IMPLEMENTING ROUGH SET TECHNIQUE IN SOFTWARE ENGINEERING PROCESS FOR INFORMATION EXTRACTION

Anirban Mitra
Faculty, Dept. of CSE
M.I.T.S., Rayagada 765017
Orissa, INDIA
mitra.anirban@gmail.com

Prasanta Kumar Padhi
Team Leader and SPOC
Sysnetglobal Technology (Pvt) Ltd
INDIA
prasanta.pdhi@gmail.com

Abstract

Our paper deals on the topic of 'Intelligent Information Retrieval'. There are many processes for extracting knowledge form a complete information system. As observed, in many cases, real life problems are uncertain. We may categorize such problem under the incomplete information system. It is difficult to deal with those systems having knowledge having 'incompleteness' (some values in dataset is missing) and 'inconsistencies' (ambiguities and contradicting values in dataset) in nature. Hence, the process of knowledge and information extraction becomes more challenging where real time data are incomplete. Rough Set can be used as a tool for clustering and handling incomplete information. In this paper we have defined and discussed on some of the known properties of Rough Set and we have implemented the tool to generate information for a software engineering process. Planning for delivery and installation for any software requires planning for procurement of hardware, software, and skilled manpower (software developers). The process of delivering the software also consists of preparing the documentation and manuals, and planning for training. Scheduling for delivery and installation within the deadline and estimated coast, on the other hand, requires the preparation of a time table for putting the system in place. It is desirable in many cases, that the new software is installed while the old system still operates, as the automation system need to be running. We have studied and analyzed our proposed technique for one such software, i.e COA. COA (Control Office Application) runs for CRIS (Centre for Railway Information System) of Indian Railway. COA mainly deals with the automation of Arrival, Departure and running of the trains. Earlier CRIS implemented and succeeded in another project i.e. FOIS (Freight Operation Information System). This is one of the Asia's biggest Networks in any organization. In our analysis, we observed that, selecting Evolutionary Model as SDLC for up gradation and modification of this process may gives a better result. It was also observed that, selection of appropriate skilled manpower (software developers) is one of the key factors for success and in-time delivery, when SDLC-Evolutionary Model is followed. During the process of selection of appropriate skilled manpower, it was found that the dataset consist of some missing, incomplete and uncertain information. Hence, we proposed and used Rough Set as a tool for classification, cluster and to generate knowledge, which can be used for software procurement planning.

Keyword:- SDLC, Evolutionary Model, Incomplete Information System and Rough Set.

I. INTRODUCTION

A. Information System

A special kind of approximation space is require, when classifying objects on basis of their properties and we identify properties with some attributes, characteristics of these objects with each attribute a set of value is associated. Description of an object is given when one value for each attribute is chosen [11], [13]. The above idea can be expressed more preciously by means of the notation of an information system introduced in Pawlak in 1981. By an information system S we mean an ordered quadruple, $S = (U, Q, V, \rho)$, where U is the set called Universe of S – Element of U are called objects; Q is a set of attributes, $V\left(=\bigcup_{q \in Q} V_q\right)$ is a set of values of attributes. V_q will be called the domain of q and $\rho : U \times Q \to V$ is a description function such that $\rho(x, q) \in V_q$ for every q ∈ Q and x∈ U. We introduce function $\rho_x : Q \to V$ such that $\rho_x(q) = \rho(x, q)$ for every q ∈ Q and x ∈ U.; ρ_x will be called the description of x in S. For sake of simplicity, function ρ_x will be written as a sequence of attribute values $v_{i_1}, v_{i_2}, \ldots\ldots, v_{i_n}$ assuming that $v_{i_j} \in V_{q_j}$, of course, the order of values in this sequence is immaterial. we say

