University of Tunis
Institut supérieur de gestion

LARODEC research laboratory

Master Thesis
STID-IGC

Expression of Queries on Evidential Databases

Elaborated by :
Fatma Ezzahra Bousnina

Supervised by :

Mme. Boutheina Ben yaghlane
M. Ludovic Liétard

Academic Year : 2014-2015

Contents

Acknowledgements **1**

Abstract **1**

Introduction **1**

1 Evidence Theory **3**
1.1 Introduction 3
1.2 Basic concepts 3
1.3 Functions of evidence theory 5
1.3.1 Belief function 5
1.3.2 Plausiblity function 6
1.3.3 Commonality function 7
1.4 Special belief functions 8
1.4.1 Vacuous belief function 8
1.4.2 Certain belief function 8
1.4.3 Bayesian belief function 9
1.4.4 Consonant belief function 9
1.5 Combination of sources 10
1.5.1 Dempster's rule of combination 10
1.5.2 Conjunctive rule of combination 11
1.5.3 Disjunctive rule of combination 12
1.6 Vacuous extension 13
1.7 Conclusion 14

2 Imperfect Databases **15**
2.1 Introduction 15
2.2 Probabilistic Databases 15

2.3 Possibilistic Databases 19
2.4 Evidential Databases 22
2.4.1 Defintion of evidential databases 22
2.4.2 Models of the evidential compact databases 23
2.5 Conclusion 27

3 Evidential database as Possible Worlds 28
3.1 Introduction 28
3.2 Definition of possible worlds 28
3.3 Querying possible worlds 32
3.4 Exprimental Study 37
3.4.1 Data Description 37
3.4.2 Experimental results 38
3.5 Conclusion 44

Conclusion 45

List of Figures

1.1 Combination of M independent sources [11] 10

3.1 Performance evaluation under focal elements' number variation 41
3.2 Performance evaluation under database size variation 42
3.3 Performance evaluation under focal elements' size variation . . 43

List of Tables

1.1 Example of application of the conjunctive rule of combination 11
1.2 Example of application of the disjunctive rule of combination . 12
1.3 An overview of the use of combination rules 13

2.1 An Example of a probabilistic database PDB 16
2.2 Probabilistic possible worlds 16
2.3 Compact Form: Probabilistic Database 17
2.4 Result of Q over the compact form of a probabilistic database 17
2.5 Result of Q over PW_1 18
2.6 Result of Q over PW_2 18
2.7 Result of Q over the possible worlds 18
2.8 An example of a possibilistic database 19
2.9 Possible worlds of the possibilistic database 20
2.10 Compact form: Possibilistic Database 20
2.11 Result of Q over the compact form of a possibilistic database . 21
2.12 Result of Q over PW_2 21
2.13 Result of Q over PW_4 21
2.14 A medical evidential table 22
2.15 Evaluation of the CL for the condition (Disease = {Anemia}) 24
2.16 T_{EDB} where (Disease = {Anemia}) 24
2.17 Result of π(ID,Blood Type) over EDB 25
2.18 Evidential table T_1 26
2.19 Evidential table T_2 26
2.20 $T_1 \times T_2$ 27
2.21 $T_1 \bowtie T_2$ 27

3.1 Example of an imprecise possible world of table 2.14 29
3.2 Imperfect worlds of EDB 29
3.3 Extension of bbas from the evidential table 2.14 30

3.4 Combination of the extended *bbas* using the conjunctive rule of combination 30
3.5 Combination of the evidential objects of EDB using the disjunctive rule of combination 31
3.6 Possible worlds of EDB 34
3.7 Comparative table for the query Q_1 39
3.8 Comparative table for the query Q_2 40
3.9 Comparative table for the query Q_3 40

Acknowledgements

I would like to express my sincere gratitude to my co-advisor, Professor Boutheina Ben Yaghlane, for introducing me to this area of research and for her valuable advices.

I would like to thank my co-advisor Doctor Ludovic Liétard from Institut universitaire des technologies de Lannion *for his patience and his valuable remarks during my intership.*

I'm so grateful to Doctor Mouna Chabbeh and Doctor Mohamed Anis Bach Tobji for their continuous guidance. I thank them for their encouragements. Their comments have been so helpful and constructive.

I would like to adress my thanks to all professors of Institut Suprieur de Gestion de Tunis. *I would also thank all members of LARODEC laboratory and IRISA laboratory.*

I owe a debt of gratitude to my family and my friends for their moral support.

Finally, i would like to acknowledge this master thesis to every person who encouraged me.

Abstract

The theory of evidence encompasses lots of powerful tools for the treatement of imperfection. Evidential databases copes with imperfect data which are modeled via the evidence theory. Existing works on such data deal only with the compact form of the database. In this master thesis, we propose a new formalism for modeling and querying evidential databases based on the possible worlds form. This work is a first step toward the definition of a strong representation system.

Keywords: Evidential databases, Dempster-Shafer theory, Strong representation system, Possible worlds.

Résumé

La théorie d'évidence englobe beaucoup d'outils puissants pour le traitement de l'imperfection. Les bases de données évidentielles stokent les données imparfaites qui sont modélisées par la théorie de l'évidence. Les travaux existants sur ce genre de données ne portent que sous la forme compacte des bases de données. Dans ce mémoire, nous proposons un nouveau formalisme pour la modélisation et l'interrogation des bases de données évidentielles sous forme des mondes possibles. Ce travail est une première étape vers la définition d'un système de représentation fort.

Mots clés: Bases de données évidentielles, Théorie de Dempster-Shafer, Système de représentation fort, Mondes possibles.

Introduction

Imperfection is present in many real life applications like medecine, industry, weather forcasting [22, 19]. That information need to be stored and treated. That's why many theories have been used to represent imperfect information. The most famous is the probability theory [9]. Other theories have been introduced to offer other models like possibility theory [30], fuzzy sets theory [29], and evidence theory [26].

The theory of evidence models and combines imperfect information through an explicit representation of uncertainty, imprecision, inconsistency and ignorance. This theory is a generalization of the bayesian inference.

In order to reach the aim of storing and managing imperfect information, specified database models, based on these theories were introduced: the probabilistic database model [9], the possibilistic database model [7] and the evidential database model [2].

In this master thesis, we consider the general framework of evidential databases where attributes' values are defined using evidence theory. This model was introduced in [2, 21]. It provides a clear and simple representation for imperfect data.

We define a new representation model of evidential databases based on *possible worlds*, contrary to existing evidential database formalisms which are based on the *compact form*, such in [2, 10, 21, 20]. Indeed, we define at first the notion of *possible worlds* for an evidential database. Then, we present the steps to generate possible worlds from evidential database and finally, we redefine some relational operators and their application over these possible worlds.

Such a new representation opens the way to the definition of a *strong representation system* for evidential databases. A database model is a strong representation system if the result of any query processed on the *compact*

form of the database is equivalent to the result of the same query over the set of its *possible worlds*. The possible worlds representation is the only way to valid the representation of the compact form.

The outline of the master dissertation is as follows:

- Chapter 1: We review some basic concepts of the theory of evidence like the belief and the plausibility functions, the vacuous extension and the combination rules.
- Chapter 2: We present three models of imperfect databases which are probabilistic, possibilistic and evidential databases. We also present the extended relational operators for the probabilistic and possibilistic models.
- Chapter 3: We propose a new formalism for modelling and querying evidential databases as possible worlds.
- Chapter 4: We present an experimental study of our approach. We demonstrate to what limit it is possible to operate with the possible worlds representation.

Chapter 1

Evidence Theory

1.1 Introduction

The theory of evidence also called *the theory of belief functions or the Dempster-Shafer theory*, was introduced by Dempster [11] in the context of statistical inference, mathematically formalized by Shafer [26] and later popularized by Smets [28] with the transferable belief model (TBM).

The theory of belief functions models and combines imperfect information. It provides an explicit representation of uncertainty, imprecision, inconsistency and ignorance. It is considered as a generalization of the bayesian inference for uncertainty treatment, i.e. It models what it is already known and what is unknown. Theory of belief functions is used in lots of applications as data fusion, clustering, classification, conflict, targets identification...

1.2 Basic concepts

In this section, basic notions of evidence theory as *the frame of discernment, the basic belief assignment, the body of evidence, the rules of combination* are presented.

Frame of discernment

In the theory of belief functions, a set $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ is a finite, non empty and exhaustive set of n elementary and mutually exclusive hypotheses related to a given problem. The set Θ is called the *frame of discernment*

or *universe of discourse.*

The *power set* 2^{Θ} is the set of all subsets of Θ defined as follows:

$$2^{\theta} = \{A : A \subseteq \Theta\} = \{\varnothing, \theta_1, \theta_2, ..., \theta_n, \{\theta_1, \theta_2\}, .., \{\theta_1, \theta_2, ..., \theta_n\}\} \quad (1.1)$$

Each element of 2^{Θ} is a *proposition* or an *event.*

Example 1. *Suppose the following frame of discernment for the attribute Disease:* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$.

Its corresponding power set is $2^{\Theta_{DE}} = \{\varnothing, \{Flu\}, \{Asthma\}, \{Anemia\}, \{Flu, Asthma\}, \{Flu, Anemia\}, \{Asthma, Anemia\}, \{Flu, Asthma, Anemia\}\}$.

Basic belief assignement

A *basic belief assignement* (*bba*), noted m, is a mapping from 2^{Θ} to the interval $[0, 1]$ that assesses a degree of belief to some elements of the power set. A bba also called *mass function* is defined such that:

$$\sum_{x \subseteq \Theta} m(x) = 1 \quad (1.2)$$

The amount $m(x)$ is called *basic belief mass* (bbm) and *mass*for short. It represents our degree of faith on the truth of the hypothesis x. The mass $m(x)$ is the degree of belief on x that is not distributed on its subsets.

Example 2. *Suppose that a doctor (an expert) assigned the following masses on a given patient's disease P:*

$m^{\Theta_{DE}}(\{Flu\}) = 0.7$
$m^{\Theta_{DE}}(\{Asthma, Anemia\}) = 0.3$

The doctor believes are: The patient P suffers from Flu with a degree of belief 0.7 *and he suffers from Asthma or Anemia with a degree of belief* 0.3.

