# Handling Uncertainty in Relational Databases with Possibility Theory - A Survey of Different Modelings

Olivier Pivert[1(✉)] and Henri Prade[2]

[1] ENSSAT-Lannion, IRISA, Rennes, France
pivert@enssat.fr
[2] IRIT – CNRS, 118, route de Narbonne, 31062 Toulouse Cedex 09, France
prade@irit.fr

**Abstract.** Mainstream approaches to uncertainty modeling in relational databases are probabilistic. Still some researchers persist in proposing representations based on possibility theory. They are motivated by the ability of this latter setting for modeling epistemic uncertainty and by its qualitative nature. Interestingly enough, several possibilistic models have been proposed over time, and have been motivated by different application needs ranging from database querying, to database design and to data cleaning. Thus, one may distinguish between four different frameworks ordered here according to an increasing representation power: databases with (i) layered tuples; (ii) certainty-qualified attribute values; (iii) attribute values restricted by general possibility distributions; (iv) possibilistic c-tables. In each case, we discuss the role of the possibility-necessity duality, the limitations and the benefit of the representation settings, and their suitability with respect to different tasks.

**Keywords:** Possibility theory · Relational databases · Uncertainty
Inconsistency · Data cleaning

## 1 Introduction

Many authors have made proposals to model and handle relational databases involving uncertain data. In particular, the last two decades have witnessed a blossoming of researches on this topic (cf. [29] for a survey of probabilistic approaches). Even though most of the literature about uncertain databases uses probability theory as the underlying uncertainty model, some approaches rather rest on possibility theory [30]. The initial idea of applying possibility theory to this issue goes back to the early 1980's [26,27]. This was short after the introduction of the idea of a "fuzzy database", for which various proposals were made, ranging from fuzzy relations (thus having weighted tuples) to ordinary relations with tuples of fuzzy values (represented by fuzzy sets), or more simply with tuples of weighted values. These different views developed by several authors

D. Ciucci et al. (Eds.): SUM 2018, LNAI 11142, pp. 396–404, 2018.
https://doi.org/10.1007/978-3-030-00461-3_30

were not necessarily referring to possibility theory; see [4] for references. Since this time, several possibilistic representations have been introduced, and it is useful to clarify their respective roles.

As we will discuss in Sect. 3, the possibilistic framework constitutes an interesting alternative to the probabilistic one, notably because of its qualitative nature. In this paper, we provide a survey of different modelings of uncertain data with possibility theory. The remainder is structured as follows. In Sect. 2, we recall some notions about uncertain databases and their interpretation in terms of possible worlds. Section 3 is devoted to a presentation of four possibilistic database models, with different levels of expressiveness. Section 4 discusses a specific topic where uncertain data management can play a role, namely data cleaning. Section 4.2 points out a sample of issues deserving further investigations. Finally, Sect. 5 concludes the paper and outlines some short-term research perspectives.

## 2 About Uncertain Databases and Possible Worlds

In the context of uncertain databases, two kinds of uncertainty are considered: tuple-level uncertainty (where the existence of some tuples in a relation is uncertain, i.e., is more or less probable/possible) and attribute-level uncertainty (where some attribute values in some tuples may be ill-known or uncertainly known). The latter case can be seen as more general than the former, since a tuple involving uncertain attribute values may be translated into a set of mutually exclusive uncertain tuples (involving only ordinary attribute values). An attribute value represented as a disjunctive weighted set can be interpreted as a probability distribution or a possibility distribution depending on the underlying uncertainty model considered. From a semantic point of view, an uncertain database $D$ can be interpreted as a set of usual databases, called possible worlds $W_1$, ..., $W_p$, and the set of all interpretations of $D$ is denoted by $rep(D) = \{W_1$, ..., $W_p\}$. Any world $W_i$ is obtained by choosing a value in each disjunctive set appearing in $D$. One of these (regular) databases is supposed to correspond to the actual state of the universe modeled. The assumption of independence between the sets of candidates is usually made and then any world $W_i$ corresponds to a conjunction of independent choices, thus the probability, or possibility, degree associated with a world is computed using a conjunction operator, namely, the product, or "min", respectively.

