I.J. Intelligent Systems and Applications, **2017, 10, 49-56**
Published Online October 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.10.06

Formal Validation of Data Warehouse Complexity Metrics using Distance Framework

Gargi Aggarwal
Information Technology Department, NSIT, New Delhi, India
E-mail: agg.gargi07@gmail.com

Sangeeta Sabharwal
Computer Science and Engineering Department, NSIT, New Delhi, India
E-mail: ssab63@gmail.com

Received: 07 February 2017; Accepted: 08 March 2017; Published: 08 October 2017

Abstract—Data Warehouse is the cornerstone for organizations that base their strategic decisions on the large scale processing of numerical data. The success of the organization depends on these decisions and hence it becomes extremely important to have a quality data warehouse. Conceptual models have been widely recognized as a key determinant of data warehouse quality during the early stages of design. Recently, metrics have been proposed by authors based on hierarchies to quantify the complexity and inturn quality of the conceptual models of data warehouse. They have formally corroborated the measures against Briand's property based framework to ensure their validity. However, Briand's set of properties for software measures are a set of necessary but not sufficient measure axioms. They are advantageous to refute software metrics but not to validate them. Thus, we focus on the theoretical validation of the data warehouse conceptual model metrics using the Distance framework whose sufficiency is ensured by the measurement theory. The results indicate that the metrics are valid measures of the complexity of data warehouse conceptual models. Besides, validation by Distance framework assures that the metrics are in the ratio scale which further aids in data analysis.

***Index Terms*—**Distance framework, metrics, theoretical validation, data warehouse quality, multidimensional models.

I. Introduction

Data warehouses (DW) store huge amounts of data integrated from one or more disparate sources [1]. They store historical and current data and are thus instrumental in supporting the management of any organization in making informed decisions. They are the backbone of decision making and strategy building for any organization. They are the key enabler for exploring past trends, predicting future business ecosystem and allowing businesses to modify, adapt and evolve. Thus, they provide a vying edge to the organizations and hence need to be very exhaustive.

Due to the ever growing intricacies of the data warehouses [1], it is paramount to ensure that their quality is given apt importance throughout the development process. This helps in avoiding issues at a later stage when it would be difficult and expensive to implement any change.

Researchers have proposed various development methodologies [2] to ensure data warehouse quality. However, development methodologies alone cannot guarantee DW quality. They need to be reinforced with processes and metrics. Metrics act as objective indicators of quality. They assist the data warehouse designers in making a choice among semantically identical schemas. Several metrics have been proposed [3] which seem to impact the understandability and efficiency of the multidimensional model of a data warehouse.

In order to have a valid set of metrics, they have to be corroborated both theoretically and empirically [4]. Theoretical validation makes certain that the metrics, indeed, quantify the quality attribute they purport to measure. It ensures their construct validity. However, theoretical validation alone does not imply the overall validity of the metrics. They need to be validated empirically as well. Empirical validation establishes the practical usefulness of the metrics. It assures that the metrics are related to some external attribute. However, since the formal validation of measures ensures their internal validity, we consider it as an essential step before the empirical validation takes place. Hence, in this paper we focus on the formal corroboration of data warehouse hierarchy measures.

Formal validation of software measures has followed two approaches as there is not yet a standard technique to formally corroborate the measures. The two paths are as under:

- Axiomatic approaches or Property based approaches as proposed by Briand et al. [5] and Weyuker [6] – These frameworks have put forth a set of properties that characterize software attributes and hence can be used to classify the software measures.

- Approaches based on Measurement Theory as proposed by Poels and Dedene [7] and Zuse [8] – These frameworks determine the scale of the measures and based on the scale, the transformations and statistical operations can be applied on the metrics.

In this paper, we have focussed on the Distance framework [7] based on measurement theory in order to corroborate the data warehouse metrics proposed by Gosain et al. [9]. The authors had formally corroborated the measures using Briand's property based framework [5] and determined that the hierarchy metrics fulfil the entire property set that characterise them as either length or size measures. However, according to Briand et al. [5], the property set proposed by them is a necessary but not sufficient set of properties as they don't ensure that the metrics which satisfy them are valid measures. They are advantageous to refute the software metrics but not to corroborate them. This motivated us to further validate the metrics formally. Thus, in this study, we have corroborated the metrics using the Distance framework, whose adequacy is guaranteed by the measurement theory.

