ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 519- 525 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

A Survey on gap between SQL and DMQL

S. POOJITA[1], P.THRIDHARA[2]

[1]Student, ASN Women's Engineering College, Tenali, sunkarapoojitarao@gmail.com

[2]Student, ASN Women's Engineering College, Tenali, , thridhara95.popuri@gmail.com

ABSTRACT

An important motivation for the development of databases is proactive use of the information significantly improve the quality of their decision making and profitability of the organization through focused actions. Query languages like SQL and DMQL plays vital role in retrieving the data from different data sources. In this paper we compare existing data mining query languages, all extensions of the standard relational query language SQL, from this point of view: how flexible are they with respect to the tasks they can be used for, and how easily can those tasks be performed? We verify whether and how these languages can be used to perform three prototypical data mining tasks in the domain of association rule mining, Clustering and Classification. We also evaluated functional gap between the SQL and DMQL.

KEYWORDS

Association, Classification, Clustering, Database, Data warehouse, DMQL, Retrieval, SQL, Transaction.

INTRODUCTION

An important motivation for the development databases is to be keepdata safe to use it later is to be stored. The data is stored in a database or a data warehouse. Database is defined as a structured set of data that is accessible in various ways. The data in the database is generally stored in form of records in tables. Each table is given a unique name through which the table is accessed. The data in the table is organized using the Primary key. The Primary Key refers to a column in the table that is unique. It used for the purposes of indexing the table which makes it much more efficient in accessing. Data warehouse is a database used to store data. It is a central repository of data extracted from various sources. The data warehouse is then used for reporting and data analysis. When both database and data warehouse are used for the same purpose then there must be some differences between them. To compare with a database is used to store data while a data warehouse is mostly used to facilitate reporting and analysis. A database stores the current data whereas data warehouse stores historical data too. A database is mostly used for Online Transactional Processing while data warehouse is used for Online Analytical Processing.

LEVELS OF DATA MANAGED IN THE ORGANISATION

Typical organization, manages three different kinds of data namely, operational data, historical data and informational data. Operational Data refers to the day to day data manipulating in the organization which is updatable, historical data is different from operational data, defines the previously committed data which is permanent and not allow any updations, final category of data is informational data, it is like historical data and used to make decisions in the organization.

DATA MODELLING AND METHODLOGIES

Data Model describes about how the data logically implemented and physically stored, often called as Design process. When we start database design the first thing to analyze is the nature of the application you are designing for, is it Transactional or Analytical. A Transactional**, i**n this kind of application, end user is more interested in creating, reading, updating, and deleting records we call these kinds of databases as OLTP.Other category of application is Analytical; **i**n these kinds of applications end user is more interested in analysis, reporting, forecasting, etc. These kinds of databases have a less number of inserts and updates. The main intention here is to fetch and analyze data as fast as possible. We refer this kind of database as OLAP.

The data can be implemented in two different formats like Two-Dimensional Structures and Multi-Dimensional Structures. These were defined with the help of different rules like Constraints, Normalization principles. Two-dimensional structures typically called as Tables and Views used in OLTP. OLAP uses the three Dimensional structures represented in terms of Cubes also referred as Materialized View, a materialized view is also like snapshot of database.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 519- 525 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

The following figures fig: 1 and fig 2, shows the representation of data in Two-Dimensional and Three Dimensional Methods

SNO	PNO	JNO	QTY
S1	P1	J1	100
S1	P2	J1	200
S2	P1	J1	100
S2	P2	J2	100

Fig 1: 2- Dimensional Representations

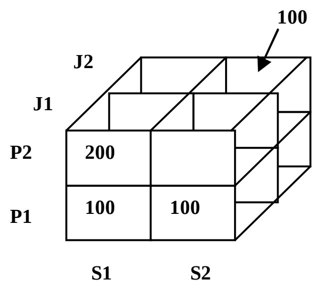

Fig 2: 3- Dimensional Representations

RETRIEVAL OF DATA

Retrieval of data from the knowledge sources varies with respect to kind of knowledge, we are fetching, like general purpose or analytical. To fetch the information from the source we may use any one of the two methods called SQL or DMQL.

Data that is stored in a database is retrieved using SQL a Structured Query Language. SQL a declarative static Structured Query Language is a combination of data definition language and data manipulation language and is used for both accessing and modifying the information in a database. The primitives for SQL in the form of a query, which specifies the following

- The set of task –relevant data to be fetched
- Kinds of Data to fetched.

