IJEMD-CSAI, 1 (2) (2022), 46 – 55 https://doi.org/ 10.54938/ijemdcsai.2022.01.2.141

International Journal of Emerging Multidisciplinaries:
Computer Science & Artificial Intelligence

Research Paper
Journal Homepage: www.ojs.ijemd.com
ISSN (print): 2791- 0164 ISSN (online): 2957-5036

A Relational Data Model for Uncertain Data

Sadam Hussain Channar[1], Muhammad Saleem Vighio[1,*]

[1] *Department of Computer Science, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh, Pakistan*
**Corresponding author*

Abstract

There are many traditional data models developed for the purpose of storing, managing, and manipulating certain data. However, data comes from different sources which may also include data which is imprecise or inexact. For an application which generates uncertain data, its storage, management, and manipulation becomes equally important as for the certain data. This study presents a relational data model for storing, managing, and manipulating uncertain data with certain data. The data model is analyzed using SQL queries to ensure the completeness of the model. Furthermore, a comparative analysis of proposed data model with existing data models is also presented.

Keywords: Uncertain data; Data model; Relational database.

1. Introduction

Data comes from different sources like mobile phones, transaction processing, and sensor nodes etc. There are many tools and technologies for storing, managing, and manipulating certain data i.e., the data which is accurate and precise. However, data may also come from sources which includes incomplete or imprecise data. For example, sensor networks, location tracking, and moving object typically generate large quantities of uncertain data (Aggarwal *et al.* 2009). The storage, management and manipulation of uncertain data is also important for applications as for the certain data. However, the procedure for handling uncertain data is much more complex than traditional databases. Therefore, it is highly needed that uncertain data may be stored, managed, and manipulated in a way which is easy and expressive (Aggarwal 2010). Traditional database management systems are not suitable for storing, manipulating, and retrieving uncertain data because such software are not developed for this purpose. As a result, several approaches have been proposed to deal with uncertainty in databases over the years. However, most of these methodologies make simplistic and restrictive assumptions concerning the kinds of uncertainties that can be represented (Deshpande *et al.* 2009). The basic difference between a traditional relational database management system

Email addresses: saleem.vighio@quest.edu.pk (M. S. Vighio)

and an uncertain relational database management system is that an uncertain relation represents a set of possible relation instances, instead of a single instance (Sarma *et al.* 2006).

The significance of uncertain data has been familiar with the quick development for collecting and managing data in numerous fields like finance, telecommunication, economy, and military of defense. Uncertainty is created in data items due to summarization of data. Another well recognized cause of uncertainty in data is heterogeneity of data. When two different databases represent dissimilar values for the similar data item, its real value is not known with confidence, and it becomes uncertain data. Furthermore, when the values for the same data items are not same it becomes very difficult to find which values are right. Uncertainties in data are categorized in two forms: Attribute Level Uncertainty (ALU) and Tuple Level Uncertainty (TLU).

Attribute Level Uncertainty

In Attribute Level Uncertainty, one is not sure about the values that an attribute of a tuple can take. An attribute may be assigned multiple values for a field in a tuple. This type of uncertainty is also called "value uncertainty" (Wang *et al.* 2013).

Tuple Level Uncertainty

In Tuple Level Uncertainty, it is uncertain whether tuple should be present in the database or not. Corresponding to every uncertain tuple, there are two possible situations: one which includes the tuple in the database while the other which does not; this type of uncertainty is also called "*existential uncertainty*" or "*May be tuple*" (Hayat and Khan, 2015).

Based on the uncertain data and its types, a question arises, how to store and manipulate uncertain data in a simple and comprehensive way? One alternate is to store values of such type of data items that are consistent; other alternate is to store the dissimilar possible values for such type of data items, identifying that there is some uncertainty related with those values. The standard relational data model doesn't have a comprehensive technique to deal with imperfect, inaccurate, and uncertain data (Sarkar and Dey, 2009). This paper presents a relational data model for storing, managing, and manipulating uncertain data with certain data in an easy and expressive way. The model is analyzed using queries for relational databases to verify that the model is complete to deal with uncertain data.