When processing a query, a naive way of doing would be to make explicit all the interpretations of $D$ in order to query each of them. Such an approach is intractable in practice and it is of prime importance to find a more realistic alternative. To this end, the notion of a representation system was introduced by Imielinski and Lipski [14]. The basic idea is to represent both initial tables and those resulting from queries in such a way that the representation of the result of a query $q$ against any database $D$ denoted by $q(D)$, is equivalent (in terms of worlds) to the set of results obtained by applying $q$ to every interpretation of $D$, i.e.: $rep(q(D)) = q(rep(D))$ where $q(rep(D)) = \{q(W) \mid W \in rep(D)\}$. If this property holds for a representation system $\rho$ and a subset $\sigma$ of the relational algebra, $\rho$ is called a *strong representation system* for $\sigma$. From a querying point of view, this property enables a direct (or compact) calculus of a query $q$, which then applies to $D$ itself without making the worlds explicit.

## 3 Possibilistic Uncertainty

We first recall some distinctive features of possibility theory before reviewing the different possibilistic representations.

### 3.1 Possibility Theory

Possibility theory departs from probability theory in several respects. Possibility theory involves two dual set functions: the possibility $\Pi$ and the necessity $N$ such that $N(A) = 1 - \Pi(\bar{A})$, while probability is self-dual, namely $P(A) = 1 - P(\bar{A})$. This provides room for modeling epistemic uncertainty, including total ignorance. Indeed, $\Pi(A) = 1$ does not prevent to have also $\Pi(\bar{A}) = 1$ in case of complete ignorance about $A$ (while $\Pi(A) = P(\bar{A})(= 1/2)$ does not distinguish situations of genuine equiprobability from situations where, due to ignorance, one applies the Insufficient Reason Principle). $\Pi$ (and $N$) are associated with a possibility distribution $\pi$, defined from a universe $U$ to a scale such as scale [0, 1], where $\forall A \subseteq U, \Pi(A) = \max_{u \in A} \pi(u)$. Due to the use of max and min operations, possibility and necessity functions are more "qualitative" than the probabilistic models involving sum and product.

Still, possibility theory may be quantitative or qualitative [8]. In the first case, the whole scale [0, 1] is used, and possibility and necessity may be thought as upper and lower bounds of an unknown probability (then conditioning is based on product rather than "min"). However, possibility theory does not require the use of the scale [0, 1], but can be defined with any linearly ordered chain (e.g., a finite subset [0, 1] including 0 and 1), or more generally any lattice, and is then qualitative. Moreover, possibility theory has a logical counterpart, namely possibilistic logic [6] (which involves only lower bounds of necessity degrees, which can be viewed as certainty levels), and generalized possibilistic logic [11] (which involves both set functions). Besides, two other set functions are of interest in possibility theory, namely the guaranteed possibility, $\Delta(A) = \min_{u \in A} \delta(u)$, and the dual set function, where $\delta$ is a possibility distribution. In bipolar representations [9], one uses a pair of possibility distributions $(\delta, \pi)$ for distinguishing between values $u$ such as $\pi(u) = 0$ that are excluded, from values $u'$ such as $\delta(u') > 0$ that are guaranteed to be possible to some extent (since, e.g., they were observed), assuming the consistency condition $\delta \leq \pi$ (expressing that what is guaranteed to be possible cannot be excluded).

### 3.2 Possibilitistic Representations

There is not a unique possibilistic data model. The existing models serve different purposes. From the least to the most expressive, we can distinguish four possibilistic models for uncertain data which have been actually proposed:

- databases with layered tuples;
- tuples involving certainty-qualified attribute values;
- tuples involving attribute values restricted by possibility distributions;
- possibilistic *c*-tables.

**Layered Tuples.** The idea, here, is just to provide a complete ordering of the tuples in the database according to the more or less strong confidence we have in their truth. This can be easily encoded by associating a possibility level with each tuple. This results in a layered database: all the tuples having the same degree are in the same layer (and only them). Those tuples having a possibility level equal to 1 may also be associated with a certainty level equal to 1, while the others with a possibility level strictly less than 1 are not certain at all; this means that any possible world database contains all the tuples at level 1, while the other tuples may or may not be present in a particular possible world; see [17] for details. This modeling is not very expressive since it provides no indication on what attribute values in the tuple are particularly uncertain. In that respect, it may be considered as a modeling that is too poor from a querying perspective. Still, it has been shown useful for design purposes by providing a setting for attaching certainty levels to functional dependencies (FDs) (through a duality relation with the possibility levels of the tuples that are violating the FDs). Then, this enables the generalization of Armstrong's axioms by attaching certainty levels, and the extension of Boyce-Codd/3rd Normal Forms approaches to database design in the presence of uncertain tuples, by taking advantage of the levels [18]. Such a possibilistic model is also useful for handling keys [15] and cardinality constraints [12,28] in presence of uncertain data.

