ROUGH SET THEORY USING SIMILARITY OF OBJECTS DESCRIBED BY ONTOLOGY

Yusei Inukai, Andreas Gehrmann, Yoshimitsu Nagai and Syohei Ishizu
Aoyama Gakuin University, 5-10-1 Fuchinobe, Sagamihara, 229-8558, Japan

ABSTRACT

Rough set theory was proposed by Pawlak Z. and is well used in the area of data mining. The main role of rough set theory is to extract important sets of attributes and decision rules based on the knowledge about objects. Rough set theory is defined by information system which is a finite information table. Although recently the concept of ontology is used in knowledge engineering, Semantic Web, etc. Since ontology can flexibly represent knowledge of the universe of discourse, rough set theory using the concept of ontology enables us to use flexible information system in ontological description. One of our main aims of this paper is to propose rough set theory applied the concept of ontology to. For flexibility, ontology often consists of complex objects. We formulate a concept of similarity which measures a degree of relationship among complex objects. The concept of similarity is useful for extracting important sets of attributes defined by flexibly represented knowledge. We present steps for finding a set of decision rule based on the proposed concepts. We demonstrate the concepts and steps by using simple examples.

Keywords: Rough set theory, ontology, OWL, information system, degree of similarity

INTRODUCTION

Rough set theory which is a useful tool for analyzing a vague description in decision situations was proposed by Pawlak.Z. , and is now used in areas of medicine, marketing and *Kansei* engineering (Mori, 2001) etc. as a tool for data mining and also is done theoretical researches for extensions (Pawlak *et al*, 2006). For analyzing knowledge representation, rough set theory applied an information system which is defined as a finite information table, the rows of which are labelled by objects, whereas columns are labelled by attributes and entries of the table are attribute-values (Pawlak *et al*, 1994). On the other hand, ontology is introduced for a specification of a conceptualization (Gruber, 2007). Main role of ontology is to naturally represent relationship among objects by flexible structure. Recently, ontology is used in Semantic Web as OWL (Web Ontology Languages) (McGuinness *et al*, 2007). In information system of usual rough set theory, attribute-value is one data. Although, there is the case that attribute-value has more than one data. Usual information system can't naturally represent the case that attribute-value has more than one data, although ontology can naturally represent. Therefore, we consider rough set theory using a concept of ontology. Our main aim of this paper is to formalize rough set theory based on ontology. Although considering on rough set theory using a concept of ontology, for ontology may be complex knowledge base, it generates overlap among objects. We propose an idea of similarity for measuring overlap, which is useful for extracting important sets of attributes defined by flexibly represented knowledge.

BASIC ROUGH SET THEORY

In this section, we overview concepts of rough set theory. In rough set theory, information system is defined as 4-tuple $S = \langle U, Q, V, f \rangle$ where U is finite set of objects, Q is finite set of attributes, $V = \bigcup \{V_q, q \in Q\}$ is finite set of attribute-values and $V_q = \{f(x, q), x \in U\}$ is range of the attribute q, and $f : U \times Q \to V$ is called information function such that $f(x,q) \in V_q$ for every $q \in Q, x \in U$. For any $P \subseteq Q$, $IND(P)$ denotes an equivalence relation and is defined as follows.

$$IND(P) = \{(x,y) \in U \times U \mid \forall q \in Q : f(x,q) = f(y,q)\}$$

For any $P \subseteq Q$ and $x \in U$, $[x]_P$ denotes an equivalence class defined as follows.

$$[x]_P = \{y \mid \forall q \in P : f(x,q) = f(y,q)\}$$

For any $P \subseteq Q$ and $Y \subseteq U$, $P^*(Y)$ denotes the P-upper approximation of Y, and $P_*(Y)$ denotes the P-lower approximation of Y are defined as follows.

$$P^*(Y) = \bigcup \{x \in U \mid [x]_p \subseteq Y\}$$

$$P_*(Y) = \bigcup \{x \in U \mid [x]_p \cap Y \neq \phi\}$$

For any set $Y \subseteq U$, $\alpha_P(Y)$ denotes an accuracy of approximation of set Y by P, and is defined as follows.

