

# DATA MINING ANALYSIS TO DRAW UP DATA SETS BASED ON AGGREGATIONS

**S.V.D.S Divya<sup>1</sup>, Sk. Nagul<sup>2</sup>**

*<sup>1</sup>M.Tech (CSE) Scholar, <sup>2</sup>Assistant Professor*

*Nalanda Institute of Engg and Tech. (NIET), Siddharth Nagar, Guntur, A.P, (India)*

## **ABSTRACT**

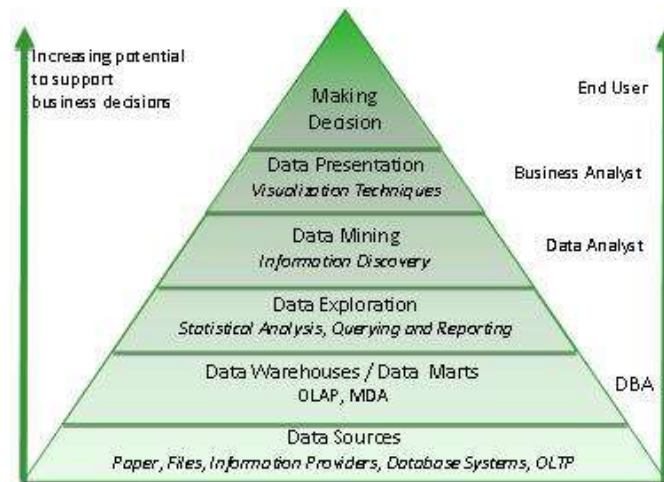
*Data mining which plays a very important role in day to day life and data mining purely depends on the data set. Data sets are those which help the application or user to sort out the data and use only that related one from the large data. For creating datasets we need to use complex query in sql which in turn increases load on the database with the connection. Here we are trying to retrieve the data in horizontal row layout which actually returns set of data instead of only a specific number, which is termed as horizontal aggregation. In our proposed work we study about the different data mining queries and their efficiency for creating a dataset that will be helpful to refine the data. The data mining queries are CASE, SPJ and PIVOT. All these three have different roles and responsibilities in their way of data mining. Various experimental setups are done using these data mining queries for the horizontal aggregated data in sql.*

**Keywords:** *Aggregation, CASE, SPJ, PIVOT, Data Mining.*

## **I. INTRODUCTION**

Data Mining is a very familiar concept in computer technology which helps user to get relevant data from the raw data. It means that by the implementation of a data mining algorithm we can mine the useful data from very huge data. As like Data Mining there is one more term called as Deep Web. Deep Web is also known as Deepnet which is not at all visible to the user; it is the World Wide Web content that is not part of the surface web. Deep web is the term which tells that the normal user can't reach the root of it i.e. not all the users are able to get the details from specific search engine only the authorized users are able to access all the details. For example, some of the universities, some of government agencies and other organizations maintain the databases of information that were not created for the general public access. Other sites which may restrict the database access to members or subscribers. Deep web mining though it is not known for many of the users but they have come across all these types of websites, the websites which will give access to their data only after a genuine authentication. For example, we here have a website "Slideshare.net" which will give data only after you verify yourself as authorized user and only after this check that user will get data, point that to be taken into consideration is that here we need not pay any amount to that organization for accessing the data. And at the same time we have Google search engine which will never ask a user to get registered for the data accessing from their database. We here need to understand a point here which when it is from Google we cannot get any data it's just a reference tool that helps user to navigate to his destiny from the Google search engine generated result page. For example, whenever we type for searching some content from Google it will give us number of links of other domains which may maintain the data of our search. But when we compare this with

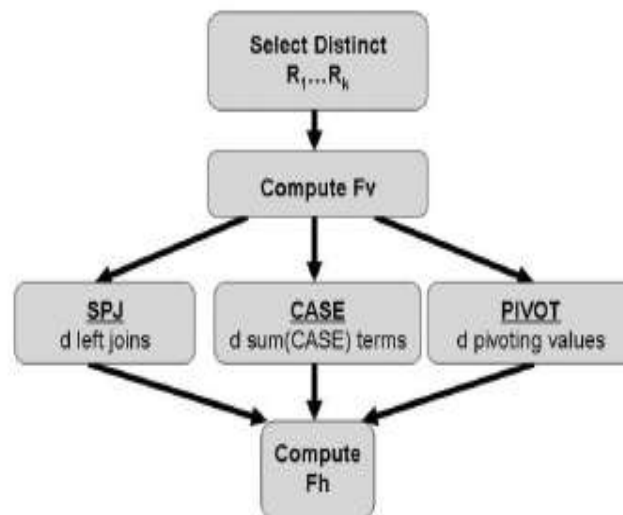
Slideshare.net, here the user will search for his required data and that domain will get the results that are closely matching to the need in this domain itself but this will not give the links of other domains from where a user can get the data required.



