ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 2, Ver. 1 (April - June 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

An Algorithm to Extend An Existing Function for Query Optimization in Rough Dataset

[I]Gargi Ray, [II]Dr S P Singh
[I]M.Tech Scholar of Computer Science & Engineering, BIT(MESRA), Noida Campus, Delhi NCR
[II]Asst. Professor, Dept of Computer Science & Engineering, BIT(MESRA), Noida Campus, Delhi NCR

Abstract
Uncertainty and incompleteness of knowledge is a challenging task in Information Technology. Rough Set theory due to its versatility can be applied to handle this challenge. Here Rouge Set theory is used for designing and querying Rough Relational Database (RRDB). Unlike traditional RDB, RRDB can contain multi-valued attributes and has an indiscernibility relation in its domain. Currently, the research on rough data querying, mainly is discussed some simple select-querying, i.e, selecting the tuples whose attribute's value is equal to a constant from a single table. It is perceived by using an encoding function to convert a multi valued attribute to a constant single valued attribute .The main idea of its implementation is to expand the original search conditions according to the indiscernibility relation in attribute's domain. Firstly encode the data of multi-valued attribute into the single valued data according to the indiscernibility relation in attributes' domain and then execute the query on the single valued derived data (equivalence class) to optimize the query search in much simpler and more efficient way and hence to reduce the response time. Here, in this paper the existing Encoding Function is extended to overcome a limitation of the method and discussed as Extended Encoding Function.

Keywords
Rough set theory, rough relational database, indiscernibility relation, equivalence class, encoding function, extended encoding function.

I. Introduction

As a matter of fact any industry that functions today has a lot of uncertain and imprecise data which needs to be compiled. Rough Set Theory [1, 2],which is a technique for managing the uncertainty and imperfection was introduced by Pawlak, which can analyze incomplete information effectively. A rough relational database [2] was later implemented to this effect.

Beaubouef.T [3] suggested the term of "Rough Relational Database Model (RRDM)" to describe the uncertain information system. Unlike Relational Database Model (RDM) where only atomic values of attributes are dealt with, in RRDM an attribute can be composed of one or more atomic values.

Previous Work includes:

The theory of RRDB's rough information entropy [4], the theory of rough relational operation theory [5], the theory of rough functional dependency [6] and the theory of rough data querying [7-9]. Currently, in [10-12] the studying of rough data querying mainly discussed some simple select querying for example selecting the tuples whose attributes value is equal to a constant from a single table.

In this paper, we extended the existing encoding function [10], [11-12] which is used to encode the data of multi valued attribute into the single valued data according to the indiscernibility relation in attribute's domain and then queries are executed on the encoded single valued data in simpler and more efficient way with less response time. However, if domains' indiscernibility relation changes, these encoding valued must be entirely recalculated which would not be feasible. To overcome this limitation, the existing Encoding Function is changed into Extended Encoding Function.

The remainder of this paper is organised as follows:

Section 2 describes related work. Section 3 reviews some basic concepts about rough set theory and rough relational database. Section 4 gives details of existing Encoding Function. Section 5 deals with our contribution. Section 6 discusses the experiment & result analysis. Finally, we conclude our work in Section 7 and discuss the future scopes in Section 8.

II. Related Work

Qiusheng An et al [7] proposed an approach based on decomposition principle and project principle which wastes time and storage space. Liang Ji Ye et al in [8] extended te SQL and got the results based on comparison between equivalence classes rather than values. Qiusheng An, Y Zhang, WX Zhang [9] studied the processing of rough data querying based on granular computing. They calculated the lower approximation and upper approximation of every atomic value in attribute's domain and got the final results by rough set operation principles. It needs scanning all the tuples of a table which takes a very long time. It also needs processing the semantics of the querying data. In [12] the authors discussed and applied the Covering based rough set and second type of covering based rough set theory for designing and querying of RRDB.

III. Basic Concept

A. Rough Set Theory

Definition 1:

Let U be a nonempty set of all tuples called universal set and R defines an equivalence relation on U and is called indiscernibility relation. The ordered pair A=(U,R) is called approximation space

Let $X \subseteq U$

The lower approximation of the set X is :

$R\ x = \{x \in U \mid [x]R \subseteq U\}$

This yields certain data.