Body of evidence

The subsets x of the frame of discernment Θ where $m^{\Theta}(x)$ is strictly positive, are named *focal elements* such that:

$$m^{\Theta}(x) > 0 \tag{1.3}$$

We denote F the set of all focal elements and the couple $\{F, m^{\Theta}\}$ the *body of evidence.*

The union of all focal elements is called *core* and is defined as follows:

$$\varphi = \bigcup_{x \in F : m^{\Theta}(x) > 0} x \tag{1.4}$$

Example 3. *Assume now that subsets* $\{Flu\}$ *and* $\{Flu, Anemia\}$ *are focal elements of* $m^{\Theta_{DE}}$ *such that:*

$F = \{\{Flu\}, \{Flu, Anemia\}\}$

The couple $(F, m^{\Theta_{DE}})$ *is the body of evidence such that:*

$\varphi_m^{\Theta_{DE}} = \{Flu\} \cup \{Asthma, Anemia\} = \{Flu, Asthma, Anemia\} = \Theta$

1.3 Functions of evidence theory

In this section, some functions of theory of evidence like the belief, the plausibility and the commonality are presented.

1.3.1 Belief function

The *belief function* noted $bel(x)$, is the degree of faith exactly committed to x [26]. It is the *minimal* degree of belief given to x and it is defined as follows:

$$bel : 2^{\Theta} \longrightarrow [0, 1]$$

$$bel(x) = \sum_{y, x \subseteq \Theta : y \subseteq x} m^{\Theta}(y) \tag{1.5}$$

Properties

- $bel(\varnothing) = 0 \quad \text{and} \quad bel(\Theta) = 1$
- $bel(x) + bel(\overline{x}) \leq 1$ (Sub additivity)
- $x \subseteq y \Rightarrow bel(x) \leq bel(y)$ (Monotonicity)
- $bel(x \cup y) \geq bel(x) + bel(y)$

Example 4. *Suppose the following mass functions:*

$m^{\Theta_{DE}}(\{Flu\}) = 0.7$
$m^{\Theta_{DE}}(\{Flu, Anemia\}) = 0.3$

The belief functions, bel, are calculated as follows:

$bel(\varnothing) = 0$
$bel(\{Flu\}) = 0.7$
$bel(\{Anemia\}) = 0$
$bel(\{Flu, Anemia\}) = 0.3$
$bel(\Theta_{DE}) = bel(\{Flu, Anemia, Asthma\}) = 0$

1.3.2 Plausiblity function

The *plausibility function* noted pl is the *maximal* amount of belief on x. It is the sum of masses relative to subsets y and not contradictory to subsets x. The plausibility function is defined as follows:

$$pl : 2^{\Theta} \longrightarrow [0, 1]$$

$$pl(x) = \sum_{y, x \subseteq \Theta : x \cap y \neq \varnothing} m^{\Theta}(y) \tag{1.6}$$

Properties

- $pl(x) = 1 - bel(\overline{x})$
- $Pl(\Theta) = 1 \quad \text{and } Pl(\varnothing) = 0$
- $Pl(x \cup y) \leq Pl(x) + Pl(y)$

- $bel(x) \leq Pl(x)$
- $x \subseteq y \Rightarrow Pl(x) \leq Pl(y)$ (Monotonicity)
- $Pl(x) + Pl(x) \geq 1$ (Sub additivity)

Example 5. *We take the same mass functions of the previous example:*

$m^{\Theta_{DE}}(\{Flu\}) = 0.7$
$m^{\Theta_{DE}}(\{Flu, Anemia\}) = 0.3$

The plausibility function pl is calculated as follows:

$pl(\varnothing) = 0$
$pl(\{Flu\}) = 1$
$pl(\{Anemia\}) = 0.3$
$pl(\{Flu, Anemia\}) = 1$
$pl(\Theta_{DE}) = 1$

1.3.3 Commonality function

The commonality function noted q of an element x represents the sum of masses allocated to supersets of y which have x in common. The commonality function represents a sort of transformation of belief function that can be useful to simplify computations.

This function is calculated as follows :

$$pl : 2^{\Theta} \longrightarrow [0, 1]$$

$$q(x) = \sum_{x \subseteq y} m^{\Theta}(y) \tag{1.7}$$

Example 6. *Suppose we have the following masses:*

$m^{\Theta_{DE}}(\varnothing) = 0.1$
$m^{\Theta_{DE}}(\{Flu\}) = 0.2$
$m^{\Theta_{DE}}(\{Asthma\}) = 0.2$
$m^{\Theta_{DE}}(\{Anemia\}) = 0.3$
$m^{\Theta_{DE}}(\{Asthma, Anemia\}) = 0.1$

$m^{\Theta_{DE}}(\Theta) = 0.1$

The commonality function q is calculated as follows:

$q(\varnothing) = 1$
$q(\{Flu\}) = 0.2 + 0.1 = 0.3$
$q(\{Asthma\}) = 0.2 + 0.1 + 0.1 = 0.4$
$q(\{Anemia\}) = 0.3 + 0.1 + 0.1 = 0.5$
$q(\{Asthma, Anemia\}) = 0.1 + 0.2 + 0.3 + 0.1 = 0.7$
$q(\Theta) = 0.1$

1.4 Special belief functions

Other types of belief functions were proposed in the literature as the vacuous belief function, the certain belief function, the consonant belief function and the bayesian belief function, that we will present in this section.

1.4.1 Vacuous belief function

A *vacuous belief function* is a normalized belief function [26] where Θ is the unique focal element. In this case, the basic belief assignement bba quantifies the *total ignorance*, such that:

$$m(\Theta) = 1 \quad \text{and} \quad m(x) = 0 \quad \text{where} \quad x \neq \Theta \tag{1.8}$$

1.4.2 Certain belief function

A certain belief function is a *bba* with only one focal element which is a singleton. That *bba* represents the total certainty such that:

$$m^{\Theta}(x) = 1 \quad \text{and} \quad m^{\Theta}(y) = 0 \quad \text{for all} \quad y \neq x \quad \text{and} \quad y \subseteq \Theta \tag{1.9}$$

Where x is a singleton from Θ

Example 7. *Let* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$ *be the frame of discernment of attribute* Disease.

$m^{\Theta_{DE}}(\{Anemia\}) = 1$ *is a certain basic belief assignment.*

1.4.3 Bayesian belief function

A bayesian belief function is a particular case of belief function. When all focal elements are singletons the *bel* function is named bayesian and it is a *probabilistic distribution* [26].

Properties

- $bel(\varnothing) = 0 \quad \text{and} \quad bel(\Theta) = 1$
- $bel(x \cup y) = bel(x) + bel(y)\ ;\ x, y \subset \Theta \quad \text{and} \quad x \cap y = \varnothing$
- $bel(x) + bel(\overline{x}) = 1\ ;\ A \subset \Theta$
- $bel = pl$

Example 8. *Assume that* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$ *is the frame of discernment of attribute Disease. We consider the following masses:*

$m^{\Theta_{DE}}(\{Asthma\}) = 0.2$
$m^{\Theta_{DE}}(\{Anemia\}) = 0.3$
$m^{\Theta_{DE}}(\{Flu\}) = 0.5$
$m^{\Theta_{DE}}(\Theta) = 0$

All focal elements are singletons, so the bba is bayesian in this case.

1.4.4 Consonant belief function

A consonant belief function is a function with all its focal elements are nested.

Properties

- $bel = (x \cap y) = min(bel(x), bel(y))$ (Necessity function)
- $pl(x \cup y) = max(bel(x), bel(y))$ (Possibility function)

Example 9. *We consider the following masses:*

$m^{\Theta_{DE}}(\{Asthma\} = 0.3$
$m^{\Theta_{DE}}(\{Asthma, Anemia\} = 0.5$
$m^{\Theta_{DE}}(\Theta) = 0.2$

In this case, all focal elements are nested so this bba is consonant.

1.5 Combination of sources

Combination rules appear as an interesting solution for having a more reliable information, specially in the presence of imperfect information (uncertain, imprecise, incomplete). The major interest of combining several sources is to get at the end one mass function that represents to the best all combined believes. Many combination rules are proposed in the framework of the theory of belief functions. In this section, we will present the most commonly used rules in literature.

1.5.1 Dempster's rule of combination

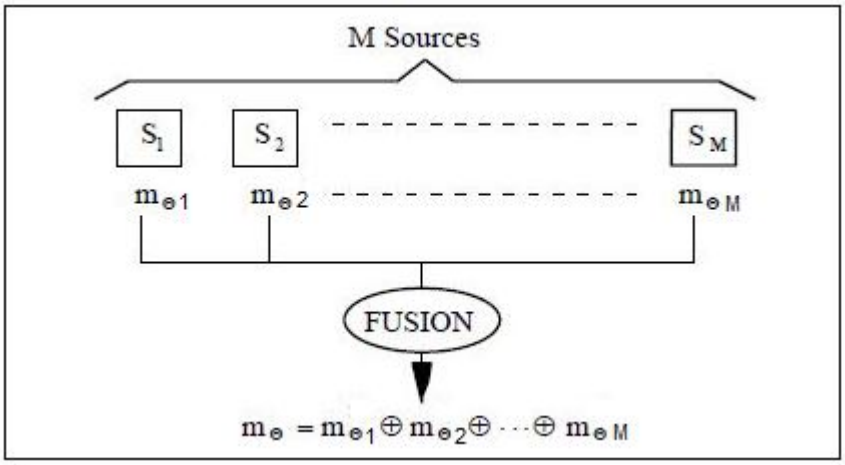
Suppose we have M *independent* and *distinct* sources of information to combine. Each source $S_j (j = 1..M)$ expresses its belief over the given problem and gives the corresponding mass m^{Θ} [11]. The combination (called the *joint mass*) is calculated as follows:

$$m^{\Theta} = m^{\Theta_1} \oplus .. \oplus m^{\Theta_M} \tag{1.10}$$

$\oplus$ is the operator of combination.

