Information Systems 63 (2017) 47–62

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Cost-benefit analysis of data warehouse design methodologies

Francesco Di Tria*, Ezio Lefons, Filippo Tangorra

Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro", Via Orabona 4, 70125 Bari, Italy

ARTICLE INFO

Article history:
Received 11 February 2015
Accepted 17 June 2016
Available online 24 June 2016

Keywords:
Schema quality
Metrics
Design effort
Evaluation
Automation

ABSTRACT

Methodologies for data warehouse design are increasing more and more in last years, and each of them proposes a different point of view. Among all the methodologies present in literature, the promising ones are the hybrid methodologies—because they represent the only way to ensure a multidimensional schema to be both consistent with data sources and adherent to user business goals—and those able to support the designer by providing some kind of automation. However, the results obtainable by the methodologies can differ substantially in terms of schema quality and required efforts. In this paper, we provide metrics for evaluating the quality of multidimensional schemata in reference to the effort spent in the design process and the automation degree of the methodology. As a case study, we apply our evaluation to the major emerging hybrid methodologies for data warehouse schema design.

1. Introduction

The main approaches to data warehouse design are the data-driven and requirement driven methodologies. Each of them presents advantages and weaknesses [1]. The data-driven approach analyzes the data source and remodels it in order to obtain a multidimensional schema. In this way, the feasibility of the data warehouse is guaranteed, but the user needs are not taken into account, going towards a possible failure. On the other hand, the requirement-driven approach considers the business goals to start with, and then produces a multidimensional schema. So, that schema is adherent to user needs but it may be not supported by the effective presence of data in the source.

To overcome the limits, several efforts are currently spent to define a design methodology that integrates the advantages of both these approaches. This research issue has led to the definition of hybrid methodologies for data warehouse design [2].

Hybrid methodologies are getting increasing attention because they allow the designer to obtain multidimensional schemata able to satisfy user requirements on the basis of data effectively available in data sources. As a counterpart, hybrid methodologies require a more complex design process due to the reconciliation of different approaches. Indeed, some methodologies have to consider simultaneously the data sources and the user requirements [3,4], while other methodologies have to integrate the data-driven approach and the requirement-driven one [5–10].

However, the advantages of adopting hybrid methodologies justify the higher efforts to be spent in the multidimensional modeling. For these reasons, the current research is devoted to introduce automatisms in order to reduce the design efforts and to support the designer in the multidimensional modeling. Automatic methodologies provide algorithms for supporting the designer in (part of) the multidimensional modeling, as to identify facts in data sources [11] and to construct multidimensional views of data [12,13], for example.

* Corresponding author.
E-mail addresses: francesco.ditria@uniba.it (F. Di Tria), ezio.lefons@uniba.it (E. Lefons), filippo.tangorra@uniba.it (F. Tangorra).

http://dx.doi.org/10.1016/j.is.2016.06.006
0306-4379/© 2004 Published by Elsevier Ltd.

In the paper, we investigate methodologies for data warehouse schema design, in order to provide a method for their evaluation. The evaluation considers the quality of the schemata produced in the design process in reference to the effort spent in that design process. To this end, we provide a set of metrics to evaluate the costs and the benefits of design methodologies.

In literature, several metrics have been defined for evaluating the quality of multidimensional schemata at the logical level [14,15] and the understandability of multidimensional schemata at the conceptual level [16]. Because of such metrics are functionally based on the number of model elements present in the schemata, their measures and results are only related to the schema complexity. Consequently, since actual metrics cannot evaluate the design efforts, the designer has no useful indicators to establish which methodology produces a better multidimensional schema by requiring the minor effort.

To this purpose, we here propose a set of metrics for measuring the quality of a multidimensional schema at the logical level in reference to the effort to be spent in the design process. The presented metrics can be applied to whatever design methodology is considered for evaluation, and independently from the underlying adopted approach —data-driven, requirement-driven, or hybrid one—for these new metrics take into account objective and general parameters, such as the number of phases and artifacts.