Apart from providing a set of sufficient and necessary measure axioms, the approach based on distance is also formal, flexible and generic. It is generic as the measures for the internal attributes of varied software products can be defined using the same set of axioms. It is flexible as alternate definitions can be proposed for an attribute of a software product. A significant effect of this relationship with measurement theory is that the measure defined using the Distance framework is in the ratio scale. Ratio scale measures permit the usage of varied data analysis techniques and hence provide flexibility to the researcher. Another significant advantage of working with distances is that the notion of similarity and dissimilarity is understood quite well and is used in day-to-day life [10]. The Distance framework also has a well defined theoretical base in measurement theory, while the property based or axiomatic techniques for the validation of measures are based on intuition, subjective experience, argumentation, etc.

The organization of the paper is as follows: The related work is presented in Section II. Section III outlines the metrics being validated. Section IV details the distance framework. Section V presents the theoretical validation of the metrics. Section VI concludes the paper.

II. Related Work

The related work has been split into two sections. The first section presents the DW conceptual model quality metrics proposed by researchers and the techniques employed for their theoretical validation while the second section focuses on the use of Distance framework in software engineering to validate the metrics.

Metrics have been proposed by the researchers in the context of DW to quantify the structural complexity of the multidimensional models and assure their quality. They aid in determining the model quality during the early stages of design. The initial proposal of measures for the DW conceptual model was given by Calero et al. [3]. These metrics began the era of objectively evaluating the multidimensional model of a data warehouse and were theoretically validated using the Zuse framework [8]. All the measures belonged to either ordinal or superior scale. Serrano et al. [11] conducted an experiment to assess the relationship between four schema level measures put forth by Calero et al. [3] and conceptual model understandability. They also corroborated the metrics using Briand's property based framework [5] and the measurement theory based frameworks of Zuse [8] and Poels and Dedene [7]. Berenguer et al. [12] defined quality objectives and corresponding measures for DW conceptual models. The metrics were classified as package or diagram level measures. The authors claimed to have formally corroborated the measures using both axiomatic and measurement theory based approaches. Gosain et al. [9] proposed five measures based on dimension hierarchies in the data warehouse conceptual models. The measures have been validated using Briand's framework. Cherfi and Prat [13] proposed measures to quantify the analyzability and simplicity of DW multidimensional models. However, neither empirical nor theoretical validation of the measures was conducted. Hence, the metrics failed to prove their practical utility. Serrano et al. [14] proposed measures for object oriented models of DW on the basis of design elements such as dimension classes, fact classes, dimension hierarchies, etc. They proposed metrics at different levels – diagram, star and class. Serrano et al. [15] formally corroborated the star scope metrics using the Distance framework. Nagpal et al. [16] proposed a comprehensive complexity metric based on the elements in the model and the relationships among them. They corroborated the metrics using Briand's framework. Sabharwal et al. [17] proposed coupling metrics and validated them using Kaner's framework.

Table 1 presents a comparison of the existing proposals of multidimensional model quality measures and techniques employed for their theoretical validation. The first column of the table cites the study where the proposal was made. The second column focuses on the quality criteria of the metrics. The third and fourth columns indicate whether formal validation was done and the technique employed for the same. From the last column we observe that Distance framework has been rarely used to corroborate measures in the context of DW. However, lots of researchers (Genero et al. [18], Tripathi et al. [19], Rossi and Fernandez [20], Bajeh et al. [21], Munoz et al. [22]) in software engineering have used the said framework to validate quality measures. This further motivated us to validate the DW conceptual model metrics using the Distance framework. Genero et al. [18] have proposed objective measures to quantify the structural properties of entity relationship diagrams.

Their measures are related to the understandability and inturn quality of the diagrams. Tripathi et al. [19] have proposed measures to evaluate the Indian e-commerce based web applications for their quality. Rossi and Fernandez [20] presented a metric suite embracing behavioural and structural aspects of distributed applications. The measures quantify the internal attributes of formal models. Bajeh et al. [21] proposed an object oriented metric to determine the complexity of software design. All the above mentioned metric proposals were successfully theoretically validated using the Distance framework.

Table 1. Summary of Data Warehouse metric proposals and their Theoretical Validation

Authors	Focus	Theoretical Validation	Technique of Theoretical Validation
Calero et al. [3]	Quality	Done	Zuse Framework
Serrano et al. [11]	Understandability and Modifiability	Done	Briand’s Framework Zuse Framework Distance Framework
Berenguer et al. [12]	Quality	Done	Based on Axiomatic and Measurement Theory
Gosain et al. [9]	Understandability	Done	Briand’s Framework
Cherfi and Prat [13]	Analyzability and Simplicity	Not Done	-
Serrano et al. [14]	Understandability and Modifiability	Not Done	-
Serrano et al. [15]	Understandability	Done	Distance Framework
Nagpal et al. [16]	Understandability	Done	Briand’s Framework
Sabharwal et al. [17]	Quality	Done	Kaner’s Framework