The figure 1.1 shows in a more clear way how to combine M independent sources using the $\oplus$ operator.

Figure 1.1: Combination of M independent sources [11]



1.5.2 Conjunctive rule of combination

The *conjunctive rule of combination* for two mass functions m_1 and m_2 defined on the same frame of discernment Θ was introduced by Smets [27] and it is defined as follows:

$$m_1^{\Theta} \textcircled{\cap} m_2^{\Theta}(z) = \sum_{x,y \subseteq \theta : x \cap y = z} m_1^{\Theta}(x).m_2^{\Theta}(y) \tag{1.11}$$

The conjunctive rule of combination merges *bbas* m_1 and m_2 provided by different sources, the result is the joint *bba* induced from the combination of m_1 and m_2.

Example 10. *Let us consider the following frame of discernment* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$. *We have two sources* S_1 *and* S_2 *giving respectively the following masses* $m_1^{\Theta_{DE}}$ *and* $m_2^{\Theta_{DE}}$ *for Patient* P*:*

$m_1^{\Theta_{DE}}(\{Flu, Anemia\}) = 0.7$

$m_1^{\Theta_{DE}}(\{Asthma, Anemia\}) = 0.3$

$m_2^{\Theta_{DE}}(\{Flu, Asthma\}) = 0.5$

$m_2^{\Theta_{DE}}(\{Asthma, Anemia\}) = 0.5$

Now, we apply the conjunctive rule of combination. Results are shown in table 1.1.

Table 1.1: Example of application of the conjunctive rule of combination

$\textcircled{\cap}$	$m_1^{\Theta_{DE}}(\{Flu, Anemia\})$ 0.7	$m_1^{\Theta_{DE}}(\{Asthma, Anemia\})$ 0.3
$m_2^{\Theta_{DE}}(\{Flu, Asthma\})$ 0.5	$\{Flu\}$ 0.35	$\{Asthma\}$ 0.15
$m_2^{\Theta_{DE}}(\{Asthma, Anemia\})$ 0.5	$\{Anemia\}$ 0.35	$\{Asthma, Anemia\}$ 0.15

$m_1^{\Theta_{DE}} \textcircled{\cap} m_2^{\Theta_{DE}}(\{Flu\}) = 0.5 * 0.7 = 0.35$

$m_1^{\Theta_{DE}} \textcircled{\cap} m_2^{\Theta_{DE}}(\{Asthma\}) = 0.5 * 0.3 = 0.15$

$m_1^{\Theta_{DE}} \textcircled{\cap} m_2^{\Theta_{DE}}(\{Anemia\}) = 0.5 * 0.7 = 0.35$

$m_1^{\Theta_{DE}} \textcircled{\cap} m_2^{\Theta_{DE}}(\{Asthma, Anemia\}) = 0.5 * 0.3 = 0.15$

1.5.3 Disjunctive rule of combination

The disjunctive rule of combination was also proposed by Smets [27]. It is based on the union of focal elements. Combining bbas m_1 and m_2 with the disjunctive rule leads to a combined *bba* which focal elements are the union of focal elements of m_1 and m_2. Combining two bbas m_1 and m_2 defined on the same frame of discernment Θ is defined as follows:

$$m_1^{\Theta} \textcircled{\cup} m_2^{\Theta}(z) = \sum_{x,y \subseteq \theta : x \cup y = z} m_1^{\Theta}(x).m_2^{\Theta}(y). \tag{1.12}$$

Example 11. *We consider the same mass function as in the previous example. Suppose now that one of the sources is reliable but we don't know which one. We apply the disjunctive rule of combination as shown in table 1.2.*

Table 1.2: Example of application of the disjunctive rule of combination

$\textcircled{\cup}$	$m_1^{\Theta_{DE}}(\{Flu, Anemia\})$ 0.7	$m_1^{\Theta_{DE}}(\{Asthma, Anemia\})$ 0.3
$m_2^{\Theta_{DE}}(\{Flu, Asthma\})$ 0.5	$\{Flu, Asthma, Anemia\}$ 0.35	$\{Flu, Asthma, Anemia\}$ 0.15
$m_2^{\Theta_{DE}}(\{Asthma, Anemia\})$ 0.5	$\{Flu, Asthma, Anemia\}$ 0.35	$\{Asthma, Anemia\}$ 0.15

$m_1^{\Theta_{DE}} \textcircled{\cup} m_2^{\Theta_{DE}}(\{Flu, Asthma, Anemia\})$ = *(0.5 * 0.7) + ((0.5 * 0.3) + (0.5 * 0.7) = 0.85*

$m_1^{\Theta_{DE}} \textcircled{\cup} m_2^{\Theta_{DE}}(\{Asthma, Anemia\})$= *0.5 * 0.3 = 0.15*

The table 1.3 represents an overview of the use of combination rules.

Table 1.3: An overview of the use of combination rules

$\textcircled{\cap}$	$\oplus$	$\textcircled{\cup}$
Independent and reliable pieces of evidence Unormalized	Independent and reliable pieces of evidence Normalized	Independent and at least one of evidence pieces is reliable Unormalized

1.6 Vacuous extension

In some cases, we need to combine two *bbas* m_1 and m_2 which are not defined on the same frame of discernment. However, all combination rules require that *bba*s have the same frame of discernment. The *vacuous extension of belief functions* [26] is a tool that defines bbas on a compatible frame of discernment. It consists in extending the frames of discernment Θ_1 and Θ_2, corresponding to the mass functions m_1 and m_2, to the joint frame of discernment Θ defined as:

$$\Theta = \Theta_1 \times \Theta_2$$

Each focal element is extended to its cylindrical extension ($A \times \Theta_2$ is the cylindrical extension of $A \subseteq \Theta_1$).

The extended mass function of any evidential value of the extended focal element A, denoted by m, is defined as follows:

$$m^{\Theta_1 \uparrow \Theta_1 \times \Theta_2}(A) = \begin{cases} m(B) & \text{where } A = B \times \Theta_2,\ B \subseteq \Theta_1 \\ 0 & \text{otherwise} \end{cases} \tag{1.13}$$

1.7 Conclusion

In this chapter, we have presented the basic concepts of evidence theory as the frame of discernment, the belief function, the plausibility function, the combination rules and the vacuous extension that we will use in the rest of our master dissertation. This theory offers efficient tools to handle imperfect data.
In the next chapter, we will present the imperfect databases and how to process some queries over them but we will focus on evidential databases.

Chapter 2

Imperfect Databases

2.1 Introduction

Imperfect databases are sets of imperfect data, they include uncertainty, imprecision, inconsistency and ignorance. There are several types of imperfect databases such as probabilistic databases [9], possiblistic databases [7], and evidential databases [2, 10, 20, 21].

We store data in the aim of querying them later, that's why imperfect data should be managed with convenient tools to get a reliable result. When a result of a query processed over an original database is the same when that query is applied over *possible worlds*, we say that the result is reliable and the model is called *strong representation system*.

The notion of possible worlds and the notion of the strong representation system will be introduced later in this chapter. We will also present the model of probabilistic databases, the model of possibilistic databases and we will focus on the model of evidential databases. We will explain through illustrative examples the application of queries over those models.

2.2 Probabilistic Databases

Probabilistic databases were introduced by Roger cavallo and Micheal Pittarelli [9], then popularized by Barbara [15]. Later, Fuhr and rolleke defined a more elaborated model of probabilistic databases [13, 14, 16, 17, 24].

A probabilistic database is an uncertain database that includes N objects

and D attributes in which the imperfect values are expressed via theory of probabilities.

We can also define a probabilistic database as a set of *possible worlds* that associated probabilities. *Probabilistic Possible worlds* are the different states generated from an imperfect database. Each probabilistic possible world is a candidate to represent the original database called *a compact form.*

Example 12. *Assume we have the following probabilistic database of medical diagnoses and we want to generate the probabilistic possible worlds.*

Table 2.1: An Example of a probabilistic database PDB

ID	Disease	Blood Type	P
1	Flu	A	0.3
2	Anemia	B	0.2
3	Asthma	{ B, O} {0.25 , 0.25}	0.5

This relation includes 3 objects, 3 attributes which are ID, $Disease$, $BloodType$ and P contains the probability of each tuple.

Table 2.2: Probabilistic possible worlds

PW_1	P
1, Flu , A	0.3
2 , Anemia , B	0.2
3 , Asthma , B	0.25

PW_2	P
1, Flu , A	0.3
2 , Anemia , B	0.2
3 , Asthma , O	0.25

The sum of probabilities of all possible worlds is 1:

- $P(PW_1)$ = *0.5*
- $P(PW_2)$ = *0.5*

The result of processing a query Q over a probabilistic database is a relation that satisfies the condition of Q and includes probabilities P.

The result of processing a query Q over possible worlds of a probabilistic database is the union of relations given by each possible world associated with their probabilities P.

Typically, a probabilistic database ranks answers of a query by their probabilities.

Example 13. *We carry on with the same example: Let query Q be the query processed over table 2.1.*

---

*Q: SELECT * WHERE (Disease = Asthma) FROM PDB;*

Table 2.3: Compact Form: Probabilistic Database

ID	Disease	Blood Type	P
1	Flu	A	0.3
2	Anemia	B	0.2
3	Asthma	{ B , O} {0.25 , 0.25}	0.5

Table 2.4: Result of Q over the compact form of a probabilistic database

ID	Disease	Blood Type	P
3	Asthma	{ B , O} {0.25 , 0.25}	0.5

Now we process the same query Q over the possible worlds of the probabilistic database.

Table 2.5: Result of Q over PW_1

PW_1	P
1, Flu , A	0.3
2 , Anemia , B	0.2
3 , Asthma , B	0.25

PW_1	P
3 , Asthma , B	0.25

Table 2.6: Result of Q over PW_2

PW_2	P
1, Flu , A	0.3
2 , Anemia , B	0.2
3 , Asthma , O	0.25

PW_1	P
3 , Asthma , O	0.25

The result of Q over the probabilistic possible worlds is the union of results as shown in the following table.

