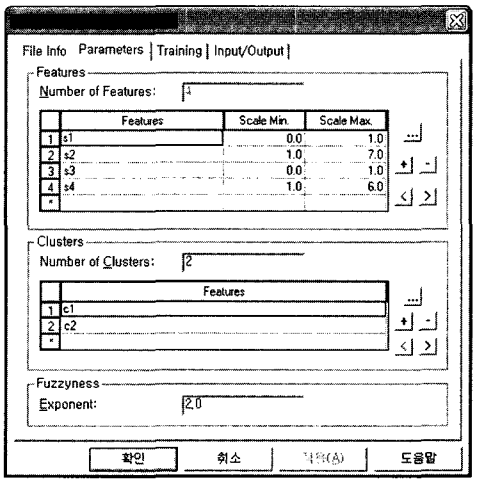


(b) Examples of web survey data



4.2 Fuzzy clustering

In this phase, we used FCM (Fuzzy C-Means) as a fuzzy clustering mechanism. Figure 3 shows the process and results of fuzzy clustering.



s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s15	C1	C2
1	3	0	2	2	4	2	5	2	1	1	0.2	0.8
1	2	0	1	5	1	1	4	10	2	0	0.0	1.0
1	4	1	4	5	3	1	5	11	2	1	1.0	0.0
1	2	0	1	30	3	2	3	5	3	1	0.0	1.0
1	2	0	1	20	2	1	3	13	1	0	0.0	1.0
1	3	0	1	10	2	1	4	10	1	1	0.1	0.9
1	2	0	1	1	3	1	4	10	2	0	0.0	1.0

Figure 3 Result of FCM (C1: Class#1, C2: Class #2)

4.3 Fuzzy membership function

Traditional fuzzy membership values computed by fuzzy membership functions were divided into three categories, such as *numeric value, linguistic value, and hybrid (combination of numeric and linguistic) value*. In this study, the theory of fuzzy sets provides a mechanism for representing linguistic constructs such as '*Low*', '*Medium*', and '*High*'. Then, each linguistic construct was induced by the bell-shaped numeric fuzzy membership function π [24]. The fuzzy membership function π, lying in the range [0, 1], with F_j was defined as follows:

$$\pi(F_j;c,\lambda)=\begin{cases} 2\left(1-\dfrac{|F_j-c|}{\lambda}\right)^2, & \text{for } \dfrac{\lambda}{2}\le|F_j-c|\le\lambda \\ 1-2\left(\dfrac{|F_j-c|}{\lambda}\right)^2, & \text{for } 0\le|F_j-c|\le\dfrac{\lambda}{2} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Where, $\lambda > 0$ is the radius of the π-function with c as the central point at which $\pi(c; c, \lambda) = 1$. Each factors and their values used to complete the fuzzy membership functions are shown in Table 2.

Table 2 Value of factors used in fuzzy membership functions (F_1: function #1, F_2: function #2)

Q	L	M	H	Q	L	M	H
Center (c) or max pref.	0	50	100	Center (c) or max pref.	1	3	5
Min	0	15	50	Min	1	1.5	3
Max	50	85	100	Max	3	4.5	5
Lambda or width (λ)	47.1	35.0	47.1	Lambda or width (λ)	1.9	1.5	1.9
$\lambda/2$	23.5	17.5	23.5	$\lambda/2$	0.9	0.8	0.9

(Q: Quantity, L: Low, M: Medium, H: High)

4.4 Fuzzy Rule Extraction

In this phase, we used C5.0 which is one of well-known ML algorithms. Table 3 shows the result of rule extraction by using ML. The rules have a form as follows:

Rule ***number predicted-value (Instance, Confidence)***
IF ***antecedent_1***
AND ***antecedent_2***
...
AND ***antecedent_n***
THEN ***predicted value***

Where, *Instance* means the number of records which contain the *antecedents* presented by the rule. The *Confidence* means the probability (%) and is computed as follows:

(1+*number of records where rule is correct*) / (2+*number of records for which the rule's antecedents are true*)

The *predicted-value* means the specific cosmetic product. In this study, we omitted the detailed name of which products.

