A MATHEMATICAL STUDY OF DISCOVER THE EXCEPTION WITHIN THE ROUGH SET FRAMEWORK APPROACH

Om Prakash Gera*

I. INTRODUCTION

The approach used by concept discovery and information re-engineering is flexible and dynamic in that the conceptual integration process can be frequent activity. As usage patterns are utilized to discover concepts further. The information re-engineering approach presented here addresses the uniqueness of each user group and allows contextual interpretation of information using terminology that is initiated and preferred by different user groups. This strategy allows users to query and retrieve information at a conceptual level determined by the user seeking information.

Concept discovery uses a discovery algorithm that can be classified as learning from observation, where database objects of legacy systems are classified into groups or clusters which can be described by a concept from a predefined concept class that is well understood within the application domain. In order to establish a cluster of objects, each object is characterized by a set of variables. The variables represent the metadata of the object and are good indicators of both structure and usage of the database object. Commonality in structure and usage pattern serves to establish similarity measures among objects that are clustered. The concept discovery algorithm partitions these objects into clusters using the values of the variable. The clusters thus formed are meaningful such that each cluster actually represents a concept common to the legacy systems in the application domain. The set of application specific generic concepts discovered in this way provides a domain abstraction and constitutes the reconstructed conceptual schema that can

* Ph. D. Research Scholar, Department of Computer Science & Engineering, Bhagwant University, Ajmer (Rajasthan).

support interoperability among the databases or information systems whose objects were clustered. This conceptual schema provides a wrapping service and serves as the middle layer to facilitates information retrieval from heterogeneous databases.

The processes of Knowledge Discovery in Databases (KDD) and information Retrieval (IR) appear deceptively simple when views from the perspective of terminological definition. Fayyad, Piatetsky-Shapiro, and Smith (1996) define KDD as "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (p. 30). The closely related process of IR is defined by Rocha (2001) as "the methods and processes for searching relevant information out of information systems that contain extremely large numbers of documents" (1.1). In execution, however, these processes are not simple at all, especially when executed to satisfy specific personal or organizational Knowledge Management (KM) requirements or as the core functionality of Knowledge Management Systems (KMS).

The potential validity or usefulness of an individual data element or pattern of data elements may change dramatically from individual to individual, organization to organization, or task to task. Relevance is a highly contextual and personal data characteristic, changing even as the

IR process is underway and information requirements are incrementally met. Making retrieved data or a description of data patterns generally understandable is also highly problematic. Data that may appear relevant and easily understandable in one retrieval context may be completely unintelligible in another, even to the same audience. KDD and IR are, in fact, highly complex processes that are strongly affected by a wide range of factors. These factors include the needs and information seeking characteristics of system users as well as the tools and methods used to search and retrieve, the structure and size of the data set or database, and the nature of the data itself.

II. KDD and IR: An Historical Perspective.

Origins

Information professionals often describe the KDD and IR processes in the context of specific types of Database Management Systems (DBMS). Devarakonda (2001) divides DBMS into four types : simple data without query, and complex data with query. An example of the first type, simple data without query, is a filling system, including files that may exist only in paper

form. The second, third and fourth types are exemplified by Relational DBMS (RDBMS), Object-Oriented DBMS (OODBMS), and Object-Relational DBMS (ORDBMS), respectively (Devarakonda, 2001, ORDBMS). The type of database that is queried significantly affects the processes of knowledge discovery (KD) and IR.

Because an RDBMS of some type forms the core of almost all KMS, improvement of RDBMS functionality for KD and IR has been a crucial part of KMS refinement for the past three decades. The relatively recent introduction of OODBMS to KMS has created many new KD and IR problem sets for researchers. These challenges have been met, thus far, primarily through the introduction of certain features of RDBMS to OODBMS. The result has been the development of a small group of ORDBMS that combine the best KD and IR features of RDBMS and OODBMS (ORDBMS).

Information processionals familiar with traditional filing systems are actually aware of the limitations imposed on KD and IR by their pre-set filing structure. Although technically a database, this type of DBMS does not lend itself to automated searching, but only to browsing or search by pre-designated subject categories and file descriptions (e.g. library card catalogs). The difficulties presented for KD and IR by simple filing structure were initially replicated in computer-supported file structures and were only alleviated with the introduction of the Relational Database Model (RDM), by E.F. Codd in 1970 (Devarakonda, 2002, RDBMS).