Table 2.7: Result of Q over the possible worlds

ID	Disease	Blood Type	P
3	Asthma	B	0.25
3	Asthma	O	0.25
			= 0.5

Since the third tuple appears in both results, the final probability for ID_3 is computed as the sum of its probabilities from PW_1 and PW_2.

A model is a **Strong Representation System** if the following relation is valid; The result of querying the *compact model* C is equivalent to the result of querying the set of *possible worlds*$\{PW_1...PW_j\}$ generated from C [5, 6], such that:

$$Q(C) = \{Q(PW_1)....Q(PW_j); PW_{1..j} \in C\} \tag{2.1}$$

The Compact form is the only feasible in practice because the possible worlds representation may be too costly in terms of memory usage. The representation of possible worlds is fundamental to validate if a model is a *strong representation system* or not.

$\Longrightarrow$ The equivalence between the compact form and possible worlds is valid for probabilistic databases.

2.3 Possibilistic Databases

Possibilistic databases were introduced by Prade [23] then were popularized by Bosc et al,. [7, 3].

A possibilistic database is an uncertain database that includes N objects, D attributes in which imperfect values are expressed via the theory of possibilities.

A possibilistic database is also defined as a set of possible worlds that associate degrees of possibilities. Each possibilistic world is a candidate to represent the compact database.

Example 14. *Suppose the possibilistic database of medical diagnoses. We want to generate the possibilistic worlds. The following relation 2.8 represents an example of a possibilistic database. It includes 4 objects and 2 attributes which are Departement and Diseae that associated degrees of possibilities.*

Table 2.8: An example of a possibilistic database

Department	Disease
D1	1/Flu + 0.7/Asthma
D2	1/Cancer + 0.3/Anemia
D3	Asthma
D4	Anemia

Table 2.9: Possible worlds of the possibilistic database

PW_1	PW_2
$(D_1, 1/Flu\ )$ $(D_2, Cancer)$ $(D_3, Asthma)$ $(D_4, Anemia)$	$(D_1, 1/Flu)$ $(D_2, 0.3/Anemia)$ $(D_3, Asthma)$ $(D_4, Anemia)$
PW_3	PW_4
$(D_1, 0.7/Asthma)$ $(D_2, Cancer)$ $(D_3, Asthma)$ $(D_4, Anemia)$	$(D_1, 0.7/Asthma)$ $(D_2, 0.3/Anemia)$ $(D_3, Asthma)$ $(D_4, Anemia)$

The result of processing a query Q over a possibilistic database is a degree of possibility that satisfies Q and including possibility degrees.

Example 15. *We carry on with the same example: Let Q be the query processed over table 2.8.*

Q: To what extent is the tuple $(D_2, Anemia)$ a possible answer ?

Table 2.10: Compact form: Possibilistic Database

Departement	Disease
D1	1/Flu + 0.7/Asthma
D2	1/Cancer + 0.3/Anemia
D3	Asthma
D4	Anemia

Table 2.11: Result of Q over the compact form of a possibilistic database

Departement	Disease
D2	1/Cancer + 0.3/Anemia

The tuple $(D_2, Anemia)$ *is a possible answer with a degree of possibility of* 0.3.

The result of processing a query Q over possible worlds of a possibilistic database is a degree of possibility that respects the set of possibilistic possible worlds [7], such that:

$$Degree(t) = Max(Min(\Pi(PW_{j_1}))..Min(\Pi(PW_{j_n})) \tag{2.2}$$

where $PW_{(j_1..j_n)}$is the number of possible worlds.

Example 16. *Now, we process the same query Q over possible worlds of the possibilistic database.*

Table 2.12: Result of Q over PW_2

PW_2
$D_2, 0.3/Anemia$

Table 2.13: Result of Q over PW_4

PW_4
$D_2, 0.3/Anemia$

The tuple $t = (D_2, Anemia)$ *appears in* PW_2 *and* PW_4 *as shown in tables 2.12 and 2.13.*

$Degree(t) = Max(Min(\Pi(PW_2)), Min(\Pi(PW_4))) = Max(0.3, 0.3) = 0.3$

The degree of possibility that the tuple is a possible answer is $t = (D_2, Anemia) =$ 0.3.

2.4 Evidential Databases

2.4.1 Defintion of evidential databases

Evidential databases (*EDB*), also called *D-S databases* are used to store both perfect and imperfect data unlike classic databases that are used to store only certain and precise data. Imperfect values in evidential databases are expressed via evidence theory presented above.

An *EDB* has N objects and D attributes. The value of an attribute c for an object l is a *bba*. It is called *evidential value* and is denoted by V_{lc} .

$$V_{lc} : 2^{\Theta_l} \rightarrow [0, 1]$$

$$\text{with } m_{lc}(\varnothing) = 0 \text{ and } \sum_{x \subseteq \Theta_l} m_l(x) = 1 \tag{2.3}$$

The set of focal elements relative to the *bba* V_{lc} is F_{lc} such that:

$$F_{lc} = \{x \subseteq \Theta / m_{lc}(x) > 0\} \tag{2.4}$$

Example 17. *Let us take an example of evidential database that stores medical diagnoses as shown in table 2.14 in below:*

Table 2.14: A medical evidential table

ID	Disease	Blood Type
1	Flu 0.7 {Asthma, Anemia} 0.3	A
2	Anemia	B 0.5 {B,O} 0.5

This relation contains two attributes Disease and BloodType. Being an evidential database, each value of each attribute may be imperfect. It is expressed through an evidence theory distribution (basic belief assignment).

Data are stored to be further queried. Some models have been proposed for this purpose. Bell et al., [2] and Choenni et al.,[10] and Lee in [21, 20] defined models of querying an evidential database using relational operators like selection, projection and join. Their models are based on a compact form of the database.

2.4.2 Models of the evidential compact databases

In this subsection we will present the compact database model as proposed in the literature.
An evidential database EDB, as defined in 2.4.1 is on its *compact form*. Table 2.14 represents the compact form of EDB for a medical diagnoses example.

Extended Relational database models

The classical relational database model handles only perfect values. That's why the *extended relational database models* were proposed to modify that feature in order to manipulate possible values for each attribute of *EDB* [20, 21, 2, 10]. It models uncertainty and imprecision in data and it is based on the theory of evidence. These models redefine the relational operators (Selection, Cartesian Product, Join, Projection, Intersection, Union) [2, 21, 20, 10].

Defintion 1. *A table T is based on Θ_1, Θ_2,..., Θ_n, and defined as $T \subseteq G_1 \times G_2 \times ... \times G_n \times CL$ where G_i is the set of a focal element and its mass function m of the domain Θ_i. The confidence level attributed for each tuple CL is represented by $CL = [b, p]$ with b, $p \in [0, 1]$; $b \leq p$ and b, p are bel and pl functions, respectively [2, 21, 20].*

Redifinition of the extended relational operators

- Selection

  The *selection* δ consists on extracting from a table T the tuples with attributes that satisfy the query and returns a new relation. In the extended database model presented in [20, 2], the select operation operates as same as the classical one but only tuples with very low confidence level values CL will be filtered out. This model requires that not every possible tuple should be kept in the result.

  A new CL is computed such that:

$$CL = [b \times bel; p \times pl] \tag{2.5}$$

with *bel* and *pl* are the values of belief and plausibility of attribute d_i.

Confidence levels CL that does not satisfy a given threshold are filtered out after.

Example 18. *To better understand how the model proposed in [2, 10] operates, we consider the example of table 2.14.*
We will apply a Selection operation over table EDB. We want to know who are the persons in our table that have Anemia as a disease.
The selection threshold is $pl \geqslant 0.5$.

---

*Q: SELECT * FROM EDB WHERE Disease = {Anemia}*

Table 2.15: Evaluation of the CL for the condition (Disease = {Anemia})

ID	Disease	Blood Type	CL
1	Flu 0.7 (Asthma, Anemia) 0.3	A	[0,0.3]
2	Anemia	B 0.5 (B,O) 0.5	[1,1]

The first tuple does not satisfy the condition $pl \geqslant 0.5$*, so it will be eliminated. The result of processing query Q over table 2.14 is the following:*

Table 2.16: T_{EDB} where (Disease = {Anemia})

ID	Disease	Blood Type
2	Anemia	B 0.5 (B,O) 0.5

- Projection

  The *projection* π consists on taking as input a relation r and giving as an output one attribute or more from attributes of the relation r. The projection of the extended model is quite similar to the projection operator in the relational model. It gives as a result the attribute values associated with the *bba* values [2, 10, 20].

Example 19. *Let us take an example of an extended projection: π(ID,Blood Type) applied over the relation 2.14.*

Q: SELECT ID, Blood Type FROM EDB

Table 2.17: Result of π(ID,Blood Type) over EDB

ID	$Blood\ Type$
1	A
2	B 0.5 (B,O) 0.5

- Cartesian Product

  The *Cartesian product* of two relations consists on matching every tuple of the first relation with every tuple of the second relation. In the extended database model, the cartesian product of two tables T_1 and T_2 is defined in the usual way except for the value of CL in each tuple of the resulting table which is the *conjunctive combination* of confidence level values from tuples of T_1 and T_2 [2, 20].

Before we take an illustrative example we remind the definition of a function that computes the confidence level $CL[bel, pl]$ of two events e_1 and e_2 as presented in [20].

