Market Basket Analysis Using Horizontal Aggregations in SQL

S.Yahya Ali Khan
M.Tech Student, Department of CSE,
Dr.K.V.S.R.I.T, Kurnool.

K. Pavan Kumar
Assistant Professor, Department of IT,
Dr.K.V.S.R.I.T, Kurnool.

Abstract

Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information. A data set is a collection of data presented in tabular form. Efficient analysis of data can be achieved by preparing a data set with columns in horizontal tabular layout. Preparing a data set is a more difficult task in data mining project as it requires many SQL queries, joining tables and aggregating columns. Traditional RDBMS manage tables with vertical format and returns one number per row. Horizontal aggregations returns a set of numbers per row with Horizontal layout as required in most of the data mining algorithms. Evaluation of horizontal aggregations is done with three methods. The methods are CASE: Exploiting the programming CASE construct; SPJ: Based on standard relational algebra operators (SPJ Queries); PIVOT: Using the PIVOT operator which is offered by some DBMSs. The data obtained from horizontal aggregations can be used for Market Basket Analysis in finding frequent item set mining by using Analysis Services of SQL Server.

Keywords— Aggregation, Data preparation, Market Basket Analysis , Pivoting, SQL

1. Introduction

Data mining is the process of extracting knowledge from large amount of data. Preparing a suitable data for data mining purposes is a time consuming task as it requires complex SQL queries, joining tables and aggregating columns [5]. Aggregation is normally associated with data reduction in relational databases. The aggregate functions available in SQL are MIN,MAX,AVG,SUM and COUNT. All these functions returns a single number as output. This is called vertical aggregation. The output of vertical aggregations is helpful in calculation. Most of the data mining operations require a data set with horizontal layout with many tuples and one variable or dimension per column. This is the case with many data mining algorithms like PCA, regression, classification, and clustering [4],[7] .

In a relational database, especially with normalized tables a significant effort is required to prepare a summary data set [8]. Every research area uses different terminology to describe a data set. In data mining the common terms are point-dimension. Statistics literature uses observation variable. Machine learning research uses instance-feature. We are introducing a new class of aggregate functions that can be used to prepare data sets in a horizontal layout for automating SQL query writing and extending SQL capabilities in this paper. To evaluate horizontal aggregations we are using three methods CASE, SPJ and PIVOT. The output obtained from horizontal aggregations is applied to Market Basket Analysis (MBA). MBA used to analyze the customer purchasing patterns by extracting associations or co-occurrences and helps in increasing the sales and maintain inventory by focusing on the point of sale transaction data. MBA process analyzes customer buying habits by finding associations between the different items that customers place in their shopping baskets.

This paper is organized as follows. The Related Works are present in Section II. Section III describes the proposed method for analyzing data sets using Market Basket Analysis , Section IV describes the Result and Analysis and Section V describes Conclusions and Future Work

2. Related Works

C.Cunningham developed two operators. They are PIVOT and UNPIVOT. The pivot and Unpivot operators are useful to transpose and transform data sets for data mining and OLAP tasks. They can quite easily be implemented inside a query processor, much like select, project, and join. Such a design provides opportunities for better performance, both during query optimization and query execution. Pivot is an extension of Group By with unique restrictions and optimization opportunities, and this makes it very easy to introduce incrementally on top of existing grouping implementations. [1].

C.Ordonez introduced two aggregation functions. These functions are vertical aggregations and horizontal aggregations. Vertical aggregations return one row for each percentage in vertical form like standard SQL aggregations. Horizontal aggregations returns each set of percentages adding 100% on the same row in horizontal layout. Experiments study different percentage query optimization strategies and compare evaluation time of percentage queries. [6]

Horizontal aggregations are capable of producing data sets that are used for data mining activities. This paper presents three horizontal aggregations methods CASE, PIVOT and SPJ. CASE is based on the SQL CASE construct, PIVOT makes use of built in pivoting facility in SQL while SPJ uses standard SQL aggregations.

Data mining is motivated by the decision support problem faced by most large retail organizations. A record in basket data typically consists of the transaction date and the items bought in the transaction. Market basket analysis helps us to discover which group of items tends to be purchased together by customers. Today the data sizes in the datasets of market basket have increased from gigabytes to terabytes or even larger due to which the complexity of analysis of huge datasets has been a major concern in almost all areas of technology in the past decade [11].