The upper approximation of the set X is :

$Rx = \{x \in U \mid [x]R \cap x \neq \Phi\}$

This yields possible data.

B. Rough Relational Database Model(RRDM)

There are several common features in rough relational database and classical relational database. Both the models contain data as a collection of relations containing tuples. These relations are sets. These tuples of a relation are unordered and non duplicated.

A tuple t_i has the form $(d_{i1}, d_{i2}, d_{i3},.....d_{im})$ where dij is a domain

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

value of a particular domain set D_j. In the classical relational model $d_{ij} \in D_j$ whereas in RRDM $d_{ij} \subseteq D_j$ and d_{ij} does not have to be a singleton $d_{ij} \neq \Phi$.

C. Rough Relational Database(RRDB)

A rough relational database is defined as follows: S = (U, A, D, R)
U is the set of all tuples.
A is the attribute set.
D is the domains of attribute set.
R is the equivalence classes on D.
In RRDB an attribute $A_i \in A$, D_{Ai} is domain of A_i, R_{Ai} is the equivalence class of A_i and $r(A_i) \subseteq D_{Ai}$.
RRDB is a special kind of multi valued information system according to the definition of information system.
Definition 2:
A rough relation is a subset of the set of cross product $P(D_1)$ X $P(D_2)$ X .....XP(D_m).
Definition 3:
An interpretation $\alpha = (a_1,a_2,a_3...am)$ of a rough tuple $t_i = (d_{i1},d_{i2},...d_{im})$ is any value assignment such that $a_j \in d_{ij}$ for all $1 \leq j \leq m$, a_j is called a sub interpretation of d_{ij}.
An real life instance of RRDB is given in Table1. This is a multi valued information system in which it is given that some organisations are working in different states in different health projects in India. The Table 1 deals with two attributes 'ORG' and 'STATE'.

Table 1

RowID	ORG	STATE
Row 1	CARE, BMGF	Bihar, MP
Row 2	CDC, UNICEF, UNFPA	Rajasthan, UP
Row 3	CORE(ADRA), CORE(CRS), CORE(PCI)	UP, MP
Row 4	Plan India	Jharkhand, UP
Row 5	MCHIP	Jharkhand, Bihar
Row 6	MI (Micronutrient Initiative)	Bihar, Uttar Pradesh
Row 7	NIPI	Rajasthan
Row 8	NIPI, UNOPS	Bihar, Madhya Pradesh
Row 9	Rotary International	Chhattisgarh, Haryana, Jharkhand, West Bengal
Row 10	UNICEF	Chhattisgarh, West Bengal, Jharkhand, Madhya Pradesh, Rajasthan
Row 11	UNICEF, UNICEF (SM-Net)	UP
Row 12	UNICEF (Health Cluster), UNICEF (SM-Net), UNICEF	Bihar
Row13	USAID, BMGF	Uttar Pradesh, Bihar

The domain and equivalence classes of the attribute ORG are D_{ORG} and R_{ORG} . Accordingly the domain and equivalence classes of the attribute STATE are D_{STATE} and R_{STATE} . These are defined as follows:
D_{ORG} = { CARE, BMGF, CDC, UNICEF, UNFPA, CORE-ADRA, CORE-CRS, CORE-PCI, Plan India, MCHIP, MI (Micronutrient Initiative), NIPI, UNOPS, Rotary International, UNICEF (Health Cluster), UNICEF (SM-Net), USAID }
R_{ORG} = {[CARE, BMGF],[CDC, UNICEF, UNFPA], [CORE-ADRA, CORE-CRS, CORE-PCI], [Plan India],[MCHIP], [MI (Micronutrient Initiative)], [NIPI], [NIPI, UNOPS], [Rotary International], [UNICEF, UNICEF (SM-Net)], [UNICEF, UNICEF (Health Cluster), UNICEF (SM-Net)], [USAID, BMGF]}
D_{STATE} = { BIHAR, RAJASTHAN, UTTAR PRADESH, JHARKHAND, MADHYA PRADESH, CHHATTISGARH, HARYANA, WEST BENGAL}
RSTATE = {[BIHAR, MP], [RAJASTHAN, UP], [UP, MP], [JHARKHAND, UP], [JHARKHAND,BIHAR], [BIHAR, UTTAR PRADESH], [BIHAR, MADHYA PRADESH], [RAJASTHAN], [JHARKHAND, CHHATTISGARH, HARYANA, WEST BENGAL], [JHARKHAND, MADHYA PRADESH, CHHATTISGARH, HARYANA, WEST BENGAL, RAJASTHAN], [UTTAR RADESH, BIHAR]}