Defintion 2. *Let e_1 and e_2 be two independent events where $CL_1[bel_1, pl_1]$ and $CL_2[bel_2, pl_2]$ are the confidence levels relative to e_1 and e_2 respectively and let f be the function that calculates the values of CL_1 and CL_2 such that:*

$f([bel_1, pl_1], [bel_2, pl_2], 0, \wedge) = [bel_1 * bel_2,\ pl_1 * pl_2]$

$f([bel_1, pl_1], [bel_2, pl_2], 0, \vee) = [1-(1-bel_1)*(1-bel_2),\ 1\text{-}(1-pl_1)*(1-pl_2)]$

In general: $f(CL_1, CL_2, k, \wedge$ *or* $\vee)$

Note that k *represents the independence factor between* e_1 *and* e_2*. When* e_1 *and* e_2 *are independent* $k = 0$*.*

The notation $\wedge$ *represents the conjunctive combination and* $\vee$ *represents the disjunctive combination.*

Example 20. *Let table 2.18 and table 2.19 be two evidential tables. The resulting table of the cartesian product* $T_1 \times T_2$ *is presented in table 2.20.*

Note that u_{d_i} *is the evidential value of attribute Disease for* ID_i *and that* u_{t_i} *is the evidential value of attribute BloodType for* ID_i*.*

Table 2.18: Evidential table T_1

ID	$Disease$	$Blood\ Type$	CL
1	u_{d_1}	u_{t_1}	C_1
2	u_{d_2}	u_{t_2}	C_2

Table 2.19: Evidential table T_2

$Disease$	$Departement$	CL
u_{d_3}	u_{p_3}	C_3
u_{d_4}	u_{p_4}	C_4

Table 2.20: $T_1 \times T_2$

ID	$Disease$	$Disease$	$BloodType$	$Departement$	CL
1	u_{d_1}	u_{d_3}	u_{t_1}	u_{p_3}	$C_x(= f(C_1, C_3, 0, \wedge))$
1	u_{d_1}	u_{d_4}	u_{t_1}	u_{p_4}	$C_y(= f(C_1, C_4, 0, \wedge))$
2	u_{d_2}	u_{d_3}	u_{t_2}	u_{p_3}	$C_z(= f(C_2, C_3, 0, \wedge))$
2	u_{d_2}	u_{d_4}	u_{t_2}	u_{p_4}	$C_g(= f(C_2, C_4, 0, \wedge))$

- Join

  The *join operation* noted $\bowtie$ in a relational database model combines tuples from two or more tables using the common values for each table. The result of a join operation is a selection on cartesian product of the joined relations. In the extended relational model, the join behaves the same way as in the relational database model except for the confidence level values which are calculated as the conjunctive combination of confidence level values of combined tuples [2, 20, 21].

Example 21. *In the following example we apply the extended join operation on T_1 and T_2 as shown in table 2.21.*

Table 2.21: $T_1 \bowtie T_2$

ID	$Disease$	$Disease$	$BloodType$	$Depart$	CL
1	u_{d_1}	u_{d_3}	u_{t_1}	u_{p_3}	$f([bel(X_{u_{d_1}} = X_{u_{d_3}}), pl(X_{u_{d_1}} = X_{u_{d_3}})],$ $0, \wedge)$
2	u_{d_2}	u_{d_4}	u_{t_2}	u_{p_3}	$f([bel(Y_{u_{d_2}} = Y_{u_{d_4}}), pl(Y_{u_{d_2}} = Y_{u_{d_4}})],$ $0, \wedge)$

2.5 Conclusion

In this chapter, we presented probabilistic, possibilistic and evidential databases on their compact forms. We also showed that possible worlds were handled in probabilistic and possibilistic databases but not introduced in evidential databases. Some extended relational operators as selection, projection, cartesian product and join were presented. In the next chapter, we will present possible worlds in evidential databases and we will redefine some relational operators.

Chapter 3

Evidential database as Possible Worlds

3.1 Introduction

Models of querying evidential databases as proposed in [2, 21, 20] are based on the compact form of the database. Those models studied the relational queries like selection, projection and join. To show if these models are strong representation systems, the defined operations should give the same result if applied on the possible worlds form.

In this chapter, we will develop our contribution. First of all, we will define *possible worlds* PW for an evidential database EDB. Then, we will show the steps of generating PW from evidential database EDB. Finally, we will apply select and a project queries over those possible worlds. This chapter is a detailed extension of [8].

3.2 Definition of possible worlds

An evidential database EDB, as defined in section 2.4 on its *compact form*, is a set of N objects (tuples) and D attributes (see table 2.14).

The *non-compact form* of evidential database EDB is a finite set of *imprecise possible worlds* such that $EDB = \{IW_1, IW_2, ..., IW_k\}$. Each imprecise possible world includes N objects, where each object contains one focal element per attribute.

The table 3.1 is an example of an imprecise possible world induced from table 2.14.

Table 3.1: Example of an imprecise possible world of table 2.14

ID	Disease	BloodType
1	Flu	A
2	Anemia	{B,O}

Possible worlds are qualified by imprecise because they include imprecise values such as $\{B, O\}$ which is the value of the attribute $BloodType$ for the second object in table 3.1.

Note that the number of all imprecise possible worlds, k, is induced from sizes of sets F_{lc} as follows:

$$k = \prod_{l=1}^{N} \prod_{c=1}^{D} |F_{lc}| \tag{3.1}$$

For example, table 3.2 includes the four imprecise possible worlds from table 2.14. The set of imprecise possible worlds defines the non-compact form of the evidential database EDB.

Table 3.2: Imperfect worlds of EDB

IW_1	IW_2
$(1, Flu, A)$ $(2, Anemia, B)$	$(1, Flu, A)$ $(2, Anemia, \{B, O\})$
IW_3	IW_4
$(1, \{Asthma, Anemia\}, A)$ $(2, Anemia, B)$	$(1, \{Asthma, Anemia\}, A)$ $(2, Anemia, \{B, O\})$

A classical database is the relational union of its objects. Analogically, an evidential database is the union of its N evidential objects [1].

Defintion 3 (Evidential object)**.** *An evidential object is a basic belief assignment issued from the combination of all evidential values of that object. Its frame of discernment is the joint frame of all attributes' domains denoted by* Θ*. The mass function* m_l^{Θ} *related to the evidential object of line* l *is the conjunctive combination of all evidential values* m_{lc} *extended to the joint frame. An evidential object is defined as follows:*

$$m_l^{\Theta} = \textcircled{\cap}_{c\in[1,D]} m_{lc}^{\Theta_c\uparrow\Theta} \text{ with } \Theta = \prod_{c\in[1,D]} \Theta_c \tag{3.2}$$

Example 22. *In order to obtain the evidential objects of table 2.14: First, their evidential values are extended to the joint frame of discernment as detailed in table 22. Then, the extended evidential values are combined using the conjunctive rule as shown in table 3.4.*

Table 3.3: Extension of bbas from the evidential table 2.14

ID	$\Theta_{DE} \uparrow \Theta$	$\Theta_{BT} \uparrow \Theta$
1	$Flu \times \Theta_{BT}$ $\{Asthma, Anemia\} \times \Theta_{BT}$	$A \times \Theta_{DE}$
2	$Anemia \times \Theta_{BT}$	$B \times \Theta_{DE}$ $\{B, O\} \times \Theta_{DE}$

Table 3.4: Combination of the extended *bbas* using the conjunctive rule of combination

$\textcircled{\cap}$
$(1, Flu, A)\ 0.7 = 1 \times 0.7 \times 1$ $(1, \{Asthma, Anemia\}, A)\ 0.3 = 1 \times 0.3 \times 1$
$(2, Anemia, B)\ 0.5 = 1 \times 1 \times 0.5$ $(2, Anemia, \{B, O\})\ 0.5 = 1 \times 1 \times 0.5$

Formally, an evidential database is the disjunctive combination of its evidential objects; hence the following definition of the non-compact form of an evidential database:

Defintion 4 (Non-compact evidential database)**.** *A non-compact evidential database is presented through a bba issued from the disjunctive combination of all its evidential objects. That bba is defined as follows:*

$$m^{\Theta} = \textcircled{\cup}_{l\in[1,N]} m_l^{\Theta} \tag{3.3}$$

Example 23. *Table 3.5 is the disjunctive combination of all evidential objects detailed in table 3.4. The combined bba is defined on a frame of discernment* $\Theta_{IW} = \{IW_1, IW_2, IW_3, IW_4\}$ *such that each* IW_i *is an imprecise possible world.*

Table 3.5: Combination of the evidential objects of EDB using the disjunctive rule of combination

$\textcircled{\cup}$	
IW_1	IW_2
$(1, Flu, A)$ $(2, Anemia, B)$ 0.35	$(1, Flu, A)$ $(2, Anemia, \{B, O\})$ 0.35
IW_3	IW_4
$(1, \{Asthma, Anemia\}, A)$ $(2, Anemia, B)$ 0.15	$(1, \{Asthma, Anemia\}, A)$ $(2, Anemia, \{B, O\})$ 0.15

Each imprecise possible world IW_i can be expanded into different precise states called possible worlds W_j, as shown in table 3.6.
A same world W_j can be a precise state of several imprecise possible worlds. For example, the possible world W_1 is derived from imprecise worlds IW_1 and IW_2.
Each possible world is a candidate to represent the evidential relation. The degree of belief on a possible world to be candidate is given from the belief of its imprecise possible worlds.

Example 24. *The expansion of the first imprecise world* IW_1 *of table 2.14 gives the first possible world* W_1*, which heritates its mass* m*. It is defined as follows:*

$$m^{\Theta_{IW}}(IW_1) = m^{\Theta_W}(\{W_1\}) = 0.35$$

We illustrate here the basic belief assignment of possible worlds IW_2*,* IW_3*,* IW_4*,* IW_5*,* IW_6 *related to table 2.14:*

$$m^{\Theta_{IW}}(IW_2) = m^{\Theta_W}(\{W_1; W_2\}) = 0.35$$

$m^{\Theta_{IW}}(IW_3) = m^{\Theta_W}(\{W_3; W_4\}) = 0.15$

$m^{\Theta_{IW}}(IW_4) = m^{\Theta_W}(\{W_3; W_4; W_5; W_6\}) = 0.15$

Note that $m^{\Theta_{IW}}$ is defined on Θ_{IW} and that m^{Θ_W} is defined on Θ_W which is the set of all possible worlds obtained from Θ_{IW} [27].

