Contemporary Engineering Sciences, Vol. 6, 2013, no. 8, 393 - 402
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2013.3843

New Roll-Up Operator for Non-Additive Numeric Measure Summarization

Anas Jebreen Atyeeh Husain

Information Systems Department
Al Al-Bayt University, Mafraq, Jordan
anasjh@aabu.edu.jo

Abstract

Summarization or aggregation is a process of computing the measures in the data cube. Summarization plays an important role in the data analysis and decision making for data mining applications. However, the potential of inaccurate summarization that could result from using improper operators may lead to inaccurate measure values that affect data analysis and the decision making process.

Efficient computation of measures and an accurate summarization process have become an important requirement in order to obtain useful results that best support the decision making process. To this purpose, a new operator for summarization based on linear goal programming is proposed. The proposed operator computes new measure values with minimum distance to its related true values in order to maintain data quality and find higher accuracy measures. Higher accuracy measures reflect more useful analysis for users and improves decision making.

The ability of the proposed operator to achieve higher accuracy when performing roll-up operations and compute new measure values that best reflect its correspondent original data values is measured and compared with average, minimum, and

maximum aggregation operators. The evaluation results demonstrated that the goal programming operator performs better than the others and produces more accurate measure values by achieving a lower distance to the related data values. We conclude that the proposed operator is able to improve the summarization process that best supports decision making by providing higher data quality to the OLAP users.

Keywords: Summarization, Measures, Roll-up, Accuracy of summarization

1.0 Introduction

Data warehousing and On-Line Analytical Processing (OLAP) are essential elements of decision support [2,15]. Decision support requires consolidation (e.g., aggregation and summarization) of data for better and faster high level analysis and decision making [5]. Summarization or aggregation is a process of computing the measures in a multidimensional data structure called a data cube which uses summarization operators. In a multidimensional model, *dimension* such as location, describes the subjects of interest; *measure* such as dollars sold, is the target of analysis in terms of dimensions, *concept hierarchies* such as city and street may exist for each dimension, allowing the analysis of data at multiple abstraction levels [8]. Concepts are defined hierarchically starting from the most general concept and ending with the most specific concept (e.g. country> state> city >street for location dimension).

Measures are multidimensional summarized information that is stored in a data cube. A measure value is computed for a given point by summarizing or aggregating the data corresponding to the respective dimension values defining the given point [5]. Many OLAP operations can compute measures in different ways to enable OLAP users (e.g. manager, executive, or analyst) [16] to analyze data at different abstraction levels. Roll-up operation is the most frequent analytical operation seen in a data warehouse [12]. The roll-up operation performs summarization on a data cube by *climbing up a concept hierarchies* for a dimension. A roll-up operation decreases the details of measures using an operator to combine more detailed data values of a specific concept into summary and more general data of more general concept[10]. For example, the daily sales data may be summarized so as to compute monthly or annual total amounts.

The process of measure computation plays an important role in the data analysis and decision making for data mining applications [9]. However, the potential for inaccurate summarization that could result from using improper operators may lead to inaccurate measure values which negatively affect data analysis and decision making.

Efficient computation of measures and the accurate summarization process have become an important requirement in order to obtain useful results that best support the analysis and decision making process [13].

The concern in this paper is the accurate summarization of non-additive measure values and therefore, we will concentrate on roll-up operations. The non-additive measure values are the data which cannot be meaningfully added with others using a summation operator [6]. The problem can be encapsulated in the following question:

- How to summarize non-additive data in such a way that attains higher accurate measures when performing a roll-up operation?

Accuracy of a measure is how close a newly computed measure value is to the actual (true) data values [7]. Thus, the newly computed measure values should be as close as possible to the actual data values, giving rise to the need for methods that minimize this difference. To this purpose, a new operator for summarization based on linear goal programming (GP) is proposed. The proposed operator computes the new measure values with minimum distance to its related true values to maintain data quality and finding higher accuracy measures. Higher accuracy measures reflect more useful analysis for OLAP users and improve decision making.

The remainder of this paper is organized as follows: The related works are reviewed in Section 2. Section 3 presents the requirements and detailed design of the proposed solution. Performance evaluation and results is introduced in Section 4. Finally, some conclusions are given in Section 5.

