Modeling and Querying Evidential Databases

Fatma Ezzahra Bousnina

THESE
Pour l'obtention du grade de DOCTEUR

DE L'ECOLE NATIONALE SUPERIEURE DE MECANIQUE ET D'AEROTECHNIQUE DE POITIERS
DE L'INSTITUT SUPERIEUR DE GESTION DE TUNIS
(Diplôme National - Arrêté du 25 mai 2016)

Ecole Doctorale : Sciences et Ingénierie pour l'information de Mathématiques
Secteur de Recherche : INFORMATIQUE ET APPLICATIONS

Présentée par :

FATMA EZZAHRA BOUSNINA

Modeling and Querying Evidential Databases

Directeurs de thèse :
HADJALI Allel
BEN YAGHLANE Boutheina

JURY

Rapporteur : MARTIN Arnaud	Professeur, Université Rennes 1
Rapporteur : BEN YAHIA Sadok	Professeur, Université de Tunis El Manar
Présidente : GHEDIRA GUEGAN Ons Chirine	Professeur, Université Lyon 3
Membres du jury :	
BEN ABDALLAH Salah	Professeur, Université de Tunis
HADJALI Allel	Professeur, ISAE-ENSMA
BEN YAGHLANE Boutheina	Professeur, Université de Carthage
BACH TOBJI Mohamed Anis	Maitre de Conférences, Université de La Manouba

Soutenue le 11/06/2019

Résumé

La théorie des fonctions des croyances offre des outils puissants pour modéliser et traiter les informations imparfaites. En effet, cette théorie peut représenter l'incertitude, l'imprécision et l'ignorance. Dans ce contexte, les données sont stockées dans des bases de données spécifiques qu'on appelle les bases de données crédibilistes. Une base de donnée crédibiliste a deux niveaux d'incertitudes: (i) l'incertitude au niveau des attributs qui se manifeste à travers des degrés de véracité sur les hypothèses des attributs; (ii) l'incertitude au niveau des tuples représentée par des intervalles de confiance sur l'existence des tuples au sein de la table en question. D'autre part, la base de donnée crédibiliste peut être modélisée sous deux formes: (i) la forme compacte caractérisée par un ensemble d'attributs et un ensemble de tuples; (ii) la forme des mondes possibles représentée par un ensemble de base de données candidates où chaque base candidate est une représentation possible de la base de donnée compacte. Interroger la représentation des mondes possibles est une étape fondamentale pour valider les méthodes d'interrogation sur la base compacte crédibiliste. En effet, un modèle de base de donnée est dit système fort si le résultat de l'interrogation de sa représentaion compacte est équivalent au résultat de l'interrogation de sa représentation des mondes possibles.

Cette thèse est une étude sur les fondements des bases de données crédibilistes. Les contributions sont résumées comme suit:

(i) La modélisation et l'interrogation de la base crédibiliste (EDB): Nous mettons en pratique le modèle compacte de la base de données (EDB) en proposant une implémentation objet-relationnelle, ce qui permet d'introduire l'interrogation de ce modèle avec les opérateurs relationnels. D'autres part, nous présentons le formalisme, les algorithmes et les expérimentations d'autres types de rêquetes: les top-k évidentiel et le skyline évidentiel que nous appliquons sur des données réelles extraites de la plateforme Tripadvisor.

(ii) La modélisation de la base de données sous sa forme des mondes possibles: Nous modélisons la forme de mondes possibles de la base de données (EDB) en traitant les deux niveaux d'incertitudes (niveau attributs et niveau tuples).

(iii) La modélisation et l'interrogation de la base de données crédibiliste (ECD): Après avoir prouvé que le modèle des bases de données (EDB) n'est pas un système de représentation fort, nous développons le modèle de la base de données crédibiliste conditionnelle nommée (ECD). Nous présentons le formalisme de l'interrogation sur les deux formes (compacte et mondes possibles) de la base de données (ECD). Finalement, nous discutons les résultats de ces méthodes d'interrogation et les spécificités du modèle (ECD).

Cette thèse inclut cinq chapitres:

Chapitre 1: Les Bases de données imparfaites: Le contexte général

Ce chapitre englobe quelques notions générales reliées au contexte général de notre thèse, à savoir les théories d'imperfection et les bases de données imparfaites. Premièrement, nous présentons les différents types d'imperfection (incertitude, imprécision et inconsistance). Ces types d'imperfection sont illustrés avec des exemples bien détaillés. Après, nous présentons quelques théories d'imperfection, introduites historiquement pour gérer l'imperfection selon ces types. Nous avons choisi la théorie des probabilités, la théorie des possibilités et la théorie des fonctions de croyances. Finalement, nous présentons quelques modèles de bases de données imparfaites basés sur ces théories, qui sont les bases de données probabilistes, les bases de données possibilistes et les bases de données crédibilistes.

