Data & Knowledge Engineering 99 (2015) 126–150

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Cardinality constraints on qualitatively uncertain data

Neil Hall [a], Henning Koehler [b], Sebastian Link [a,*], Henri Prade [c], Xiaofang Zhou [d,e]

[a] Department of Computer Science, The University of Auckland, New Zealand
[b] School of Engineering & Advanced Technology, Massey University, New Zealand
[c] IRIT, CNRS Université de Toulouse III, France
[d] School of Information Technology and Electrical Engineering, The University of Queensland, Australia
[e] Soochow University, Suzhou, China

ARTICLE INFO

Article history:
Received 22 March 2015
Accepted 9 June 2015
Available online 2 July 2015

Keywords:
Data and knowledge visualization
Data models
Database semantics
Management of integrity constraints
Requirements engineering

ABSTRACT

Modern applications require advanced techniques and tools to process large volumes of uncertain data. For that purpose we introduce cardinality constraints as a principled tool to control the occurrences of uncertain data. Uncertainty is modeled qualitatively by assigning to each object a degree of possibility by which the object occurs in an uncertain instance. Cardinality constraints are assigned a degree of certainty that stipulates on which objects they hold. Our framework empowers users to model uncertainty in an intuitive way, without the requirement to put a precise value on it. Our class of cardinality constraints enjoys a natural possible world semantics, which is exploited to establish several tools to reason about them. We characterize the associated implication problem axiomatically and algorithmically in linear input time. Furthermore, we show how to visualize any given set of our cardinality constraints in the form of an Armstrong sketch. Even though the problem of finding an Armstrong sketch is precisely exponential, our algorithm computes a sketch with conservative use of time and space. Data engineers may therefore compute Armstrong sketches that they can jointly inspect with domain experts in order to consolidate the set of cardinality constraints meaningful for a given application domain.

1. Introduction

1.1. Background

The notion of cardinality constraints is fundamental for understanding the structure and semantics of data. In traditional conceptual modeling, cardinality constraints were introduced in Chen's seminal paper [7]. They have attracted significant interest and tool support ever since. Intuitively, a cardinality constraint consists of a set of attributes and a positive integer b, and holds in an instance if there are no $b + 1$ distinct objects in the instance that have matching values on all the attributes of the constraint. For example, bank customers with no more than 5 withdrawals from their bank account per month may qualify for a special interest rate. Traditionally, cardinality constraints empower applications to control the occurrences of certain data, and therefore have significant applications in data cleaning, integration, modeling, processing, and retrieval.

1.2. Motivation

Traditional conceptual modeling was targeted at certain data for applications such as accounting, inventory and payroll. Modern applications, such as information extraction, radio-frequency identification (RFID), scientific data management, data cleaning, and

* Corresponding author.
E-mail addresses: hall.neil@gmail.com (N. Hall), h.koehler@massey.ac.nz (H. Koehler), s.link@auckland.ac.nz (S. Link), prade@irit.fr (H. Prade), zxf@itee.uq.edu.au (X. Zhou).

http://dx.doi.org/10.1016/j.datak.2015.06.002
0169-023X/

financial risk assessment produce large volumes of uncertain data. For example, RFID can track movements of endangered species of animals, such as the Indiana bat in Georgia, USA. For such an application, data comes in the form of objects associated with some discrete level of confidence in the signal reading; for example based on the quality of the signal received. More generally, uncertainty can be modeled qualitatively by associating objects with the degree of possibility (p-degree) that the object is perceived to occur in the instance. Fig. 1 shows such a possibilistic instance (p-instance), where each object is associated with an element from a finite scale of p-degrees: $\alpha_1 > \ldots > \alpha_{k+1}$. The top degree α_1 is reserved for objects that are 'fully possible', the bottom degree α_{k+1} for objects that are 'impossible' to occur. Intermediate degrees are used as required and linguistic interpretations attached as preferred, such as 'quite possible' (α_2) and 'somewhat possible' (α_3).

As this scenario is typical for a broad range of applications, we investigate in this article how cardinality constraints can benefit from the p-degrees assigned to objects. More specifically, we investigate cardinality constraints on uncertain data, where uncertainty is modeled qualitatively in the form of p-degrees.