IV. The Encoding of Rough Data

A. The Existing Encoding Function

Suppose the rough relation R and a multi valued attribute $a \in A$, Da is attribute
a's domain, R_a is attribute a's equivalence class and it can be denoted that:
$D_a / R_a = \{[x]_{Ra} / x \in D_a\} = \{B_1,B_2,.....B_t\}$, and $t = | D_a / R_a |$
Definition 4:
Let ENCODE be a mapping function,
ENCODE: ENCODE (a,v) $\rightarrow$b1b2....bt
Where a is a multivalued attribute and $v \subseteq Da$
$Bi = 1$ if $x \in v \wedge x \in Bi$ otherwise bi = 0.
In order to store the encoding data, a new field is needed to add for each attribute to store the encoded values. Here, the relation schema of Table 2 is as follows :

Table 2

RowID	ORG	ORG_BIN	STATE	STATE_BIN

According to the definition 4, Table 3 can be obtained from Table 2.

B. Algorithm (Encoding Function)

Step 1: Calculate the number of equivalence classes defined for each attribute ai defined over a domain d(ai).
Step2: Assign as many number of bits as calculated in Step 1.
Eg, Considering Rstate defined above the total number of equivalence classes is 8 and hence number of bits required for the encoding function will be 8, i.e, initially all the bits will be assigned 0; i.e,00000000.
Step 3: For any given value of a tuple ti for an attribute ai, check in which equivalence class the value is present. Assign 1 to the position corresponding to the class if present else 0.
Eg, Consider Uttar Pradesh, present in 3rd equivalence classes so its encoding function will have a bit 1 at the 3rd position ,i.e, 00100000.

Step 4: Repeat the above step2 & step3 and compute the encoded value for all the attributes in Di.
Step 5: For multi valued attribute encoding value is computed by OR operation of the individual encoded value of the elements.
Eg, For {Bihar, UP}, the encoded value is get by applying OR on the encoded value of the individual elements {Bihar} & {UP}, i.e, 10100000.

Table 3

RowID	ORG	ORG_BIN	STATE	STATE_BIN
Row1	CARE, BMGF	1100000000000	Bihar, MP	10001000
Row2	CDC, UNICEF, UNFPA	0011100000000	Rajasthan, UP	01100000
Row3	CORE(ADRA), CORE(CRS), CORE(PCI)	0000010000000	UP, MP	00101000
Row4	Plan India	0000001000000	Jharkhand, Bihar	10010000
Row5	MCHIP	0000000100000	Jharkhand, UP	00110000
Row6	MI (Micronutrient Initiative)	0000000010000	Bihar, Uttar Pradesh	10100000
Row7	NIPI	0000000001000	Rajasthan	01000000
Row8	NIPI, UNOPS	0000000001100	Bihar, Madhya Pradesh	10001000
Row9	Rotary International	0000000000010	Chhattisgarh, Haryana, Jharkhand, West Bengal	00000111
Row10	UNICEF	0001000000000	Chhattisgarh, West Bengal, Jharkhand, Madhya Pradesh, Rajasthan	01011101
Row11	UNICEF, UNICEF (SM-Net)	0001000000000	UP	00100000
Row12	UNICEF (Health Cluster), UNICEF (SM-Net), UNICEF	0001000000000	Bihar	10000000
Row13	USAID, BMGF	0100000000001	Uttar Pradesh, Bihar	10100000

C. Limitations

Above discussed Encoding Function is even more useful when the total number of elements in an equivalence class set is large. However, if domain's indiscernibility relation changes, these encoding values must be entirely recalculated which would not be feasible. To overcome this limitation the existing Encoding Function is modified into an Extended Encoding Function.

