J.T. Yao, J. P. Herbert, Web-based Support Systems with Rough Set Analysis,
Proceedings of International Conference on Rough Sets and Emerging Intelligent System
Paradigms (RSEISP'07), June 28 -30, 2007, Warsaw, Poland, LNAI 4585, pp 360-370.

Web-based Support Systems with Rough Set Analysis

JingTao Yao   Joseph P. Herbert

Department of Computer Science
University of Regina, Regina
Canada S4S 0A2
[jtyao,herbertj]@cs.uregina.ca

Abstract. Rough sets have been applied to many areas where multi-attribute data is needed to be analyzed to acquire knowledge for decision making. Web-based Support Systems (WSS) are a new research area that aims to support human activities and extend human physical limitations of information processing with Web technologies. The applications of rough set analysis for WSS is looked at in this article. In particular, our focus will be on Web-Based Medical Support Systems (WMSS). A WMSS is a support system that integrates medicine practices (diagnosis and surveillance) with computer science and Web technologies. We will explore some of the challenges of using rough sets in a WMSS and detail some of the applications of rough sets in analyzing medical data.

1 Introduction

Web-based Support Systems (WSS) are a completely new frontier for computerized support systems [16]. It can be understood as extensions of existing research in two dimensions. It can also be viewed as natural extensions of decision support systems with the use of the Web to support more activities. In the technology dimension, WSS use the Web as a new platform for the delivery of support with new advances in technology can lead to further innovations in support systems. Along the application dimension, the lessons and experiences from DSS can be easily applied to other domains.

Research on information retrieval support systems [17], research support systems [4, 13], decision support systems [5, 11], and medical support systems [1, 12] are just some of the recent investigations for moving support systems to the Web platform [14, 15].

Rough set theory is a way of representing and reasoning imprecision and uncertain information in data [9]. It deals with the approximation of sets constructed from descriptive data elements. This is most helpful when trying to discover decision rules, important features, and minimization of conditional attributes. The beauty of rough sets is how it creates three regions, namely, the positive, negative and boundary regions. The boundary regions are useful for a undeterminable cases.

Researchers have used rough sets for diagnosing cancer [8], brain disorders [3], lung disease [7], and others. These applications of rough sets to data analysis may be included in a Web-based Medical Support System (WMSS).

This paper will focus on the issues of migrating the rough set model for use in Web-based support systems. The organization of this paper is as follows. Section 2 will discuss rough set theory and an extended probabilistic model that incorporates risk. Section 3 will provide WSS applications with rough sets and introduce Web-based medical support systems with rough set functionality. Finally, we conclude this paper in Section 4.

2 Rough Set Models

2.1 Algebraic Rough Set Model

Approximation is used to characterize a set $A \subseteq U$ [9], where U is a finite, non-empty universe. It may be impossible to precisely describe A given a set relation B. Equivalence classes are simply objects in U in which we have information. Definitions of lower and upper approximations follow:

$$\begin{aligned} \underline{apr}(A) &= \{x \in U | [x] \subseteq A\}, \\ \overline{apr}(A) &= \{x \in U | [x] \cap A \neq \emptyset\}. \end{aligned} \tag{1}$$

The lower approximation of a set A, denoted $\underline{apr}(A)$, is the union of all elementary sets that are included (fully contained) in X. The upper approximation of a set A, denoted $\overline{apr}(A)$, is the union of all elementary sets that have a non-empty intersection with A. This allows us to approximate unknown sets with known objects. We can now define notions of positive, negative, and boundary regions [9] of A:

$$\begin{aligned} POS(A) &= \underline{apr}(A), \\ NEG(A) &= U - \overline{apr}(A), \\ BND(A) &= \overline{apr}(A) - \underline{apr}(A). \end{aligned} \tag{2}$$

2.2 Probabilistic Rough Set Model

The algebraic method has very little flexibility for determining the classification regions. It may not be useful or applicable when majority cases are undeterminable. More flexible models include some probabilistic approaches, namely, variable precision rough sets [20] and decision-theoretic rough sets [18, 19].