that objects $x, y \in U$ are indiscernible with respect to $q \in Q$ in A, iff $\rho_x(q) = \rho_y(q)$, and we shall write $x_{\tilde{q}} y$, certainly $\tilde{q}$ is an equivalence relation. Objects $x, y \in U$ are indiscernible with respect to $P \subset Q$ in S, in symbols $\tilde{P} = \bigcap_{p \in P} \tilde{p}$. In particular, if $P = Q$ we say that x and y are indiscernible in S and write $x_{\tilde{S}} y$ instead of $x_{\tilde{Q}} y$. Obviously, P is an equivalence relation, thus each information system $S = (U, Q, V, \rho)$ defines uniquely an approximation space $A_S = (U, \tilde{S})$, where $\tilde{S}$ is the indiscernibility relation generated by the information system. If $x \in U$ and ρ_x is the description of x in S, then we assume that ρ_x is also the description of the equivalence class of the relation $\tilde{S}$ containing x. we say that the subset $X \subset U$ is describable in S iff X is definable in A_S, if X is undefinable in A_S, X will be called nondescribable in S. Description of a describable set in S consists of all description of its elementary sets. Description of an empty set is denoted by ψ.

B. Data Representation and its Relationship

A data set can be represented by a table where each row represents, for instance, an object, a case, or an event. Every column represents an attribute, or an observation, or a property that can be measured for each object; it can also be supplied by a human expert or user. This table is called an information system.

The choice of attributes is subjective (they are often called conditional attributes) and reflect our intuition about factors that influence the classification of objects. The chosen attributes determine in turn primitive descriptors that provide intensions of primitive concepts.

In many cases the target of the classification, that is, the family of concepts to be approximated is represented by an additional attribute d called decision. Information systems of this kind are called decision systems and they are written down as triples $A = (U, A, d)$.

II. DEFINITION & PROPERTIES OF ROUGH SET

The concept of rough set is another approach to deal with imperfect knowledge. It was introduced by **Z. Pawlak in 1982** ([59]). From a philosophical point of view, rough set theory is a new approach to deal vagueness and uncertainty, and from a practical point of view, it is a new method of data analysis [2].

This method has the following important advantages:

- It provides efficient algorithms for finding hidden patterns in data;
- It finds reduced set of data (data reduction);
- It evaluates significance of data;
- It generates minimal set of decision rules from data;
- It is easy to understand;
- It offers straightforward interpretation of results;
- It can be used in both qualitative and quantitative data analysis; and
- It identifies relationship that would not be found by using statistical methods.

Rough set theory overlaps with many other theories, such as fuzzy sets, evidence theory and statistics. Nevertheless, it can be viewed in its own right as an independent, complementary and non-competing discipline. The rough set methodology has found many real life applications in various domains. It seems that the rough sets approaches can also be used in legal reasoning, particularly in drawing conclusions from factual data.

According to the interpretation of Pawlak, knowledge about an universe can be considered as one's capability to classify objects of the universe. By classification of an universe U, we mean a set of subsets $\{C_i,\ i = 1, 2, \ldots, n\}$ of U such that $C_i \cap C_j = \varphi$ for $i \neq j$ and $\bigcup_{i=1}^{n} C_i = U$.

Let $R \subseteq U x U$ denote an equivalence relation on U, that is, R is a reflexive, symmetric and transitive relation. The equivalence class of an element $x \in U$ with respect to R is the set of elements $y \in U$ such that xRy. If two elements x, y in U belong to the same equivalence class then we say that x and y are indistinguishable with respect to relation R. The pair $aprA = (U, R)$ is called an approximation space. It is well known that an equivalence relation induces a partition of U into disjoint equivalence classes. Also, corresponding to every partition on U there is an equivalence relation, which has these partitions as its equivalence classes. It defines the quotient set U/R consisting of all equivalence classes of R. The equivalence class $[x]R$, containing x plays dual roles. It is a subset of U if considered in relation to the universe, and an element of U/R if considered in relation to the quotient set. The empty set φ and the equivalence classes are called the elementary sets. The union of one or more elementary sets are called a compound sets. The family of all compound sets is denoted by *Comp(apr)*. It is

a sub-algebra of Boolean algebra $2U$ formed by the power set of U [1],[5],[6].