3.3 Querying possible worlds

Let Q be the query processed on each possible world W_j. Querying each possible world W_j (noted $Q(W_j)$) gives a possible answer R_u:

$$R_u = Q(W_j)$$

Defintion 5 (Mass function on result sets)**.** *Let be the mass function m^{Θ_W}, defined over possible worlds of an evidential database EDB, where F^{Θ_W} is the set of focal elements. The mass function m^{Θ_R} is derived m^{Θ_W}, and is defined on the result sets R_u returned by a query Q:*

$$m^{\Theta_R}(\{R_u/R_u = Q(W_j), W_j \in F_j, F_j \in F^{\Theta_W}\} = m^{\Theta_W}(F_j/F_j \in F^{\Theta_W}) \quad (3.4)$$

- Selection

Example 25. *We consider the following selection query and its results on each possible world of table 2.14.*

Q_1 *:* SELECT * FROM EDB WHERE (Disease = {Anemia})

$Q_1(W_1) = \{(2, Anemia, B)\} = R_1$
$Q_1(W_2) = \{(2, Anemia, O)\} = R_2$
$Q_1(W_3) = \{(2, Anemia, B)\} = R_1$
$Q_1(W_4) = \{(1, Anemia, A); (2, Anemia, B)\} = R_3$
$Q_1(W_5) = \{(2, Anemia, O)\} = R_2$
$Q_1(W_6) = \{(1, Anemia, A); (2, Anemia, O)\} = R_4$

From example 25 and example 23 we deduce that:

The mass m defined on Θ_R for the first result R_1 is 0.35. It is presented as follows:

$m^{\Theta_R}(\{R_1\}) = 0.35$

Masses of R_2, R_3, R_4 are presented in below:

$m^{\Theta_R}(\{R_1; R_2\}) = 0.35$

$m^{\Theta_R}(\{R_1; R_3\}) = 0.15$

$m^{\Theta_R}(\{R_1; R_2; R_3; R_4\}) = 0.15$

Now, we can assign to each tuple t belonging to the union of sets R_u an interval limited by a belief and a plausibility measures, called the confidence level. The belief and the plausibility of a tuple t quantify the amount of belief and plausibility that t is a part of the query result. These measures are derived from the belief and the plausibility of results' sets that returned t.

Defintion 6 (Confidence Level)**.** *For each tuple t belonging in the set of responses R_u, a confidence level CL is calculated such that:*

$$t.CL = [bel(S), pl(S)] \text{ where } S = \{R_u / t \in R_u\} \tag{3.5}$$

Tuples, that satisfy a query Q, belong to the union of results' sets R_u. Assume they compose the set $R = \bigcup R_u$. Each tuple $t \in R$ is a potential result of the query Q applied on evidential database EDB.
To quantify satisfaction belief and plausibility of each one, we use definition 6.

Example 26. *The union of results' sets, denoted by $R = \bigcup R_u$, includes following tuples:*

$R = R_1 \cup R_2 \cup R_3 \cup R_4 = \{(1, Anemia, A); (2, Anemia, B); (2, Anemia, O)\}$

We compute the confidence level of each tuple in R:

- $t_1 = (1, Anemia, A)$ *is the answer appearing in R_3 and R_4. Thus, $t_1.CL = [bel(S), pl(S)]$; with $S = \{R_3; R_4\}$.*

Consequently, $t_1.CL = [0, 0.3]$.

- $t_2 = (2, Anemia, B)$ *is the answer appearing in* R_1 *and* R_3. *Thus,* $t_2.CL = [bel(S), pl(S)]$; *with* $S = \{R_1; R_3\}$.

  Consequently, $ID_2.CL = [0.5, 1]$.

- $t_3 = (2, Anemia, O)$ *is the answer appearing in* R_2 *and* R_4. *Thus,* $t_2.CL = [bel(S), pl(S)]$; *with* $S = \{R_2; R_4\}$.

  Consequently, $ID_2.CL = [0, 0.5]$.

Table 3.6: Possible worlds of EDB

IW_1
W_1
$(1, Flu, A)$ $(2, Anemia, B)$

IW_2	
W_1	W_2
$(1, Flu, A)$ $(2, Anemia, B)$	$(1, Flu, A)$ $(2, Anemia, O)$

IW_3	
W_3	W_4
$(1, Asthma, A)$ $(2, Anemia, B)$	$(1, Anemia, A)$ $(2, Anemia, B)$

IW_4			
W_3	W_4	W_5	W_6
$(1, Asthma, A)$ $(2, Anemia, B)$	$(1, Anemia, A)$ $(2, Anemia, B)$	$(1, Asthma, A)$ $(2, Anemia, O)$	$(1, Anemia, A)$ $(2, Anemia, O)$

Now, we present a different type of query application which is the projection. Then, study the results of the examples we will illustrate in below.

Example 27. *We consider the following projection query and its results on each possible world of table 2.14.*

Q_2 : SELECT *Disease*, *BloodType* FROM *EDB*

$$
\begin{aligned}
Q_2(W_1) &= \{(Flu, A); (Anemia, B)\} = R_1 \\
Q_2(W_2) &= \{(Flu, A); (Anemia, O)\} = R_2 \\
Q_2(W_3) &= \{(Asthma, A); (Anemia, B)\} = R_3 \\
Q_2(W_4) &= \{Anemia, A); (Anemia, B)\} = R_4 \\
Q_2(W_5) &= \{(Asthma, A); (Anemia, O)\} = R_5 \\
Q_2(W_6) &= \{(Anemia, A); (Anemia, O)\} = R_6
\end{aligned}
$$

From example 27 and example 23 we deduce the masses m^{Θ_R} *over responses* R_u*:*

$$m^{\Theta_R}(\{R_1\}) = 0.35$$

$$m^{\Theta_R}(\{R_1; R_2\}) = 0.35$$

$$m^{\Theta_R}(\{R_3; R_4\}) = 0.15$$

$$m^{\Theta_R}(\{R_1; R_2; R_3; R_4\}) = 0.15$$

Responses that satisfy the projection query Q_2 *belong to the union of sets* R_u*.*

$$R = R_1 \cup R_2 \cup R_3 \cup R_4 = \{(Flu, A); (Anemia, A); (Anemia, B); (Anemia, O); (Asthma, A)\}$$

Now we compute the confidence level CL *including the belief and the plausibility for each response* R *satisfying* Q_2*:*

- (Flu, A) *is the answer appearing in* R_1 *and* R_2*. Thus,* $CL_1 = [bel(S), pl(S)]$*; with* $S = \{R_1; R_2\}$*.*

  Consequently, $CL_1 = [0.7, 0.7]$*.*

- $(Anemia, B)$ *is the answer appearing in* R_1*,* R_3 *and* R_4*. Thus,* $CL_2 = [bel(S), pl(S)]$*; with* $S = \{R_2; R_3; R_4\}$*.*

  Consequently, $CL_2 = [0.5, 1]$*.*

- *$(Anemia, O)$ is the answer appearing in R_2, R_5 and R_6. Thus, $CL_3 = [bel(S), pl(S)]$; with $S = \{R_2; R_5; R_6\}$.*

  Consequently, $CL_3 = [0, 0.5]$.

- *$(Asthma, A)$ is the answer appearing in R_3 and R_5. Thus, $CL_4 = [bel(S), pl(S)]$; with $S = \{R_3; R_5\}$.*

  Consequently, $CL_4 = [0, 0.3]$.

- *$(Anemia, A)$ is the answer appearing in R_4 and R_6. Thus, $CL_5 = [bel(S), pl(S)]$; with $S = \{R_4; R_6\}$.*

  Consequently, $CL_5 = [0, 0.3]$.

We will take now an example of a projection query with a selection condition.

Example 28. *We consider an other projection query with the following selection condition and we present its results on each possible world of the same table 2.14.*

Q_3 : SELECT *Disease*, *BloodType* FROM *EDB*
WHERE (BloodType = B) OR (BloodType = O)

$Q_3(W_1) = \{(Anemia, B)\} = R_1$
$Q_3(W_2) = \{(Anemia, O)\} = R_2$
$Q_3(W_3) = \{(Anemia, B)\} = R_1$
$Q_3(W_4) = \{(Anemia, B)\} = R_1$
$Q_3(W_5) = \{(Anemia, O)\} = R_2$
$Q_3(W_6) = \{(Anemia, O)\} = R_2$

From example 28 and example 23 we deduce the masses m^{Θ_R} over responses R_u:

$m^{\Theta_R}(\{R_1\}) = 0.5$

$m^{\Theta_R}(\{R_1; R_2\}) = 0.5$

Responses that satisfy this query Q_3 belong to the union of sets R_u.

$R = R_1 \cup R_2 = \{(Anemia, B); (Anemia, O)\}$

Now we compute the confidence level CL including the belief and the plausibility for each response R satisfying Q_3:

- *$(Anemia, B)$ is the answer appearing in R_1 . Thus, $CL_1 = [bel(R_1), pl(R_1)]$. Consequently, $CL_1 = [0.5, 1]$.*

- *$(Anemia, O)$ is the answer appearing in R_2. Thus, $CL_2 = [bel(R_2), pl(R_2)]$. Consequently, $CL_2 = [0, 0.5]$.*

3.4 Exprimental Study

After we presented our formalism to model and query the possible worlds generated from the compact form of an evidential database. we will present an experimental study for the implemented model. The implementation was carried out for two reasons:

- The first reason is to study the accuracy of the theoretical results implemented in practice and to express the possible worlds representation in real-world applications.
- The second reason is to demonstrate to what limit it is possible to operate with possible worlds representation.

This chapter is dedicated to present the experimental results in order to meet these aims.

3.4.1 Data Description

In order to validate and test our approach described in the previous chapter, we consider a database structure with N tuples , D attributes and NF focal elements. Other parameters will be used like SF which represents the size of focal elements and U that represents the percentage of uncertainty in data.

$$U = \frac{Imperfect\ \ tuples}{N}.$$

The database model is evaluated using artificial data set. The approach that we implemented in NetBeans with Java programming language.

3.4.2 Experimental results

In the experimental phase, we performed tests using a data set on an i5-560 CPU(2.67) with 8 GB of RAM running Windows 8 Professional operating system.

The implementation treats the several contributions presented in the previous chapter:

- Generate the imprecise worlds IW from the original compact table EDB and their beliefs.
- Generate the possible W worlds from imprecise worlds and their heritated beliefs.
- Process a selection query Q with simple condition B over the compact form of EDB.
- Process a selection query Q with simple condition B over the non compact form of EDB.

In order to study the efficiency of the algorithm we will first of all present the results of testing some queries over the compact form of a medical table 2.14 and over its possible worlds.