The degrees of possibility are a natural source for extending the expressivity of traditional cardinality constraints. In fact, our use of p-degrees enjoys a natural possible world semantics, as illustrated on the running example in Fig. 1. Here, the world w_1 contains the RFID readings of high quality only, that is, all the objects with p-degree α_1. The world w_2 contains RFID readings of high or good quality, that is, all the objects with p-degree α_1 or α_2. Finally, world w_3 contains RFID readings of high, good, or low quality, that is, all the objects with p-degree α_1, α_2 or α_3. This possible world semantics enables us to express traditional cardinality constraints with different degrees of certainty. The certainty by which a traditional cardinality constraint holds is derived from the possible worlds in which it holds.

For example, we can express that for all low, good, and high quality readings, there are at most three readings recorded in the same zone, by declaring the cardinality constraint $card(Zone) \leq 3$ to be 'fully certain'. That is, $card(Zone) \leq 3$ must hold in the largest possible world w_3, and therefore also in all the worlds it contains. Similarly, we can express that for all good and high quality readings, at most two bats are recorded in the same zone at the same time, by declaring the cardinality constraint $card(Zone, Time) \leq 2$ to be 'quite certain'. That is, $card(Zone, Time) \leq 2$ must hold in the second largest possible world w_2, but not necessarily in the largest world w_3. Finally, we can express that for all high quality readings, the zone and time together identify the bat, by declaring the cardinality constraint $card(Zone, Time) \leq 1$ to be 'somewhat certain'. That is, $card(Zone, Time) \leq 1$ must hold in the smallest possible world w_1, but not necessarily in the worlds w_2 or w_3.

1.3. Contributions

Our objective is to apply possibility theory from artificial intelligence to establish qualitative cardinality constraints (QCs) as a fundamental tool to control the occurrences of uncertain data. Our contributions can be summarized as follows:

- **Modeling**. We introduce qualitative cardinality constraints as a class of integrity constraints on uncertain data. Here, uncertainty is modeled qualitatively by assigning to each object a degree of possibility with which it occurs in the instance. The p-degrees bring forward a nested chain of possible worlds, with each world being a classic instance that has some possibility. Hence, the higher the possibility of a world the fewer objects it contains. This empowers us to assign degrees of certainty to cardinality constraints, stipulating to which possible worlds they apply. The degrees of certainty (c-degree) are usually denoted by $\beta_1 > \ldots > \beta_k > \beta_{k+1}$, where β_{k+1} denotes the bottom c-degree reserved for constraints that are satisfied by any p-instance. Cardinality constraints that apply to the largest possible world hold with 'full certainty', denoted by the top c-degree β_1, while cardinality constraints that apply to the smallest possible world are only 'somewhat certain' to hold, denoted by the c-degree β_k. Fig. 1 shows the possible

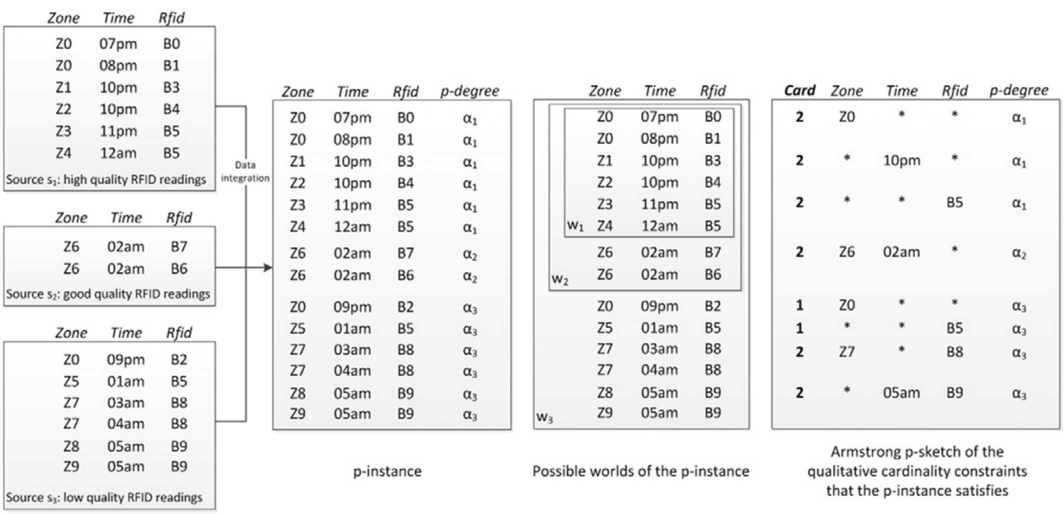

Fig. 1. P-instance and its possible worlds as the result of integrating RFID readings of different qualities; Armstrong p-sketch of the qualitative cardinality constraints that the p-instance satisfies.