Chapitre 2: La théorie des fonctions de croyances

Dans ce chapitre, nous présentons quelques concepts de bases reliés à la théorie des fonctions de croyances. Des notions comme les règles de combinaison, l'extension à vide, l'indépendance cognitive et évidentielle sont illustrées avec des exemples détaillés afin de gérer les données modélisées à travers cette théorie.

Chapitre 3: Les bases de données crédibilistes: La forme compacte

Ce Chapitre est à propos la forme compacte des bases de données évidentielles (EDB) et il est divisé en trois parties: Dans la première, nous présentons l'unique modèle existant sur les bases de données évidentielles (Bell et al., 1996; Lee, 1992b; Lee, 1992a). Dans la deuxième partie, nous définissons le méta-modèle, ainsi que l'implémentation du modèle des base de données (EDB) (Bousnina et al., 2016). Dans la dernière partie, nous nous focalisons sur l'interrogation de la forme compacte de (EDB). En effet, nous rappelons les opérateurs relationnels évidentiels étendus (Bell et al., 1996) et nous appliquons l'opérateur de sélection-projection étendu on se basant sur l'algorithme présenté. Finalement, nous discutons deux opérateurs préférentiels; (i) le top-k évidentiel (Bousnina et al., 2017a) qui permet de donner les k meilleurs résultats avec leurs intervalles de confiance et (ii) le skyline évidentiel (Elmi et al., 2014; Bousnina et al., 2017b) qui permet de donner les meilleurs résultats qui ne sont pas dominés par aucun autre résultat dans la base de données. Pour appliquer le skyline évidentiel, nous avons collecté des données réelles de la plateforme *TripAdvisor*. Ces données ont été traitées et modélisées à travers les outils de la théorie des fonctions des croyances.

Chapitre 4: Les bases de données crédibilistes: La forme des mondes possibles

Ce chapitre est à propos la modélisation et l'interrogation des bases de données évidentielles sous la forme des mondes possibles. Premièrement, nous définissons les systèmes de représentation et leurs propriétés. Après, nous présentons le modèle des bases de données évidentielles (EDB) sous sa forme non compacte, c'est à dire, les mondes possibles. La génération des mondes possibles nécessite le traitement de l'incertitude niveau attributs et niveau tuples. En effet, elle se déroule sur trois étapes intermédiaires: (1) la génération des mondes imprécis à partir de la base compacte; (2) la génération des mondes incertains à partir des mondes imprécis; (3) la génération des mondes possibles à partir des mondes incertains. Une fois nous avons les deux représentations équivalentes de la base de données (EDB), nous définissons le processus d'interrogation. Interroger les deux formes permet d'évaluer les méthodes d'interrogation appliquées sur ce modèle afin de déterminer la nature de son système de représentation. Nous avons prouvé, à la dernière partie de ce chapitre, que le modèle (EDB) ne représente pas un système fort.

Chapitre 5: Les bases de données crédibilistes Conditionnelles

Ce chapitre est à propos la modélisation et l'interrogation des bases de données évidentielles conditionnelles, appelées ec-tables. En effet, nous utilisons les points forts des bases de données conditionnelles classiques et ceux de la théorie des fonctions des croyances pour introduire un nouveau modèle de bases de données évidentielles appelé les bases de données évidentielles conditionnelles (ECD). Tout d'abord, nous présentons ce modèle sous ces deux formes: la forme compacte et la forme des mondes possibles. La génération des mondes possibles pour le modèle (ECD) se déroule sur deux étapes intermédiaires: (1) la génération des mondes incertains à partir de la base compacte; (2) la génération des mondes possibles à partir des mondes incertains. A travers ces étapes nous traitons l'incertitude niveau tuple. Une fois nous avons les deux représentations nous présentons le processus d'interrogation et nous appliquons l'opérateur de sélection-projection. En effet, les méthodes d'interrogation sont utilisées pour vérifier l'équivalence entre les réponses issues de chacune des deux représentations. En plus, nous montrons que chaque base de données crédibiliste (EDB) peut être transformée en base de données crédibiliste (ECD). Cette transformation peut être un moyen de rendre le modèle (EDB) un système de représentation fort. Finalement, nous discutons les spécificités des bases de données conditionnelles qui présente une base solide pour la définition d'un système de représentation fort dans le cadre crédibiliste.

Abstract

The theory of belief functions (a.k.a, the Evidence Theory) offers powerful tools to model and handle imperfect pieces of information. Thus, it provides an adequate framework able to represent conjointly uncertainty, imprecision and ignorance. In this context, data are stored in a specific database model called evidential databases. An evidential database includes two levels of uncertainty: (i) the attribute level uncertainty expressed via some degrees of truthfulness about the hypotheses in attributes; (ii) the tuple level uncertainty expressed through an interval of confidence about the existence of the tuple in the table. An evidential database itself can be modeled in two forms: (i) the compact form represented as a set of attributes and a set of tuples; (ii) the possible worlds' form represented as a set of candidate databases where each candidate is a possible representation of the imperfect compact database. Querying the possible worlds' form is a fundamental step in order to check the querying methods over the compact one. In fact, a model is said to be a strong representation system when results of querying its compact form are equivalent to results of querying its non compact form.