$$\alpha_P(Y) = card(P_*(Y)) / card(P^*(Y))$$

Let $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ be a partition of U, and $P \subseteq Q$. $\gamma_p(\mathbf{Y})$ is called the quality of approximation of partition $\mathbf{Y}$ by a set of attributes P, defined as fallows.

$$\gamma_p(\mathbf{Y}) = \sum_{i=1}^{n} card(P_*(Y_i)) / card(U)$$

ROUGH SET THEORY USING A CONCEPT OF ONTOLOGY

Ontology Information System: Information System Using a Concept of Ontology

In this section, we introduce a concept of rough set theory using ontology (Ishizu *et al*, 2007). In rough set theory, an information system is based on the decision table, which is theoretically same as relational database. Ontology has more flexible information structure. Ontology system is defined as $\langle U, Q, C, Dom, Range, rel \rangle$. Where U is a finite set of individuals, Q is a finite set of property names. Let C be a finite class of subsets of U. This means if $c \in C$, then $c \subset U$. Each property name has domain, range, and relation (e.g., $Dom : Q \to C$, $Range : Q \to C$, $rel(p) \subset Dom(p) \times Range(p)$ for any $p \in Q$). Note $rel(p)$ may not be function, but relation. In ontology, class hierarchy is important concept. Class hierarchy is easily represented by the form of $\langle C, \subseteq \rangle$, where $\subseteq$ is set theoretical inclusion. In order to compare usual rough set theory with ontology system directly, we introduce some conditions. 1) Every property has same domain. 2) Individuals are divided by domain and union of the ranges of properties. 3) Every relation is functional. Under the conditions, the following notations are introduced. Since every property has same domain, so $U = Dom(q)$ where $q \in Q$. Let $V_q = Range(q)$ where $q \in Q$, then V_q is finite set of individuals. Let $V = \bigcup \{V_q \mid q \in Q\}$. Let $f : U \times Q \to V$

be defined as $f(x,q) = \{y \mid (x,y) \in rel(q)\}$. Since every relation is functional, then f is function. Then $\langle U, Q, V, f \rangle$ can be regarded as information system. Conversely ontology is free from these conditions, and the conditions 1), 2), 3) may characterize flexible features of ontology information system comparing with information system in rough set theory. In ontology information system, domain of properties may not same. There may be some individuals, which involved in both range and domain of properties. For some $p \in Q$,$rel(p)$ may not be function, and $rel(p)(x)$ may be null or multiple values. Since usual ontology information system may not satisfy the conditions 1), 2), 3), we can't apply rough set theory directly to ontology system. We need to define formulate of rough set theory using ontology. Next we define information system using a concept of ontology. We call it is ontology information system. From the nature of relation rel , i.e. rel is not functional, we introduce extend information function. Extended information function $\tilde{f} : U \times Q \to Pow(U)$ is defined by $\tilde{f}(x,q) = \{y \mid (x,y) \in rel(q)\}$. Where $Pow(U)$ denotes power set of U . Note that $\tilde{f}(x,q)$ is a subset of U .We define ontology information system as $\langle U, Q, C, \tilde{f} \rangle$. Where U' is a set of interested objects. We show in Table 1 which is definition using ontology information system along each concept of rough set theory. We define equivalence class, upper and lower approximation, accuracy, and dependency same as rough set theory. Let U' be a subset of U and $\Upsilon = \{Y_1, Y_2, \ldots, Y_n\}$ be a partition of U' . We can define quality of approximation based on ontology information system.

Table 1 Relationship between the concept of rough set and rough set using a concept of ontology

