International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X;pISSN:2321-2241
Volume: 1; Issue: 2; June –2013

Extension of Conventional Database Query Processing

Mr. Srinivash Reddy
Associate Professor, Dept. of ISE
NMIT, Bangalore, India

Mrs. Suhasini. S
Asst. Professor Dept. of Computer Science
CIT, Bangalore, Karnataka,India

Abstract : The formulation of different data models has been evolved for efficient data retrieval. The traditional data models deal with data which is crisp and atomic. However, in real life scenario data is often ambiguous or vague or multivalued , hence NRDM i.e. Natural Relational Data Model, as an Efficient formal system for intelligent processing as well as retrieval of required information more efficiently, and which requires a set of operators and tools for manipulating data and describing their relationships.

Conventional relational databases are effective in handling crisp data in large volumes but fail to capture the real world data which is closer to human understanding. In order to realize this concept of real world data, it is required to process data which is natural and similar to the human mode of understanding of data[3]. Hence, the concept of Natural Relational Data Model i.e. NRDM which consists of real world data which is not just singular, crisp or numeric but also multivalued or interval based or multivalued with priority.

KEYWORDS
Natural Relational Algebra, degree of similarity, relational algebra, set theory,

INTRODUCTION

A query language is a communication channel through which a user requests information from the database present in the system. Query languages can be categorized into two types as Procedural or Nonprocedural query language. In a procedural query language, the user instructs the system to perform a sequence of operations on the database to gather or compute the desired result. In a nonprocedural query language, the user describes only the desired result but not the operations to be performed in order to acquire the desired result.[1]The Relational Algebra is a pure procedural query language. It is mainly used in the Relational databases to solve queries based on the users' requirements. Since the relational data model has its foundations in Set theoretic concepts, the Relational algebra includes the set theoretic approaches towards solving queries.[1]

In order to retrieve information in the Real world databases as per users' requirements, the concepts of existing conventional relational algebra can be applied. Since Real world databases consist of natural data which are not always crisp, but wherein data can be interpreted in terms of degree of similarity existing; the relational algebra operations have to be altered to work on the notion of degree of similarity. These modified concepts of Relational algebra applicable to the real world databases are collectively called '***Natural Relational Algebra*** '.

The operations are explained along with the conventional Relational algebra operations in order to enhance the modifications based on the degree of similarity which is the basis of Natural **Relational Algebra.**

The fundamental operations called ***'Select'*** and ***'Project '*** are further modified and are based on the degree of similarity existing between the user's query and the data present in the database relation given. These operations therefore require the computation of degree of similarity which is proposed in the next section. Based on the value of similarity computed, the working of the **Select** and **Project** operations are explained further.

Computation of degree of similarity $\mathbf{d_s\{ x , y\}}$***:***

For the computation of degree of similarity between any two multivalued entities X & Y, the following similarity measures were applied based on the data type of the attribute values [2].

Case 1: If X & Y are Pure multivalued entities :

Let $X = \{ x_1, x_2, x_3, x_4, \ldots\ldots\ldots\ldots x_n \}$
Let $Y = \{ y_1, y_2, y, y_4, \ldots\ldots\ldots\ldots y_n \}$

If $\{X \cap Y\} = \varnothing$

Sim (X, Y) = 0;

Else,

$$Sim(X \rightarrow Y) = \mathbf{1} - \left\{ \frac{| \{X\} - \{X \cap Y\} |}{|\{X \cup Y\} - \{X \cap Y\}|} \right\}$$

$$Sim(Y \rightarrow X) = \mathbf{1} - \left\{ \frac{| \{Y\} - \{X \cap Y\} |}{|\{X \cup Y\} - \{X \cap Y\}|} \right\}$$

Example:

Let X = {reading, writing, sports}
Y={reading,music}

$\{X \cap Y\}$ = {reading} i.e. $|\{X \cap Y\}| = 1$;
$\{X\} - \{X \cap Y\}$ = {writing, sports}
i.e. $|\{X\} - \{X \cap Y\}| = 2$;
$\{Y\} - \{X \cap Y\}$ = {music}
i.e. $|\{Y\} - \{X \cap Y\}| = 1$;

$\{X \cup Y\} - \{X \cap Y\}$ = {writing, sports, music}
i.e. $|\{X \cup Y\} - \{X \cap Y\}| = 3$
Sim (X->Y)=1-(1/3) = 0.33

Sim (Y->X)=1-(2/3) = 0.66

Case 2: If X & Y are ' Priority based multivalued' entities :

Let $X = \{ x_1, x_2, x_3, x_4, \ldots\ldots\ldots\ldots x_n \}$ and their associated weights are

$W_x = \{ w_1, w_2, w_3, w_4, \ldots\ldots\ldots\ldots w_n \}$ respectively,

Let $Y = \{ y_1, y_2, y, y_4, \ldots\ldots\ldots\ldots y_m \}$ and their associated weights are

$W_y = \{ w_1, w_2, w_3, w_4, \ldots\ldots\ldots\ldots w_m \}$ respectively.