Introduction of the RDM resulted in rapid adoption of RDMS for information organization and control across a broad range of commercial and social organization as well as the development of increasingly effective data collection and storage technologies. DBMS permitted much more flexibility in data organization and retrieval than traditional data filing systems, but traditional IR Methods did not permit flexibility in the characterization of user needs or the delineation of search parameters (Rocha, 2001, 1.2). The result, of course, was increasing numbers of organizations that possessed very large and continually growing databases but only rudimentary tools for KD and IR. Two areas of research focus in information management developed in response to this problem : data warehousing and data mining.

Data warehousing, defined by Fayyad et al. as “Collecting and ‘cleaning’ transactional data to make it available for online analysis and decision support” (2001, p. 30), focuses on the methodical collection and pre-processing of data for specific analytical uses. The data is subject-

oriented, time-stamped, and integrated to permit interactive analysis in support of decision-making processes. A data warehouse normally integrates data from a variety of sources, "thus enriching the data and broadening the context and issued of the information" (Rauber et al., 2002, Data Warehousing ……..

Data mining, defined as "the application of specific algorithms to a data set for the purpose of extracting data patterns" (p. 28) Formses on improving the utility of large data sets as well as IP response. Data mining, in particular the algorithms used in data mining, has received a lion's share of attention in the development of Decision Support Systems (DSS) and RDMS research because results are often immediately applicable in high-payoff decision making industries such as insurance, sales, and financial and medical services.

III Inspirations and Intentions for the Technology

Rocha describes the ultimate goal of IR as the production of recommendation of relevant information to users (2001, 1.2). We can ascribe the same motivation to the development of KDD systems and methods in general, particularly in regards to the refinement of DBMS. Research in data collection, storage, and retrieval has focused on issues specifically related to the increment of KD and functionality. Among the topics given and attention have been data translation, change detection, integration, duplication, summarization, aggregation and defines (Widom, 1995).

Research has also focused on the need to improve automation ID and IR, especially in the areas of data selection and pre-messing, data transformation and data interpretation and utilization (Fayyad et al, 1996, p. 28). However, increased information in KD and IR requires increased attention to the methods used for data collection and storage as well as the statistical foundations of the search and retrieval processes (p. 29). Despite this complication, however, it is clear that manual analysis of billions of records and hundreds of fields is impractical and that automated data handling will be even more in demand as requirements for on-the-fly analysis and more flexible presentation of search results increase (p. 28).

IV KDD and IR : Application to KMS

Technological Systems and Processes

A commonly used RDBMS is Microsoft Access, he existence of a standard query language allows data to be executed easily from one RDBMS to another (Devarakonda, 2002,.).

Although the structure of RDBMS renders them capable of handling complex data types such as spatial data, images or number arrays without the use of BLOBs, it does permit data access and large storage capacities.

AIM OF THIS RESEARCH

The aim of this research work is to discover the exception by using the rough set approach and to structure/represent the exceptions in the form of rule pair, a knowledge structure that consist of commonsense rule and exception rule. Knowledge structures are compact representation of rules and increase the comprehensibility.

REVIEW OF LITERATURE

"The KDD process for extracting useful knowledge from volumes of data", By U.M. Fayyad, G.P. Shapiro and P. Smith

This paper sort out some difficulties of some traditional method of turning data into knowledge. Such method relies on manual analysis and interpretation of data sets. Such manual turning is slow, expensive and subjective. In this paper the term CD that refer to the overall process of discovering useful knowledge from data. Data mining is a particular step in whole process of KDD that apply specific algorithms extracting patterns or Model. In this paper KDD is defined as the nontrivial process of Identifying valid, novel, potentially useful and ultimately understandable patters in data.

"A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery", By Alex A. Freitas.

This paper discusses the use of evolutionary algorithms, particularly genetic algorithms and genetic programming, in data mining and knowledge discovery. The focus of paper is on the data mining task of classification. In addition, some preprocessing and post processing steps of the knowledge discovery are discussed. Knowledge discovery process, focusing on attributes selection and pruning of an ensemble of classifiers. This paper shows how the requirements of data mining and knowledge discovery influence the design of evolutionary algorithms.

