International Journal of Approximate Reasoning 96 (2018) 95–111

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

An incremental approach for data quality measurement with insufficient information

A. Bronselaer *, J. Nielandt, G. De Tré

Department of Telecommunications and Information Processing, Ghent University, Sint-Pietersnieuwstraat 41, B9000, Ghent, Belgium

ARTICLE INFO

Article history:
Received 20 December 2017
Received in revised form 16 March 2018
Accepted 17 March 2018
Available online 20 March 2018

Keywords:
Data quality measurement
Uncertainty modelling
Insufficient information
Possibility theory

ABSTRACT

Recently, a fundamental study on measurement of data quality introduced an ordinal-scaled procedure of measurement. Besides the pure ordinal information about the level of quality, numerical information is induced when considering uncertainty involved during measurement. In the case where uncertainty is modelled as probability, this numerical information is ratio-scaled. An essential property of the mentioned approach is that the application of a measure on a large collection of data can be represented efficiently in the sense that (i) the representation has a low storage complexity and (ii) it can be updated incrementally when new data are observed. However, this property only holds when the evaluation of predicates is clear and does not deal with uncertainty. For some dimensions of quality, this assumption is far too strong and uncertainty comes into play almost naturally. In this paper, we investigate how the presence of uncertainty influences the efficiency of a measurement procedure. Hereby, we focus specifically on the case where uncertainty is caused by insufficient information and is thus modelled by means of possibility theory. It is shown that the amount of data that reaches a certain level of quality, can be summarized as a possibility distribution over the set of natural numbers. We investigate an approximation of this distribution that has a controllable loss of information, allows for incremental updates and exhibits a low space complexity.

1. Introduction

Since the introduction of the dimensional model for data quality [1], there has been an increasing interest in models for assessment of data quality. Among other things, this research has produced several ways in which the numerical quantification of data quality can be grounded in measurement theory [2]. In this paper, we consider a framework in which quality of data is essentially expressed on an ordinal scale, but where uncertainty about the precise outcome of the measurement process is explicitly taken into account [3]. One of the key issues of this framework is that, under absence of uncertainty, a large set of measurements can be represented very efficiently by means of a histogram. When uncertainty comes into play, this property does no longer apply, as each number in the histogram must be replaced by a distribution that has linear space complexity in terms of the number of data items. In the current paper, we investigate this problem more closely in the specific case where uncertainty about predicates is caused by insufficient information. Because of this, we assume possibility theory as the uncertainty model in place [4,5]. Under this assumption, the aforementioned distribution for each level of quality is a convex fuzzy number. We then develop a strategy to approximate this fuzzy number by taking

* Corresponding author.
E-mail addresses: antoon.bronselaer@ugent.be (A. Bronselaer), joachim.nielandt@ugent.be (J. Nielandt), guy.detre@ugent.be (G. De Tré).

https://doi.org/10.1016/j.ijar.2018.03.007
0888-613X/

k α-samples such that the different α values are uniformly spread over the unit interval. We show that this sample has linear space complexity in terms of k, has an approximation error of at most $1/(k-1)$ and allows for efficient, incremental updates.

The remainder of this paper is structured as follows. In Section 2, a review of the literature on data quality measurement is provided as well as a review of techniques for efficient and summarized representation of data. In Section 3, some preliminary notions concerning quality measurement and possibility theory are introduced. Next, in Section 4, we introduce the problem of efficiently representing a large set of measurements under circumstances of uncertainty. We first show how to *represent* this knowledge adequately and then introduce an efficient and incremental method of *constructing* such a representation. In Section 5, we present some evaluations to show the advantage of the incremental update method over some naive methods. Finally, the most important contributions of this paper are summarized in Section 6.

2. Related work

The modern view on data quality has been installed by Wang et al. [1,6] and Redman [7]. In their respective works, they observed that "quality" is a complex property of data that can not be measured directly. Instead, one should consider the different facets (i.e., *dimensions*) of quality that are relevant to a specific application and develop measurement procedures for those dimensions accordingly. This understanding has led to several approaches for assessment of data quality dimensions. In [8], a procedure for assessing completeness and accuracy has been proposed. Later work by the same authors focused on currency [9]. Completeness has also been investigated in the setting of distributed querying [10], where the usefulness of one or more sources is evaluated through their completeness. Ballou et al. have studied the trade-off between concurrent quality dimensions in a decision making context [11,12]. They argue that, in an economics-driven scenario, a lack of perfect data requires a mechanism that can select the best possible data by making a trade-off between several, possibly conflicting, quality dimensions of the data. They studied the balance between accuracy and timeliness [11] and the balance between consistency and completeness [13,12].

Whereas the previous methods are fairly simple procedures, there have been a number of attempts to develop measurement procedures for data quality that are rooted in Representational Measurement Theory [2]. A first attempt can be found in [14], where a procedure is described that expresses correctness and completeness on a ratio scale. However, this measurement procedure has some issues. For example, in their construction of a standard sequence, the authors choose a row in a table as the elementary unit. This choice is however not well motivated and it leaves aside the question of how to measure quality on the level of attributes. This problem is not easily solved because if we would choose an attribute as the elementary unit, then the difference in granularity of information between attributes is a non-trivial problem. One way of solving this problem, is to treat quality as uncertainty. More specifically, in [15], currency of data is measured as the probability that this data is still up-to-date. This approach relies on the rigour of well established frameworks of measurement (i.e., probability theory) and leads to well-defined procedures of measurement. In [16], this approach is modified to situations where parameter estimation is complicated by *Big Data issues* (heterogeneity of data structures, high volume...), resulting in a fuzzy rule-based approach.

In general, the parallel between quality and uncertainty can not always be drawn so easily. When measuring consistency for example, a model of uncertainty seems inapplicable. One could just want to evaluate some rules on the data instead. In [3], a general framework has been proposed in which data quality is essentially measured on an ordinal scale, but uncertainty on the level of quality is explicitly taken into account. This framework allows to combine rule-based approaches (e.g., for consistency and completeness) with uncertainty based approaches (e.g., for currency). In general, this means that the quality of data is modelled as an uncertainty distribution over a set of totally ordered quality levels. Needless to say, such a distribution conveys much more information than a single number. While this is an important benefit of the framework, the question rises how such information can be consolidated and/or summarized to higher levels.

The problem of aggregating multiple distributions into a summary, has been investigated under the umbrella of *linguistic summaries*. A fuzzy logic based technique using linguistically quantified propositions for data summarization has been initially proposed by Yager [17,18] and further developed in [19,20]. An example of a summary obtained with this technique is, e.g., '*all Andalusian cities have moderate winter temperatures*'. Later, the technique has been generalized using protoforms [21] and using interval-valued fuzzy sets [22]. Specific works focusing on the linguistic summarization of time series are [23,24] and [25]. Linguistic summaries have also been studied from the perspective of predictive statements like if-then rules [26]. Other related work includes research on the exploitation of multidimensional data characteristics in summarization [27], on the use of ontologies for generating data summaries [28] and on the use of fuzzy sets and incremental algorithms [29]. Of special interest is the study of a reduction algorithm for generated data summaries [30]. More recently, some researchers have studied linguistic summaries in the context of large data collections. In [31], it is studied how linguistic summaries can be extracted from graph databases. It is pointed out that the specific structure of graph databases requires specific types of linguistic summaries and different types of summaries are proposed. In [32], an overview of techniques for fuzzy quantification is presented. Recent research has also focused on consistency of linguistic summaries that are generated on very large datasets [33,34].