The decision-theoretic approach may lend itself to a more Web-friendly application for two reasons. First, it calculates approximation parameters by obtaining easily understandable notions of risk or loss from the user [19]. This allows for simpler user involvement instead of having parameters being arbitrarily provided. This is important when users are not qualified to set the parameters and just wish to perform analysis. Second, many types of WSS could make use of cost or risk annotations. We present a slightly reformulated decision-theoretic rough set model in this section, as reported in [18, 19].

The Bayesian decision procedure allows for minimum risk decision making based on observed evidence. Let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be a finite set of m possible actions and let $\Omega = \{w_1, \ldots, w_s\}$ be a finite set of s states. Let $P(w_j|\mathbf{x})$ be the conditional probability of an object x being in state w_j given the object description $\mathbf{x}$. Let $\lambda(a_i|w_j)$ denote the loss, or cost, for performing action a_i when the state is w_j.

Object classification with approximation operators can be fitted into this framework. The set of actions is given by $\mathcal{A} = \{a_P, a_N, a_B\}$, where a_P, a_N, and a_B represent the three actions to classify an object into $POS(A)$, $NEG(A)$, and $BND(A)$ respectively. Let $\lambda(a_\diamond|A)$ denote the loss incurred for taking action $a_\diamond$ when an object belongs to A, and let $\lambda(a_\diamond|A^c)$ denote the loss incurred by taking the same action when the object belongs to A^c. This can be given as loss functions $\lambda_{\diamond 1} = \lambda(a_\diamond|A)$, $\lambda_{\diamond 2} = \lambda(a_\diamond|A^c)$, and $\diamond = P$, N, or B.

If we consider the loss function inequalities $\lambda_{P1} \leq \lambda_{B1} < \lambda_{N1}$,that is, the loss incurred by λ_{N1} (false-negative) is more than the losses incurred by both a correct classification (λ_{P1}) and an indeterminant classification (λ_{B1}) we can formulate decision rules based on this division of the universe. The corresponding inequalities $\lambda_{N2} \leq \lambda_{B2} < \lambda_{P2}$, that is, a false-positive (λ_{P2}) has a greater cost than a correct classification (λ_{N2}) and an indeterminant classification (λ_{B2}), can further tell us how the universe is divided. We can formulate the following decision rules (P)-(B) [18] based on the set of inequalities above:

$$
\begin{array}{lll}
\text{(P)} & \text{If } P(A|[x]) \geq \gamma \text{ and } P(A|[x]) \geq \alpha, & \text{decide } \; POS(A), \\
\text{(N)} & \text{If } P(A|[x]) \leq \beta \text{ and } P(A|[x]) \leq \gamma, & \text{decide } \; NEG(A), \\
\text{(B)} & \text{If } \beta \leq P(A|[x]) \leq \alpha, & \text{decide } \; BND(A),
\end{array}
$$

where,

$$
\begin{aligned}
\alpha &= \frac{\lambda_{P2} - \lambda_{B2}}{(\lambda_{B1} - \lambda_{B2}) - (\lambda_{P1} - \lambda_{P2})}, \\
\gamma &= \frac{\lambda_{P2} - \lambda_{N2}}{(\lambda_{N1} - \lambda_{N2}) - (\lambda_{P1} - \lambda_{P2})}, \\
\beta &= \frac{\lambda_{B2} - \lambda_{N2}}{(\lambda_{N1} - \lambda_{N2}) - (\lambda_{B1} - \lambda_{B2})}.
\end{aligned} \tag{3}
$$

