Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 298 (2016) 4–21
FUZZY sets and systems
www.elsevier.com/locate/fss

A graded approach to database repair by context-aware distance semantics

Ofer Arieli [a,*], Anna Zamansky [b]

[a] *School of Computer Science, The Academic College of Tel-Aviv, Israel*
[b] *Department of Information Systems, University of Haifa, Israel*

Received 9 October 2014; received in revised form 23 May 2015; accepted 14 June 2015
Available online 18 June 2015

Abstract

The problem of inconsistent information in databases often arises in the context of data integration and data exchange. In these areas the common assumption is that the real world is consistent, thus an inconsistent database does not correspond to any reliable state and it needs to be "repaired" according to a chosen policy. Many of these policies have to deal with the problem of an exponential blowup in the number of possible repairs. For this reason, recent approaches advocate more flexible and fine-grained policies based on the user's preference. In this paper we take a further step towards more personalized inconsistency management by incorporating ideas from context-aware systems. The idea is to employ grades of different repairs according to their relevance for a particular user. The outcome is a graded framework for inconsistency maintenance in database systems, controlled by context-aware and distance-based considerations.

1. Introduction

Inconsistency handling in constrained databases is a primary issue in the context of consistent query answering, data integration, and data exchange. The general view in such cases is that the inconsistent database does not provide a faithful description of its domain of discourse and therefore it should be "repaired" so that its consistency will be restored. The standard approaches to this issue are usually based on the principle of minimal change, aspiring to achieve consistency via a minimal amount of data modifications (see, e.g., [4,12,13,22]). A key question in this respect is how to *choose* among the different possibilities of restoring the consistency of a database (i.e., 'repairing' it).

Earlier approaches to inconsistency management were based on the assumption that there should be some fixed, pre-determined way of repairing a database. Recently, there has been a paradigm shift towards user-controlled inconsistency management policies. Works taking this approach provide a possibility for the user to express some *preference* over all possible database repairs, preferring certain repairs to others (Some examples are [35,41,46,60];

* Corresponding author.
E-mail addresses: oarieli@mta.ac.il (O. Arieli), annazam@is.haifa.ac.il (A. Zamansky).

http://dx.doi.org/10.1016/j.fss.2015.06.007

See [56] for a survey of other related works). While such approaches provide the user with flexibility and control over inconsistency management, in reality they entail a considerable technical burden on the user's shoulders of calibrating, updating and maintaining preferences or policies. Moreover, in many cases these preferences should be *dynamic*, changing quickly on the go (e.g., depending on the user's geographical location). In the era of ubiquitous computing, with database systems practically everywhere, database users have little technical background, and even less time to dwell on the technical details of inconsistency management. As a consequence, there is a frequent demand for *easy* – and sometimes even *fully automatic* – inconsistency management solutions with little cognitive load, which are still expected to be *personalized* and tuned for particular needs. This leads to the idea of introducing *context-awareness* into inconsistency management.

Context-awareness is defined as the use of contexts to provide task-relevant information and services to the user (see [1]). We believe that inconsistency management has natural relations to the concept of context. Accordingly, the goal of this paper is to incorporate notions and techniques that have been studied by the context-aware computing community (see, e.g., [20] and [54]) into consistency management in database systems. For this, we use a logical approach for preferring repairs, which combines the following two grading ingredients:

- *Distance-based semantics* for restoring the consistency of inconsistent databases according to the principle of minimal change, and
- *Context-awareness considerations*, based on graded ranking, for incorporating user preferences.

Example 1. Let us consider the following simple database instance:

empNum	name	address	salary
1	John	Tower Street 3, London, UK	70K$
1	John	Herminengasse 8, Wien, AT	80K$
2	Mary	42 Street 15, New York, US	90K$

A functional dependency that may be violated in this case is empNum → salary, stating that the salary of an employee is uniquely determined by the employee's number. Thus, assuming that this database contains information coming from several equally reliable sources, one has to resolve the inconsistency in it, although each source could have provided a completely consistent data. Minimal change considerations (which will be expressed in what follows by distance functions) imply in this case that it is enough to delete either the first or the second tuple for restoring consistency. Now, the decision which tuple to delete may be *context-dependent*. For instance, for tax assessments tuples with higher salaries may be preferred, while tuples with lower salaries may have higher priority when loans or grants are considered.

The choice between the first two tuples may also be determined by other, more dynamic considerations. For instance, it is quite possible that two different employees called John were assigned the same employee number by mistake. Alternatively, the same employee (John) may have two different addresses in two different countries, but the salary information associated with at least one of them is erroneous. In either cases, a user located in Austria is most probably interested in the Austrian address (or the Austrian employee), while a user located in the UK will make the most out of the other address (or employee).

The rest of this paper is organized as follows. In Section 2 we review some of the basic definitions of database concepts and distance-based semantics. In Section 3 we show how context-awareness can be modeled in our framework using a graded approach, and incorporate context-aware considerations into distance-based semantics. In Section 4 we consider some applications of our approach and in Section 5 we discuss some future work and conclude.[1]

2. Inconsistent databases and distance semantics

To simplify of the presentation, in this paper we remain on the propositional level and reduce first-order databases to our framework by grounding them. In the sequel, $\mathcal{L}$ denotes a propositional language with a *finite* set of atomic formulas $\mathsf{Atoms}(\mathcal{L})$. An *$\mathcal{L}$-interpretation* I is an assignment of a truth value in $\{T, F\}$ to every element in $\mathsf{Atoms}(\mathcal{L})$.

[1] This paper is an extension of the work presented in the 35th Linz Seminar on Fuzzy Set Theory, dedicated to Graded Logical Approaches and their Applications (Linz, Austria, February 2014). A short version of this paper was also published in [64].