Rough set	Rough set using a concept of ontology
Information system $<U, Q, V, f>$ $f : U \times Q \to V$ $V = \bigcup \{V_q \mid q \in Q\}$ $V_q = \{f(x,q) \mid x \in U\}$	Ontology system $<U, Q, C, Dom, Range, rel>$ Ontology information system $<U, Q, C, \tilde{f}>$ $\tilde{f} : U \times Q \to Pow(U)$
Equivalence relation $IND(P) = \{(x,y) \in U \times U \mid$ $\forall q \in P : f(x,q) = f(y,q)\}$	Equivalence relation $IND(P) = \{(x,y) \in U \times U \mid$ $\forall q \in P : \tilde{f}(x,q) = \tilde{f}(y,q)\}$
Equivalence class $[x]_P = \{y \mid \forall q \in P : f(x,q) = f(y,q)\}$	Equivalence class $[x]_P = \{y \mid \forall q \in P : \tilde{f}(x,q) = \tilde{f}(y,q)\}$
Upper and lower approximation $P^*(Y) = \bigcup \{x \in U \mid [x]_P \subset Y\}$ $P_*(Y) = \bigcup \{x \in U \mid [x]_P \cap Y \neq \phi\}$	Upper and lower approximation $P^*(Y) = \bigcup \{x \in U \mid [x]_P \subset Y\}$ $P_*(Y) = \bigcup \{x \in U \mid [x]_P \cap Y \neq \phi\}$
Accuracy and dependency $\alpha_P(Y) = card(P_*(Y)) / card(P^*(Y))$ $IND(P) \subset IND(P')$	Accuracy and dependency $\alpha_P(Y) = card(P_*(Y)) / card(P^*(Y))$ $IND(P) \subset IND(P')$

Quality of approximation	Quality of approximation
$\gamma_p(\mathrm{Y}) = \sum_{i=1}^{n} card(P_*(Y_i))/card(U)$	$\gamma_p(\mathrm{Y}) = \sum_{i=1}^{n} card(P_*(Y_i))/card(U')$

A Concept of Similarity

In basic rough set theory, function is defined as $f: U \times Q \to V$. If $f(x,q) \neq f(y,q)$, then relationship between x and y is discernible relation. On the other hand, in ontology information system function is defined as $\tilde{f}: U \times Q \to Pow(U)$. Even if $\tilde{f}(x,q) \neq \tilde{f}(y,q)$, there may be overlap $\tilde{f}(x,q) \cap \tilde{f}(y,q)$. We consider if $\tilde{f}(x,q) \cap \tilde{f}(y,q)$ is large, then a similarity between x and y is great. We introduce similarity relation and degree of similarity for measuring degree of overlap in relationship among objects.

For every $q \in P$ and $x_i, x_j \in U'$, similarity relation is defined as follows.

$$H(x_i, x_j)_P = \frac{2 \cdot card(\prod_{\forall q \in P} \tilde{f}(x_i, q) \cap \prod_{\forall q \in P} \tilde{f}(x_j, q))}{card(\prod_{\forall q \in P} \tilde{f}(x_i, q)) + card(\prod_{\forall q \in P} \tilde{f}(x_j, q))} \quad (1)$$

Character of similarity relation is, for any condition attribute sets, to measure relationship among objects, focussing on overlap of attribute-values. If attribute-values between $\tilde{f}(x_i,q)$ and $\tilde{f}(x_j,q)$ is same for every $q \in P$, then similarity relation is 1.00. If there is no overlap between $\tilde{f}(x_i,q)$ and $\tilde{f}(x_j,q)$ for every $q \in P$, then similarity relation is 0.00. To illustrate the concept of similarity relation, let us consider a simple example of sneakers. We show ontology information system (Table 2).

Table 2 Ontology information system for sneakers

	color (q_1)	model (q_2)	closing mechanism (q_3)	material (q_4)	decision attribute (q_5)
sneaker1(x_1)	white	low	lace	textile	o
sneaker2(x_2)	black	high	lace, Velcro fastener	textile, leather	o
sneaker3(x_3)	colorful	low	lace	textile	×
sneaker4(x_4)	black	low	-	textile	×
sneaker5(x_5)	white	high	lace	textile, leather	×
sneaker6(x_6)	colorful	low	-	leather	o

For example, let us consider for similarity relationship between object x_2 and x_5, for $P = \{q_3, q_4\}$.

$\prod_{\forall q \in P} \tilde{f}(x_2, q) =$ {(lace, textile), (lace, leather), (Velcro fastener, textile), (Velcro fastener, leather)}.

$\prod_{\forall q \in P} \tilde{f}(x_5, q) =$ {(lace, textile), (lace, leather)}.