"A belief driven method for discovering unexpected betters, By Padmanabhan and Tuzhilin.

In this article a method for discovering unexpected rules is proposed this method is a directed approach for finding exceptions. This method can be regarded as discovering a kind of exceptions because it defines unexpectedness as a logical contradiction of in rule to a set of beliefs (commonsense rule). Let XA represent X/A and given a commonsense rule X->Y, the method first discovers all rule XA->B each of which satisfied the condition for association rules that of support and confidence are greater then their thresholds and B contradicts Y. Next the method obtains more general and more unexpected rules X'A->B by generalizing X

Objective

The objective of this research work is to discover or find the exception.. within the rough set approach and to structure the algorithms. in the form of rule pair knowledge structure.

An exception mining is important as exceptions challenge the data mining knowledge and lead to the growth of knowledge in new research, Exception also improves the quality of decision making. Begin set generate a large number of rules so that there manual Insertion is very difficult as some rules are redundant and this algorithm. also remove redundant rules and hence improves the apprehensibility.

Methodology

The methodology for this research work includes the following :

- Survey of literature on data mining, exceptions, rough set.
- Selection of rough set tools.
- Selection of data sets from diverse application domains.
- Generating decision rules on data sets.
- Finding exception in the form of rule pair from generated set. by implementing required algorithms.
- Estimating the predictive accuracy on test data.
- Concluding the whole work.
- Rough Set Theory are described.

Information System

In information system containing a set of objects. Each object has a number of attributes with attribute value related to it. The attributes are the same for all objects, but the attribute value may differ. An information system is thus more or less the same as a relational databse.

Definition 3.1.1 (Information System) An information System (IS) is an ordered pair A = (U; A) where U is a nonempty finite set of objects called Universe, and A is a nonempty, finite set of elements called Attributes. The elements of the Universe will be referred as the Objects in the following.

Every attribute a □ A is a total function a : U □ Va, where Va is the set of allowed values for the attribute (its range). For an example consider the table 3.1.

Information System

	Headache	**Muscle Pain**	**Temperature**
P1	No	Yes	High
P2	Yes	No	High
P3	Yes	Yes	Very High
P4	No	Yes	Normal
P5	Yes	No	High
P6	No	Yes	Very High

An information system can be represented by an information table, where the rows in the table are objects in the universe and the columns correspond to the attributes. Consider, as an example, the information table 3.1, where U = {P1, P2, P3, P4, P5, P6} is a set of patients and A = {Headache Muscle Pain, Temperature} are the attributes corresponding to the symptoms of a patient. Every row can be seen as information about a specific patient. For example, patient P5 is characterized by the attribute value set {(headache, yes), (Muscle pain, no), (temperature, high)}.

An information table can be seen as a set of training examples in machine learning. Each training example is then connected with a decision that classified the example into a predefined attribute shows the diagnosis of a patient, i.e. whether or not the patient has the disease flu [3, 4, 21, 41, 42].

Decision Class

Let IS = (U, A, D) be a decision system. Every di □ D partitions the universe U in| Vdi | classes $X_1, \ldots, X_k$. Each class X_j (j □ {1, . . . ,| Vdi |}) is called a decision class [4, 22].

Indiscernibility

Objects that have the same values of the conditional attributes are called indiscernible (inseparable). Patients, for example, can have the same set of symptoms but different diagnoses. For instance, patients P2 and P5 in decision table 3.2 ae examples of such a situation. Rough set theory takes into account indiscernibility between objects through the notion of an indiscernibility relation. The indiscernibility relation is used to a describe the fact that it may not be possible to separate certain objects in the universe by using the information given by the attributes [4, 22, 41, 42].

So an information system is extended with a set of decision attributes [3, 4, 21, 41, 42].

Decision System

The information system IS extended with a set of decision distributes D, such that IS = (U, A, D) and D □ A = □, is called decision system. In a decision system, the attributes in A are called conditional attributes. Decision attributes may take several values, though binary outcomes are rather frequent. Decision systems are often represented by decision tables.