Overview: The rest of the paper is organized as follows: Section 2 presents methodology for implementing and analyzing uncertain data model. Section 3, gives details of data model for storing, managing, and manipulating uncertain data using relational database approach. This section also provides the analysis of the data model for a case study of students' admission process. Section 4 gives a comparative analysis of existing data models for uncertain data with our proposed data model. Finally, Section 5 gives concluding remarks.

2. METHODOLOGY

As shown in Figure 1, the first step is to create a database for an application generating uncertain data. In the next step uncertain data is separated from certain data using vertical partitioning approach. After the separation of uncertain data, the data is simplified. After the simplification certain and uncertain data will be merged by join operators for relational database. At the final step data will be analyzed for certain and uncertain data to prove the completeness of the data model.

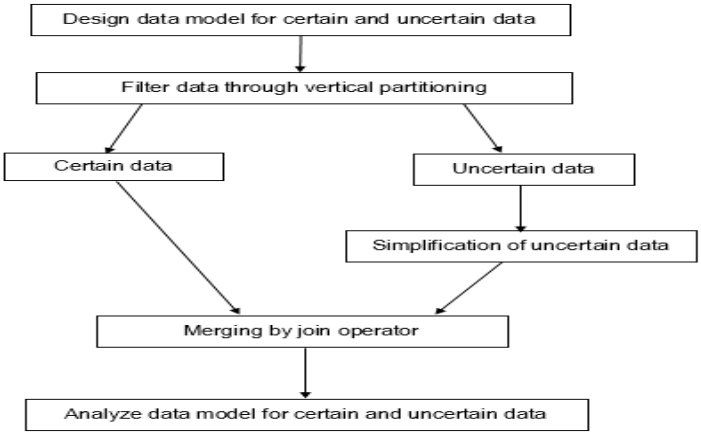

Figure 1: Proposed Research Process

3. DATA MODEL FOR UNCERTAIN DATA

To implement and analyze uncertain data model, a case study of an admission process is considered in which a number of candidates fill admission forms in graduate programs. Each admission form contains different fields like id, name of the candidate and last degree completed. The data included in these forms is to be stored in the database. As an example, we create ten records of the candidates, some records contain certain data and others contain uncertain data as shown in Figure 2. Due to the importance of data, data filled in the forms is to be stored in the database which may also include some uncertain data about the appropriate entries for some records of the forms as shown in Table 1. In these forms, there are two types of uncertainty cratered by the candidates. In some forms they tick two options out of four options which in Attribute Level Uncertainty, whereas, in some forms they tick all options or do not select any option which results in Tuple Level Uncertainty. In the database, ALU is annotated with vertical line "|" and TLU is annotated with a question mark "?".

Based on the data provided in the admission forms as given in Figure 1, a relational database has been created. As it can be seen from Table 1, a candidate with id 3 has ticked two options, whereas the candidate with id 8 has selected no option. We are uncertain about the choice of degree programs these students are applying for.