III. Metrics for DW Conceptual Model

In this paper, we have formally corroborated the metrics proposed by Gosain et al. [9] using the Distance framework. A brief description of the metrics is summarized below:

- NMH – Number of multiple hierarchies in the schema
- NLDH – Number of levels in dimension hierarchies of the schema
- NAPMH – Number of alternate paths in multiple hierarchies of the schema
- NDSH – Number of dimensions involved in shared hierarchies of the schema
- NSH – Number of shared hierarchies of the schema

The metrics are exemplified using the conceptual schema shown in Fig. 1. The values of the metrics for the said schema are calculated in Table 2.

IV. Distance Framework

Poels and Dedene had put forward a conceptual framework based on measurement theory. It is called the Distance framework [7]. It can be used to formally validate software measures. It puts forth the requirements that have to be fulfilled so that mathematical functions could be used as “measures”. Distance framework proposes a method which helps to verify if those requirements are fulfilled by the software measures. It presents a procedure for measure construction which models the features of software artefacts and defines the corresponding software measures. Object properties are defined by the framework as distances between them and other objects that act as reference point for the concerned measurement. According to the framework, the larger the distance between these objects, the more they are characterised by the properties. Due to such a definition of object properties they can be quantified by functions called “metrics” in maths which fulfil the metric axioms. Metric axioms are sufficient and necessary to define distance measures (as defined by measurement theory). This ensures that the formal validation of the measures that have been obtained with distance is proven within the framework of measurement theory. There are five steps in the distance-based measure construction procedure. A brief description of the same is as under:

1. For an internal attribute, x and the corresponding set of software entities, S, pick a set of software entities M that can be considered as measurement abstractions such that they give prominence to the internal attribute x, and describe a function abs: S → M.
2. Define a set of elementary transformations, Trans, on M that is constructively and inverse constructively complete. The set Trans is then used to determine the distances among measurement abstractions.
3. Quantify distances between measurement abstractions defining a metric $\delta: M \times M \rightarrow \Re$ such that (M, δ) is a metric space.
4. Select a reference model $r \in M$ that is the software abstraction for which it holds such that for all $s \in S$ with $abs(s) = r$, s has the lowest value of x.
5. Define a function $\mu: S \rightarrow \Re$ such that for all $s \in S$,

μ(s) = δ(abs(s), r) which is a measure of distance from abs(s) to r.

The next section outlines the results of theoretical validation of the metrics using Distance framework.

V. Validation Results

A. *Validation of the NMH Metric*

To validate the NMH metric we will follow the steps put forward by the Distance framework. To better understand the procedure we will be using the models represented in Fig. 2, as an example.

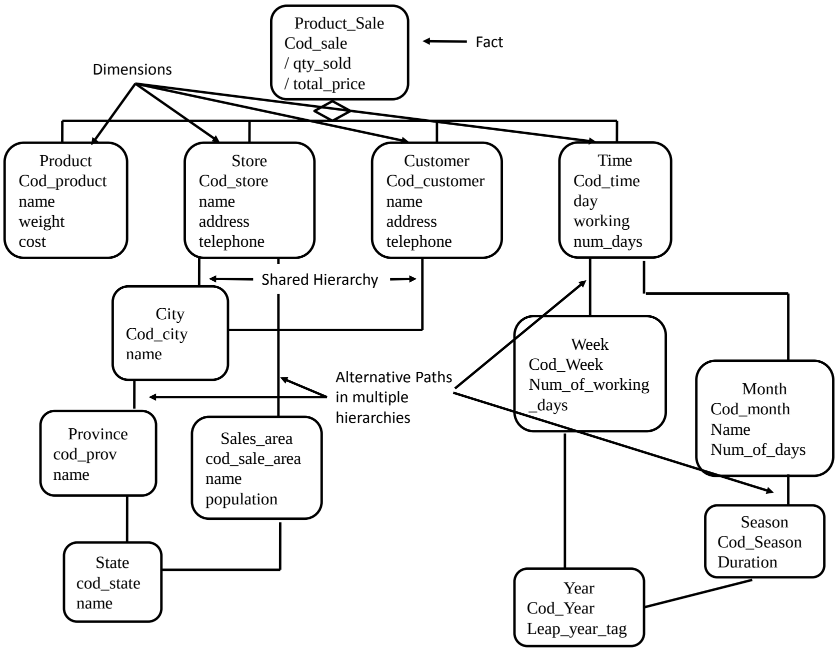

Fig.1. Example Multidimensional Model [9]