Comparision between results of EDB and results of PW

We tested three types of queries. The query Q_1 is selection query with a simple condition. The second Q_2 is projection query and the query Q_3 is a projection query with selection condition. Results of tables 3.7, 3.8 and 3.9 show an equivalence between the calculated confidence levels in the compact form and in possible worlds of EDB for each query. .

Table 3.7: Comparative table for the query Q_1

Q_1: **SELECT * FROM EDB WHERE (Disease = Anemia)**	
Compact Form of EDB	**Possible Worlds of EDB**
$t_1.CL(Anemia) = [0, 0.3]$	$t_1.CL(Anemia) = [0, 0.3]$
$t_2.CL(Anemia) = [1, 1]$	$t_2.CL(Anemia) = [1, 1]$

Table 3.8: Comparative table for the query Q_2

Q_2: **SELECT Disease, BloodType FROM EDB**	
Compact Form of EDB	**Possible Worlds of EDB**
$CL(Flu, A) = [0.7, 0.7]$	$CL(Flu, A) = [0.7, 0.7]$
$CL(Anemia, B) = [0.5, 1]$	$CL(Anemia, B) = [0.5, 1]$
$CL(Anemia, O) = [0, 0.5]$	$CL(Anemia, O) = [0, 0.5]$
$CL(Asthma, A) = [0, 0.3]$	$CL(Anemia, O) = [0, 0.3]$
$CL(Anemia, A) = [0, 0.3]$	$CL(Anemia, A) = [0, 0.3]$

Table 3.9: Comparative table for the query Q_3

Q_3: **SELECT Disease, BloodType FROM EDB WHERE (BloodType = B) OR (BloodType = O)**	
Compact Form of EDB	**Possible Worlds of EDB**
$CL(Anemia, B) = [0.5, 1]$	$CL(Anemia, B) = [0.5, 1]$
$CL(Anemia, O) = [0, 0.5]$	$CL(Anemia, O) = [0, 0.5]$

In order to study to what limit is possible to operate with possible worlds

we will try to vary each time one parameter from N, NF, SF and fix the two other parameters.

Performance evaluation under focal elements' number variation

First of all we will start by increasing the number of focal elements NF in the database. The number of objects (tuples) is ($N = 2$), the size of focal elements is a combination between one or two focal elements ($SF = 1$) or ($SF = 2$), the number of attributes is ($D = 2$) and the percentage of imperfection is ($U = 100\%$) which means that all tuples contains imperfect values. Results are shown in figure 3.1.

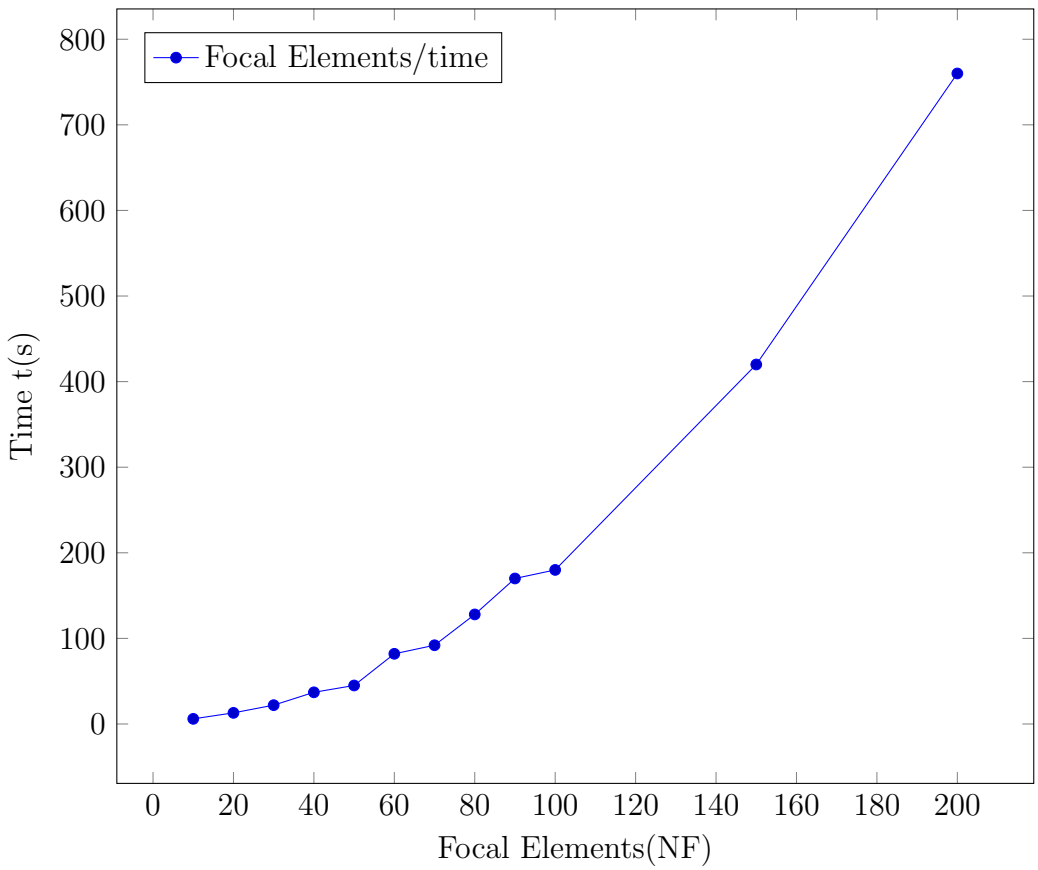


Figure 3.1: Performance evaluation under focal elements' number variation

This figure shows that we can increase the number of focal elements up to $NF = 200$ from this number the generation of possible worlds will exceed

$t = 30$ minutes and answering a query Q will take too long.

Performance evaluation under database size variation

Now, we will increase the number of tuples N in the database. The number of focal elements NF is four focal elements per tuple ($NF = 4$), the size of focal elements is a combination between one or two focal elements ($SF = 1$) or ($SF = 2$), the number of attributes is $D = 2$ and the percentage of imperfection is ($U = 100\%$); All tuples contains imperfect values. Results are shown in figure 3.2.

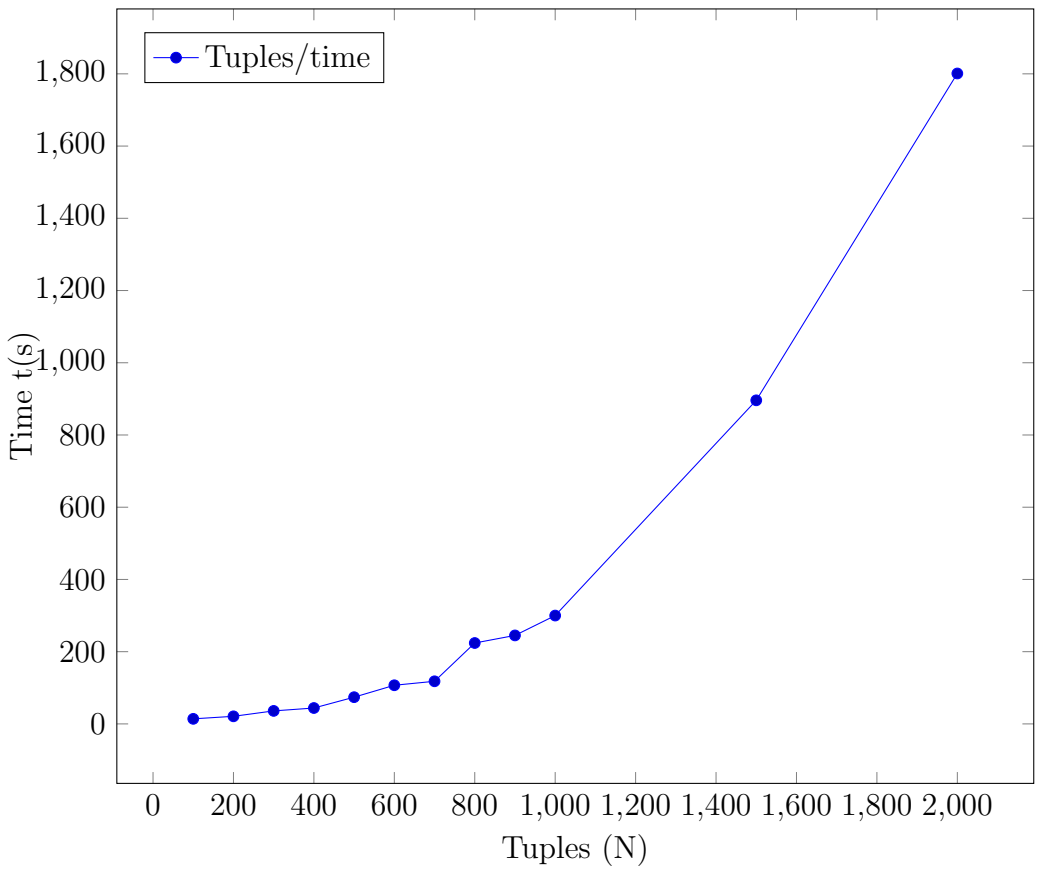


Figure 3.2: Performance evaluation under database size variation

This figure shows that we can increase the number of tuples up to $N = 2000$ from this number the generation of possible worlds will exceed $t = 30$

minutes and answering a query Q will take too long.