3. Proposed Method

Horizontal aggregations propose a new class of functions that aggregate numeric expressions and the result are transposed to produce data sets with a horizontal layout. Horizontal aggregations represents an extended form of SQL, returns a set of values in horizontal tabular layout. Horizontal Aggregation is evaluated using three methods.

A. *Case Method*

For this method we use the CASE programming construct available in SQL. The case statement returns a value selected from a set of values based on boolean expressions. From a relational database theory point of view this is equivalent to doing a simple projection/aggregation query where each monkey value is given by a function that returns a number based on some conjunction of conditions. We propose two basic sub-strategies to compute F_H. In a similar manner to SPJ, the first one directly aggregates from F and the second one computes the vertical aggregation in a temporary table F_V and then horizontal aggregations are indirectly computed from $F_{V.}$

B. *SPJ Method*

The SPJ method is based on relational operators only. The basic idea is to create one table with a vertical aggregation for each result column, and then join all those tables to produce F_H. We aggregate from F into d projected tables with d Select- Project-Join-Aggregation queries (selection, projection, join, aggregation). Each table F_I corresponds to one sub grouping combination and has {L_1, . . ., Lj} as primary key and an aggregation on A as the only non-key column. It is necessary to introduce an additional table F_0, that will be outer joined with projected tables to get a complete result set. We propose two basic sub-strategies to compute F_H. The first one directly aggregates from F. The second one computes the equivalent vertical aggregation in a temporary table F_V grouping by L_1, . . ., L_j, R_1, . . .,R_k. Then horizontal aggregations can be instead computed from F_V, which is a compressed version of F, since standard aggregations are distributive.

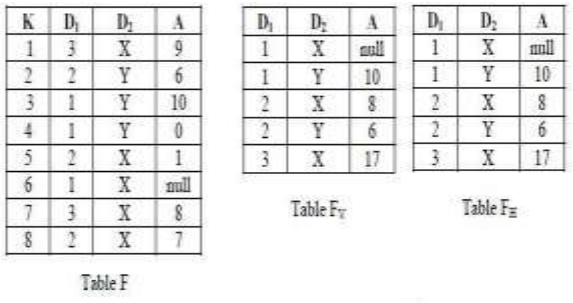
In a horizontal aggregation there are four input parameters to generate SQL code (i) the input table F (ii) the list of GROUP BY columns L_1, …… ,L_j (iii) the column to aggregate (A) and (iv) the list of transposing columns R_1, … ,R_k. We extend standard SQL aggregate functions with a transposing BY clause followed by a list of columns (i.e. R_1, … ,R_k) to produce a horizontal set of numbers instead of one number.

Proposed syntax is as follows.

SELECT (L_1, …… ,L_j), H(A BY R_1, … ,R_k)
FROM F
GROUP BY (L_1, …… ,$L_{j)}$;

C. *PIVOT Method*

PIVOT operator which is a built-in operator in a commercial DBMS. Since this operator can perform transposition it can help in evaluation horizontal aggregations. The PIVOT method internally needs to determine how many columns are needed to store the transposed table and it can be combined with the GROUP BY clause.



K	D_1	D_2	A
1	3	X	9
2	2	Y	6
3	1	Y	10
4	1	Y	0
5	2	X	1
6	1	X	null
7	3	X	8
8	2	X	7

Table F

D_1	D_2	A
1	X	null
1	Y	10
2	X	8
2	Y	6
3	X	17

Table F_V

D_1	D_2	A
1	X	null
1	Y	10
2	X	8
2	Y	6
3	X	17

Table F_H

Fig 1. Vertical Aggregation F_V and Horizontal Aggregation F_H

D. Performance Analysis of CASE, SPJ and PIVOT

CASE method has similar speed to the PIVOT operator and it is much faster than the SPJ method. In general, the CASE and PIVOT methods exhibit linear scalability, whereas the SPJ method does not.