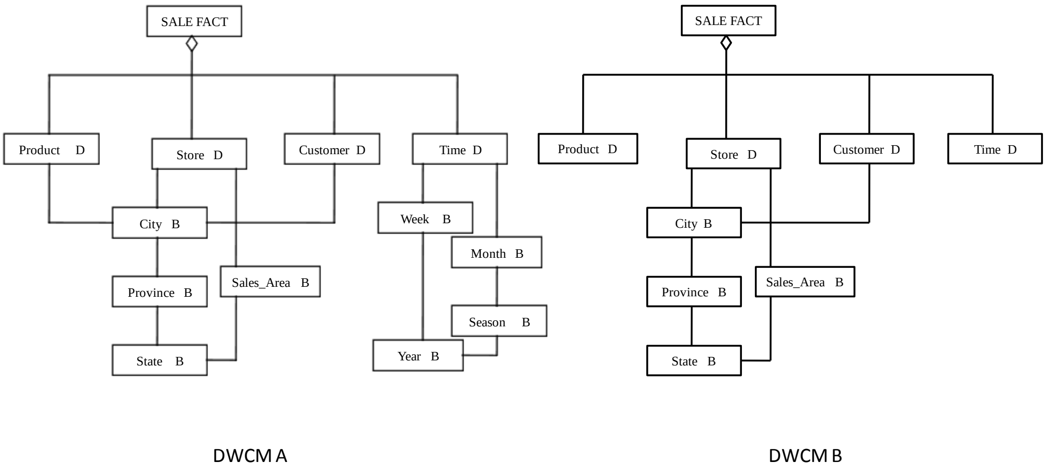

Fig.2. Two examples of Multidimensional Models of Data Warehouse

Step 1. In this step, we identify a measurement abstraction for the attribute of interest. Since we are working with data warehouses, the set of software entities S, in our case is the Universe of DW Conceptual Models (UDWCM). UDWCM must be pertinent for some Universe of Discourse and s is a DW Conceptual Model (DWCM) ie. s ∊ S. Number of multiple hierarchies in a DWCM is the attribute of interest. Let UMH be the Universe of Multiple Hierarchies admissible to the Universe of Discourse. The set of multiple hierarchies within a conceptual model DWCM, called MH(DWCM) is then a subset of UMH. All the sets of multiple hierarchies within the conceptual models of UDWCM are elements of the power set of UMH, depicted by P(UMH).

Table 2. Values of Metrics for the model in Fig. 1

Metric	Value
NMH	2
NLDH	4
NAPMH	4
NDSH	2
NSH	1

Thus, we can associate the set of measurement abstractions M to P(UMH) and define the abstraction function as:

$$\mathrm{abstraction(NMH): UDWCM} \rightarrow \mathrm{P(UMH): DWCM} \rightarrow \mathrm{MH(DWCM)} \quad (1)$$

This abstraction function determines the extent to which a DWCM is characterized by the number of multiple hierarchies. By comparing such abstractions we can deduce whether a conceptual model is more, equally or less characterized by the number of multiple hierarchies. For better understanding, an example has been taken in which we have the set of multiple hierarchies of DWCM A and DWCM B (Fig. 2):

$$\mathrm{DWCM\ A:\ abstraction(NMH) = MH(DWCM\ A) = \{Store\_hierarchy, Time\_hierarchy\}} \quad (2)$$

$$\mathrm{DWCM\ B:\ abstraction(NMH) = MH(DWCM\ B) = \{Store\_hierarchy\}} \quad (3)$$

Step 2. In this step we model the distances among the elements of measurement abstractions. It is essential to determine a set of elementary transformations for P(UMH) so that any set of multiple hierarchies from P(UMH) can be transformed into any other set of multiple hierarchies by a finite sequence of such transformations. Since the elements of P(UMH) are set of multiple hierarchies, the set Trans can contain elementary transformations of only two types - one for the inclusion of a multiple hierarchy to a set and the other for the removal of a multiple hierarchy from a set. The smallest series of elementary transformations are eligible to be considered as models of distance. Thus, Trans = {t_1, t_2}, where t_1 and t_2 are defined as under:

$$\forall mh \in P(UMH): \mathrm{t}_1(mh) = mh \cup \{m\} \quad where\ m \in UMH \quad (4)$$

$$\forall mh \in P(UMH): \mathrm{t}_2(mh) = mh - \{m\} \quad where\ m \in UMH \quad (5)$$

where mh is a set of P(UMH) and can be transformed into any other set of P(UMH) by the addition and removal of corresponding hierarchies.