This thesis focuses on foundations of evidential databases in both modeling and querying. The main contributions are summarized as follows:

(i) Modeling and querying the compact evidential database (EDB): We implement the compact evidential database (EDB) using the object-relational design which allows to introduce the querying of the database model under relational operators. We also propose the formalism, the algorithms and the experiments of other types of queries: the evidential top-k and the evidential skyline that we apply over a real dataset extracted from *TripAdvisor*.

(ii) Modeling the possible worlds' form of (EDB): We model the possible worlds' form of the evidential database (EDB) by treating both levels of uncertainty (the tuple level and the attribute level).

(iii) Modeling and querying the evidential conditional database (ECD): After proving that the evidential database (EDB) is not a strong representation system, we develop a new evidential conditional database model named (ECD). Thus, we present the formalism of querying the compact and the possible worlds' forms of the (ECD) to evaluate the querying methods under relational operators. Finally, we discuss the results of these querying methods and the specificities of the (ECD) model.

Acknowledgements

I would like to thank the members of my thesis committee: *Pr. Arnaud martin* (Université de Rennes 1, France), *Pr. Sadok Ben Yahia* (Université de Tunis El Manar, Tunisia), *Pr. Ons Chirine Ghedira Guegan* (Université de Lyon 3, France) and *Pr. Salah Ben Abdallah* (Université de Tunis, Tunisia) for accepting to review my dissertation.

I am grateful to my supervisors *Pr. Allel HADJALI* and *Pr. Boutheina BEN YAGHLANE* for allowing me to be a part of *LIAS* and *LARODEC* laboratories.

I am deeply grateful to my co-supervisors *Dr. Mohamed Anis BACH TOBJI* and *Dr. Mouna CHEBBAH* for their continuous support and guidance. I thank them for their valuable discussions and for their patience all long this work.

Many thanks to all members of *LARODEC* laboratory and *LIAS* Laboratory.

I would like to thank my mother *Ines* and my father *Habib* for their endless support and encouragement. I am deeply thankful to my beloved husband *Ahmed* for his encouraging words throughout all the tough moments. I would like also to thank my siblings *Rayhane* and *Iheb* for their love and support.

Finally, I am so thankful to my friends in Tunisia and in France and to my family members. Their words and prayers meant a lot for me.

This hole experience was very challenging mentally, emotionally and financially for me, but at the same time, it was so rich and deep. I am so proud to have fully accomplish it.