$\prod_{\forall q \in P} \widetilde{f}(x_2,q) \cap \prod_{\forall q \in P} \widetilde{f}(x_5,q) =$ {(lace, textile), (lace, leather)}.

Therefore, its result is as follow. $H(x_2,x_5)_P = 2 \cdot 2/6 = 0.67$. Similarity relation shows the rate of same objects of attribute-values between x_2 and x_5 for $P = \{q_3, q_4\}$ to all is 67 percent. We do this calculation among every $x \in U'$, for $P = \{q_3, q_4\}$. We show result of $P = \{q_3, q_4\}$ in Table 3.

Table 3 Relation similarity for $P = \{q_3, q_4\}$

$i \backslash j$	1	2	3	4	5	6
1	1.00	0.40	1.00	0.00	0.67	0.00
2	0.40	1.00	0.40	0.00	0.67	0.00
3	1.00	0.40	1.00	0.00	0.67	0.00
4	0.00	0.00	0.00	1.00	0.00	0.00
5	0.67	0.67	0.67	0.00	1.00	0.00
6	0.00	0.00	0.00	0.00	0.00	1.00

Next we introduce a concept of degree of similarity. Let $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ be a partition of U', for every $q \in P$ and $x \in U'$, degree of similarity $\chi_P(\mathbf{Y})$ is defined as follows.

$$\chi_P(\mathbf{Y}) = \sum_{\alpha=1}^{n} \beta_P(Y_\alpha)/n \tag{2}$$

$$\text{where } \beta_P(Y_\alpha) = \sum_{x_i \in Y_\alpha, x_j \notin Y_\alpha} H(x_i, x_j)_P / card(Y_\alpha) \cdot card(Y_\alpha{}^c)$$

$\beta_P(Y_\alpha)$ expresses the ratio that Y_α and $Y_\alpha{}^c$ are compounded by how same attribute-values for P. $\chi_P(\mathbf{Y})$ expresses distinction of each Y_α in the ontology information system, that is to say, if value of degree of similarity is low, then it means distinction of each Y_α is high . And also if $\mathbf{Y} = \{Y_1, Y_2\}$, then $\chi_P(\mathbf{Y}) = \beta_P(Y_\alpha)$, for symmetry. This definition corresponds to the quality of approximation. The quality of approximation is defined by lower approximation and cardinal subset of universe (i.e. quality of approximation is defined by $\gamma_p(\mathbf{Y}) = \sum_{i=1}^{n} card(P_*(Y_i))/card(U')$). On the other hand, the degree of similarity has no approximation and focuses on overlap of attribute-values among Y_α. We illustrate the concept of degree of similarity to use result of Table 3. A degree of similarity for $P = \{q_3, q_4\}$ is as follows.

$$\beta_P(Y_1) = \sum_{x_i \in Y_\alpha, x_j \notin Y_\alpha} H(x_i, x_j)_P / card(Y_1) \cdot card(Y_1{}^c)$$

$$= (1.00 + 0.00 + 0.67 + 0.40 + 0.00 + 0.67 + 0.00 + 0.00 + 0.00)/3 \cdot 3$$

$$= 0.30$$

$$\chi_P(\mathbf{Y}) = 0.30$$

We show analysis of quality of approximation of partition and degree of similarity of partition in Table 4.

Table 4 Analysis of quality of approximation and degree of similarity

P	quality of approximation	degree of similarity
q_1,q_2,q_3,q_4	1.00	0.00
q_1,q_2,q_3	1.00	0.00
q_1,q_2,q_4	1.00	0.00
q_1,q_3,q_4	1.00	0.07
q_2,q_3,q_4	0.67	0.19
q_1,q_2	0.67	0.11
q_1,q_3	0.67	0.11
q_1,q_4	1.00	0.15
q_2,q_3	0.33	0.30
q_2,q_4	0.33	0.33
q_3,q_4	0.17	0.30
q_1	0.00	0.33
q_2	0.00	0.56
q_3	0.17	0.48
q_4	0.17	0.63

Next we illustrate how to use the degree of similarity. We consider in the case that degree of similarity is high or low, for quality of approximation is same. Let's consider in the case of Table 5 and Table 6 as example. Quality of approximation for each of these examples is same (i.e. 1.00).