Performance evaluation under focal elements' size variation

Now, we will increase the size of focal elements SF in the database. The number of tuples is setted to two ($N = 2$), the number of focal elements NF is four focal elements per tuple ($NF = 4$), the number of attributes is ($D = 2$) and the percentage of imperfection is ($U = 100\%$); All tuples contains imperfect values. Results are shown in figure 3.3

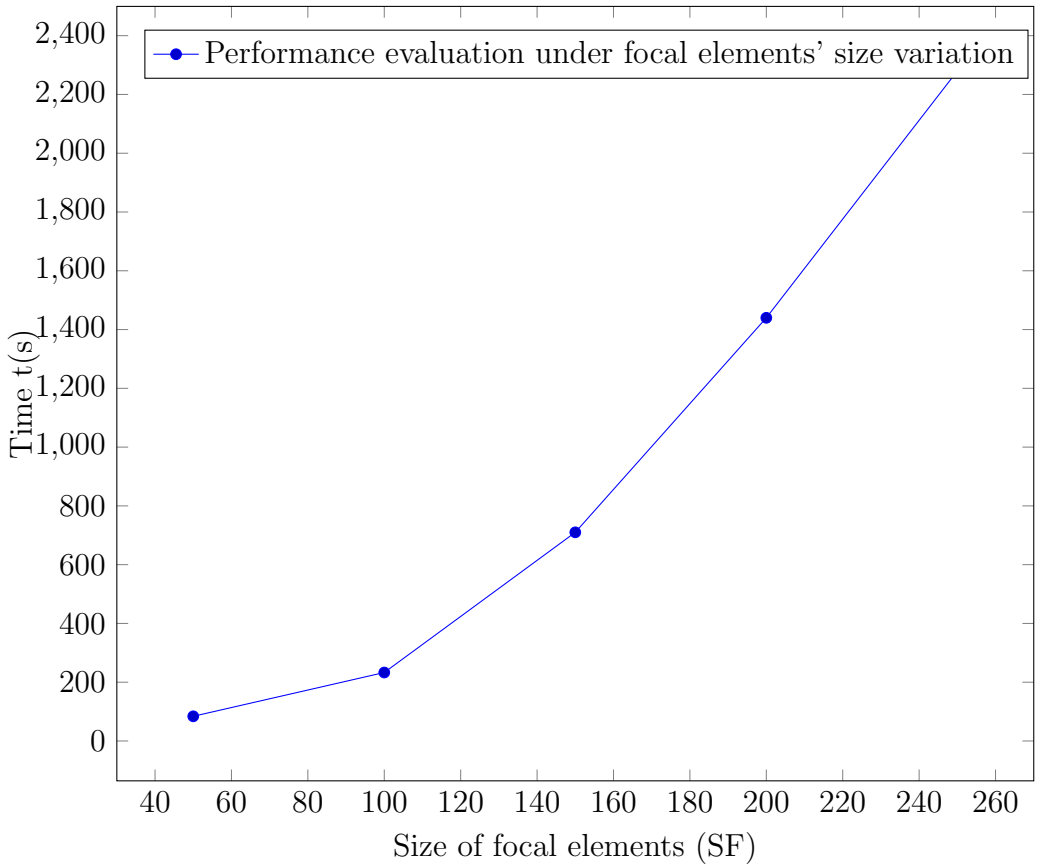


Figure 3.3: Performance evaluation under focal elements' size variation

This figure shows that we can increase the size of focal elements up to $SF = 250$, from this number the generation of possible worlds will exceed

$t = 40$ minutes and answering a query Q will take too long.

3.5 Conclusion

In this chapter we presented our new approach for modelling and querying evidential databases based on possible worlds.
At first, we used Dempster-Shafer theory tools, mainly the conjunctive and disjunctive rules to quantify the uncertainty about possible worlds generated from a compact evidential database. Uncertainty quantifies the degree of belief on each possible world to be a candidate to the evidential database. This result is very important in the literature since it impacts on all previous works on querying and analysing evidential data. At the best of our knowledge, all contributions about querying [2, 10], analysing [12] and mining [1, 25, 18, 4] evidential data are based on the compact form. Models cited above are not validated in comparison to the possible worlds form. In other words, it has not been proved that they constitute strong representation systems. Then, we formalized queries of evidential databases by the way of possible worlds form. The definition we introduced was about a generic relational query Q. For short, the examples we gave were about selection and projection operators only.

Finally, we putted our approach into tests to demonstrate to what limit the representation of possible worlds is usable. We noticed through tests that increasing one dimension of the evidential database among several dimensions each time generates the exponential increase in execution time. Certainly, the compact form is the only feasible representation but models of compact databases can not be validated without the possible worlds representation.

Conclusion

Many theories have proposed models to handle imperfect data. Probability theory is the oldest. It's drawback is being inefficient when it comes to represent the total ignorance.

Theory of evidence is a generalization of the bayesian inference. It offers a very simple representation to deal with uncertainty, imprecision, inconsistency and ignorance. Evidential databases are databases that handle imperfect data via the theory of evidence.

Data is stored to be later managed and queried. The existing models of evidential databases are based on the compact form and there is no model as our knowledge that represents an evidential database as its possible worlds.

The representation of possible worlds is fundamental to validate those compact models, that's why we proposed a new formalism for modelling and querying possible worlds of an evidential databases.

We redefined some relational operators as the selection and the projection operators [8].

As future works, all relational evidential operators may be detailed. In fact, we focused on the selection and the projection operators. However, all relational evidential operators namely the join, union and intersect may also be detailed in a future wok.

In addition, previous works on evidential databases introducing new relational operators (selection, projection, join, union and intersection); new analysis operators (such as skyline operator) and new data mining methods should be reviewed and validated with regard to the possible worlds form.

Finally, implementation of the rest of relational operators on possible worlds is a matter of a future work. Scalability is the main problem to solve,

because the number of possible worlds increases dramatically when sizes and numbers of focal elements increase.

Bibliography

[1] M. A. Bach Tobji, B. Ben Yaghlane, and K. Mellouli. A new algorithm for mining frequent itemsets from evidential databases. In *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1535–1542, Málaga, Spain, 2008.

[2] D. A. Bell, J. W. Guan, and S. K. Lee. Generalized union and project operations for pooling uncertain and imprecise information. *Data & Knowledge Engineering*, 18:89–117, 1996.

[3] F. Ben Abdelaziz, J. C. Pomerol, and A. Telmoudi. The evidential database model. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 110–113, 2002.

[4] S. Ben Hariz and Z. Elouedi. New dynamic clustering approaches within belief function framework. *Intelligent Data Analysis*, 18(3):409–428, 2014.

[5] P. Bosc, L. Lietard, and O. Pivert. About querying operations against possibilistic databases. In *Proceedings of the 5th World Multiconference on Systemics, Cybernetics and Informatics*, page 335342, 2001.

[6] P. Bosc, L. Lietard, and O. Pivert. Extending the possibilistic data model with one-level nested relations. In *Proceedings of the 11th International IEEE Conference on Fuzzy Systems*, 2002.

[7] P. Bosc and O. Pivert. About projection-selection-join queries addressed to possibilistic relational databases. *IEEE Transactions on Fuzzy Systems*, 13(1):124–139, 2005.

[8] F. E. Bousnina, M. A. Bach Tobji, M. Chabbeh, L. Lietard, and B. Ben Yaghlane. A new formalism for evidential databases. In *22nd International Symposium on Methodologies for Intelligent Systems*, Lyon, France, 2015.

[9] R. Cavallo and M. Pittarelli. The theory of probabilistic databases. In *Proceedings of the thirteenth VLDB Conference, Brighton*, pages 71–81, 1987.

[10] S. Choenni, H. E. Blok, and E. Leertouwer. Handling uncertainty and ignorance in databases: A rule to combine dependent data. In *Database Systems for Advanced Applications, 11th International Conference, DASFAA 2006, Singapore, April 12-15, 2006, Proceedings*, pages 310–324, 2006.

[11] A. P. Dempster. Upper and lower probabilities induced by a multiple valued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[12] S. Elmi, K. Benouaret, A. HadjAli, M. A. Bach Tobji, and B. Ben Yaghlane. Computing skyline from evidential data. In *8th International Conference on Scalable Uncertainty Management*, pages 148–161, Oxford, UK, 2014.

[13] N. Fuhr. A probabilistic framework for vaque queries and imprecise information in databases. In *Proceedings of the 16 th International Conference on very large Databases*, pages 696–707, Los Altos, USA, 1990.

[14] N. Fuhr. Integration of probabilistic fact and text retrieval. In *Proceedings of the 15 th Annual International Conference on Research and Developement in Infermation Retrieval*, pages 211–222, ACM, New York, 1992.

[15] N. Fuhr. The management of probabilistic data 4. In *IEEE Transactions on Knowledge and Data Engineering*, pages 487–502, 1992.

[16] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35:243–255, 1992.

[17] N. Fuhr. A probabilistic relational model for the integration of ir and databases. In *Proceedings of the 16 th Annual International Conference on Research and Developement in Infermation Retrieval*, pages 309–17, ACM, New York, 1993.

[18] K. Hewawasam, K. Premaratne, and M.-L. Shyu. Rule mining and classification in a situation assessment application: a belief theoretic approach for handling data imperfections. *IEEE Transactions on Systems, Man and Cybernetics*, 37(6):1446–1459, 2007.

[19] D. Kanjanatarakul.O, Sriboonchitta.S. Forecasting using belief functions: an application to marketing econometrics.

[20] S. K. Lee. An extended relational database model for uncertain and imprecise information. In *Proceedings of the 18th Conference on Very Large Data Bases*, pages 211–220, Canada, 1992.

[21] S. K. Lee. Imprecise and uncertain information in databases : an evidential approach. In *Proceedings of the 8th International Conference on Data Engineering*, pages 614–621, 1992.

[22] B. Lelandais, I. Gardin, L. Mouchard, P. Vera, and S. Ruan. Dealing with uncertainty and imprecision in image segmentation using belief function theory. pages 376–387, Canada, 2014.

[23] T. C. Prade, H. Generalizing database relational algebra for the treatment of incomplete/uncertain information and vague queries. *Information Sciencea*, (34):115–143, 1984.

[24] T. Rolleke. Equivlences of the probabilistic relational algebra. Technical report, University of Dortmund, Germany, 1994.

[25] A. Samet, E. Lefévre, and S. Ben Yahia. Mining frequent itemsets in evidential database. In *Proceedings of the Fifth International Conference Knowledge and Systems Engineering 2013*, pages 377–388, 2013.

[26] G. Shafer. A mathematical theory of evidence. *Princeton University Press*, 1976.

[27] P. Smets. Belief functions: The disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[28] P. Smets. The application of the transferable belief model to diagnostic problems. *International Journal of Intelligent Systems*, 13(2-3):127–157, 1998.

[29] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[30] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.