Decision System

	Headache	**Muscle Pain**	**Temperature**	**Flue**
P1	No	Yes	High	Yes
P2	Yes	No	High	Yes
P3	Yes	Yes	Very High	Yes

P4	No	Yes	Normal	No
P5	Yes	No	High	No
P6	No	Yes	Very High	Yes

In decision table 3.2, we extend information table 3.1 with the decision attribute flu, i.e. D = {flu}. The value of the decision.

Indiscernibility Relation

* Let IS = (U, A) be an information system and let B is subset of A. The indiscernibility relation INDI (B) is defined as :

* INDi (B) = {(x, x') □ U^2 | for all a □ B, a(x) = a(x')}.

* If (x, x') □ IND_1 (B), then x and x' are indiscernible with respect to the attributes in B. The subscript I in IND_1 (B) is ofter omitted if it is clear which information system we have in mind. The indiscernibility relation is reflexive, i.e. an object in U is indiscernible from itself. It is also symmetric, i.e. if (x, x') □ IND (B) then (x', x) □ IND (B). Moreover, it is transitive, i.e. if (x, x') □ IND (B) and (x', x'') □ IND (B) then (x, x'') □ IND (B) relations with these characteristics are called equivalence relations. The equivalence class of an object x □ U consists o all objects y □ U such that (x, y) □ IND (B). The equivalence classes obtained from IND (B) are denoted by $[X]_B$, with x □ U from information table 3.1 we have that :

* IND ({headache}) = {{P1, P4, P6}, {P2, P3, P5}},

* IND ({muscle pain}) = {{P1, P3, P4, P6}, {P2, P5}},

* … IND ({headache, muscle pain, temperature}) = {{P1}, {P2}, {P5}, {P3}, {P4}, {P6}}.

* In the last case above, the patients P2 and P5 and P5 are indiscernible regarding all the conditional attributes. However, their values for the decision attribute are different. A decision system that has indiscernible objects with different values of the decision attributes is called inconsistent [4, 22].

References

1. [Aho96] C. Apte, S. J. Hong, "Predicting equity returns from securities data", Advances in knowledge discovery and data mining, American association for artificial intelligence, P. 541-560, 19960.
2. [Ais93] R. Agrawal, T. lmielinski, A. Swami, "Mining association rules between sets of items in large databases", Proceeding of the 1993 ACM SIGMOD conference Washington DC, USA, P. 207-216, May 1993.
3. [Ayu98] C.C. Aggarwal, P. S. Yu, "Data mining techniques for associations, clustering and classification", Lecture notes in computer science 1574, P. 13-23, 1998.
4. [Eab98] J. Elder, D. Abbott, "A comparison of leading data mining tools", Fourth international conference on knowledge discovery and data mining, New York, NY, 1998.

   http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf
5. [Fay98] U. Fayyad, "Mining databases : Towards algorithms for knowledge discovery", Bulletin of the IEEE Computer society technical committee on data engineering, Vol. 21, No. 1, P. 39-48, March 1998.
6. [Fdw96] U. Fayyad, S. G. Djorgovski, N. Weir, "Automating the analysis and cataloguing of sky surveys", Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, P. 471-493, 1996.

7. [Ffa99] C. Fertig, A. Freitas, L. Arruda, C. Kaestner, "A Fuzzy Beam-Search Rule Induction Algorithm", Lecture Notes in Artificial Intelligence 1704, Berlin, Springer, P. 314-347, 1999.

8. [Fps96] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery : An overview", Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, P. 1-34, 1996.

9. [Ggr99] M. Goebel, L. Gruenwald, "A survey of data mining and knowledge discovery software tools", SIGKDD Explorations, Vol. 1, No. 1, P. 20-33, June 1999.

10. [Gor98] B. Gray, M. Orlowska, "CCAIIA" Clustering categorical attributes into interesting association rules", Lecture notes in artificial intelligence 1394, P. 132-143, Berlin Springer, 1998.

11. [Han99] D. Hand, "Statistics and data mining intersecting disciplines", SIGKDD Explorations, Vol. 1, No. 1, P. 16-19, June 1999.

[Hec96] D. Heckerman, "Bayesian Networks for knowledge discovery", Advances in knowledge discovery and data mining.