**Fig.1 Data Mining Process**

As we can see in the above image, data mining will involve many steps for getting the accurate results from a very complex data. Querying will be done for the purpose of fetching the relevant data from the database only by applying many queries i.e. connecting to that database many times where we can fetch the data which we actually needed. In this process of fetching data that is what we required we are connecting to database many times and thereby reducing the efficiency of the application. In any of the real time application, the connections that are made to the database must be minimized to the extent possible by the developer. To overcome this drawback of connecting to database too many times for the normal queries we propose three types of data mining queries in this paper They are: CASE, PIVOT and SPJ (Select Project Join). In general, the CASE and the PIVOT will exhibit linear scalability, whereas the SPJ is not that scalable when compared with the CASE and PIVOT. There are many methods that can be used to extract the data from the database but here in this paper we proposed a new technique called Horizontal Aggregation. This is the technique that displays data in horizontal tabular format with all the required set of columns for a better understanding. The horizontal aggregations will be extended form of the SQL aggregations which will return a horizontal tabular row instead of a single value. The advantages of the proposed system are, it will represent a template that can generate a SQL code from data mining tool used. This sql code will completely reduce the manual work that is to be made in the data presentation phase of data mining process. Generated sql code will be more efficient as it will automatically be generated rather than the sql queries that are written number of times to fetch the data. Datasets will be created in very easy and fast manner taking the help of these queries.

Whenever we want to have horizontal aggregations, we can use other techniques that are of sql database i.e. taking the help of order by or group by clauses. The problem that can be seen is using these clauses is that the user needs to connect to database table containing the required data number of times until the requirement is fulfilled. The output will remain same if we can utilize the technology or not but whenever it is seen from a resource allocation point of view, it will be a total damage. That is the main reason why a user who has to perform the manipulations on various tables that are available in the database will do it using data mining queries rather than that of using the plain simple sql queries. Whenever the data mining queries will be applied the application will effectively use the resources and will create the data sets.



**Fig. 2 Architecture of Data Mining Based On Aggregation**

As we see in the above block diagram that process initially will take the data in the form of a request which we can generally call as user request. Later this data or request which will come in from of the user is manipulated and then accordingly all the data sets are created based upon the available methods i.e. CASE, SPJ and PIVOT. Here we consider a small example for understanding the implementation process in very easier manner. Consider a small locality with one sub-station office that will be serving near about a 100 houses. So now if we consider it as application wise, there will be only one admin who will take care of this complete locality through online. And once at the end of every year, the admin who wanted to know all the details of a specific customer and in the backend all those details that are of the customer that are segregated. It becomes difficult for fetching the data at a time, so here the developer may need to write too many queries depending upon the user requirement. Here to overcome drawback of writing many queries, the data mining queries are employed which will be in turn create datasets. The data will be specifically not correspond to admin; and it could be the user who also actually wants to know their yearlong status in a very simple manner.

## II. RELATED WORK

Here in our implementation we will be given an option to select the fields from available options and based upon user selection, application will generate the code in sql depending upon data mining view that is selected by the user. We here have three views that are made available to the user i.e. CASE, PIVOT and SPJ. CASE is operator that will be utilized for the sql queries; similarly SPJ and PIVOT are used to fetch the results as per the definitions given by the operators. One thing is very clear between a normal query and data mining operator employed query i.e. is that the query size differs. Queries with a normal operator like select, update, delete are small and also queries involving the keywords order by will not have much effect in the size. Only when these operators belonging to the aggregation the size of the query may be long but the result fetched is comparatively far better and so reduces the overhead process on the database thereby increasing the efficiency of this application and giving appropriate results as that expected by the user.