V. Extended Encoding Function

In Encoding Function, the number of equivalence classes for each attribute is calculated and that number of 0s are initially assigned for the equivalence class field ("_BIN") of each attribute making them static for any change in indiscernibility relation of the RRDB. In our Extended Encoding Function, the number of bits assigned for the equivalence classes of the attribute is not fixed as follows:
Step 1: The size of equivalence classes ("_BIN") of each attribute is not assigned a fixed value rather it is kept sufficiently large.
Step2: Initially all the bits are assigned 0 for each equivalence class.
Step 3: For any given value of a tuple ti for an attribute ai, check in which equivalence class the value is present. Assign 1 to the position corresponding to the class if present else 0.
Step 4: Repeat the above step2 & step3 and compute the encoded value for all the attributes in Di.
Step 5: For multi valued attribute encoding value is computed by OR operation of the individual encoded value of the elements.

A. Advantages

If domain's indiscernibility relation changes, for example new elements are added to any attributes resulting formation of new equivalence classes there is no need for recalculation of encoding values to a sufficiently large limit.
Like Encoding Function, Extended Encoding Function is also useful when the number of elements in an equivalence class set is too large.

VI. Experiments & Results

Table 4

RowID	ORG	ORG_BIN	STATE	STATE_BIN
Row1	CARE, BMGF	1100000000000000000000000000	Bihar, MP	1000100000000000
Row2	CDC, UNICEF, UNFPA	0011100000000000000000000000	Rajasthan, UP	0110000000000000
Row3	CORE-ADRA, CORE-CRS, CORE-PCI	0000010000000111000000000000	UP. MP	0010100000000000
Row4	Plan India	0000001000000000000000000000	Jharkhand, Bihar	1001000000000000
Row5	MCHIP	0000000100000000000000000000	Jharkhand, UP	0011000000000000
Row6	MI (Micronutrient Initiative)	0000000010000000000000000000	Bihar, Uttar Pradesh	1010000000000000
Row7	NIPI	0000000001000000000000000000	Rajasthan	0100000000000000
Row8	NIPI, UNOPS	0000000001100000000000000000	Bihar, Madhya Pradesh	1000100000000000
Row9	Rotary International	0000000000010000000000000000	Chhattisgarh, Haryana, Jharkhand, West Bengal	0000011100000000

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Row10	UNICEF	0001000000000 00000000000000 00	Chhattisgarh, West Bengal, Jharkhand, Madhya Pradesh, Rajasthan,Gujrat	01011101 10000000 00
Row11	UNICEF, UNICEF (SM-Net)	0001000000000 00010000000000 00	Uttar Pradesh	00100000 00000000 00
Row12	UNICEF (Health Cluster), UNICEF (SM-Net), UNICEF	0001000000000 00011000000000 00	Bihar	10000000 00000000 00
Row13	USAID, BMGF	0100000000001 00000000000000 00	Uttar Pradesh, Bihar	10100000 00000000 00

In Table 3, three different organisations of CORE and three parts of UNICEF are assigned same equivalence class; however, they are different and work independently. So it is needed to separate then which was not possible using existing Encoded Function. Using the Extended Encoded Function, these required changes have done easily which is shown in Table 4.

A. Querying Rough Data using Encoded Value

The rough relation is composed of a low approximation or those tuples which are certain responses to the query and an upper approximation, tuples which are possible responses to the query [9]. According to that, rough data querying are commonly divided into two kinds: "Certain Data Querying" & "Possible Data Querying".

Certain data querying is that search those objects fully matching the querying condition and the querying results are obtained by the lower approximation of attribute values. And Possible data querying is rough querying and the querying results are obtained by the upper approximation of attributes values.