The α, β, and γ parameters define our regions, giving us an associated risk for classifying an object. The α parameter can be considered the division point between the POS region and BND region. Likewise, the β parameter is the division point between the BND region and the NEG region. When $\alpha > \beta$, we get $\alpha > \gamma > \beta$ and can simplify the rules (P-B) into (P1-B1):

$$
\begin{array}{lll}
\text{(P1)} & \text{If } P(A|[x]) \geq \alpha, & \text{decide } \; POS(A); \\
\text{(N1)} & \text{If } P(A|[x]) \leq \beta, & \text{decide } \; NEG(A); \\
\text{(B1)} & \text{If } \beta < P(A|[x]) < \alpha, & \text{decide } \; BND(A).
\end{array}
$$

When $\alpha = \beta = \gamma$, we can simplify the rules (P-B) into (P2-B2) [18]:

(P2)	If $P(A\|[x]) > \alpha$,	decide $POS(A)$;
(N2)	If $P(A\|[x]) < \alpha$,	decide $NEG(A)$;
(B2)	If $P(A\|[x]) = \alpha$,	decide $BND(A)$.

These minimum-risk decision rules offer us a basic foundation in which to build a rough set risk analysis component for a WSS. They give us the ability to not only collect decision rules from data, but also the calculated risk that is involved when discovering (or acting upon) those rules.

3 Web-based Support Systems with a Rough Set Component

For our future purposes of using rough sets for a WSS, we will look at a particular probabilistic approach that allows us to calculate associated risk for a partitioning of the object universe. The decision-theoretic rough set model [19] allows us to enhance the traditional data mining component of a WSS by adding a risk element to the decision process. Using this risk element, users of a WSS can make more informed decisions based on the rule-based knowledge base. Based on the three regions (POS, BND, and NEG), there are two types of decisions or support that the rough set component can offer the user:

1. **Immediate Decisions** (Unambiguous) - These types of decisions are based upon classification within the POS and NEG regions. The user can interpret the findings as:
   (a) Classification in the POS region is a definitive "yes" answer, for instance, the symptoms or test results indicate a patient suffers breast cancer.
   (b) Classification in the NEG region is a definitive "no" answer, for instance, the symptoms indicate that a patient does *n*ot suffer breast cancer.
2. **Delayed Decisions** (Ambiguous) - These types of decisions is based upon classification within the BND region. Since there is some element of uncertainty in this region, the user of the WSS should proceed with a "wait-and-see" agenda. Rough set theory may be meaningless when the "wait-and-see" cases are too large and unambiguous rules are scarce. Two approaches may be applied to decrease ambiguity:
   (a) Obtain more information [9]. More lab tests will be conducted in order to diagnose whether a patient suffers a disease, i.e., introduce more attributes of the information table. Conduct further studies to gain knowledge in order to make an immediate decision from the limited data sets.
   (b) A decreased tolerance for acceptable loss [18–20]. The probabilistic aspects of the rough set component allows the user to modify the (acceptable) loss functions in order to increase certainty. However, this may also increase the risk of "false-positives" and "false-negatives". For instance, a doctor may diagnose a patient with a lung infection with a

simple cough symptom and prescribe an antibiotic for treatment. The risk to both patient and doctor of the wrong diagnosis is relatively low compared to a conclusion of lung cancer and treated with chemotherapy. The decision-theoretic rough set model is adapted to consider the risk factor and calculate the tolerance level for a WSS.

These types of decisions could greatly influence the effectiveness of the knowledge base derived from the rough set component. The risk element provided by the decision-theoretic rough set model provides the user with the ability to customize the knowledge base to suit their priorities.

3.1 Binding Rough Sets with Web-based Support Systems

Both algebraic and probabilistic rough sets provide the user with methods to derive rules. These rules can then be used to support decision making. There are many types of WSS that support some form of decision making, including but not limited to Web-based decision support systems. Therefore, it follows that an important extension of rough sets should be the migration to the Web.

Looking at rough sets from a data mining perspective, it is one of many knowledge discovery methods that are available to the users. Given a depository of data, rough sets can be used to perform analysis of this data. The end result being a set of decision rules that can be used to describe, extend, or predict the domain in which the data was derived [10]. For example, a time-series data set describing stock index prices can be analyzed with rough sets in order to obtain decision rules that aid in forecasting the market [2].