The following figure fig 3 shows the process involved in retrieval process for SQL

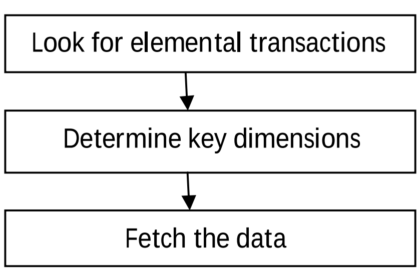

Fig 3: Retrieval Process in SQL

General syntax for SQL is as follows

SELECT <ATTRIBUTE LIST>
FROM <SOURCES>
WHERE <CONDITION>
GROUP BY <LIST OF ATTRIBUTES>
HAVING <CONDITION>
ORDER BY <LIST OF ATTRIBUTES> [ASC/DESC];

Data that is mined in prior is retrieved using DMQL, dynamic Data Mining Query Language used to mine data from larger databases and allows the ad hoc mining of several kinds of knowledge data from various relational data bases and data warehouses at multiple abstraction levels. A data mining query language provides necessary primitives that allow users to communicate with data mining systems. The primitives for defining a data mining task in the form of a data mining query. The primitives specify the following:

- The set of task-relevant data to be mined
- The kind of knowledge to be mined
- The background to be mined
- The background knowledge to be used in the discovery process
- The interestingness measures and thresholds for pattern evolution
- The expected representation for visualizing the discovered patterns

The following figure fig 4 shows the retrieval process by using DMQL

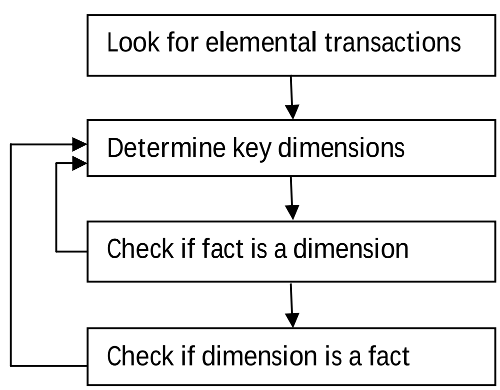

Fig 4 : Retrieval process in DMQL

General Syntax for DMQL as follows

Use database (database_name) | use data warehouse(data_warehouse_name)

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 519- 525 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

```
|<use hierarchy (hierarchy_name>for
(attribute_or_dimension)}

<Mine_Knowledge_Specification>
In relevance to {attribute_or_dimension_list}

From (relation (s) /cube(s))
[Where (condition)]
[Order by (order list)]
[Group by (grouping_list)]
[Having (condition)]
[With [(interest_measure_name)]
threshold={threshold_value)

[For (attribute(s))]
(Mine_Knowledge_Specification)::=(Mine_Char)
|(Mine_Discr) |( Mine_Assoc) |(Mine_Class)
{Mine_Char)::= mine characteristics [as
(pattern_name)]
Analyze {measure(s))
<Mine_Discr>::= mine comparison
[as(pattern_name)]
For (target class) where (tarhet_condition)
(Verses (contrast_class_i)
(contrast_condition_i)}
Analyze (measure(s))
{Minc_Assoc}::= mine associations
[as{pattern_name}]
[Matching {met pattern)]
{Mine_Class}::= mine classification
[as{pattern_name}]
Analyze {classifying_attribute_or_dimension)
(Concept_Hierarchy _Dennition_Statement)::=
Define hierarchy (hierarchy_name)
[for (attribute_or_dimension}]
On (relation _or_cube_or_hierarchy)
As (hierarchy_description)
[Where (condition}]
{Visualization_and_Presentation}:=
Display as (result_form) |
{MultUevel_Manipulation)}
(Multilevel _Manipulation)::= roll up on
(attribute_or_dimension}
| drill down on (attribute_or_dimension)
| add (attribute_or_dimension)
| Drop (attribute _or_dimension)
In relevance to (attribute_or_dimensionlist):
attributes or dimensions
Order by (order list):
Group by (grouping_list): (Grouping Attribute)
Having (condition):<Condition>
```

FUNCTIONAL GAP BETWEEN SQL AND DMQL

It is usually assumed that standard query languages such as SQL will not suffice for this; and indeed, SQL offers no functionality for, for instance, the discovery of frequent item sets. We can overcome this shortcoming by using Data Mining Query Language. DMQL can helpful in determining "Characterisation", "Discovery", "Discrimination", "Association", "Clustering" and "Classification". An other limitation for SQL is representation of data, by SQL we will report the fetched information as a report only, but by using DMQL, we can represent it as "Pivot Table", "Chart" or "Cubes".