Given an arbitrary set $A \subseteq U$ it may not be possible to describe 'A' precisely in the approximation space $aprR = (U, R)$ Instead one may only characterise 'A' by a pair of lower and upper approximations. This leads to the concept of rough sets. We define,

$\underline{RA} = \cup\{Y \in U/R : Y \subseteq A\}$; and

$\overline{RA} = \cup\{Y \in U/R : Y \cap A \neq \varphi\}$. Where,

$\overline{RA}$ and $\underline{RA}$ are respectively called the ***R-lower*** and ***R-upper approximation*** of A with respect to R [2]. It can be noted that $\underline{RA} = \{x \in U : [x_R] \subseteq A\}$ and

$\overline{RA} = \{x \in U : [x]_R \cap X \neq \varphi\}$.

The set $BN_R(A) = \overline{RA} - \underline{RA}$ is called the R-boundary of A. The set RA consists of all those elements of U which can with certainty be classified as elements of A, employing the knowledge R. The set $\overline{RA}$ consists of all those elements of U which can possibly be classified as elements of A, employing the knowledge R. Set $BNR(A)$ is the set of elements which cannot be classified as either belonging to A or belonging to $-A$ having the knowledge R. We say that a set A is R-definable if and only if $\underline{RA} = \overline{RA}$. Otherwise A is said to be R-rough [7], [8], [10], [12].

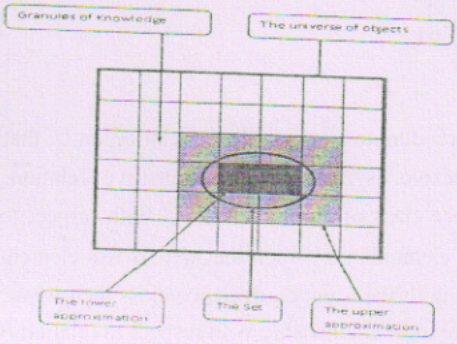

Fig: Lower and Upper Approximation of RS

III. COA (CONTROL OFFICE APPLICATION)

COA stands for **Control Office Application**; here this application is totally projected by the main IT vendor **Centre for Railway Information System** (CRIS) of Indian Railway. This COA project is to computerize the Arrival/Departure of the train. The information generated is used for various purposes by Indian Railways. This project is presently implemented in 67 Division of the Railway. One of the main operations of this project is to generate chart of the trains arrival and departure and other related details information. These information are reflected graphically. Decisions related to Fright and rail controlling are made as per the generated chart of COA. COA has also feature of route and schedule prediction for rails and frights for a given period of time under a division.

IV. INTRODUCTION TO SOFTWARE ENGINEERING

Software Engineering is the process of developing and maintaining of software in a systematic approach. The process from birth to retirement of software is divided vaguely in six phases; they are requirement analysis, designing phase, coding phase, testing phase, implementation and maintenance phase. Various ***SDLC*** Models (Software Life Cycle Development Model) deal with these phases in certain sequential manner. Every SDLC model, namely, Waterfall model or Spiral Model or Evolutionary model has its own advantages and disadvantages. We, in our work will concentrate of Evolutionary Model [3],[4],[9].

A. Evolutionary model

The software requirement is first broken down into several modules that can be incrementally constructed. The core is first developed. Core modules are those module, which do not need services from the other modules. This initial product skeleton is refined into increasing levels of capability by adding new functionalities in successive versions. The layers are built, one above the other in a way that each successive version of the product is a fully functional software capable of performing more work that its previous version.

We have chosen this model for some of its advantages. In this model, the user or client gets a chance to experiment with partially developed software much before the complete version of the system is released.

B. Analysis of the Problem

Software like **COA** needs time to time maintenance and enhancement. Our observation shows that evolutionary model gives a better result, when implemented in a prototype. The phase of maintenance depends on the feedback and involvement of the technical man power. It was further observed that selection of appropriate skilled

manpower (software developers) is one of the key factors for success and in-time delivery, when SDLC-Evolutionary Model is followed. The selection can be done from the pool of people who are associated with the COA. On basic of some of the parameters, we have classified the people into classification for appropriate selection. During the process of selection of appropriate skilled manpower, it was found that the dataset consist of some missing, incomplete and uncertain information. Hence, we proposed and used Rough Set as a tool for classification, cluster and to generate knowledge, which can be used for software procurement planning. The following sub section discuss on our classification and selection process.

For simplicity, we have assumed that, we will make the selection process among the 50 employees, who are serving the organization.