2.0 Related Works

The function of the aggregation and summarization operators is to approximate the computation of measures which plays a central role in data warehousing and data mining issues[5]. Several classical operators are oriented to give a kind of summary for data values. The most common operator used to summarize measures is the summation operator and it can be considered the default aggregate operator for rolling-up the data [6]. However, in many instances, using the sum operator to summarize data cannot be meaningfully applied due to the non-additive nature of data. Temperature, blood pressure and car speed are examples of data values that cannot be handled using summation operators.

Because it is often useful to store non-additive data in a data warehouse, another class of operators to approximate the computation of non-additive measures exists.

Operators like averages, minimum, and maximum [1][5] can be meaningfully applied to summarize such non-additive data values. For example; average blood pressure of the sample of patients with same medical history and between ages of 40 and 50 comes out to 140/110.

However, the summarization process plays an important role in maintaining data quality that supports a better decision making process. Therefore, it makes sense for such an application to seek higher accuracy results and compute the closest measure values to its related true data values. Therefore, methods that minimize the difference between measure values and its related true data values are needed. Indeed, the concept of minimizing of such difference when computing measures is not considered by the existing operators and the issue of accuracy and data quality is neglected. Finding higher accuracy measures reflect higher data quality and are a more useful analysis for OLAP users.

Therefore, a new aggregation operator for a roll-up operation is proposed based on the linear goal programming. The purpose of the operator is to compute new measure values that best represent its corresponding data values when performing a roll-up operation by minimizing the difference among them. Indeed, the proposed GP operator seeks the measure values with the minimum distance to its related true values.

GP is a form of linear programming for multiple goals developed by Charnes and Cooper [3]. It is a multi-criteria satisfying methodology that seeks a solution that best fits or satisfies the desired set of multi-criteria in a problem situation. GP can and has been used to model different problems from business like economics, finance, management, and marketing, and operations such as industry [11]. The GP problem formulations basically consist of three elements: goal constraints, an objective function, and other requirements. The constraints are the goals and objectives in the problem. The solution selection will be based on the deviations from these constraints. Therefore, the objective function is to minimize the overachievement and underachievement of all constraints.

3.0 The Proposed Operator

Suppose that a specific dimension D_i consists of concept hierarchies such as $D_i=C_{i1},C_{i2},\ldots..C_{im}$ where the C_{im} is the most specific concept , $C_{i(m-1)}$is more general concept and C_{i1} is the most general concept. Each concept hierarchy consists of *x* data

values such as $C_j = V_{j1}, V_{j2}, \ldots . V_{jx}$. Data values of C_m can be grouped to *n* number of sets of related data values according to the different values of C_{m-1}. Each set consist of *k* related values V_{mrh} that can be summarized to compute a more general (higher) concept value $V_{(m-1)r}$ where $h=1,\ldots,k$ and $r=1,\ldots,n$. Accordingly , data can be summarized when performing a roll-up operation by applying an operator to each set of related data values of a specific concept hierarchy to produce a new and higher general concept value and so forth for all hierarchies in all dimensions. Figure 1 shows the proposed GP operator that is responsible for such a process by climbing up the concept hierarchy and moving from C_m to C_{m-1} computing new measure values.

$$\text{Minimize } T = \sum_{r=1}^{n}\sum_{h=1}^{k} d^{+}_{mrh} + d^{-}_{mrh}$$

$$\text{Subject to } \sum_{r=1}^{n}\sum_{h=1}^{k} V_{(m-1)r} - d^{+}_{mrh} + d^{-}_{mrh} = V_{mrh}$$

$$r = 1,\ldots, n;\ h = 1,\ldots, k;$$

$$V, d^{+}, d^{-} \geq 0$$

Figure 1

The *n* values $V_{(m-1)r}$ for each dimension are the decision variables of the problem which are the expected output of the GP operator and represent the new (higher level) measure values. The GP operator seeks an equal or closest measure value $V_{(m-1)r}$ to each related value V_{mrh}, which serves as the goal constraints and is presented in the second line. The d_{mrh+} and d_{mrh-} are deviational overachievement and underachievement variables, respectively. The objective function is shown in first line of the Figure which is to minimize the gap or distance between each new measure values $V_{(m-1)r}$ and each it's related values V_{mrh} to achieve higher accuracy results**.** The gap is represented by the overachievement ($d_{mrh}+$) and underachievement ($d_{mrh}-$) variables.