The WSS framework utilizing rough sets would be connected to the components *knowledge base*, *database*, *interface*, as well as the other components [16]. This is shown in Fig. 1. Taking on the duties of the *data mining* component, rough sets would perform analysis on the data within the *database* component. It would derive decision rules based on this data. These rules would be captured by the *knowledge management* component, which would index it into the *domain specific knowledge base*.

Some derivations of WSS could make use of the cost or loss annotations provided by the decision-theoretic rough set models. This may include Web-based decision support systems where a decision made in conjunction with a decision rule could have some perceived implications portrayed by the loss functions. These λ_{P2} and λ_{N1} errors, or "false positive" and "false negative" errors respectively can be provided to the decision maker so that he or she can be better informed on which decision to make. Fig. 1 can be modified so that the *domain specific knowledge base* contains information regarding the $\lambda_{\diamond 1}$ and $\lambda_{\diamond 2}$ values corresponding to the decision rules used by the user.

The use of the decision-theoretic rough set model in WSS distances itself from the uses of the traditional rough sets. Rough set analysis is transformed into decision-theoretic rough analysis. Rules that are normally formed through rough set analysis are transformed into a "risk analysis" pair (decision rules with their respective costs). The decision making performed with the traditional

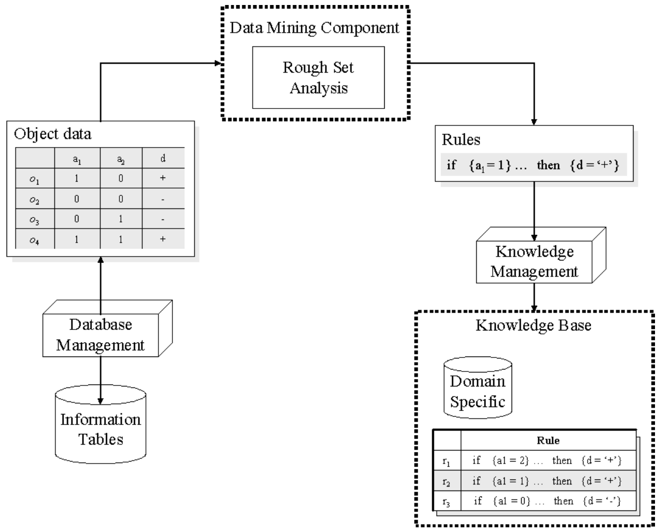

Fig. 1. Sub-architecture with Rough Set Analysis as a data mining component.

rule set can now be thought of a decision making with minimum cost tasks. For example, a set of rules governing the diagnosis of cancer would also have a set of risks that indicate the potential loss for a false positive or false negative.

3.2 Web-Based Medical Support Systems

In this section, we will describe a Web-based medical support system. A WMSS contains many components whose duties range from scheduling of appointments to maintaining a knowledge base of symptoms and diseases [12]. We focus on the decision support aspect [6] of a WMSS. This system will use rough sets to perform analysis on compatible medical data. A WMSS has a primary goal of supporting decisions of an expert (doctor, primary or secondary diagnostician).

For our purposes of using rough sets for a WMSS, we will look at a probabilistic approach that allows us to calculate associated risk for a partitioning of the object universe. The decision-theoretic rough set model allows us to enhance the traditional data mining component of a WSS by adding a risk element to the decision process. The architecture is shown in Fig 2. The individual components are described as follows:

Patient Database The patient database contains data pertaining to patient symptoms. This is gathered by the users of the system by a number of questions and trials performed on the patients. The rough set component and information retrieval component access this database regularly.

Database Management System The DBMS is a major component in any modern system. This is middleware that provides access to the patient database. The rough set component communicates with the DBMS for tuple data.