ILLUSTRATION

To illustrate the functional gap between the SQL and DMQL, we extracted the table called diabetic database extracted from the UCL machine learning. The dataset contains 768 record samples, each having 8 attributes. We used this dataset for our classification exercise, as the data is complete.

The following Table 1 illustrates the description about the Dataset used.

S.No	Attribute	Type
1	Number of times pregnant	Continuous
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Continuous
3	Diastolic blood pressure (mm Hg)	Continuous
4	Triceps skin fold thickness (mm)	Continuous
5	2-Hour serum insulin (mu U/ml)	Continuous
6	Body mass index (kg/m)^2)	Continuous
7	BMI type	Discrete
8	Diabetes pedigree function	Continuous
9	Age (years)	Continuous
10	Class Variable(0,1)	Discrete

The performances of the SQL and DMQL were examined by two different factors like levels of data retrieved and represented. To represent these features, we used the method called discrimination. To illustrate the results, we used software's like MS-Access for SQL and Tanagra for Data Mining Query Language (DMQL).

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 519- 525 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

The following SQL query represents fetching of information about the different levels of HBA1C Index values with relevance to the attributes Glucose, BP and BMI.

SELECT DISTINCTROW Sheet1.[HBA1C INDEX],count(Sheet1.gloucose) as Glucose, count(Sheet1.bp) as BP, count(Sheet1.bmi) as BMI

FROM Sheet1

GROUP BY Sheet1.[HBA1C INDEX];

The following table 2 describes number of people struggle with relative factors.

Table 2: List fetched through the SQL Query

Sheet1 Query			
HBA1C INDEX	Glucose	BP	BMI
Diabetic and Good Control	32	32	32
High Stage of Diabetic	175	175	175

Sheet1 Query			
HBA1C INDEX	Glucose	BP	BMI
No Diabetic	439	439	439
Very High Diabetic	122	122	122

We define the association between the HBA1C Index with Glucose, BP and BMI Levels with minimum support 33% and Confidence 100% as following manner

Use DIABETIES
Mine Characteristics As HBA1C INDEX
In Relevance GLOUCOSE, BP, BMI
From DIANETIC_SOURCE
GROUP BY HBA1C INDEX

Analyze (measure(s))
With SUPPORT = 33 AND CONFEDENCE =100

Table 3: Information from DMQL (Association (Support 33% and Confidence 100%)
(A: Attribute, B: Test Value, C:Group, D: Over All)

Results

Description of "HBA1C INDEX"

HBA1C INDEX=High Stage of Diabetic			
Examples		[22.8 %] 175	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	0.54	138.21 (8.30)	120.89 (31.97)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	0.20	33.55 (7.41)	31.99 (7.88)

HBA1C INDEX=No Diabetic			
Examples		[57.2 %] 439	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Diastolic blood pressure (mm Hg)	-0.11	67.00 (18.67)	69.11 (19.36)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	-0.16	30.74 (7.95)	31.99 (7.88)
Plasma glucose concentration a 2	-0.54	103.76 (12.94)	120.89 (31.97)

HBA1C INDEX=Very High Diabetic			
Examples		[15.9 %] 122	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	1.67	174.30 (12.98)	120.89 (31.97)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	0.40	35.11 (6.90)	31.99 (7.88)

HBA1C INDEX=Diabetic and Good Control			
Examples		[4.2 %] 32	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Diastolic blood pressure (mm Hg)	-0.27	63.94 (20.59)	69.11 (19.36)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	-0.41	28.80 (8.09)	31.99 (7.88)
Plasma glucose concentr	-1.98	57.69 (26.	120.89 (31.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 519- 525 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Diastolic blood pressure (mm Hg)	0.14	71.83 (21.26)	69.11 (19.36)
Discrete attributes : [Recall] Accuracy			

hours in an oral glucose tolerance test			
Discrete attributes : [Recall] Accuracy			

Diastolic blood pressure (mm Hg)	0.26	74.12 (17.18)	69.11 (19.36)
Discrete attributes : [Recall] Accuracy			

ation a 2 hours in an oral glucose tolerance test		19)	97)
Discrete attributes : [Recall] Accuracy			

Similarly, we define patterns for clustering and classification as follows

Table4: Information from DMQL (K Mean Clustering with Number of Clusters: 3)
(A: Attribute, B: Test Value, C:Group, D: Over All)

Results

Description of "HBA1C INDEX"

HBA1C INDEX=High Stage of Diabetic			
Examples		[22.8 %] 175	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	8.15	138.21 (8.30)	120.89 (31.97)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	2.97	33.55 (7.41)	31.99 (7.88)
Diastolic blood pressure (mm Hg)	2.12	71.83 (21.26)	69.11 (19.36)
Discrete attributes : [Recall] Accuracy			