In the example that we have considered, the distance between abstraction(NMH) for DWCM A and DWCM B can be determined by a series of transformations from the set Trans. MH(DWCM A) can be transformed to MH(DWCM B) simply by one elementary transformation i.e. the removal of Time hierarchy from MH(DWCM A). The two sets can be made equal by other sets of elementary transformations also but since this is the shortest sequence, we have considered it.

Step 3. In this step, the distance between two sets of multiple hierarchies in P(UMH) is quantified. This distance is determined by the measure of the smallest series of elementary transformations that make both the sets equal. Given two sets, mh, mh' ∊ P(UMH), if an element is contained in either mh or mh' but not both then exactly one transformation is needed to make them equal. Thus, the distance between the sets is equivalent to the cardinality of the symmetric difference between mh and mh'.

$$\forall mh, mh' \in P(UMH): \delta_{MH}(mh, mh') = |mh - mh'| + |mh' - mh| \quad (6)$$

The symmetric difference model, for our example, gives a value of 1 for the distance among the set of multiple hierarchies of DWCM A and DWCM B. Formally,

$$\delta_{NMH}(abstraction_{NMH}(DWCM\ A), abstraction_{NMH}(DWCM\ B)) = |\{Store\_hierarchy, Time\_hierarchy\} - \{Store\_hierarchy\}| + |\{Store\_hierarchy\} - \{Store\_hierarchy, Time\_hierarchy\}| = |\{Time\_hierarchy\}| + |\{\}| = 1 \quad (7)$$

Step 4. In this step, a reference abstraction is identified for the attribute of interest. The void set of multiple hierarchies would form the reference abstraction in this study. A DWCM will have the lowest value for the NMH metric if it has no multiple hierarchies. Thus, we can define the following function:

$$\mathrm{RefNMH:\ UDWCM} \rightarrow \mathrm{P(UMH): DWCM} \rightarrow \emptyset \quad (8)$$

Step 5. The software measure is defined in this step. The number of multiple hierarchies of a DWCM can be determined by the distance between its set of multiple hierarchies i.e. MH(DWCM) and the empty set. It is the smallest series of elementary transformations between MH(DWCM) and Ø. Thus, the NMH metric can be perceived as a function that returns for any DWCM ∊ UDWCM the value of the measure δ_{NMH} for the sets MH(DWCM) and Ø :

$$\forall DWCM \in UDWCM: NMH(DWCM) = \delta_{NMH}(MH(DWCM), \emptyset) = |MH(DWCM) - \emptyset| + |\emptyset - MH(DWCM)| = |MH(DWCM)| \quad (9)$$

This proves that the NMH metric is theoretically valid.

B. Validation of the NAPMH Metric

Step 1. The measurement abstraction for the attribute of interest i.e. the number of alternate paths in multiple hierarchies (NAPMH) can be defined as:

$$\mathrm{A}bstraction(NAPMH): UDWCM \rightarrow P(UAPMH): DWCM \rightarrow SAPMH(DWCM) \quad (10)$$

This abstraction function ascertains to what extent a DWCM is characterized by the set of alternate paths in multiple hierarchies. Again as an example we will consider the set of alternate paths in multiple hierarchies of DWCM A and DWCM B (Fig. 2):

$$\text{DWCM A}: \text{abstraction(NAPMH)} = \text{SAPMH(DWCM A)} = \{\text{City\_Province\_State}, \text{Sales\_Area\_State}, \text{Week\_Year}, \text{Month\_Season\_Year}\} \quad (11)$$

$$\text{DWCM B}: \text{abstraction(NAPMH)} = \text{SAPMH(DWCM B)} = \{\text{City\_Province\_State}, \text{Sales\_Area\_State}\} \quad (12)$$

Step 2. The definition of the set Trans of elementary transformation types on P(UAPMH) that is both constructively and inverse constructively complete is :

Trans = {t_1, t_2}, where

$$\forall apmh \in P(UAPMH): t_1(apmh) = apmh \cup \{m\} \text{ where } m \in UAPMH \quad (13)$$

$$\forall apmh \in P(UAPMH): t_2(apmh) = apmh - \{m\} \text{ where } m \in UAPMH \quad (14)$$

where apmh, is a set of P(UAPMH) and can be transformed into any other set of P(UAPMH) by the addition and removal of corresponding alternate paths from the hierarchies.

In the example that we have considered, the shortest sequence of elementary transformations that can determine the distance between the abstraction(NAPMH) for DWCM A and DWCM B is the removal of the paths Week_Year and Month_Season_Year from SAPMH(DWCM A).