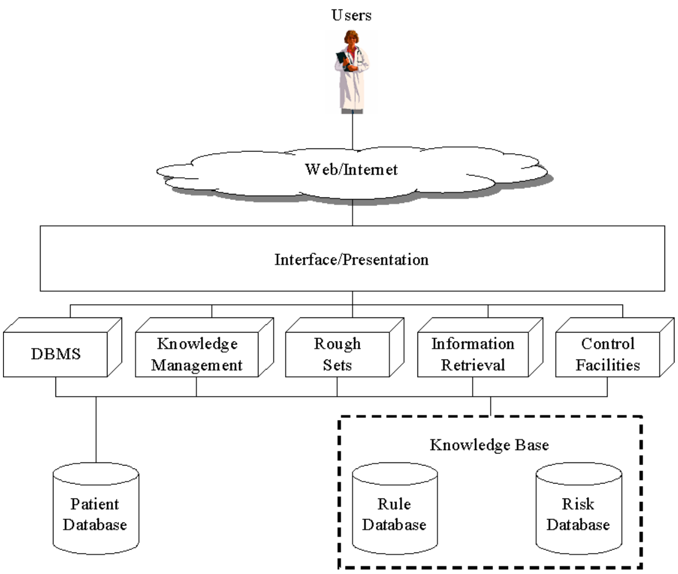

Fig. 2. A Partial Architecture of a Web-based Medical Support System.

Knowledge Management The knowledge management middleware component manages the knowledge base and provides access to the rule database and associated risk database. It acquires the risk analysis pairs from the rough set analysis component and indexes them accordingly.

Rough Set Component The rough set component in this particular system makes use of the decision-theoretic rough set model to acquire knowledge (rules) and the associated risks of using that knowledge. It provides the users of the system with timely information to support their decision making.

Information Retrieval The information retrieval component provides search and indexing functionality. Rough sets can play a role here [21]. This component is directly interfaced and has primary communication with the interface / presentation layer. Users of the system will access this component to retrieve patient data, information from knowledge base and other tasks.

Other Control Facilities Other control facilities include a robust security and permission component. Since patient data is very sensitive and with the Web functionality of the entire system, security is a major concern.

Knowledge Base The knowledge base component contains two major subcomponents: the rule database and associated risk database. The rule database is an index of the knowledge derived from the rough set analysis component. The associated risk database contains risk values for accepting a decision implied by the rule database.

Interface/Presentation This component is an entire layer of user interfaces and server-side form request handlers. This layer presents the users of the system with a clean and efficient web interface for entering patient data, searching, and obtaining decision support.
Users The users of the system include general practitioners, primary doctors, secondary diagnosticians, etc. The users access the WMSS via Interface / Presentation layer through the Internet.

The users may take the information provided and make an *unambiguous* decision. This represents a definitive "yes" or "no" diagnosis for a particular set of symptoms. This corresponds to those patients classified in either the POS or NEG regions. For those cases in the BND region, a "wait-and-see" decision is used. The support system would suggest that the users either decrease their tolerance (loss functions) or acquire additional data on the subject.

3.3 Web-based Medical Support Systems with Risk Analysis

To see how a decision-theoretic rough set analysis component effects decisions in a WMSS, let us consider two diagnosis scenarios. The risk or cost is defined as consequences of the wrong diagnosis. Based on our common sense, the cost of wrong diagnosis of a flu is lower than that of a wrong diagnosis of cancer. The cancer diagnosis tolerance levels of either a false-negative or false-positive are very low. Patients may sacrifice their lives when a false-negative level is high as they may miss the best treatment time. They may suffer consequences of chemotherapy for non-existent cancer when a false-positive is high.