We assume that these 50 employees are distributed over four criteria (say: knowledge to handle the computer, knowledge on Rail movement, knowledge to handle COA and was part of COA up-gradation team).

Apart from regular responsibilities, some of the employee have assign extra and responsibilities to look after the smooth running of particular areas (say: computerized ticket booking counter related, Divisional MIS related, Signal system related and so on.)

Let the Fifty (50) employees distributed over four criteria are mentioned as a_i where: $i = 1,2,3,...,50$. Hence, If we consider the group of employee as Universal set U then $U = \{a_1, a_2, a_3, ..., a_{49}, a_{50}\}$ such that $a_i \to a_j$, if they belong to same criteria or of belong to same group.

We assume some arbitrary Conditions, which are very much comparable with the real World: as, some of the employees are also involved are assigned with some added responsibility. It implies that the employees name can be found in more than one section or with more than one group.

But we restrict employees grouping on the basic of criteria that is every employee's reference will be grouped in under only one criterion. However, when grouped on the basic of added responsibilities, it may consist of employees belonging to different criteria. For example, responsibility related to 'computerized ticket booking counter related' may consist of employees, who belongs to 'knowledge to handle the computer' and 'knowledge on Rail movement'.

Let, in our problem, we have four (04) criteria. Let it be mentioned as D_1, D_2, D_3 *and* D_4. Such that $D_1 \cup D_2 \cup D_3 \cup D_4 = U$, and $D_i \cap D_j = \phi$, for any $\{i, j\} = 1,2,3,4$.

The above mathematical interpretation shows that no employees are belonging to more than one criterion. Now, as mentioned, the 50 employees are distributed over four criterion, let us assume, our distribution of employees are: $D_1 = \{a_1, a_2, a_3, ..., a_{15}\}$ $D_2 = \{a_{16}, a_{17}, a_{18}, ..., a_{30}\}$, $D_3 = \{a_{31}, a_{32}, a_{33}, ..., a_{45}\}$ and $D_4 = \{a_{46}, a_{47}, a_{48}, a_{49}, a_{50}\}$. Let C_1, C_2, C_3 *and* C_4 be the group of employees, working for the smooth running of the added responsibilities like 'computerized ticket booking counter related, Divisional MIS related, Signal system related' and so on.

Dividing the employees, on basic of responsibilities, assuming that we presently there are four (04) different type of responsibility, as:

$C_1 = \{a_1, a_2, a_9, a_{13}, a_{14}, a_{19}, a_{39}, a_{49}\}$,

$C_2 = \{a_9, a_{16}, a_{17}, a_{19}, a_{29}, a_{30}, a_{39}, a_{49}\}$,

$C_3 = \{a_9, a_{19}, a_{31}, a_{32}, a_{39}, a_{45}, a_{46}, a_{49}\}$ and

$C_4 = \{a_9, a_{19}, a_{39}, a_{46}, a_{47}, a_{48}, a_{49}, a_{50}\}$.

The physical interpretation of a Lower Approximation of a set $C_i, i = 1,2,3,4$ here is that, it provides the fact whether all the employees belonging to a particular Criteria, works for a particular responsibility or not. Similarly, the Upper Approximation of these set provides the fact whether – there exist any Employee on any particular Criteria, who is involved in any of the Responsibility Group or not. Considering an example from above, $\overline{RC}_1 = U \Rightarrow$ that employees from Every (Four) Criteria's are involved in that particular Responsibility. Similarly, $\underline{RC}_4 = D_4 \Rightarrow$ all the employees belonging to the D_4 is involved for the Responsible group: C_4.

An efficient searching method is required to process queries like: "employee, who had worked in COA and had knowledge in Divisional MIS". Such types of queries need a proper classification to give an efficient search result. As, we not only assume, but have also observed that the profile and other details of the employees are sometime remains uncertain or consist of missing information/data. In such cases, a traditional tool for classifications fails. Classification through Rough Set provides a better result. With having flexible boundary of Lower and Upper Approximation, exact match and possible matched items can be classified to make the result more acceptable and satisfactory. With such generalized concept, an employer / manager can choose the parameters for selection of employees from the pool. A better and suitable selection of employee will make the SDLC process faster and

qualitatively better. As observed, software like COA, need input from the people who have the knowledge of rail movement and knowledge of MIS. A employee having the knowledge of building COA type application without the technical details of the rail movements will not be sufficient enough for enhancement and implementation of such type of application in more qualitatively and scheduled way. A trained user will reduce the time and cost of training and will be quick assets in giving proper feedback, which will be used for developing the next version of such application.