HBA1C INDEX=No Diabetic			
Examples		[57.2 %] 439	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Diastolic blood pressure (mm Hg)	-3.48	67.00 (18.67)	69.11 (19.36)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	-5.09	30.74 (7.95)	31.99 (7.88)
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	-17.15	103.76 (12.94)	120.89 (31.97)
Discrete attributes : [Recall] Accuracy			

HBA1C INDEX=Very High Diabetic			
Examples		[15.9 %] 122	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	20.10	174.30 (12.98)	120.89 (31.97)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	4.76	35.11 (6.90)	31.99 (7.88)
Diastolic blood pressure (mm Hg)	3.12	74.12 (17.18)	69.11 (19.36)
Discrete attributes : [Recall] Accuracy			

HBA1C INDEX=Diabetic and Good Control			
Examples		[4.2 %] 32	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Diastolic blood pressure (mm Hg)	-1.54	63.94 (20.59)	69.11 (19.36)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	-2.34	28.80 (8.09)	31.99 (7.88)
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	-11.42	57.69 (26.19)	120.89 (31.97)
Discrete attributes : [Recall] Accuracy			

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 519- 525 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Table5: Information from DMQL (Classification)
(A: Attribute, B: Test Value, C:Group, D: Over All)

Group characterization 1 (CS-RT)

Results

Description of "HBA1C INDEX"

HBA1C INDEX=High Stage of Diabetic			
Examples		[22.8 %] 175	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	8.15	138.21 (8.30)	120.89 (31.97)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	2.97	33.55 (7.41)	31.99 (7.88)
Diastolic blood pressure (mm Hg)	2.12	71.83 (21.26)	69.11 (19.36)
Discrete attributes : [Recall] Accuracy			

HBA1C INDEX=No Diabetic			
Examples		[57.2 %] 439	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Diastolic blood pressure (mm Hg)	-3.48	67.00 (18.67)	69.11 (19.36)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	-5.09	30.74 (7.95)	31.99 (7.88)
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	-17.15	103.76 (12.94)	120.89 (31.97)
Discrete attributes : [Recall] Accuracy			

HBA1C INDEX=Very High Diabetic			
Examples		[15.9 %] 122	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	20.10	174.30 (12.98)	120.89 (31.97)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	4.76	35.11 (6.90)	31.99 (7.88)
Diastolic blood pressure (mm Hg)	3.12	74.12 (17.18)	69.11 (19.36)
Discrete attributes : [Recall] Accuracy			

HBA1C INDEX=Diabetic and Good Control			
Examples		[4.2 %] 32	
A	B	C	D
Continuous attributes : Mean (StdDev)			
Diastolic blood pressure (mm Hg)	-1.54	63.94 (20.59)	69.11 (19.36)
Body mass index (weight in kg/(height in m)^2) insulin (mu U/ml)	-2.34	28.80 (8.09)	31.99 (7.88)
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	-11.42	57.69 (26.19)	120.89 (31.97)
Discrete attributes : [Recall] Accuracy			

Form the Tables 2 through 5, we observe the information was analyzes widely in DMQL as compared with SQL.

CONCLUSION

The Structured Query Language and Data Mining Language were commonly used mechanisms for informational retrieval from the data source. The SQL has the limitation in terms of static and restricted representation. But in the case of DMQL analysed through the three data mining primitives called Clustering, Association and Classification we found DMQL is very flexible with respect to the tasks they can be used for, and these tasks can easily performed.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 519- 525 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

REFERENCES

[1].A Practical Comparative Study Of Data Mining Query Languages, Hendrik Blockeel, Toon Calders, ´Elisa Fromont, Bart Goethals, Adriana, Prado, and C´eline Robardet

[2] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Elsevier

[3] Jean-Francois Boulicautand Cyrille Masson, Data Mining Query Languages, Data Mining and Knowledge Discovery Handbook 2nd ed, Springer, 2010

[4 J. Han, Y. Fu, W. Wang, K. Koperski, and O. Zaiane. DMQL: a Data Mining query languagefor relational databases. In R. Ng, editor,Proc. ACM SIGMOD Workshop DMKD'96,Montreal, Canada, 1996.

[5] R. Meo, G. Psaila, and S. Ceri. An extension to SQL for mining association rules.Data Mining and Knowledge Discovery, 2(2):195–224, 1998.

[6] L. De Raedt, M. Jaeger, S. Lee, and H. annila. A theory of inductive query answering. In Proc. IEEE ICDM'02, pages 123–130, 2002.