Contents

Introduction **1**

I Background Material **9**

1 Imperfect Databases: General Context **11**

1.1 Introduction 12

1.2 Typology of Imperfection 12

1.3 Theories of Imperfection 14

1.3.1 The Theory of Probabilities 15

1.3.2 The Possibility Theory 15

1.3.3 The Evidence Theory 17

1.4 Imperfect Databases 17

1.4.1 Probabilistic Databases 18

1.4.2 Possibilistic Databases 21

1.4.3 Evidential Databases 23

1.5 Conclusion 27

2 Theory of Evidence **29**

2.1 Introduction 30

2.2 Basic Concepts 30

2.2.1 Frame of discernment 30

2.2.2 Basic belief assignment 31

2.2.3 Body of evidence 32

2.2.4 Belief and plausibility functions 32

2.2.5 Special belief functions 38

2.3 Combination Rules 42

2.3.1 Dempster's rule of combination 43

2.3.2 Conjunctive rule of combination 44

2.3.3 Disjunctive rule of combination 45

2.4 Discounting 46

2.5 Compatible Frames of Discernments 47

2.5.1 Vacuous Extension 47

2.5.2 Refinement and Coarsening 49

2.6 Cognitive and Evidential Independences 50

2.6.1 Cognitive Independence 50

2.6.2 Evidential Independence 51

2.7 Decision Making 52

2.7.1 Pignistic Probability 52

2.7.2 Maximum of Credibility 53

2.7.3 Maximum of Plausibility 54

2.8 Conclusion 54

II Evidential Database Models 55

3 Evidential Databases: The Compact Form 57

3.1 Introduction 58

3.2 Evidential Databases Model 58

3.3 Implementing Evidential Databases 59

3.3.1 Meta-model 60

3.3.2 Object-Relational Implementation 60

3.4 Querying Evidential Databases EDB 62

3.4.1 Evidential Relational Queries 63

3.4.2 Evidential Top-k Query 68

3.4.3 Evidential Skyline Query 79

3.5 Conclusion 83

4 Evidential Databases: The Possible Worlds' Form 87

4.1 Introduction 88

4.2 Representation Systems 88

4.3 Evidential Database as Possible Worlds 90

4.3.1 Imprecise Possible Worlds 92

4.3.2 Uncertain Possible Worlds 93

4.3.3 Possible Worlds 98

4.4 Querying Possible Worlds 106

4.5 Generalizing Probabilistic Databases 109

4.6 EDB: What Kind of Representation System? 110

4.7 Conclusion 114

5 Evidential Conditional databases 117

5.1 Introduction 118

5.2 Conditional Tables 118

5.3 Evidential Conditional Tables 120

5.3.1 Modeling ec-tables: Compact form 120
5.3.2 Modeling ec-tables: Possible Worlds' Form 121
5.3.3 Querying ec-tables: Possible worlds form 124
5.3.4 Querying ec-tables: Compact form 126
5.3.5 Checking the equivalence 128
5.4 Discussion 128
5.5 Conclusion 140

Conclusion **143**

Bibliography **147**

List of Tables

1 Notations ix

1.1 Recapitulation of imperfection types 14
1.2 An example of a probabilistic table PDB 19
1.3 Possible worlds of the probabilistic table PDB 19
1.4 An example of a possibilistic database 22
1.5 Possible worlds of the possibilistic database 22
1.6 An evidential table EDB 25
1.7 Possible worlds of the evidential table EDB 25

2.1 A medical diagnosis for two patients 35
2.2 Dempster's Rule of combination over two different diagnoses of a same patient 44
2.3 An application of the conjunctive rule of combination 45
2.4 An application of the disjunctive rule of combination 46
2.5 An overview of combination rules properties 46
2.6 Diagnoses of two patients 48
2.7 The vacuous extension of Table 2.6 49
2.8 $pl^{\Theta_O \uparrow \Theta}$ and $pl^{\Theta_G \uparrow \Theta}$ 50
2.9 $bel^{\Theta_O \uparrow \Theta}$ and $bel^{\Theta_G \uparrow \Theta}$ 52

3.1 A Medical Evidential Table 59
3.2 Result of query Q_1 64
3.3 Result of query Q_2 64
3.4 EDB_1 65
3.5 EDB_2 65
3.6 Cartesian Product: $EDB_1 \times EDB_2$ 65
3.7 Join: $EDB_1 \bowtie EDB_2$ 65
3.8 Contribution of caches to queries re-execution 68
3.9 Books Appreciations' Table: BAT 69
3.10 Evidential Score per Item 71
3.11 ETop-k Naive Algorithm 76
3.12 ETop-k Optimized Algorithm 77
3.13 Impact of the database size for methods: NaiTopK and OptTopK 78

3.14 Reviews about Hotels 80
3.15 Construction of mass functions 80
3.16 Discounting of mass functions 82
3.17 Combination of *bba*s about the Price of h_2 82
3.18 The obtained Evidential Database from $TripAdvisor$ 83

4.1 Medical Evidential Table about Blood Types 88
4.2 Vacuous extension of the focal elements of the evidential Table 4.1 . . . 91
4.3 Combination of the extended *bbas* using the conjunctive rule of combination 92
4.4 An Imprecise World 92
4.5 Combination of the evidential objects of EDB using the disjunctive rule of combination 94
4.6 Uncertain Possible worlds of EDB 95
4.7 Uncertain Possible worlds of EDB 97
4.8 Uncertain Possible worlds of EDB 97
4.9 Possible Worlds 99
4.10 Possible Worlds with their CLs 102
4.11 An Evidential Table EDB 111
4.12 Imprecise Worlds of EDB 112
4.13 Possible Worlds of EDB 112

5.1 Examples of Representation Systems of Imperfect Databases 119
5.2 Example of a c-table 120
5.3 Possible worlds of the c-table of Table 5.2 120
5.4 A medical ec-table 121
5.5 Uncertain possible worlds of a medical ec-table 122
5.6 Possible worlds of ec-table 123
5.7 Compact result of query Q: $Q(ECD)$ 126
5.8 The possible worlds of the compact result of query Q 126
5.9 An Evidential table EDB 129
5.10 Obtained Conditional Evidential table 129
5.11 A medical evidential database EDB 133
5.12 Imperfect worlds of EDB 133
5.13 Uncertain worlds of EDB 134
5.14 Uncertain worlds of EDB 134
5.15 Possible worlds of EDB 135
5.16 Table result Q(EDB) 135
5.17 The Obtained evidential database ECD after translation 137
5.18 Uncertain worlds of ECD 137
5.19 Possible worlds of EDB 138
5.20 The compact result Q(ECD) 138

List of Figures

1 Modeling and Querying evidential databases EDB: Context and Contributions . . . 4
2 Modeling and Querying ec-tables: Context and contributions . . . 5