Using Table 1, let us form two hypothetical scenarios of patient diagnoses. First, a diagnosis of low severity with a low cost for a wrong diagnosis. This could be testing for a patient's minor allergies. An allergy test would be looking for positive indicators for symptoms $S = \{S_1, S_2, S_3\}$ and the diagnosis decision $D = \{\text{Decision}\}$. Below is a typical sample of loss functions for this situation:

$$\lambda_{P2} = \lambda_{N1} = 1u, \qquad \lambda_{P1} = \lambda_{N2} = \lambda_{B1} = \lambda_{B2} = 0, \tag{4}$$

where u is a unit cost determined by the individual administration. In this scenario, the administration has deemed that a false-positive (λ_{P2}) and false-negative (λ_{N1}) diagnosis has some form of cost whereas indeterminant diagnoses (λ_{B1}, λ_{B2}) and correct diagnoses (λ_{P1}, λ_{N2}) have no cost.

A diagnosis of high severity could have a high cost for a wrong diagnosis. This could be testing for whether a patient has a form of cancer. In Table 1, the cancer test would be looking for positive indicators for symptoms $S = \{S_1, S_4, S_5\}$ and the diagnosis decision $D = \{\text{Decision}\}$. Patient o_3 having symptoms $s_{3,1}$, $s_{3,2}$, and $s_{3,3}$ would give a decision $d_{3,1}$ for allergy tests. Below is a typical sample of loss functions for this situation:

$$\lambda_{P2} = \lambda_{N1} = 2u, \qquad \lambda_{B1} = \lambda_{B2} = 1u, \qquad \lambda_{P1} = \lambda_{N2} = 0, \tag{5}$$

where u is a unit cost determined by the individual administration. In this scenario, the administration has deemed that a false-positive (λ_{P2}) and false-negative (λ_{N1}) diagnoses is twice as costly as indeterminant diagnoses (λ_{B1} and λ_{B2}). Correct diagnoses (λ_{P1} and λ_{N2}) have no cost.

Patient	S_1	S_2	S_3	S_4	S_5	Decision
o_1	$s_{1,1}$	$s_{1,2}$	$s_{1,3}$	$s_{1,4}$	$s_{1,5}$	d_1
o_2	$s_{2,1}$	$s_{2,2}$	$s_{2,3}$	$s_{2,4}$	$s_{2,5}$	d_2
o_3	$s_{3,1}$	$s_{3,2}$	$s_{3,3}$	$s_{3,4}$	$s_{3,5}$	d_3
o_4	$s_{4,1}$	$s_{4,2}$	$s_{4,3}$	$s_{4,4}$	$s_{4,5}$	d_4
o_5	$s_{5,1}$	$s_{5,2}$	$s_{5,3}$	$s_{5,4}$	$s_{5,5}$	d_5
o_6	$s_{6,1}$	$s_{6,2}$	$s_{6,3}$	$s_{6,4}$	$s_{6,5}$	d_6

Table 1. An Information Table

Using the loss functions in (4) and calculating the parameters using the formulas in (3), we obtain $\alpha = 1$, $\gamma = 0.5$, and $\beta = 0$. When $\alpha > \beta$, we get $\alpha > \gamma > \beta$. We use the the simplified decision rules (P1-B1) to obtain our lower and upper approximations:

$$\begin{aligned} \underline{apr}_{(1,0)}(A) &= \{x \in U | P(A|[x]) = 1\}, \\ \overline{apr}_{(1,0)}(A) &= \{x \in U | P(A|[x]) > 0\}. \end{aligned} \tag{6}$$

Using the loss functions in (5) and calculating the parameters using the formulas in (3), we obtain $\alpha = \beta = \gamma = 0.5$. When $\alpha = \beta = \gamma$, we use the simplified decision rules (P2-B2) to can obtain our new lower and upper approximations:

$$\begin{aligned} \underline{apr}_{(0.5,0.5)}(A) &= \{x \in U | P(A|[x]) > 0.5\}, \\ \overline{apr}_{(0.5,0.5)}(A) &= \{x \in U | P(A|[x]) \geq 0.5\}. \end{aligned} \tag{7}$$

The approximations in (6) mean that we can definitely class patient x into diagnosis class A if all similar patients are in diagnosis class A. The low loss functions (4) have indicated that users of the system can have high certainty when dealing with this class of patient. The approximations in (7) mean that we can definitely class patient x into diagnosis class A if strictly more than half of similar patients are in diagnosis class A. These examples use the loss functions to determine how high the level of certainty regarding a patient's symptoms needs to be in order to minimize cost.