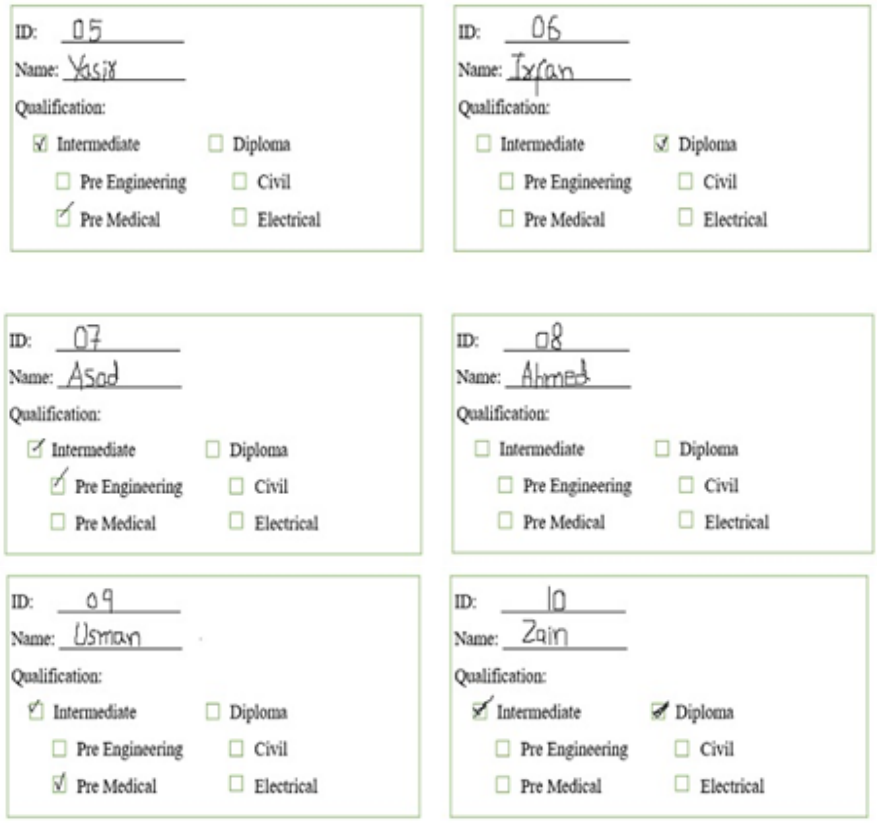

Figure 2: Filled Admission Forms

ID	Name	Qualification
01	Ali	PE
02	Aslam	CE
03	Bilal	PE | PM
04	Talha	EL
05	Yasir	PM
06	Irfan	CE | EL
07	Asad	PE
08	Ahmed	?
09	Usman	PM
10	Zain	?

Table 1: adm relation

After the database has been created, the next step is to separate certain data from uncertain data using vertical partitioning approach. Based on the primary key attribute, incomplete columns are separated using the following query:

select id, qualification into uncertain from adm

Attributes containing uncertain data are separated from the original relation and are put in a new relation named as **"uncertain",** see Table 2. Since the qualification attribute contains the uncertain data so it is put in a new relation along with the corresponding primary key attribute.

ID	Qualification
01	PE
02	CE
03	PE | PM
04	EL
05	PM
06	CE | EL
07	PE
08	?
09	PM
10	?

Table 2: Relation of uncertain data

As a next step, each distinct element of uncertain data contained in "**uncertain**" relation now becomes the attributes of "**simplified**" relation. For simplification of uncertain data, a function has been created for converting separated uncertain data into Boolean values of 1 and 0. Simplification query is given below:

```
declare @loop int = 0
while @loop < 10

begin
    execute orset @loop;
    set @loop = @loop + 1;
end
```

Above query declares a loop which is initialized by value 0. The loop runs 10 times in which the function **"orset"** is executed. Using the **"orset"** function, uncertain data attributes are given Boolean values of 1 and 0 as shown in Table 3.

ID	PE	PM	CE	EL

01	1	0	0	0
02	0	0	1	0
03	1	1	0	0
04	0	0	0	1
05	0	1	0	0
06	0	0	1	1
07	1	0	0	0
08	0	0	0	0
09	0	1	0	0
10	0	0	0	0

Table 3: Simplified relation

The Boolean values of 1 and 0 shown in Table 3 represent presence and absence of attributes in simplified relation.

After the simplification of uncertain data, certain and uncertain data is merged in a new relation named as **"resultant"** using join operator and based on the primary key attribute. Query for merging the data is shown below.

```
select id, name, PE, PM, CE, EL
into resultant from adm


join  qualified
on adm.id = q_id
```

In above query, id and name are selected from "**adm**" relation which is certain data and PE, PM, CE and EL are selected from "**simplified**" relation which represents uncertain data. The join operator is used to merge two relations based on common key attribute.

Following query retrieves all the attributes of the "**resultant**" relation which represent simplified uncertain and certain data as shown in Table 4.

```
select * from resultant
```

ID	Name	PE	PM	CE	EL
01	Ali	1	0	0	0
02	Aslam	0	0	1	0
03	Bilal	1	1	0	0
04	Talha	0	0	0	1
05	Yasir	0	1	0	0
06	Irfan	0	0	1	1
07	Asad	1	0	0	0
08	Ahmed	0	0	0	0
09	Usman	0	1	0	0
10	Zain	0	0	0	0

Table 4: Resultant relation

Analysis of the Data Model: In the last phase, we analyze certain and uncertain data using queries on the relational database in order to

ensure that the data stored in the database is complete and precise.