Step 3. This step determines the metric space (P(UAPMH),δ). The distance between any two sets of alternate paths in multiple hierarchies in P(UAPMH) can be quantified by the smallest series of elementary transformations that makes both the sets equal.

$$\forall apmh, apmh' \in P(UAPMH): \delta_{APMH}(apmh, apmh') = |apmh - apmh'| + |apmh' - apmh| \quad (15)$$

The symmetric difference model, for our example, gives a value of 2 for the distance among the set of alternate paths in multiple hierarchies of DWCM A and DWCM B. Formally,

$$\delta_{NAPMH}(abstraction_{NAPMH}(DWCM\ A), abstraction_{NAPMH}(DWCM\ B)) = |\{\text{City\_Province\_State}, \text{Sales\_Area\_State}, \text{Week\_Year}, \text{Month\_Season\_Year}\} - \{\text{City\_Province\_State}, \text{Sales\_Area\_State}\}| + |\{\text{City\_Province\_State}, \text{Sales\_Area\_State}\} - \{\text{City\_Province\_State}, \text{Sales\_Area\_State}, \text{Week\_Year}, \text{Month\_Season\_Year}\}| = |\{\text{Week\_Year}, \text{Month\_Season\_Year}\}| + |\{\}| = 2 \quad (16)$$

Step 4. This step determines the reference abstraction RefNAPMH ∊ P(UAMPH) for the number of alternate paths in multiple hierarchies. There exists conceptual models with no multiple hierarchies and hence, no alternate paths in multiple hierarchies. Thus, the RefNAPMH is the void set.

Step 5. The software measure is defined in this step. The number of alternate paths in multiple hierarchies of a DWCM can be determined by the distance between its set of alternate paths in multiple hierarchies i.e. APMH(DWCM) and the empty set. It is the smallest series of elementary transformations between APMH(DWCM) and Ø. Thus, the NAPMH metric can be perceived as a function that returns for any DWCM ∊ UDWCM the value of the measure δ_{NAPMH} for the sets APMH(DWCM) and Ø:

$$\forall DWCM \in UDWCM : NAPMH(DWCM) = \delta_{NAPMH}(APMH(DWCM), \emptyset) = |APMH(DWCM) - \emptyset| + |\emptyset - APMH(DWCM)| = |APMH(DWCM)| \quad (17)$$

This proves that the NAPMH metric is theoretically valid.

C. Validation of the NDSH Metric

Step 1. The measurement abstraction for the attribute of interest i.e. the number of dimensions participating in shared hierarchies (NDSH) can be defined as:

$$abstraction(NDSH) : UDWCM \to P(UDSH): DWCM \to SDSH(DWCM) \quad (18)$$

As an example we will consider the set of dimensions participating in shared hierarchies of DWCM A and DWCM B (Fig. 2):

$$\text{DWCM A}: \text{abstraction(NDSH)} = \text{SDSH(DWCM A)} = \{\text{Product}, \text{Store}, \text{Customer}\} \quad (19)$$

$$\text{DWCM B}: \text{abstraction(NDSH)} = \text{SDSH(DWCM B)} = \{\text{Store}, \text{Customer}\} \quad (20)$$

Step 2. The definition of the set Trans of elementary transformation types on P(UDSH) that is both constructively and inverse constructively complete is:

Trans = {t_1, t_2}, where

$$\forall dsh \in P(UDSH): t_1(dsh) = dsh \cup \{m\} \text{ where } m \in UDSH \quad (21)$$

$$\forall dsh \in P(UDSH): t_2(dsh) = dsh - \{m\} \text{ where } m \in UDSH \quad (22)$$

where dsh is a set of P(UDSH) and can be transformed into any other set of P(UDSH) by the addition and removal of corresponding dimensions from shared hierarchies.

In the example that we have considered, the shortest sequence of elementary transformations that can determine the distance between the abstraction(NDSH) for DWCM A and DWCM B is the removal of the Product dimension from SDSH(DWCM A).

Step 3. This step determines the metric space (P(UDSH), δ). The distance between any two sets of dimensions participating in shared hierarchies in P(UDSH) can be quantified by the smallest series of elementary transformations that makes both the sets equal.

$$\forall dsh, dsh' \in P(UDSH): \delta_{DSH}(dsh, dsh') = |\, dsh - dsh'| + |\, dsh' - dsh| \qquad (23)$$

The symmetric difference model, for our example, gives a value of 1 for the distance among the set of dimensions participating in shared hierarchies of DWCM A and DWCM B. Formally,

$$\begin{gathered}\delta_{NDSH}(abstraction_{NDSH}(DWCM\ A), \\ abstraction_{NDSH}(DWCM\ B)) \\ = |\{\text{ Product, Store, Customer}\} - \{\text{Store, Customer}\}| \\ +|\{\text{ Store, Customer }\} - \\ \{\text{Product, Store, Customer}\}| = \\ |\,\{\text{Product}\}| + |\,\{\}| = 1 \qquad (24)\end{gathered}$$

Step 4. This step determines the reference abstraction RefNDSH ∊ P(UDSH) for the number of dimensions participating in shared hierarchies. There exists conceptual models with no shared hierarchies and hence, no dimensions participating in shared hierarchies. Thus, the RefNDSH is the void set.