4 Conclusion

We further explain the importance of Web-based support systems. A decision-theoretic rough set model can be used as the data mining component for a WSS. This extended model allows the component to provide additional decision support to the users. The two types of decision the users can make, immediate and delayed, are now fully supported by the rough set component. We reiterate this by detailing a Web-based medical support system framework that incorporates risk analysis through loss functions. The rough set component builds and maintains a risk database to assist the users in assessing the knowledge provided by the rule database.

References

1. Grzymala-Busse, J.W., Hippe, Z.S.: Data mining methods supporting diagnosis of melanoma. In: 18th IEEE Symposium on Computer-Based Medical Systems CBMS'05. (2005) 371–373
2. Herbert, J., Yao, J.T.: Time-series data analysis with rough sets. In: CIEF'04. (2005) 908–911
3. Hirano, S., Tsumoto, S.: Rough representation of a region of interest in medical images. International Journal of Approximate Reasoning **40** (2005) 23–34
4. Ishikawa, T., Klaisubun, P., Honma, M., Qian, Z.: Remarkables: A web-based research collaboration support system using social bookmarking tools. In: WSS'06, Proceedings of WI-IAT 2006 Workshops. (2006) 192–195
5. Khosla, R., Lai, C.: Mediating human decision making with emotional attitudes in web based decision support systems. In: WSS'06, Proceedings of WI-IAT 2006 Workshops. (2006) 204–207
6. Krause, P., Fox, J., O'Neil, M., Glowinski, A.: Can we formally specify a medical decision-support system? IEEE Expert-Intelligent Systems & Their Applications **8** (1993) 56–61
7. Kusiak, A., Kern, J.A., Kernstine, K.H., Tseng, B.T.L.: Autonomous decision-making: A data mining approach. IEEE Transactions on Information Technology In Biomedicine **4** (2000) 274–284
8. Mitra, P., Mitra, S., Pal, S.K.: Staging of cervical cancer with soft computing. IEEE Transactions on Biomedical Engineering **47** (2000) 934–940
9. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences **11** (1982) 341–356
10. Peters, J.F., Skowron, A.: A rough set approach to knowledge discovery. International Journal of Intelligent Systems **17** (2002) 109–112
11. Power, D.J., Kaparthi, S.: Building web-based decision support systems. Studies in Informatics and Control **11** (2002) 291–302
12. Stalidis, G., Prentza, A., Vlachos, I.N., Maglavera, S., Koutsouris, D.: Medical support system for continuation of care based on xml web technology. International Journal of Medical Informatics **64** (2001) 385–400
13. Yao, J.T.: Supporting research with weblogs: A study on web-based research support systems. In: WSS'06, Proceedings of WI-IAT 2006 Workshops. (2006) 161–164
14. Yao, J.T., Lingras, P., eds.: Proceedings of 2003 WI/IAT Workshop on Applications, Products and Services of Web-based Support Systems (WSS'03), Halifax, Canada (2003)
15. Yao, J.T., Raghvan, V.V., Wang, G.Y., eds.: Proceedings of the Second International Workshop on Web-based Support System (WSS'04), Beijing, China (2004)
16. Yao, J.T., Yao, Y.Y.: Web-based support systems. In: [14]. (2003) 1–5
17. Yao, Y.Y.: Information retrieval support systems. In: Proceedings of FUZZ-IEEE'02. (2002) 773–778
18. Yao, Y.Y.: Decision-theoretic rough set models. In: Proceedings of RSKT'07, LNAI. 4481 (2007) 1–12
19. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. International Journal of Man-machine Studies **37** (1992) 793–809
20. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences **46** (1993) 39–59
21. Ziarko, W., Fei, X.: VPRSM approach to web searching. In: RSCTC'02, LNAI. 2475 (2002) 514–521