If $\{X \cap Y\} = \varnothing$

Sim (X, Y) = 0 ;

Else,

$$Sim(X \rightarrow Y) = \mathbf{1} - \left\{ \frac{\forall a \in \{X \cap Y\} \Sigma (| w_a^X - w_a^Y |)}{\{ | X | \}} \right\}$$

$$Sim(Y \rightarrow X) = \mathbf{1} - \left\{ \frac{\forall a \in \{X \cap Y\} \Sigma (| w_a^X - w_a^Y |)}{\{ | Y | \}} \right\}$$

Example :

Let X = {painting, books, food, movies}

W_x = {1, 0.75, 0.5, 0.25} be the weightages associated with x.

Let Y= { music,painting}

W_y = {1, 0.5 } be the weightages associated with Y.

$\{X \cap Y\}$ = {painting},| {x}|= 4 and | {y} | = 2.

Sim(X-> Y) = 1- (0.5 / 4) = 0.875
Sim (Y -> X) = 1- (0.5/2) = 0.75.

The similarity values here are asymmetric since | X| != |Y| always. Further, in a realistic sense these values represent similarity between two people who have multiple favorites but with different degrees of liking.

NOTATIONS USED IN NATURAL RELATIONAL ALGEBRA

s, r → Relations .

S, R → Relation Schema.

t_r → tuple of relation r.

t^i_r → i^{th} tuple of relation r.

k → (r *op* s)
where *op* can be $\cap$, $\cup$, - , X ,etc.

A_i → i^{th} attribute of relation

$v_{Ai}(t^j_r)$ → values in the domain of the i^{th} attribute of the j^{th} tuple of relation r.

n → number of tuples in the relation.

d → number of attributes in the relation.

$d_s\{ x , y\}$ → degree of similarity between x and y values.

$d_s\{ x \rightarrow y\}$ → degree of similarity of x w.r.t y values (Asymmetric value) i.e. Sim (X-> Y)

δ → threshold value of similarity.

FUNDAMENTAL OPERATIONS

The fundamental Query operations **select** and **project,** discussed here**.** These Operations have their functions and names unaltered from the conventional relational algebra, but they are ***dependent on the degree of similarity existing between the entity value and the query*** rather than matching the crisp data values of the entities as existing in the conventional relational algebra.

THE SELECT OPERATION (σ)

The main function of the select (σ) operation is to select tuples in the specified relation say r which satisfy a given predicate 'ρ'[1]. It is denoted as

$$\sigma_\rho (r).$$

In conventional relational databases , the select operation can be realized as follows .

For example :

$$\sigma_{amt > 1000} (\text{loan}).$$

Here, all the tuples which have 'amt' attribute value greater than '1000' are selected in the relation 'loan'.

In Natural Relational algebra , The select operation can be realized in two ways :

1. Attribute value based.
2. Hypothesis based.

Attribute value based select operation:

The select operation is expressed as follows

$$\sigma^{\delta}_{\rho} (r) = \{ t / t \in r \text{ and } d_s\{ t , \rho\} \geq \delta \}$$

It implies
Select all those tuples ' t' in relation ' r' which have values which share a degree of similarity greater than or equal to 'δ' with predicate 'ρ' given.

Example : Given a relation as follows:

Age	Hobbies	Experience(yrs)
24	R,M, A	2-3
26	R, Sc	1-2
23	R, N , M, A	0-1
31	R, N .	3-4

Query:

Select entities with hobbies as { R, M,N } with similarity greater than 0.5 (or 50 %) .

The tuples retrieved along with their degrees of similarity are

Age	Hobbies	Experience(yrs)	d_s
23	R,N,M, A	0-1	1
24	R, M,A	2-3	0.66
31	R, N ,	3-4	0.66

2. Hypothesis based Select operation:

Given any attribute, there may be a set of general terms associated which can be derived from building hypothesis.

For example, if attribute is 'salary ', the generic terms can be (Low, medium, high).

The hypothesis can be

H1 (salary –high) : 8000 ≤ value of salary ≤ any maximum value .

H2(salary-medium) : 4000 ≤ value of salary < 8000.

H3(salary –low) : any minimum value ≤ value of salary < 4000.