1.1 Numerical and Ordinal interpretations of the Possibility Theory . . . 16
1.2 A Probabilistic Database . . . 18
1.3 Querying process of the Probabilistic Databases (where PWs refers to possible worlds) . . . 20
1.4 A Possibilistic Database . . . 21
1.5 Querying process of the Possibilistic Databases . . . 23
1.6 An Evidential Database . . . 24
1.7 Querying process of the Evidential Databases . . . 26

2.1 Credal and Pignistic Levels of the Evidence theory . . . 31
2.2 Bayesian mass function . . . 39
2.3 Consonant mass function . . . 40
2.4 Simple support mass function . . . 41
2.5 Combination of M independent sources (Dempster, 1967) . . . 43
2.6 Vacuous Extension . . . 48
2.7 Coarsening and Refinement . . . 49

3.1 Meta-Model of Evidential Databases . . . 60
3.2 Experimentation Results . . . 84
3.3 Comparison of Evidential Scores . . . 85
3.4 Comparison of performance of NaiETopK and OptETopK . . . 85
3.5 Computation of points in *TripAdvisor* and their corresponding levels . 86
3.6 Estimated Reviewers' Reliabilities . . . 86

4.1 An imperfect database DB and its possible worlds Rep(DB) . . . 89
4.2 Querying formalism towards a SRS . . . 90
4.3 Process of producing the possible worlds form . . . 104
4.4 Illustrative example for generating possible worlds of an EDB . . . 105
4.5 The process of querying both forms of EDB . . . 111
4.6 The querying process the compact form of EDB . . . 113

4.7 The querying process of the non compact form of EDB 113
4.8 Checking the equivalence between possible answers issued from both forms of EDB 114

5.1 ECD Model 124
5.2 The first tuple's interpretation from EDB to ECD 130
5.3 From the compact ec-table to its possible worlds 131
5.4 The complete process: From EDB to ECD 132

Table 1: Notations

$Notation$	$Meaning$
DB	an imperfect database
EDB	an evidential database
CD	a conditional database
ECD	an evidential conditional database
t_l	a tuple/object
a	an attribute
N	number of tuples of an EDB
D	number of attributes of an EDB
m^Θ	mass function
$\mathcal{F}$	set of focal elements of a *bba*
φ_{m^Θ}	the core of a *bba* m^Θ
I	number of imprecise possible worlds
J	number of uncertain possible worlds
P	number of possible worlds
V_{ta}	evidential value of the tuple t $(1 \leq t \leq N)$ for the attribute a $(1 \leq a \leq D)$
IW	an imprecise possible world
UW	uncertain possible world
W	a possible world
Θ, Ω	frames of discernment
Θ_{IW}	the set of imprecise possible worlds
Θ_{UW}	the set of uncertain possible worlds
Θ_W	the set of possible worlds
CL	a confidence level composed of two values $[bel; pl]$
bel	belief measure
pl	plausibility measure
x, y	hypothesis
A, B	focal elements
M	independent sources
S_j	a source $\in \{1..M\}$
R_i	a response of a given query Q
R_u	a possible response derived from the possible worlds' form
R'_s	a possible response derived from the compact form
$\%IR$	imperfect rate
nfe	number of focal elements
sfe	size of focal elements
d	size of attribute domain
λ	preference degree quantification
$S(R_i)$	score of evidential answer R_i

Introduction

Imperfect information is prevalent in real life fields. For example in the medical domain doctors can provide more than one hypothesis about a patient's diagnosis. In weather forecasting also, experts often give approximations about the weather, etc. In fact, several types of imperfection can be encountered, depending on the domain, like imprecision, uncertainty, ignorance, ambiguity, inconsistency, etc. Therefore, lots of theories were introduced to handle some types of these imperfection. One of the most known and used is the *probability theory* (Laplace, 1812). Other theories were later introduced to offer different models like the *possibility theory* (Zadeh, 1978), the *fuzzy sets theory* (Zadeh, 1965), the *rough sets theory* (Pawlak, 1982) and the *belief functions theory* (Shafer, 1976; Dempster, 1967). The latter, called also the *evidence theory*, models and combines imperfect information through an explicit representation of uncertainty, imprecision and ignorance. It offers several powerful tools to aggregate information coming form one or more sources in order to improve the decision making.

Managing imperfect information requires storing them under specified database models. Therefore, based on the mentioned theories above, different database models were introduced: such as the *probabilistic databases* (Cavallo and Pittarelli, 1987), the *possibilistic databases* (Bosc and Pivert, 2005) and the *evidential databases* (Bell et al., 1996; Lee, 1992b; Lee, 1992a), etc. In the case of an evidential database, two equivalent forms can be modeled: (i) *the compact form* (Lee, 1992b; Lee, 1992a; Bell et al., 1996) where attributes' values are distributions (ii) *the possible worlds' form, also named the non compact form* (Bousnina et al., 2015; Bousnina et al., 2018a) where the database is a distribution of candidate databases.