Table 3 Example of fuzzy inference rules

Rule 1 CI (8, 0.20)
IF L6 = Low THEN CI

Rule 2 DF (2, 0.50)
IF L2 = Low AND L5 = Low AND L6 = Medium THEN DF

Rule 3 HR (1, 0.67)
IF L3 = Low AND L4 = High AND L5 = Medium AND L6 = Medium THEN HR

:

Rule 7 KR (5, 0.43)
IF L3 = High AND L4 = High AND L5 = Low AND L6 = Medium THEN KR

Rule 8 LG (2, 0.50)
IF L1 = High AND L4 = Low AND L6 = Medium THEN LG

(* L1: Good Advertisement, L2: Brand Image, L3: Good Design, L4: Skin Fitness, L5: Preference for Low-Price, L6: Fashion)

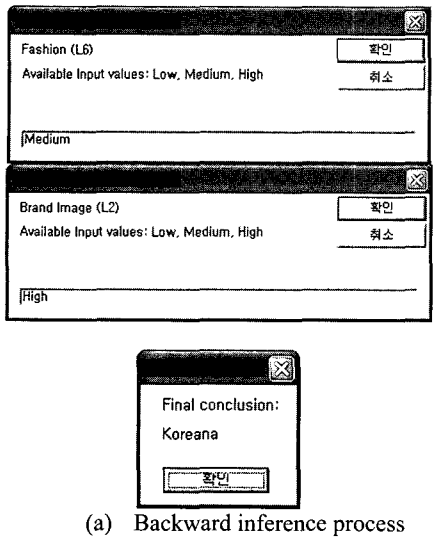
Using these fuzzy rules we examined our experimental data. As a result, which concerned to the CRM, we could find the *sustainability* and *changeability* of customers. The *changeability* means the probability of changing product from specific firm's product to another (competitive) firm's product. In contrast with *changeability*, *sustainability* means the capability of being maintained as a specific product. Table 4 shows the result of experiments.

Table 4 Result of inference
(Sustainability .vs. Changeability)

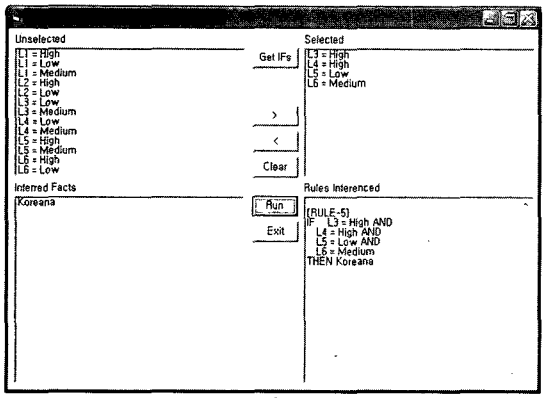
Brand (products)	Sustainability (%)		Changeability (%)	
CI	100.0		0.0	
DF	100.0		0.0	
HR	41.7		58.3	
KR	70.0	25.0	30.0	75.0
LG	50.0		50.0	
MI	66.7		33.3	
SL	33.3		66.7	

4.5 Inference in RDB

Figure 4 shows the backward inference (Figure 4(a)) and forward inference (Figure 4(b)) simultaneously by using UFIS.



(a) Backward inference process



(b) Forward inference process

Figure 4 Inference by using RDB

In backward inference, the system gives simple queries with available input (selectable) values to end users to find relevant inference rules. In comparison with backward inference, in forward inference process, the system shows all selectable (unselected) input values to users as shown in Figure 4(b). Then upper right side window in Figure 4(b) shows the selected input values. With these input values the system find all matched inference rules and shows final inference result.

5. Conclusion

In this study, we proposed unified fuzzy rule-based knowledge inference systems UFIS based on fuzzy clustering, machine learning inference rule, RDB, and SQL. The fuzzy classification and use of conventional SQL queries-based inference provide ease-to-use functionality for knowledge extraction and inference in ES. For the implementation of UFIS the prototype based on Microsoft Visual Basic and MS-Access was developed. After the implementation and experiment with UFIS we found that the framework was effective to find the hidden knowledge from web DB and inference by using fuzzy rules, RDB and SQL. Nevertheless, elaborate design of RDB & SQL-based inference engine and simplified process for knowledge base expansion are remained as further research topics.

References

[1] Abdennadher, S. and Schuetz, H. (1998), Flexible query language, *Proceedings of Flexible Query Answering System Conference*, Roskilde, Denmark, 1-14.