**Certainty-Qualified Attribute Values.** In this model [23], attribute values (or disjunctions thereof) are associated with a certainty level (which is the lower bound of the value of a necessity function). This amounts to associating each attribute value with a simplified type of possibility distribution restricting it[1]. Different attributes in a tuple may have different certainty levels associated with their respective values. Then a tuple may be associated with a certainty level, which is the minimum of the certainty levels associated with the attribute values of the tuple, in agreement with the minitivity of necessity functions. Still this global certainty level should not be confused with the possibility level of the

[1] Then the attribute value, or more generally the disjunction of possible values is/are considered as fully possible, while any other value in the attribute domain is all the less possible as the certainty level is higher. In case of full certainty these other values are all impossible. This is a particular case of the certainty qualification of a fuzzy set, here reduced to a singleton, or in any case to a classical subset. There are other basic qualifications of a fuzzy set in possibility theory, for instance in terms of guaranteed possibility (rather than in terms of necessity as in certainty qualification), or which lead to enlarge the core, or to reduce the support of the fuzzy set, see [7] for the four canonic transformations; see also [20] for hybrid transformations combining enlargement with uncertainty.

previous approach. In terms of possible worlds, a tuple associated with such a certainty level correspond to *several* tuples with a possibility level. Indeed consider the simple example of a tuple made of two attribute values $a$, and $b$, associated respectively with certainty $\alpha$ and $\beta$: this yields as possible worlds $\langle a, b\rangle$ with possibility 1, $\langle a', b\rangle$ with possibility $1 - \alpha$, $\langle a, b'\rangle$ with possibility $1 - \beta$, $\langle a', b'\rangle$ with possibility $\min(1 - \alpha, 1 - \beta)$, where $a'$ (resp. $b'$) is any value distinct from $a$ (resp. $b$) in the attribute domain to which $a$ (resp. $b$) belongs.

This model has some advantages with respect to querying: (i) it constitutes a strong representation system for the whole relational algebra (up to some minor restrictions); (ii) it does not require the use of any lineage mechanism and the query complexity is close to the classical case; (iii) the approach seems more robust with respect to small changes in the value of degrees than a probabilistic handling of uncertainty (see the last section of [23]). Moreover, there exists a simplified version of this model, see [24], that uses a scale with only three certainty levels ("completely certain","somewhat certain", "not at all certain"). This makes the assessment of certainty particularly easy. Besides, another approach with the same formal type of modeling, but where certainty is evaluated in terms of subsets of sources (together with their reliability level) makes it possible to rank-order the answers to a query also on such a basis [22].

**Attribute Values Restricted by General Possibility Distributions.** In this "full possibilistic model" [3], any attribute value can be represented by any possibility distribution. Moreover, representing the result of some relational operations (in particular the join) in this model requires the expression of dependencies between candidate values of different attributes in the same tuple, which leads to the use of nested relations. In [3], it is shown that this model is a strong representation system for selection, projection and foreign-key join only. The handling of the other relational operations requires the use of a lineage mechanism as in the probabilistic approaches. This model makes it possible to compute not only the more or less certain answers to a query (as in the previous model), but also the answers which are only possible to some extent.

**Possibilistic *c*-tables.** This model is outlined in [25]. The possibilistic extension of *c*-tables preserves all the advantages of classical *c*-tables (for expressing constraints linking attribute values) while the attribute values are restricted by any kind of possibility distribution. This model generalizes the two previous ones. In fact, possibilistic *c*-tables, as probabilistic *c*-tables, can be encompassed in the general setting of the semiring framework proposed by Val Tannen *et al.*

## 4 Data Cleaning

This section first provides a brief overview of two approaches that respectively (i) allow you to query inconsistent databases, and (ii) take advantage of a possibilistic modeling for cleaning the data, before suggesting new lines of research.

### 4.1 Some Existing Approaches

In the presence of inconsistent data, two points of view may be taken. The first one consists in cleaning the database so as to make it consistent, either by means of an automated process [13], or by an interactive approach. The second one, such as Consistent Query Answering (CQA) approaches [2], takes into account the inconsistencies at query processing time.

An approach corresponding to this second line of thought is described in [21]. It aims at warning the user about the presence of suspect answers in a selection query result, in the context of a classical database (that may include data inconsistent with some functional dependencies). Roughly speaking, the idea is that such elements can be identified inasmuch as they can also be found in the result of negative associated queries. The notion of a suspect answer can be refined by introducing some gradedness in terms of cardinality (number of functional dependency violations in which the tuple is involved) or similarity (by relaxing the equality constraint of a functional dependency into an approximate equality). However, this approach, for the moment, does not involve any uncertainty degree associated with attribute values or tuples. In other words, it handles only inconsistency but not uncertainty.