(i) In the compact form, the uncertainty is expressed in two levels:

1. The attribute level uncertainty where an evidential distribution is assigned to each hypothesis in the imperfect attribute. This assignment is called a mass.
2. The tuple level uncertainty where an interval is assigned to each tuple that reflects the confidence on the existence of the tuple in the database. This interval is named the confidence level.

(ii) The possible worlds, is the other representation where the evidential database is modeled through a distribution of candidate databases.

The compact representation is the only feasible model in practice because the number of the generated possible worlds can increase exponentially. Added to that it is quasi-impossible to store and query all of them. However, defining possible worlds is fundamental to evaluate and validate representation systems. According to (Imielinski and Lipski, 1984) and (Abiteboul et al., 1995b), a model is said to be a *strong representation system* when querying the compact form is *equivalent* to querying its possible worlds.

Many researchers have been interested in modeling and querying the compact model of an evidential database (Lee, 1992b; Lee, 1992a; Bell et al., 1996; Bousnina et al., 2016). Indeed, this model was used in real evidential databases like the amphiphilic chemical database (Samet and Dao, 2015) and the TripAdvisor reviews' database (Bousnina et al., 2017b). However, to the best of our knowledge, no model was proved to be a strong representation system for evidential databases.

The main purpose of this thesis is to study the evidential databases from a modeling and querying point of view. Thus, we address the following research questions: Is the most used evidential database (Lee, 1992b; Lee, 1992a; Bell et al., 1996) a strong representation system? How to model its possible worlds' form? How to treat the double uncertainty levels? How to query evidential databases? Is there another database model that represents and queries efficiently the evidential data?

Our main contributions to respond to these research issues are the following:

(i) Modeling and querying the compact Evidential DataBase (EDB):

- Object-Relational Evidential Implementation (Bousnina et al., 2016): We give an object relational model for the evidential database on its compact form (Lee, 1992b; Lee, 1992a; Bell et al., 1996). Then we present the object relational implementation using SQL3 and Java. In fact, we take advantage of Oracle Database Management characteristics to evaluate two evidential relational operators: the select and the project.

- Evidential Top-K query (Bousnina et al., 2017a; Bousnina et al., 2018b): We introduce a new ranking query under the evidential framework. Thus, we present a formalism to select the best k responses when querying the evidential database on its compact form. Then, the semantic aspect and the experimental study are detailed. The evidential Top-k query evaluation is based on the object-relational evidential implementation.

- Evidential skyline query (Bousnina et al., 2017b): We face the problem of aggregating information coming from different sources to construct an evidential database where information are coming from a real application platform (*the TripAdvisor platform*). Then, the evidential skyline query, as introduced in (Elmi et al., 2014), is applied.

(ii) Modeling and querying the possible worlds of the Evidential DataBase (EDB):

Modeling evidential databases as Possible worlds (Bousnina et al., 2018a): We model the evidential database (Bell et al., 1996) on its non compact form, i.e, the possible worlds' form. This contribution is based on a previous work (Bousnina et al., 2015). When modeling the evidential database into possible worlds we manage the two levels of uncertainties (the tuple level uncertainty and the attribute level uncertainty). This work allows the evaluation of querying methods in order to check if the EDB (Lee, 1992b; Lee, 1992a; Bell et al., 1996) is a strong representation system or not.

(iii) Modeling and querying the Evidential Conditional Database (ECD):

Evidential Conditional Tables: We prove that the most used evidential database EDB (Lee, 1992b; Lee, 1992a; Bell et al., 1996) is not a strong representation system. Indeed, we introduce a new evidential database model, named ec-tables, ECD for short. In fact, we present the mathematical formulation of this database model on its compact form and then on its possible worlds' form. This model, i.e, the ec-tables, represents a further step towards strong representation systems under relational operators. We use several detailed examples to illustrate the presented formalism.

The context of this thesis and our contributions are graphically detailed in Figure 1 and Figure 2.