[2] Afuah, A. and Tucci, C.L. (2001), *Internet business models and strategies: Text and cases*, McGraw-Hill, Inc.

[3] Agrawal, R. and Srikant, R. (1994), Fast algorithms for mining association rules, *Proceedings of the International Conference on Very Large Databases*, 487-499.

[4] Agrawal, R., Imielinski, T., & Swami, A. (1993), Mining association rules between sets of items in large databases, *Proceedings of ACM SIGMOD International Conference on Management of Data*, 207-216.

[5] Arotaritei, D. and Mitra, S. (2004), Web mining: A survey in the fuzzy framework, *Fuzzy Sets and Systems*, 148, 5-19

[6] Bellma, M. and Vojdani, N. (2000), Fuzzy prototypes for fuzzy data mining, *Studies in Fuzziness and Soft Computing*, 39, 175-286.

[7] Bezdek, J.B., Keller, J., Krinapuram, R., & Pal, N.R. (1999) *Fuzzy models and algorithms for pattern recognition and image processing*, Kluwer Academic Publishers, Boston, MA.

[8] Blanco, I., Cubero, J., Pons, C., & Vila, A. (2000), An implementation for fuzzy deductive relational databases, *Studies in Fuzziness and Soft Computing*, 53, 183-208.

[9] Borgodna, G., Leporati, A., Lucarella, D., & Pasi, G. (2000), The fuzzy object-oriented database management system, *Studies in Fuzziness and Soft Computing*, 53, 209-236.

[10] Bosc, P. and Pivert, O. (2000), SQLf query functionality on top of a regular relational database management system, *Studies in Fuzziness and Soft Computing*, 39, 171-191.

[11] Dubois, D., Nakata, M., & Prade, H. (2000), Extended divisions for flexible queries in relational databases, *Studies in Fuzziness and Soft Computing*, 39, 105-121.

[12] Kacprzyk, J. and Zadrozny, S. (2000a), Data mining via fuzzy querying over the internet, *Studies in Fuzziness and Soft Computing*, 39, 211-233.

[13] Kacprzyk, J. and Zadrozny, S. (2000b), On combining intelligent querying and data mining using fuzzy logic concepts, *Studies in Fuzziness and Soft Computing*, 53, 67-84.

[14] Kaufmann, A. (1986), On the relevance of fuzzy sets for operations research, *European Journal of Operational Research*, 25(3), 330-335.

[15] Lake, M. (1998), *The new megasites: All-in-one Web superstores*, PCWorld.com.

[16] Martin-Bautista, M.J., Sanchez, D., Chamorro-Martinez, J., Serrano, J.M., & Vila, M.A. (2004), Mining web documents to find additional query terms using fuzzy association rules, *Fuzzy Sets and Systems*, 148, 85-104.

[17] Mitra, S. & Pal, S.K. (1994), Logical Operation Based Fuzzy MLP for Classification and Rule Generation, *Neural Networks*, 7(2), 353-373.

[18] Ozer, M. (2005), Fuzzy *c*-means clustering and Internet portals: A case study, *European Journal of Operational Research*, 164, 696-714.

[19] Quinlan, J. (1993), *C4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA.

[20] Quinlan, J. (1996), Improved use of continuous attributes in C4.5, *Journal of Artificial Intelligence Research*, 4, 77-90.

[21] Setnes, M. (2000), Supervised fuzzy clustering for rule extraction, *IEEE Transactions on Fuzzy Systems*, 8(4), 416-424.

[22] Veryha, Y. (2005), Implementation of fuzzy classification in relational databases using conventional SQL querying, *Information and Software Technology*, [online] available www.sicencedirect.com.

[23] Zadeh, L. (1989), Knowledge representation in fuzzy logic *IEEE Transactions on Knowledge and Data Engineering*, 1, 89-100.

[24] Zadrozny, S. and Kacprzyk, J. (1998), Implementing fuzzy querying via the internet/www: java applets active X controls and cookies, *Proceedings of Flexible Query Answering System Conference*, Roskilde, Denmark, 382-392.

Jin Sung Kim

He has been assistant professor of MIS at the School of Business Administration, Jeonju University, South Korea. His current research interests are in fuzzy logic and AI-based intelligent decision support systems, neural networks, e-business, and Web-based negotiation support systems.

TEL: +82-63-220-2932
FAX: +82-63-220-2787
E-mail: kimjs@jj.ac.kr