4.0 Performance Evaluation and Result

The accuracy of summarization and measure computation for non-additive numerical data values when performing a roll-up operation is an issue in this paper. Therefore, the objective is to compute the closest measure values to its related true data values in order to maintain data quality that leads to better analysis and decision making. Thus, the ability of the proposed GP operator to achieve higher accuracy when performing a roll-up operation and to produce summarized measures that best reflect the original data values need to be measured and compared with the other common aggregation operators namely: average, minimum, and maximum. The average operator can compute measure values through the following equation:

$$\sum_{r=1}^{n}\sum_{h=1}^{k} V_{(m-1)r} = (V_{mrh}) / k$$

The accuracy of summarized measures can be computed by calculating the distance between the obtained measure values and its related data values for p dimensions as shown in Eq.(1). Distance is the gap or deviation degree of a selected value among a set of related values. The less the distance produced means higher accuracy achieved.

$$\textit{Distance of measure values to its related true values} = \sum_{i=1}^{p}\sum_{r=1}^{n}\sum_{h=1}^{k} V_{i(m-1)r} - V_{imrh} \qquad \textbf{(1)}$$

Where p is the number of the dimensions, n is the number of sets for the most specific concept hierarchy m for each dimension, and k is the number of related values in each set. Since the applications of data mining are widely different in many terms (e.g. number of dimensions or concept hierarchies), three different scenarios with different parameters are proposed and used in the evaluation test as shown in Table 1. Each scenario consists of a different number of data values that need to be summarized according the number of dimensions and concept hierarchies for each dimension.

For example, daily temperature needs to be rolled-up as a monthly temperature which requires an operator to be applied to a large data set that contains data value for each day in the year. Thus the number of data values will be 365 for each year.

Table 1: Parameters of accuracy tests

Scenario number	Number of dimensions	Number of data values corresponding to each dimension	Total number of data values
1	3	200	600
2	4	250	1000
3	5	400	2000

We ran out tests by applying the GP operator to each set of related data values to find a new higher level measure value that best represents its correspondent data values. LINGO solver [14] is an application that can run GP models and is used to run the proposed operator in these tests. For comparison purposes, average, minimum, and maximum operators are used and applied to same data values. The performance results for all operators in all scenarios are present in Table 2 and figure 2.

Table 2: distance of measure values to its related true values

Scenario	AVG	MIN	MAX	GP
1	14734.99	21856.96	36943.04	13945.7
2	27985.42	57785.62	41214.38	25962.66
3	115613.6	239080	150920	108739.3
Average	52778.02	106240.9	76359.15	49549.23

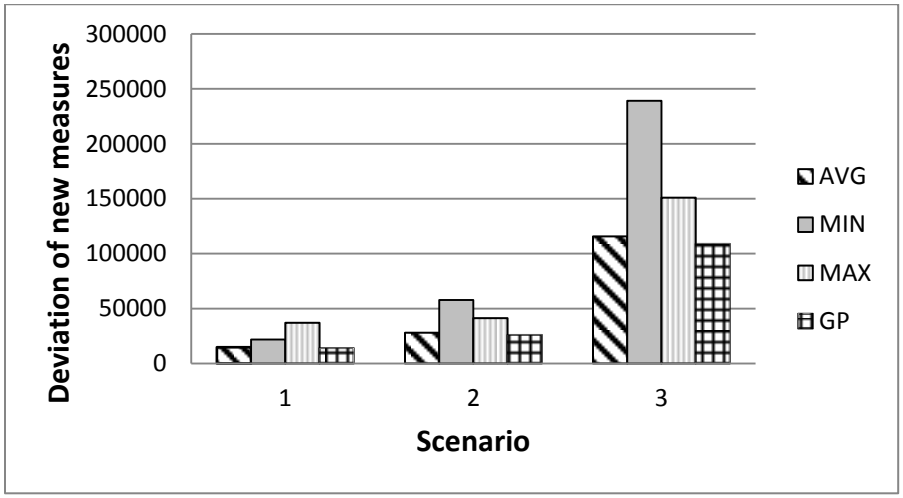

Figure 2

The results showed that the GP operator performs better than the other operators and produces more accurate measure values by achieving a lower distance to the related data values. Using the efficiency standard equation [4], the efficiency of GP operator over average operator was 6.11 % and over maximum operator was 35.11% and over minimum operator was 53.36% for all scenarios.