For example, we query RowID from Table4 whose state is UttarPradesh or UP, i.e, "Select RowID from Table4 where STATE=[UP]RSTATE ;".

Method:

1. Compute the encoded value for UP, eg, ENCODE (STATE, 'UP') = 001000000000000000.
2. To get the results of Certain Data, the following SQL is executed :
   Select RowID from Table4 where STATE_BIN=001000000000000000;
   Eg, Certain Dataset is obtained: RUP = {Row3, Row11, Row13}
3. To get the results of Possible Data, the following SQL is executed :

Select RowID from Table4 where STATE_BIN >= 001000000000000000AND STATE_BIN & 001000000000000000 = 001000000000000000;

Eg, Possible Dataset is obtained: RUP = {Row3, Row6, Row11, Row16}

VII. Conclusion

In this paper we extended the existing Encoding Function for querying rough data based on encoding. Under the circumstances of no change in the domain's indiscernibility relation, the querying rough data using Encoding Function is much simpler and more efficient, even for the equivalence class set having large number of elements. However, when domain's indiscernibility relation changes these values must be entirely recalculated which will cost much time and not feasible. Our extension over the existing Encoding Function, the Extended Encoding Function overcomes the limitation to a large extent sufficiently, efficiently in simpler way with reduced response time.

VIII. Future Scopes

Further research could be carried out to introduce incremental modification of the encoded values if the domain's indiscernibility relation changes.

References

[1] *Pawlak Z."Rough Sets".International Journal of Computer and Information science, 1982, 11(5): 341-356.*
[2] *PawlakZ. "Rough sets - theoretical aspects of reasoning about data". Dordrecht: Kluwer Academic Publishers, 1991, pp. 68-162.*
[3] *T. Beaubouef, "Uncertainty processing in a relational database model via a rough set representation". University Microfilms International, A Bell&Howell Information Company, PhD. dissertation, 1994, pp. 67-76.*
[4] *T. Beaubouef, Petry F, Aroar G. "Information theoretic measures of uncertainty for rough sets and rough relational databases". Information Science, 1998, 109:185-195.*
[5] *T. Beaubouef, F. Petry, and B. Buckles. "Extension of the relational database and its algebra with rough set techniques". Computational Intelligence, 1995, 11(2):233-245.*
[6] *Nakata M, Murai T. "Data Dependencies over Rough Relational Expressions". In: IEEE Intl. Fuzzy Systems Conf, 2001, pp. 1543- 1546.*
[7] *Qiusheng An, Guoyin Wang, Junyi Shen, Jiusheng Xu. Querying Data from RRDB Based on Rough Sets Theory. LNAI2639, Springer-Verlag, 2003, pp. 342-345.*
[8] *Fuyuan Cao, Jiye Liang."The Rough Data Query Based on SQL language",Computer Science, 2004,VOL.31No.2.*
[9] *Qiusheng An, Yusheng Zhang, Wenxiu Zhang. "The study of rough relational database based on granular computing". Granular Computing, 2005 IEEE International Conference on Granular Computing, July 2005, VOL. 1: 108~111.*
[10] *Wei, Ling-ling, Zhang, Z. " A method for rough relational database transformed into relational database", 2009 IITA International Conference on Services Science, Management and Engineering. 978-0-7695-3729-0/09, IEEE DOI 10.1109/SSME.2009.*
[11] *P Prabhavathy, BK Tripathy. " An Efficient Rough Set Approach in Querying Covering Based Relational Databases", May 2013, International Journal of Computer Science and Business Informatics(IJCSBI) - ijcsbi.org.*
[12] *Efficient Approach for Query Optimization in Rough Data S Hiremath, P Chandra - International Journal, June 2013, International Journal of Science and Research (IJSR) - ijsr.net.*

Author

I. Gargi Ray, MCA, Scholar of MTech (CS), BIT(MESRA), Noida Campus, Delhi NCR, INDIA, gargiray2010@gmail.com
II. Dr. S. P. Singh, Ph.D, M.Tech(CS), Assistant Professor, Noida Campus, Delhi NCR, INDIA, spsinghbit@bitmesra.ac.in