Table 5 Example 1

sample no.	Material	decision attribute
sneaker1(x_1)	leather	_
sneaker2(x_2)	suede	_
sneaker3(x_3)	leather, suede	_
sneaker4(x_4)	textile	_
degree of similarity = 0.34		

Table 6 Example 2

sample no.	Material	decision attribute
sneaker1(x_1)	leather, suede	_
sneaker2(x_2)	suede	_
sneaker3(x_3)	leather	_
sneaker4(x_4)	textile	_
degree of similarity = 0.17		

We waver in judgment of which table (Table 5 or Table 6) to extract rules from. Although, considering that degree of similarity is low (i.e. overlap among set Y_α is small), we find to be able to extract brief decision rule from table 5 than do from table 6. The rules from Table 5 are the following:

Rule 1: IF 'material' = leather and not suede THEN _,

Rule 2: IF 'material' = not leather and suede THEN _,

Rule 3: IF 'material' = leather and suede THEN _,

Rule 4: IF 'material' = textile THEN _.
The rules from Table 6 are the following:

Rule 1: IF 'material' = suede THEN _,

Rule 2: IF 'material' = not leather and suede THEN _,

Rule 3: IF 'material' = textile THEN _.

Let's compare different example (Table 7 and Table 8). Quality of approximation for Table 8 is 1.00.

Table 7 Example 3(=Example 1)

sample no.	Material	decision attribute
sneaker1(x_1)	leather	_
sneaker2(x_2)	suede	_
sneaker3(x_3)	leather, suede	_
sneaker4(x_4)	textile	_
degree of similarity = 0.34		

Table 8 Example 4

sample no.	Material	decision attribute
sneaker1(x_1)	leather, nylon	_
sneaker2(x_2)	nylon, suede	_
sneaker3(x_3)	leather, suede	_
sneaker4(x_4)	textile	_
degree of similarity = 0.25		

Instinctively, we will extract decision rules from Table 6. But, we can extract compact decision rules from Table 8 than Table 7. That is to say, degree of similarity is useful for extracting from which condition attribute set.
The rules from Table 7 are same as Table 5. The rules from Table 8 are the following:

Rule 1: IF 'material' = nylon THEN _,

Rule 2: IF 'material' = leather and suede THEN _,

Rule 3: IF 'material' = textile THEN _.

CONCLUSION

One of our main aims of this paper is to propose rough set theory applied the concept of ontology to. We propose rough set theory using a concept of ontology and formulate it by defining ontology information system. Function $\tilde{f}: U \times Q \rightarrow Pow(U)$ is defined by $\tilde{f}(x,q) = \{y \mid (x,y) \in rel(q)\}$ in ontology information system. Even if $\tilde{f}(x,q) \neq \tilde{f}(y,q)$, there may be overlap $\tilde{f}(x,q) \cap \tilde{f}(y,q)$. We consider if $\tilde{f}(x,q) \cap \tilde{f}(y,q)$ is large, then a similarity of between x and y is great. For measuring degree of overlap in relationship among objects, we propose a concept of similarity relation and degree of similarity. We show simple examples to demonstrate how to extract rules, considering degree of similarity. Consequently we show that degree of similarity is useful for extracting decision rules.

REFERENCES

Gruber, T. (2007). What is an Ontology?, http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Ishizu, S., Gehrmann, A., Nagai, Y., and Inukai, Y. (2007). Rough set: extension of Ontologies by rough sets, in *Methods, Techniques and Tools in Information Design*, (Smith, M.J., and Salvendy, G.ed.), 456-462, Springer-Verlag, Berlin.

McGuinness, D.L., and van Harmelen, F. (2007). OWL Web Ontology Language Overview, http://www.w3.org/TR/owl-features/.

Mori, N. (2001). Rough Set and Kansei Engineering, *Journal of Japan Society for Fuzzy Theory and Systems*, 13(6):600-607. (in Japanese).

Pawlak, Z., and Slowinski, R. (1994). Rough set approach to multi-attribute decision analysis, *European J. of OR.*, 72(3):443-459.

Pawlak, Z., and Skowron, A. (2007).Rough set: Some extensions, *Information Sciences*, 177(1):28-40.