Moreover, the results showed that the accuracy of the computed measures could be lower when summarizing larger numbers of data values and the need for accurate summarization when performing a roll-up operation to produce higher quality results increased for applications that have a large number of data values to be summarized. Nevertheless, the GP operator performs better than the others in such cases and produces more accurate measure values by achieving a lower distance to the related true data values. Making decisions for such problems would be better when using GP operator for summarization purpose.

5.0 Conclusion and Future Work

Accurate summarization of non-additive numerical data values when performing roll-up operations have been addressed in this paper. A new operator for summarization based on linear goal programming is proposed to maintain data quality and improve the decision making process. The GP operator is able to compute more accurate measure values that best represents its correspondent lower level data values, especially when summarizing larger numbers of data values.

In conclusion, higher accuracy was achieved when using the proposed GP operator. Comparatively, in other related operators, there was no data quality criterion used or the distance to the true values considered when computing higher level measures. Indeed, computing higher level measure values when performing roll-up operations using the GP operator attains higher accuracy of data than the other related approaches that based on average, minimum, and maximum operators. Different numbers of data values are used and the GP performs better in all scenarios.

The proposed GP operator seeks to find the closest measure value to its related lower level data values for each dimension, and so forth for all dimensions. On the other hand, the other related method does not consider the distance or deviation between the data values and the computed measure values. Minimizing this distance resulted in higher data quality in terms of accuracy which reflects a better analysis and decision making process. In future work, this solution may be extended to cover the additive numerical data values, and to support other data mining operations.

References

[1] A. Hassan, F. Ravat, O. Teste, R. Tournier and G. Zurfluh, Differentiated multiple aggregations in multidimensional databases, 14th International Conference on Data Warehousing and Knowledge Discovery, Vienna, 2012, 93-104 .

[2] A. Tripathy, K. Das, A descriptive approach towards *data* warehouse and OLAP technology: An overview , communications in computer and information science, 2nd international conference on advances in communication, network, and computing– springer, 2011, 409-411.

[3] A. Charms, W.W.Cooper, Management Models and Industrial Applications of Linear Programming, John Wiley & Sons, New York, 1961.

[4] A.J. Husain, Utility based policy management for virtual organization using goad programming and bee behavior, PHD thesis, 2010.

[5] J. Han, M. Kamber, Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann Publishers, San Francisco, 2011

[6] J. Horner, I.Y. Song and P.P. Chen, An Analysis of Additivity in OLAP Systems, Proceedings of the 7th ACM international workshop on Data warehousing and OLAP, 2004,83-91.

[7] JCGM 200:2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM).

[8] L. Chou, X. Zhang, Computing complex iceberg cubes by multiway aggregation and bounding, Data Warehousing and Knowledge Discovery: Sixth International Conference, 2004, 108-117.

[9] L. Geng and H. J. Hamilton, Interestingness measures for data mining: A survey. ACM Computing Surveys, 38(3), (2006).

[10] M. Ceci, A. Cuzzocrea and D. Malerba, Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering, Journal of Intelligent Information Systems, (2013), 1-25.

[11] M. J. Schniederjans, J. L. Hamaker and A.M. Schniederjans, Information Technology Investment Decision Making Methodology, World Scientific Publishing, River Edge, USA , 2004.

[12] O. Baltzer, F. Dehne, A. Rau-Chaplin, OLAP for moving object data, International Journal of Intelligent Information and Database Systems, 7, (2013), Pages 79-112

[13] R. Tiwari , M.P. Singh, Correlation-based Attribute Selection using Genetic Algorithm, International Journal of Computer Applications, 4(2010) ,28–34,

[14] Schrage, L.Optimization Modeling with Lingo, Lindo Systems Inc., 2008.

[15] V. Sreenivasarao, V.S. Pallamreddy, Advanced data warehousing techniques for analysis, interpretation and decision support of scientific data , communications in computer and information science, 1st international conference on advances in computing and information technology-springer, 2011, 162-174 .

[16] Z. Nimg, Concept hierarchy-based cube aggregation for ETL process in Matriculation warehouse, 14th WSEAS International Conference on Computers, Greece, 2010, 510-515.

Received: August 5, 2013