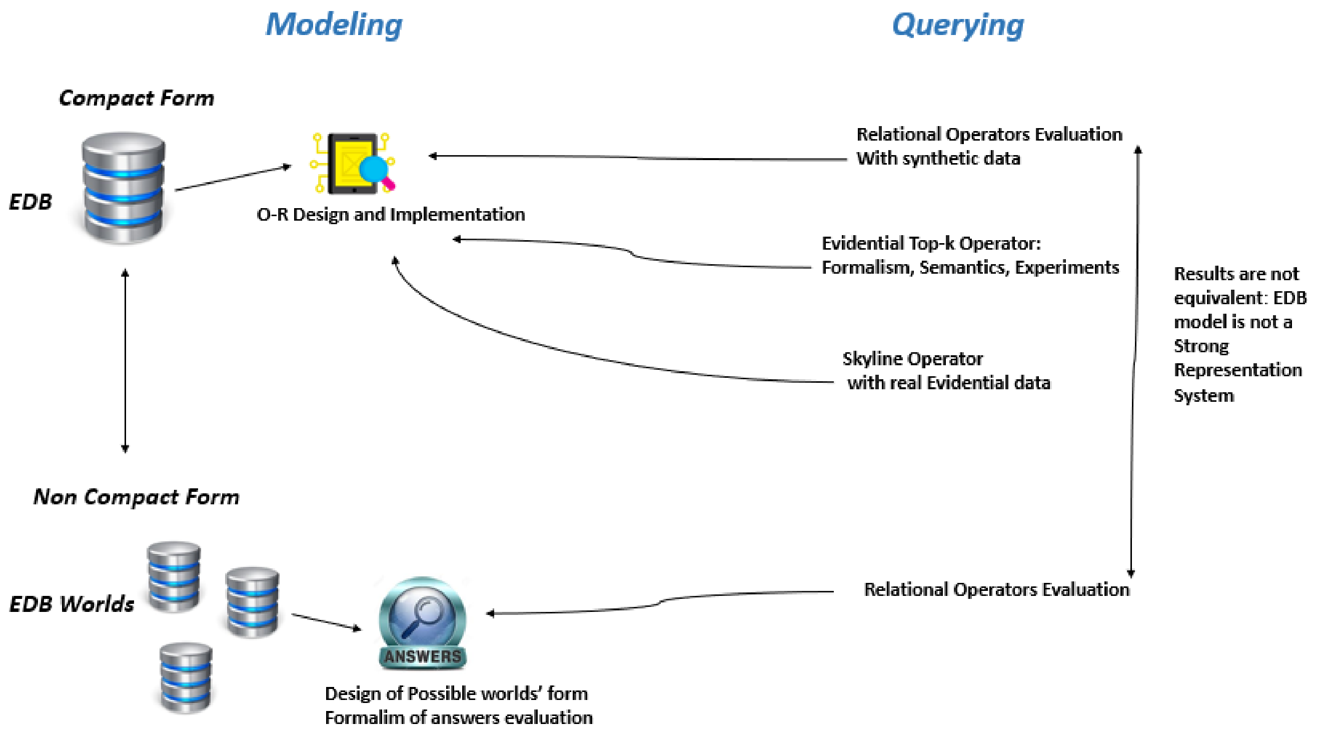

Modeling
Querying
Compact Form
EDB
O-R Design and Implementation
Relational Operators Evaluation
With synthetic data
Evidential Top-k Operator:
Formalism, Semantics, Experiments
Skyline Operator
with real Evidential data
Results are not equivalent: EDB model is not a Strong Representation System
Non Compact Form
EDB Worlds
ANSWERS
Design of Possible worlds' form
Formalim of answers evaluation
Relational Operators Evaluation

Figure 1: Modeling and Querying evidential databases EDB: Context and Contributions