A possibilistic approach to data cleaning has been recently proposed in [16]. This approach belongs to the research trend aimed at restoring a form of consistency in the database. Still, the approach identifies tuples that are suspect or even fraudulous. This is done independently from any particular query. This relies on a model closely related to the layered-tuple-based model reviewed above. However, it is used in the reverse way, since it starts with certainty-valued constraints (called business rules) from which one computes the confidence levels associated with the tuples (on a qualitative scale: “normal”/“suspect”/“fraud”), by solving a minimal possibilistic vertex cover (taking into account the number of violations in which the tuples are involved). Here these are the possibility levels of the tuples that are revised in order to restore the (graded) consistency.

### 4.2 Some Issues Deserving Further Investigations

A first extension we may think of is to introduce certainty degrees in the first of the two approaches reviewed in the preceding section (reference [21]). This means extending the querying method keeping the data as they are and indicating which answers are suspect, to the setting of the certainty-based model described in Sect. 3. In the original model, an answer is suspect as soon as there exists a repair (w.r.t. a functional dependency) of the query result to which it does not belong. In the extended context, the notion of repair becomes naturally graded[2], as well as the concept of suspiciousness (now appreciated both in terms of the certainty degrees attached to the values of the concerned tuples and in terms of the number of functional dependencies violated by the tuples).

[2] For example, assume Peter has two ages, each with a certainty level, the levels being denoted by $\alpha$ and $\beta$ respectively. Then the FD $\mathit{name} \rightarrow \mathit{age}$ is violated with a certainty degree that is equal to $\min(\alpha, \beta)$.

Another interesting issue is to unify the above view with the possibilistic approach to data cleaning reviewed in the previous section (reference [16]). We can observe that, although the outputs of the two approaches are quite similar (tuples assigned with a certainty degree expressing different levels of suspiciousness), the inputs are completely different: in one case, constraints with certainty levels, in the other case, attribute values with certainty levels. However, it seems clear that the approach [21] can also be extended by introducing functional dependencies with certainty levels and keeping all of the attribute values completely certain (rather than the opposite as suggested in the paragraph above). Then, this will make the two approaches easier to compare.

## 5 Conclusion

In this brief survey, we have tried to make clear that there exist different possibilistic models, with different levels of expressiveness, but also dedicated to different database tasks (design, data cleaning, querying). Other worth mentioning issues are the modeling of null values [1] and the extrapolation of missing data [5]. Two kinds of tasks, in our opinion, are particularly worth investigating: (i) a practical comparison of the certainty-based model (which offers a rather good simplicity/expressivity compromise) with probabilistic approaches; (ii) the comparison and the cooperation between different possibilistic data cleaning tools and probabilistic ones. Another line of thought which, we think, might be of interest, is to consider causality issues for evaluating the responsibility in inconsistencies, for which AI probabilistic models have been considered in a database perspective [19], while there also exist possibilistic counterparts to these AI models [10].

## References

1. Arrazola, I., Plainfossé, A., Prade, H., Testemale, C.: Extrapolation of fuzzy values from incomplete data bases. Inf. Syst. **14**(6), 487–492 (1989)
2. Bertossi, L.E.: Database Repairing and Consistent Query Answering. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, San Rafael (2011)
3. Bosc, P., Pivert, O.: About projection-selection-join queries addressed to possibilistic relational databases. IEEE Trans. Fuzzy Syst. **13**(1), 124–139 (2005)
4. Bosc, P., Prade, H.: An introduction to the fuzzy set and possibility theory-based treatment of flexible queries and uncertain or imprecise databases. In: Motro, A., Smets, P. (eds.) Uncertainty Management in Information Systems. From Needs to Solutions, pp. 285–324. Kluwer Academic Publishers, Dordrecht (1997)
5. De Tré, G., De Caluwe, R.M.M., Prade, H.: Null values in fuzzy databases. J. Intell. Inf. Syst. **30**(2), 93–114 (2008)
6. Dubois, D., Lang, J., Prade, H.: Automated reasoning using possibilistic logic: semantics, belief revision, and variable certainty weights. IEEE Trans. Knowl. Data Eng. **6**, 64–71 (1994)
7. Dubois, D., Prade, H.: What are fuzzy rules and how to use them. Fuzzy Sets Syst. **84**(2), 169–185 (1996)