E. Market Basket Analysis

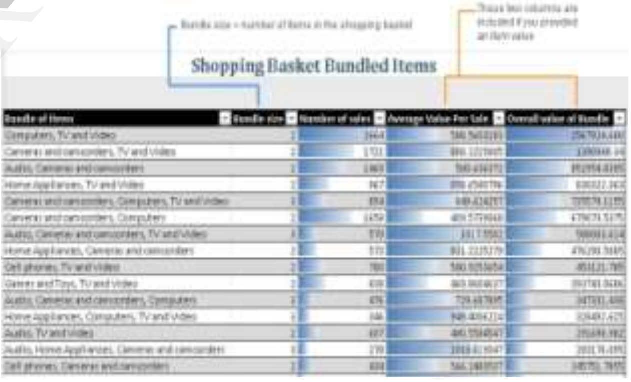
Market basket analysis requires the analysis and mining of large volumes of transaction data for making business decisions. If customers are buying milk, how likely is that they also buy bread? Such rules help retailers to plan the shelf space: by placing milk close to bread they may increase the sales provide advertisements/recommendation to customers that are likely to buy some products put items that are likely to be bought together on discount, in order to increase the sales. Market Basket Analysis can be done by using Shopping Basket Analysis Tool of SQL Server Analysis services. The Shopping Basket Analysis tool helps you to find associations in your data. Associations can be used for tasks as analyzing products that are frequently purchased together. Based on the results, we can also recommend and promote related products. For example, an online store can use Shopping Basket Analysis to make recommendations such as, "People who bought this product frequently bought these products also."

Data for the Shopping Basket Analysis needs to be an Association dataset. To use shopping basket analysis, the items that you want to analyze must be related by some sort of transaction ID. For example, if you are analyzing all the orders received through a Website, each order would have an order ID or transaction ID that is associated with one or more purchased items. So you would choose the Order ID as the transaction ID-NOT the customer ID or product ID-because you want to analyze the associations that are found within orders. The Shopping Basket Analysis tool creates two complementary reports.

Shopping Basket Bundled Items - Lists all the item sets that were found.

Shopping Basket Recommendations - Lists the rules, or inferences about products that belong together, that can be made based on the data.

The list of item sets is useful for exploring your data, while the list of rules is more useful for making predictions and recommendations. The worksheet contains a list of the items that frequently appear together in transactions. The worksheet also contains statistics to help you understand the significance of the results. We recommend that you add a price column to the analysis, if the data is available, because it provides the sum of the value of all the related items, which is helpful in understanding the value of the transactions.



Bundle size = number of items in the shopping basket

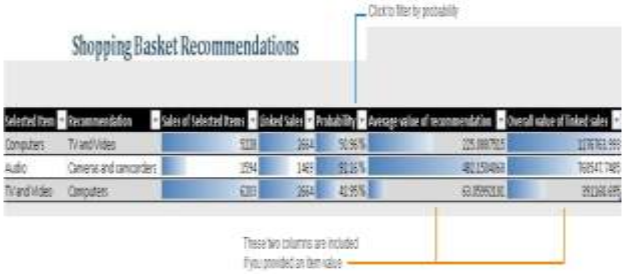
These two columns are included if you provided an item value

Shopping Basket Bundled Items

Bundle of Items	Bundle size	Number of sales	Average Value Per Sale	Overall value of Bundle
Computers, TV and Video	2	1664	[illegible]	[illegible]
Cameras and camcorders, TV and Video	2	1721	[illegible]	[illegible]
Audio, Cameras and camcorders	2	[illegible]	[illegible]	[illegible]
Home Appliances, TV and Video	2	967	[illegible]	[illegible]
Cameras and camcorders, Computers, TV and Video	3	854	[illegible]	[illegible]
Cameras and camcorders, Computers	2	[illegible]	[illegible]	[illegible]
Audio, Cameras and camcorders, TV and Video	3	[illegible]	[illegible]	[illegible]
Home Appliances, Cameras and camcorders	2	[illegible]	[illegible]	[illegible]
Cell phones, TV and Video	2	[illegible]	[illegible]	[illegible]
Games and Toys, TV and Video	2	[illegible]	[illegible]	[illegible]
Audio, Cameras and camcorders, Computers	3	476	[illegible]	[illegible]
Home Appliances, Computers, TV and Video	3	[illegible]	[illegible]	[illegible]
Audio, TV and Video	2	[illegible]	[illegible]	[illegible]
Audio, Home Appliances, Cameras and camcorders	3	[illegible]	[illegible]	[illegible]
Cell phones, Cameras and camcorders	2	[illegible]	[illegible]	[illegible]

Fig 2. Shopping Basket Bundled Items Report

The Shopping Basket Recommendations Report shows recommendations that can be made based on the analysis. Recommendations are based on rules derived from the source data, and are ordered in the report based on probability. If the recommendation has a high probability score, it means that the items were frequently purchased together, so it makes sense to recommend the related items to someone who put just one of the items in their shopping basket.
You can filter and sort on the columns in the report. For example, you can filter out item sets that don't have at least 3 products, or you can order the item sets by their value, or you can filter out recommendations that have a lower probability score.