IV. CONCLUSION

We have discussed on 'Intelligent Information Retrieval'. There are many processes for extracting knowledge form a complete information system but for incomplete information system, it is difficult to deal with those systems having knowledge having '**incompleteness**' (some values in dataset is missing) and '**inconsistencies**' (ambiguities and contradicting values in dataset) in nature. Thus, the process of knowledge and information extraction becomes more challenging where real time data are incomplete. We have used Rough Set as a tool for clustering and handling incomplete information. In our work, we have defined and discussed on some of the known properties of Rough Set and we have implemented the tool to generate information for a software engineering process.

Here, we have studied and analyzed our proposed technique for one such software, i.e COA. In our analysis, we observed that, selecting Evolutionary Model as SDLC for up gradation and modification of this process tends to give a better result. Selection of appropriate skilled manpower (software developers) from the pool of employee, that's too when the dataset consist of some missing, incomplete and uncertain information becoming challenging. Using Rough Set, having the boundary of lower approximation and upper approximation for classification, gives a better result for selection of employee, when various criteria and responsibilities were concerned.

V. ACKNOWLEDGEMENT

About **COA**: For COA project, main vendor is CRIS. HP is the second vendor for the development of application and maintenance. This particular application was developed by WIPRO. DBA, Database maintenance and supported is carried out by Sysnetglobal Technology (Pvt) Ltd. Sysnetglobal is the third vendor and is guided by HP. This project is also known as HP_CRIS. The author: Prasant Kumar Padhi is presently working on HP_CRIS under Sysnetglobal.)

VI. REFERENCES

[1] Novotny, M., Pawlak, Z.: On Rough Equalities, Bulletin of Pol. Aca. Sci. Math., 1985.

[2] Tripathy, B. K.: On Some Basic Properties of Rough Sets and Applications, FLATEM-2004, IIT Kharagpur, May 21-22 (2004).

[3] Rajiv Mall, "Fundamental of Software Engineering," PHI Publications, Third Edition, pp. 44-48

[4] Nicolas Guelfi, "A technical report on Software Engineering", European Software Engineering Conference and ACM SIGSOFT Symposium on Foundations of Software Engineering, pp. 143, 2005.

[5] Banerjee. M, & Chakroborty M.K.: Rough Consequences & Rough Algebra in Rough Sets, RSKD 1993.

[6] Boniokowski. Z: A Certain Conception of the Calculus of Rough Sets , Norte Dame Journal of Formal Logic, 1992.

[7] Chakroborty M.K., Banerjee. M,: Rough Dialogue and Implication Lattices, Fundamenta Informetica, 2007

[8] Novotny, M., Pawlak, Z.: Characterization of Rough Top Equalities and Rough Bottom Equalities, Bulletin of Pol. Aca. Sci. Math., 1985.

[9] P. Loucopoulos; W. Robinson, "Requirements Engineering – An Introduction" Edited Volume. Kluwer Publications,1999.

[10] Advances in Intelligent and Soft Computing, Vol. 32 Hoffmann, F., Köppen, M., Klawonn, F., Roy, R. (Eds.), 2001.

[11] Huangbin, Zhou Xianzhong, "Extension of rough sets Model based on connection degree under incomplete information systems", Systems Engineering—Theory & Practice, 2004, 1(1), pp. 88-92.

[12] I. Duentsch, G. Gediga: Rough set data analysis. In Encyclopedia of Computer Science and Technology, Marcel Dekker (to appear).

[13] Bazan, J.G., Nguyen, Son Hung, Nguyen, Tuan Trung, Skowron, A., Stepaniuk, J., "Decision rules synthesis for object classification", In: E. Orlowska (ed.), Incomplete Information: Rough Set Analysis, Physica - Verlag, Heidelberg, 1998, pp. 23–57.