In addition, we also propose metrics to evaluate the automation degree of the methodology to be analyzed.

In order to check the metrics validity, we present the evaluation of the category of hybrid and (semi-)automatic methodologies for data warehouse design as a case study.

The paper is organized as follows. Section 2 discusses about metrics present in literature for data warehouse evaluation. Section 3 introduces the framework we used to define metrics. Section 4 presents the cost-benefit metrics. Section 5 shows the theoretical validation of the new metrics. Section 6 briefly describes the methodologies to be compared and shows the case study data warehouse. Then, it reports the experimental results. Section 7 shows how to evaluate and compare methodologies falling into different categories (the semi-automatic and the manual ones). Section 8 concludes the paper with some our remarks.

2. Related work

The work presented in [14] aims at evaluating the quality of a data model. The authors provide a set of metrics to measure the complexity of a relational schema, by counting the numbers of model elements, such as the numbers of attributes of a table and the numbers of foreign keys present in that schema. Moreover, the authors validate the metrics on the basis of the formal framework proposed by Zuse [17], in order to show that such metrics can be used for measuring the complexity of a data warehouse schema. Experimental results of the application of the metrics are furnished in [18]. Similarly, the work presented in [16] is devoted to the evaluation of the quality of data models. Here, the design abstraction level is the conceptual schema and, then, the proposed metrics allow measuring the complexity of a data warehouse schema formalized in UML. In detail, the authors investigate how the understandability of a schema is affected by its complexity.

Pighin and Ieronutti propose a set of metrics for measuring statistical aspects of data, such as the percentage of null values [15]. On the basis of these metrics, the quality of a schema is evaluated by checking whether each attribute of a relational schema has been correctly identified as a cube measure or as a dimensional attribute.

In [19], Kesh proposes a framework for evaluating the performance and the quality of conceptual schemata. To this end, the framework includes a set of seven criteria to check the structure of the schema— *ie*, the entities and their relationships— and the content—*ie*, the attributes included in each entity. As to the content, for example, the criteria evaluate the *completeness* and the *validity*. Then, each criterion is estimated by assigning a score. It is worth noting that objective metrics are defined only for some criteria. These can be computed automatically, since they count the number of elements present in the schema. For the remaining criteria, subjective metrics are used. For example, the *soundness* of a schema must be evaluated by a technical group not involved with the project. This group assigns 1 to 5 points and the schema is assumed to be correct if the score falls over an acceptance value. Also Moody [20] proposes criteria which agree with [19,21] and adopts both a set of objective metrics that aim at counting the conceptual elements in a given a schema—*ie*, the number of missing items—and a set of subjective metrics that aim at providing qualitative information—*ie*, which items are missing.

2.1. Discussion

In reference to the related work, we provide the comparison Table 1 in order to highlight the common evaluation criteria and the corresponding concepts. In addition, Table 1 can also be used to solve inconsistencies in the terminology, as the authors use different names to represent the same concept(s). Among the others, we focus on those criteria that have been considered by the most part of the authors—namely criteria #1, #2, #4, and #5.

On the basis of these four criteria, we can state that a schema is of good quality if and only if it is *correct*, *complete*, *minimal*, and *easily understandable.*

As to the evaluation criteria, we choose objective metrics for they can be computed automatically and provide unbiased assessment. The selected metrics are reported in Table 2.

It is worth noting that Moody in [20] uses the *complexity* in order to evaluate the *simplicity*. This led us to consider that if a schema is *complete* and *correct*, then it is of good quality if and only if it is *minimal*, that is, it has the minimum number of elements as possible.

This means that when comparing two schemata which are both *correct* and *complete*, the one having the minor number of elements shows a better quality.

On the other hand, Serrano *et al.* in [16] use the *complexity* in order to evaluate the *understandability*, because they deem that the number of elements in a schema