The query to retrieve all the attributes of the **resultant** relation which represent the records of PE is given below.

```
Select * from resultant
where PE = 1 and PM = 0 and
CE = 0 and EL = 0
```

The result of the above query is shown in Table 5 which shows that only two candidates have selected Pre-Engineering. This shows the certainty of the data.

ID	Name	PE	PM	CE	EL
01	Ali	1	0	0	0
02	Asad	1	0	0	0

Table 5: Result for PE

The query for retrieving records of students who have selected PM is given below.

```
select * from resultant
where PE = 0 and PM = 1 and
CE = 0 and EL = 0
```

ID	Name	PE	PM	CE	EL
05	Yasir	0	1	0	0
09	Usman	0	1	0	0

Table 6: Result for PM

Table 6 shows that only two candidates have selected Pre-Medical. This is also certain data.

The queries for retrieving records of candidates who have selected CE and EL are also written in the same way.

Following query retrieves the records of students who have selected both PM and PE disciplines.

```
select * from resultant
where PE = 1 and PM = 1 and
CE = 0 and EL = 0
```

The corresponding result of the above query is shown in Table 7.

ID	Name	PE	PM	CE	EL
03	Bilal	1	1	0	0

Table 7: Result for PE and PM

Since a candidate can select only one discipline whereas Table 7 shows a record of the candidate who has selected two disciplines so this data is considered as uncertain data and the corresponding uncertainty is called Attribute Level Uncertainty.

As another example, following query retrieves the records of students who have selected no discipline.

select * from resultant
where PE = 0 and PM = 0 and
CE = 0 and EL = 0

ID	Name	PE	PM	CE	EL
08	Ahmed	0	0	0	0
10	Zain	0	0	0	0

Table 8: Result for TLU

The result of the above is query is shown in Table 8. This type of uncertainty is called Tuple Level Uncertainty.

Table 9 shows the final resultant relation in which certain and uncertain data are stored. In this relation, there are six records of certain data and two records represent the Attribute Level Uncertainty and other two records represent Tuple Level Uncertainty. This shows that our presented data model handles both type of uncertainties Attribute Level Uncertainty as well as Tuple Level Uncertainty. Furthermore, because of the Boolean values for the presence and absence of data, the data model become simple to understand, manage and implement.

ID	Name	PE	PM	CE	EL
01	Ali	1	0	0	0
02	Aslam	0	0	1	0
03	Bilal	1	1	0	0
04	Talha	0	0	0	1
05	Yasir	0	1	0	0
06	Irfan	0	0	1	1
07	Asad	1	0	0	0
08	Ahmed	0	0	0	0
09	Usman	0	1	0	0
10	Zain	0	0	0	0

Table 9: Resultant relation

4. Comparative Analysis of Previous and Presented Data Model

This section provides comparative analysis of different data models like world set decomposition, U-relations, and c-table data models with our presented data model for uncertain data. These models are compared from the aspects of easiness, expressiveness, and completeness. World set decomposition (WSD) model provides the idea of world set relation, which is based on vertical partitioning approach (Antova *et al.* 2007). In this technique, a world set relation is decomposed in a manner that the Cartesian product of decomposed relation forms a world set relation again. It is an advanced model in terms of expressiveness. It is also easy and understandable model to implement. World set decomposition model supports only Attribute Level Uncertainty. The main problem in this model is that it does not support Tuple Level Uncertainty. However, to implement efficient data model for uncertain data Tuple Level Uncertainty must also be implemented as is done in our presented data model.

U-relation data model is a precise relational data model for uncertain databases (Antova *et al*. 2008). An important feature of U-relation data model is whether it shows uncertainty at the attribute level or at the tuple level. U-relation data model combines the advantages of World set decomposition and Uncertainty lineage databases. U-relation data model uses the concept of vertical partitioning methodology. U-relation data model is more succinct than both World set decomposition and Uncertainty lineage databases. The relational operations like selection, join and projection are used in same way as used in relational data model. This model is considered as complete and succinct model, the major drawback of the model is that it is not expressive. On the other hand, our presented data model is more expressive as the presence and absence of data is represented with Boolean values.