Shopping Basket Recommendations

Selected Item	Recommendation	Sales of Selected Items	Linked Sales	Probability	Average value of recommendation	Overall value of linked sales
Computers	TV and Video	5228	2664	50.96%	225.0887025	1178763.993
Audio	Cameras and camcorders	1594	1463	91.16%	481.1504868	768547.7485
TV and Video	Computers	6203	2664	42.95%	63.0595101	293166.695

Click to filter by probability
These two columns are included
if you provided an item value

4. Result and Analysis

Market Basket Analysis done with the help of Shopping Basket Analysis Tool for Horizontal Aggregations in SQL is helpful in analysing frequent patterns, associations, correlations, or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories. Fig 2 and Fig 3 shows the two reports Shopping Basked Bundled Report and Shopping Basket Recommendations Report.

5. Conclusion and Future Work

The data obtained from horizontal aggregations is analyzed with the help of Shopping Basket Analysis Tool and reports are generated. Horizontal Aggregations can be extended for Association Rules by applying Apriori Algorithm.

6. References

[1] C. Cunningham, G. Graefe, and C.A. Galindo-Legeria, "PIVOT AND UNPIVOT: Optimization and Execution Strategies in an RDBMS," Proc: 13th Int'l Conf. Very Large Data Bases (VLDS'04), pp.998-1009, 2004.

[2] G. Graefe, U. Fayyed, and S. Chaudhuri, "On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases, "Proc. ACM Conf. Knowledge Discovery and Data Mining (KDD'98), pp. 204-208, 1998.

[3] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh, "Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab and Sub-Total, "Proc: Int'l Conf. Data Eng., pp. 152- 159, 1996.

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Second edition. Morgan Kaufmann, 2006.

[5] C. Ordonez and Z. Chen. Horizontal aggregations in SQL to prepare data sets for data mining analysis.IEEE Transactions on Knowledge and Data Engineering (TKDE), 24(4), 2012.

[6] C. Ordonez, "Vertical and Horizontal Percentage Aggregations," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD'04), pp. 866-871,2004.

[7] C. Ordonez, "Statistical Model Computation with UDFs," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 12, pp. 1752-1765, Dec. 2010.

[8] C. Ordonez, "Data Set Preprocessing and Transformation in a Database System," Intelligent Data Analysis, vol. 15, no. 4, pp. 613-631, 2011.

[9] C. Ordonez and S. Pitchaimalai, "Bayesian Classifiers Programmed in SQL," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 1, pp. 139-144, Jan. 2010.

[10] S. Sarawagi, S.Thomas, and R. Agrawal, "Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications, "Proc ACM SIGMOD Int'l Conf. Mnagement of Data (SIGMOD '98) , pp. 343-354, 1998.

[11] R.Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. ternational Conference on Very Large Databases (VLDB). 1994.

[12] Jamie MacLennan, ZhaoHui Tang and Bogdan Crivat, Data Mining with SQL Server 2008 Wiley Publishing, Inc, 2009.

About the Authors

S.Yahya Ali Khan, recieved his M.C.A. degree from Osmania University, Hyderabad, India in the year 1995. He is currently pursuing M.Tech in Computer Science and Engineering from Dr. K.V.S.R.I.T, Kurnool, India. His research interests include Data Mining, Software Engineering and Cloud Computing.

K.Pavan Kumar, received his B.Tech degree in Computer Science and Engineering from JNTU, Hyderabad, India in the year 2006 and M.Tech in Computer Science from JNTU Hyderabad, India, in the year 2009. He is currently working as a Assistant Professor at Dr.K.V.S.R.I.T, Kurnool, India. His research interests include Mobile Ad-hoc Networks, Computer Networks and Cloud Computing.