Given a relation say employee-info,if
The query is

'Select employees with low salary'.

Here the tuples retrieved will have their 'salary' attribute values sharing similarity with hypothesis H3 as mentioned previously.

i.e. all tuples 't' in the relation 'r' with salary value in the range of 4,000 and 8,000 will be retrieved.
Note: Salary attribute can be crisp numeric or interval valued.

Hence, given a hypothesis H (t) on the tuple't',

The select operation is denoted as follows

$$\sigma^{\delta}{}_{\rho}\,(r) = \{\, t \;/\; t \in r \text{ and } d_s\{H(t), t\} \geq \delta \,\}.$$

THE PROJECT OPERATION (Π)

The main function of the project (Π) operation is to list the values present in the domain of the mentioned attributes A_1, A_2, A_3,…. A_n in the specified relation say r[1]. It is denoted as

$$\Pi_{A1, A2, A3, \ldots An}(r).$$

In conventional relational databases, the project operation can be realized as follows.
For example:

$$\Pi_{\text{name, amt}}\ \textbf{(loan).}$$

Here, all the 'names' of the people along with their respective 'amt' attribute values are projected / listed as existing in the loan relation. Duplicate data values are eliminated in project operation.

In Natural Relational Algebra, the project operation is denoted as

$$\Pi^{\delta}_{Ai1}(r) = \begin{cases} v_{Ai1}(t^i_r) & \text{where } t^i_r \in r \text{ and } d_s\{t^i_r, t^j_r\} < \delta \\ v_{Ai1}(t^i_r) \cup v_{Ai1}(t^j_r) & \text{otherwise,} \\ & \text{Where } 1 \leq i,j \leq n \text{ and } i \neq j \end{cases}$$

It implies, Project values of attribute A_{i1} of relation (r) with δ being the threshold value of similarity.

This can be achieved in 2 ways

1. If degree of similarity between the i^{th} tuple and j^{th} tuple A_{i1} attribute Values are less than $_{\delta}$, then list the values of i^{th} tuple attribute.

2. If 2 different tuples say t^i_r and $t^j{}_r$ have a similarity greater than or equal to δ ,then their attribute values can be united as one set to eliminate redundant data Values.

Example : Given a relation 'student-info '

Students	Groups	Terms
Anil	a,b,c	x,y
Vijay	b,c	y
Sunil	d,e,f	x,y
Madhu	d,e,f	y,z
Sushma	a,b,c	x,y

Query:

$$\Pi^{1}_{\text{groups, terms}}\ \textbf{(student-info)}$$

Retrieved tuples are :

Groups	Terms
a,b,c	x,y
b,c	y
d,e,f	y,z

Note: since δ is given as 1, exact copies are eliminated i.e. tuples t3 and t5 are eliminated.

CONCLUSION

The research work proposed in this paper aims at extending the conventional Relational algebra theory into the field of NRDM i.e. **Natural Relational Data Model**. The work therefore aims at exploring the issues of designing of operations on the NRDM based on the degrees of similarity similar to the working methods in conventional database theory. The research work is aimed at defining the relational algebra operations such as ***Select and Project*** which is the two most fundamental operations required to evaluate and solve most of the queries performed on data in the NRDM. The modified version of these operations is based on the standard definitions and operation of the existing conventional relational algebra theory.

The crux of this research work lies in designing the operators mentioned above which are an extension of the conventional relational algebra operations. These concepts are referred as Natural Relational Algebra Operations. The design of these operations is dependent on the similarity values between the entities or between the data values and the query, and is based on the corresponding similarity measures designed and presented in the research paper [2].

FUTURE SCOPE

The research work will further be extended to cover all the different relational algebra operations both fundamental and extended operations which will result in formulation of an extended and complete relational algebra theory for NRDM called the Natural Relational Algebra. These concepts will therefore serve as a basis for designing a query language to process and retrieve data similar to the human understanding and query processing from the NRDM i.e. Natural Relational Data Model.

REFERENCES

[1]Abraham Silberschatz, Henry.F.Korth, S.Sudarshan- *Database System concepts*, Fifth edition. McGraw Hill International Edition, 2006.

[2] Suhasini.M and Nagalakshmi.H.S - *Similarity Measures For Real world Data Mining*, proceedings of the 3rd National conference on computing for nation development, pg 687 -690, Indiacom-2009.

[3] Jain A.K. Murthy M.N and Flynn P.J, 1999, "Data Clustering: A review", ACM computing surveys, Vol 31, No 3, No 3. pp 264-324.

[4] Duda.R.O,Hart.P.E and Stork D.G. 2000, Pattern Classification, John Wiley & Sons, Edition No 2,pp550-554