Step 5. The software measure is defined in this step. The number of dimensions participating in shared hierarchies of a DWCM can be determined by the distance between its set of dimensions participating in shared hierarchies i.e. DSH(DWCM) and the empty set. It is the smallest series of elementary transformations between DSH(DWCM) and Ø. Thus, the NDSH metric can be perceived as a function that returns for any DWCM ∊ UDWCM the value of the measure δ_{NDSH} for the sets DSH(DWCM) and Ø:

$$\begin{gathered}\forall\, DWCM \in UDWCM : NDSH(DWCM) \\ = \delta_{NDSH}(DSH(DWCM), \emptyset) \\ = |DSH(DWCM) - \emptyset\,| + |\,\emptyset - DSH(DWCM)\,| \\ = |DSH(DWCM)| \qquad (25)\end{gathered}$$

This proves that the NDSH metric is theoretically valid.

The measure construction and theoretical validation process of NSH and NLDH is analogous to that of the NMH, NAPMH and NDSH metrics and is summarized in Table 3. Since the Distance framework has been used to define the measures, they can all be described as distances. This guarantees that they are all characterised by the ratio scale and hence are formally sound software measures.

VI. Conclusion

In this paper, we have used the measurement theory based Distance framework to formally validate the data warehouse hierarchy metrics. The said metrics had previously been validated using Briand's property based framework. However, it offers a preferable set of properties to validate the software metrics which is not sufficient. Thus, we have employed Distance framework to validate the hierarchy metrics which offers a set of sufficient and necessary measure axioms. The measures validated using this framework are above the ordinal scale and hence a wide range of data analysis techniques can be used to analyse them. All the five hierarchy measures (NMH, NAPMH, NLDH, NSH and NDSH) have been successfully corroborated using the Distance framework. Thus, they are valid measures of data warehouse conceptual model complexity.

Table 3. Abstraction functions for the remaining hierarchy metrics

Metric	Abstraction Function
NSH	$absNSH : UDWCM \rightarrow P(USH) : DWCM \rightarrow SSH(DWCM)$ (26) where UDWCM is the Universe of Data Warehouse Conceptual Models USH is the Universe of Shared Hierarchies relevant to a UoD SSH(DWCM) ⊆USH is the set of shared hierarchies in a DWCM
NLDH	Metric NLDH is represented at the class level as: $absNLDH: UC \rightarrow P(UC) : C \rightarrow LongestPath(C)$ (27) where UC is the Universe of classes LongestPath(C)⊆UC is the set of classes that are a part of dimension hierarchy When multiple hierarchies are considered, only the classes in the longest hierarchy are taken into consideration Metric NLDH is the largest value of NLDH computed for all the classes present in DWCM

REFERENCES

[1] W. H. Inmon, *Building the Data Warehouse*. Wiley, 2005.

[2] N. T. Debevoise, *The data warehouse method*. Prentice Hall, 1998.

[3] C. Calero, C. Pascual, M. Piattini, and M. A. Serrano, “Towards Data Warehouse Quality Metrics,” in *Proceedings of the International Workshop on Design and Management of Data Warehouses*, 2001, pp. 1–10.

[4] C. Calero, M. Piattini, and M. Genero, “Method for Obtaining Correct Metrics,” in *Third International Conference on Enterprise Information Systems*, 2001, pp. 779–784.

[5] L. C. Briand, S. Morasca, and V. R. Basili, “Property-based software engineering measurement,” *IEEE Trans. Softw. Eng.*, vol. 22, no. 1, pp. 68–86, 1996.

[6] E. J. Weyuker, “Evaluating Software Complexity Measures,” *IEEE Trans. Softw. Eng.*, vol. 14, no. 9, pp. 1357–1365, 1988.

[7] G. Poels and G. Dedene, “Distance-based software measurement: Necessary and sufficient properties for software measures,” *Inf. Softw. Technol.*, vol. 42, no. 1, pp. 35–46, 2000.

[8] H. Zuse, *A Framework of Software Measurement*. Walter de Gruyter, 1998.