C-table (Conditional table) data model contains two types of conditions: one is local condition and other is global condition which are used in relations (Imieliñski and Lipski, 1984). The scope of local condition is only applied in local relation and global condition is applied in all relation, both conditions are Boolean formulae. The fundamental significance of the model is that this data model is considered as complete model, it may retrieve the outcomes of a query formalism and this data model has a higher degree of expressiveness. This model is complex to understand and implement. It also deals only Attribute Level Uncertainty.

5. Conclusion

In this paper, a relational data model for uncertain data is presented. The data model has been implemented and tested on a case study of admission process to store, manage and retrieve uncertain data along with certain data. This data

model can deal both types of uncertainties i.e., Attribute Level Uncertainty and Tuple Level Uncertainty for better management of uncertain data. To ensure the completeness of the model, it has been analyzed using queries for relational database. It has been shown that the data model is complete, expressive and allows implementation of scenarios involving uncertain data in an easier way. Another feature of this data model is that it also represents uncertainty in query results. Furthermore, a comparative analysis of previous data models and presented data model is also provided.

Acknowledgement

This work is based on our master's thesis defended in 2017.

References

Aggarwal, C.C., and Yu, P.S. (2009), A Survey of Uncertain Data Algorithms and Applications, IEEE Transactions on Knowledge and Data Engineering, Volume. 21, No. 5, pp. 605-623.

Aggarwal Charu C. (2010), "An Introduction to Uncertain Data Algorithms and Applications", Managing and Mining Uncertain Data, ISBN: 978-0-387-09689-6, Chapter 1, pp 1-8, Springer.

Anish Das Sarma, Omar Benjelloun, Alon Halevy, Shubha Nabar and Jennifer Widom (2009), "Representing Uncertain Data: Models, Properties, and Algorithms", The International Journal Very Large Database, Springer, pp 989-1019.

Antova, L., Jansen, T., Koch, C., and Olteanu, D. (2008), "Fast and Simple Relational Processing of Uncertain Data", 24th IEEE International Conference on Data Engineering, pp. 983-992, Mexico.

Antova Lyublena, Koch Christoph, and Olteanu Dan (2007), "World-set Decompositions: Expressiveness and Efficient Algorithms", 11th International Conference, Barcelona, pp. 194-208, Springer, Spain.

Deshpande Amol, Getoor Lise and Sen Prithviraj (2009), "Graphical Models for Uncertain Data", Managing and Mining Uncertain Data, ISBN: 978-0-387-09689-6, Chapter 4, pp 77-105, Springer.

Hayat Umar and Khan Muhammad Usman Ghani (2015), "An improved data model for uncertain data" Mehran University Research Journal of Engineering & Technology, Volume 35, Issue 1, pp. 83-94.

Imieliñski, T., and Lipski, Jr, W. (1984) "Incomplete Information in Relational Databases", Journal of the ACM, Volume 31, No. 4, pp. 761-791, New York, USA.

Sarkar Sumit and Dey Debabrata (2009), "Relational Models and Algebra for Uncertain Data", Managing and Mining Uncertain Data, ISBN: 978-0-387-09689-6, Chapter 11, pp 45-76.

Sarma, A. Das, Benjelloun Omar, Halevy A., and Widom, J. (2006) "Working Models for Uncertain Data", Proceedings of 22nd IEEE International Conference on Data Engineering, USA.

Wang Yijie, Li Xiaoyong, Li Xiaoling and Wang Yuan (2013), "A Survey of Queries Over Uncertain Data" Knowledge and information systems, Springer, Volume 37, Issue 3, pp. 485–530.

Zhang, W., Yue, K., and Liu, W. (2011), "Learning Uncertain Knowledge from Uncertain Data", Journal of Information and Computational Science, Volume 8, No. 6, pp. 933-940, Hong Kong.