CASE is the operator that can be used with all the combinations i.e. it can be used with a select query, can be used with order by clause, can be used with an update statement, can be used with a set statement, and can also be used in the query with having clause. SQL query is written taking the order by operator into count and gives

the result in the sorted manner as written by the programmer putting user in a perspective. We also have other operator which works similar to that of the order by i.e. group by, this operator groups all the items belonging to the similar kind and thereby making it easy for the user to understand the process of with the output shown. The query appears in the below form when the group by the operator is used,

## 2.1 Select Name from Register Group by Users

In above written query we can see that the is query clearly makes us understand the intension behind it means to say that the user here is trying to group the names from the register table. Grouping here is done only because of the operator used in the query. In this here case if we want to modify the search and want to produce the results those are to be shown in a systematic manner then here we need to employ order by operator. So with this we can understand the usage of the various operators in the sql queries. The case operator can be used and this operator involves condition checking while fetching the data from the database.

Data mining queries are implemented in our proposed work are written below,

To understand the implementation, here consider the example of same electricity department.

CASE Query,

```
“select meterid,sum(case when month='jan' then paidamt else null end) jan,sum(case when month='feb' then paidamt else null end ) feb,sum(case when month='mar' then paidamt else null end ) mar,sum(case when month='apr' then paidamt else null end ) apr,sum(case when month='may' then paidamt else null end ) may,sum(case when month='jun' then paidamt else null end ) jun,sum(case when month='jul' then paidamt else null end)jul,sum(case when month='aug' then paidamt else null end)aug,sum(case when month='sep' then paidamt else null end)sep,sum(case when month='oct' then paidamt else null end)oct,sum(case when month='nov' then paidamt else null end)nov,sum(case when month='dec' then paidamt else null end ) dec from tblpayment group by meterid”
```

The above written query where we can see that case query is here involving various condition checks and at last we can see that the result is segregated using the keyword group by.

In Oracle 10g XE, the Pivot operator is not directly made available for the users. And Pivot in actual says that fetching the relevant data from the very complex data. This functionality will be shown in the implementation taking the help of sql queries. And in the new version of Oracle i.e. Oracle 11g, pivot operator is made available for the users to mine the data in an efficient manner.

And coming with the spj in SQL, it also helps users input data from more than one table and also to display results as required by the third party to have a simplified look and moreover gives better results to the end user. For better understanding consider the below written query to mine the data from two different tables for better understanding.

```
“select registration2.meterid ,tblpayment.month,updatebills.paidamt from registration2 left outer join tblpayment on registration2.meterid=tblpayment.meterid left outer join updatebills on registration2.meterid=updatebills.meterid”
```

Here in the above written query, registration2 and tbln payment are also two different tables from which the data is retrieved to make user understand the complex data in a very easy manner. It can be understood that the first table is a kind of reference where the user details are taken and based on that, payment details are fetched to

have a clear idea on a specific meter id. From the above written queries it is very clear that the operation is very simple but here the queries written are complex. It must be understood that using the data mining queries the user will connect to database only once and can perform the action but if it is a no data mining operation, here user will connect to database number of times and perform the action.

Drawback of this is that there is a possibility of the connection to reset with database when ever more number of times we connect to database and this could be a very big damage to organization as their clients will not be able to retrieve the information they actually wanted. So here each and every developer here needs to take care about the resources for being utilized as the application need to be flexible and user friendly. Moreover here it must support any type of action.

### III. CONCLUSIONS

Our proposed model will give a better reliability and performance when is compared to the normal SQL operations. Thus we here can say that the aggregations in sql helps the application to be more and more flexible for the user input as it will connect the database at a single instance and will create datasets which will give the information to user about the complex data in a very easy manner. Aggregations in database studied in this work are CASE, SPJ and PIVOT; all three have their own significance in their way.

### IV. ACKNOWLEDGMENT



The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

### REFERENCES

- [1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
- [2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
- [3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
- [4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.
- [5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
- [6] (2002) The IEEE website. [Online]. Available: <http://www.ieee.org/>
- [7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: <http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/>
- [8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
- [9] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
- [10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

## AUTHOR PROFILE

	<p><b>S.V.D.S Divya</b> is currently pursuing M.Tech in the Department of Computer Science &amp; Engineering, from Nalanda Institute of Technology (NIT), siddharth Nagar, Kantepudi(V), Sattenapalli (M), Guntur (D), Andhra Pradesh , Affiliated to JNTU-KAKINADA.</p>
	<p><b>Shaik Nagul</b> working as Assistant Professor at Nalanda Institute of Technology (NIT), siddharth Nagar, Kantepudi(V), Sattenapalli (M), Guntur (D), Andhra Pradesh , Affiliated to JNTU-KAKINADA.</p>