[9] A. Gosain, S. Nagpal, and S. Sabharwal, “Validating dimension hierarchy metrics for the understandability of multidimensional models for data warehouse,” *IET Softw.*, vol. 7, no. 2, pp. 93–103, 2013.

[10] P. Suppes, M. Krantz, R. Luce, and A. Tversky, *Foundations of Measurement*. New York: Academic Press, 1989.

[11] M. A. Serrano, C. Calero, H. A. Sahraoui, and M. Piattini, “Empirical studies to assess the understandability of data warehouse schemas using structural metrics,” *Softw. Qual. J.*, vol. 16, no. 1, pp. 79–106, 2008.

[12] G. Berenguer, R. Romero, J. Trujillo, M. Serrano, and M. Piattini, “A set of quality indicators and their corresponding metrics for conceptual models of data warehouses,” in *Data Warehousing and Knowledge Discovery*, 2005, pp. 95–104.

[13] S. S. Cherfi and N. Prat, “Multidimensional Schemas Quality : Assessing and Balancing Analyzability and Simplicity,” in *Proceedings of ER Workshops, Springer LNCS*, 2003, pp. 140–151.

[14] M. Serrano, C. Calero, J. Trujillo, S. Lujan, and M. Piattini, “Empirical validation of metrics for conceptual models of data warehouse,” in *16th International Conference on Advanced Information Systems Engineering (CAISE’04)*, 2004, pp. 506–520.

[15] M. Serrano, J. Trujillo, C. Calero, and M. Piattini, “Metrics for data warehouse conceptual models understandability,” *Inf. Softw. Technol.*, vol. 49, no. 8, pp. 851–870, 2007.

[16] S. Nagpal, A. Gosain, and S. Sabharwal, “Theoretical and empirical validation of comprehensive complexity metric for multidimensional models for data warehouse,” *Int. J. Syst. Assur. Eng. Manag.*, vol. 4, no. 2, pp. 193–204, 2013.

[17] S. Sabharwal, S. Nagpal, and G. Aggarwal, “Coupling metrics for object-oriented data warehouse design,” in *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*, 2015, pp. 918–922.

[18] M. Genero, G. Poels, and M. Piattini, “Defining and validating metrics for assessing the understandability of entity-relationship diagrams,” *Data Knowl. Eng.*, vol. 64, no. 3, pp. 534–557, 2008.

[19] P. Tripathi, M. Kumar, and N. Shrivastava, “Theoretical validation of quality metrics of Indian e-commerce domain,” in *2009 2nd International Conference on Computer, Control and Communication*, 2009, pp. 1–7.

[20] P. Rossi and G. Fernandez, “Definition and validation of design metrics for distributed applications,” in *Proceedings. 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (IEEE Cat. No.03EX717)*, 2003, pp. 124–132.

[21] A. O. Bajeh, S. Basri, and L. T. Jung, “A theoretical validation of the number of polymorphic methods as a complexity metric,” in *2014 International Conference on Computer and Information Sciences (ICCOINS)*, 2014, pp. 1–6.

[22] L. Muñoz, J. N. Mazón, and J. Trujillo, “A family of experiments to validate measures for UML activity diagrams of ETL processes in data warehouses,” Inf. Softw. Technol., vol. 52, no. 11, pp. 1188–1203, 2010.

Authors’ Profiles

Gargi Aggarwal received her B.Tech degree in Computer Science from Indira Gandhi Delhi Technical University for Women (IGDTUW) in 2006. She received her M. Tech degree in Information Systems from University of Delhi in 2012. Ms. Aggarwal is a Teaching cum Research Fellow in the Division of IT at Netaji Subhas Institute of Technology, affiliated to University of Delhi. She is presently pursuing her PhD in Data Warehouse Quality from University of Delhi. Her research interests include Data Warehouse, Machine Learning and Software Quality Management.

Sangeeta Sabharwal is currently working as a Professor in the department of Computer Engineering at Netaji Subhas Institute of Technology in Delhi, India. Her areas of interest are Software Engineering, Meta modeling, Object Oriented Analysis, Software Testing, and Data warehousing. Currently she is actively involved in applying different Soft Computing Techniques to different areas of software engineering. She has published papers in several international journals and conferences. She has also written a book on software engineering. A number of students are pursuing their Ph.D. under her guidance.

How to cite this paper: Gargi Aggarwal, Sangeeta Sabharwal, "Formal Validation of Data Warehouse Complexity Metrics using Distance Framework", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.10, pp.49-56, 2017. DOI: 10.5815/ijisa.2017.10.06