8. Dubois, D., Prade, H.: Possibility theory: qualitative and quantitative aspects. In: Gabbay, D.M., Smets, P. (eds.) Quantified Representation of Uncertainty and Imprecision. Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 1, pp. 169–226. Kluwer, Dordrecht (1998)
9. Dubois, D., Prade, H.: An overview of the asymmetric bipolar representation of positive and negative information in possibility theory. Fuzzy Sets Syst. **160**(10), 1355–1366 (2009)
10. Dubois, D., Prade, H.: A glance at causality theories for artificial intelligence. In: A Guided Tour of Artifial Intelligence, vol. 1: Knowledge Representation, Reasoning and Learning. Springer (2018)
11. Dubois, D., Prade, H., Schockaert, S.: Generalized possibilistic logic: foundations and applications to qualitative reasoning about uncertainty. Artif. Intell. **252**, 139–174 (2017)
12. Hall, N., Köhler, H., Link, S., Prade, H., Zhou, X.: Cardinality constraints on qualitatively uncertain data. Data Knowl. Eng. **99**, 126–150 (2015)
13. Ilyas, I.F., Chu, X.: Trends in cleaning relational data: consistency and deduplication. Found. Trends Databases **5**(4), 281–393 (2015)
14. Imielinski, T., Lipski, W.: Incomplete information in relational databases. J. ACM **31**(4), 761–791 (1984)
15. Koehler, H., Leck, U., Link, S., Prade, H.: Logical foundations of possibilistic keys. In: Fermé, E., Leite, J. (eds.) JELIA 2014. LNCS (LNAI), vol. 8761, pp. 181–195. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11558-0_13
16. Köhler, H., Link, S.: Qualitative cleaning of uncertain data. In: Mukhopadhyay, S., et al. (eds.) Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, 24–28 October 2016, pp. 2269–2274. ACM (2016)
17. Link, S., Prade, H.: Possibilistic functional dependencies and their relationship to possibility theory. IEEE Trans. Fuzzy Syst. **24**(3), 757–763 (2016)
18. Link, S., Prade, H.: Relational database schema design for uncertain data. In: Mukhopadhyay, S., et al. (eds.) Proceedings of the 25th ACM International conference on Information and Knowledge Management, CIKM 2016, Indianapolis, 24–28 October, pp. 1211–1220 (2016)
19. Meliou, A., Roy, S., Suciu, D.: Causality and explanations in databases. PVLDB **7**(13), 1715–1716 (2014)
20. González, A., Marín, N., Pons, O., Vila, M.A.: Qualification of fuzzy statements under fuzzy certainty. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.) IFSA 2007. LNCS (LNAI), vol. 4529, pp. 162–170. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72950-1_17
21. Pivert, O., Prade, H.: Detecting suspect answers in the presence of inconsistent information. In: Lukasiewicz, T., Sali, A. (eds.) FoIKS 2012. LNCS, vol. 7153, pp. 278–297. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28472-4_16
22. Pivert, O., Prade, H.: Querying uncertain multiple sources. In: Straccia, U., Calì, A. (eds.) SUM 2014. LNCS (LNAI), vol. 8720, pp. 286–291. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11508-5_24
23. Pivert, O., Prade, H.: A certainty-based model for uncertain databases. IEEE Trans. Fuzzy Syst. **23**(4), 1181–1196 (2015)
24. Pivert, O., Prade, H.: Database querying in the presence of suspect values. In: Morzy, T., Valduriez, P., Bellatreche, L. (eds.) ADBIS 2015. CCIS, vol. 539, pp. 44–51. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23201-0_6

25. Pivert, O., Prade, H.: Possibilistic conditional tables. In: Gyssens, M., Simari, G. (eds.) FoIKS 2016. LNCS, vol. 9616, pp. 42–61. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30024-5_3
26. Prade, H.: Lipski's approach to incomplete information databases restated and generalized in the setting of Zadeh's possibility theory. Inf. Syst. **9**(1), 27–42 (1984)
27. Prade, H., Testemale, C.: Generalizing database relational algebra for the treatment of incompleteuncertain information and vague queries. Inf. Sci. **34**, 115–143 (1984)
28. Roblot, T.K., Link, S.: Possibilistic cardinality constraints and functional dependencies. In: Comyn-Wattiau, I., Tanaka, K., Song, I.-Y., Yamamoto, S., Saeki, M. (eds.) ER 2016. LNCS, vol. 9974, pp. 133–148. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46397-1_11
29. Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic Databases. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, San Rafael (2011)
30. Zadeh, L.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. **1**, 3–28 (1978)