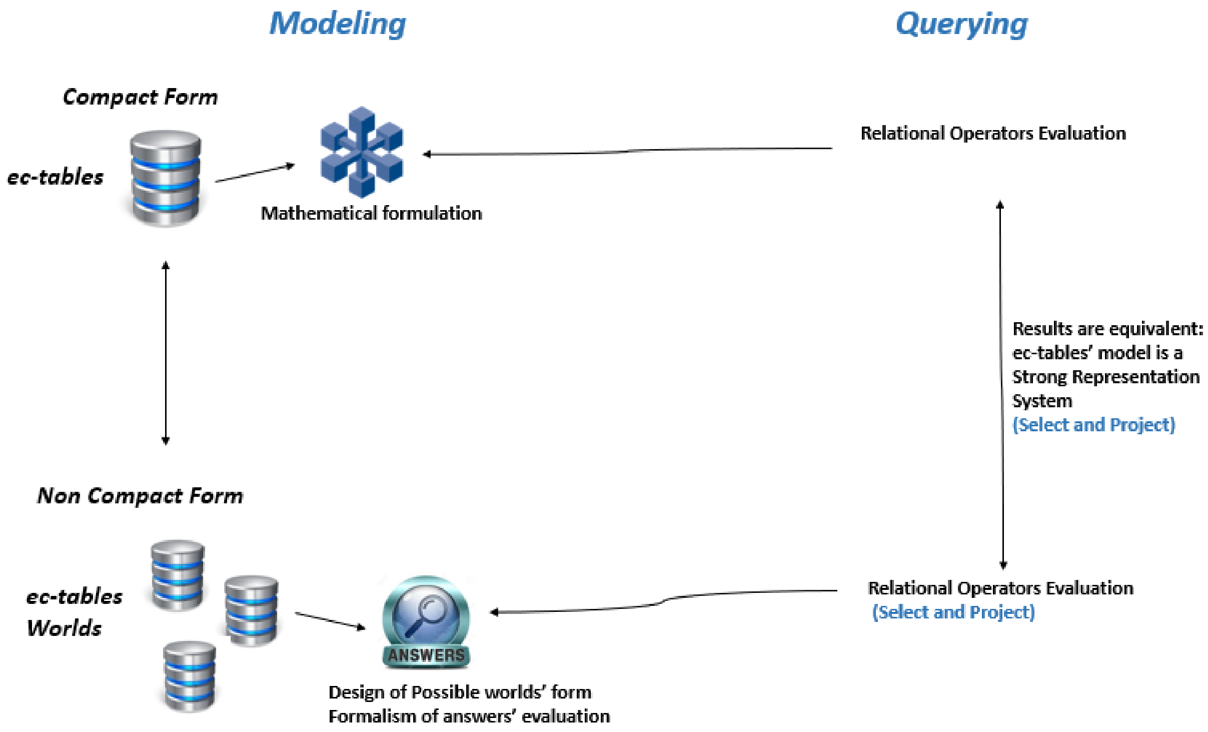


Figure 2: Modeling and Querying ec-tables: Context and contributions

The remainder of this thesis is organized as follows:

- In chapter 1, we present a survey about imperfection. First, we recall the different types of imperfect information and we give several illustrative examples. Then, we present the most used and known theories of imperfection: the probability theory, the possibility theory and the evidence theory. Finally, we present the imperfect database models, based on these theories, into their both representations (the compact and the possible worlds' forms).
- In chapter 2, we provide the background material related to the belief functions theory. This theory provides strong mathematical assets to model, manage and combine imperfect data coming from multiple sources. Thus, it models uncertain and imprecise information. Added to that, it handles partial and total ignorance. The evidence theory provides the appropriate tools for the decision making, when data are imperfect.
- In chapter 3, we recall the definition of an evidential database EDB on its compact form. Then, we present the object-relational implementation of this database model. Querying data comes as a natural step after the storage. Thus, we first of all present the evidential relational queries; the evidential select and the evidential project are evaluated using the object-relational implementation. After that, we introduce the evidential top-k query that we apply over the compact form. Finally, we use the evidential skyline query to evaluate the real data extracted from the TripAdvisor platform.
- In chapter 4, we define representation systems as the way of modeling imperfect databases from the compact form to the possible worlds' form. In fact, we introduce the formalism of generating possible worlds from an evidential database EDB. We consider the tuple level uncertainty and the attribute level uncertainty which imply several intermediate forms: the imprecise possible worlds, the uncertain possible worlds and finally the possible worlds. We discuss the similarities between the probabilistic and the evidential models. The last part is an investigation about the nature of the evidential database model EDB. Indeed, we prove that this model is not a strong representation system.
- Chapter 5 is dedicated to modeling and querying evidential conditional databases. This model comes to rectify the issues discussed in the previous chapter. Indeed, the evidential conditional database ECD, named also ec-tables is a new model that handles evidential data by conditioning the tuples. The model handles also

the two levels of uncertainty and can transform any evidential database modeled as an EDB to an ECD. We prove that this model is very promising to be a strong representation system under relational operators.

Part I

Background Material

1
Imperfect Databases: General Context

Contents

1.1 Introduction **12**

1.2 Typology of Imperfection **12**

1.3 Theories of Imperfection **14**

1.3.1 The Theory of Probabilities 15

1.3.2 The Possibility Theory 15

1.3.3 The Evidence Theory 17

1.4 Imperfect Databases **17**

1.4.1 Probabilistic Databases 18

1.4.2 Possibilistic Databases 21

1.4.3 Evidential Databases 23

1.5 Conclusion **27**

Summary

This chapter encompasses notions about the general context related to the subject of this thesis. It presents theories of imperfection and imperfect databases. First, we present the types of data imperfection associated with illustrative examples. Then, we present some imperfect theories, historically introduced to manage imperfect data according to their types. Finally, we give a brief review about the database models that store and query the imperfect data based on the presented theories.

1.1 Introduction

Imperfect information can be detected in various domains like meteorology, medicine, industry, aeronautics, etc. Having imperfect data in such domains may generate defective and unreliable results that may lead to the wrong decisions. This is why imperfect information should be efficiently treated. Indeed, theories of imperfection like the theory of probability (Laplace, 1812), the theory of possibility (Zadeh, 1978), the fuzzy sets theory (Zadeh, 1965), the rough sets theory (Pawlak, 1982) and the belief functions theory (Shafer, 1976; Dempster, 1967) were introduced. Based on the typology and the domain of the imperfect information, these theories offer several tools to model and manage such data (Dubois et al., 2004; Hong and May, 2004; Willink, 2006; Pfeiffer, 2013; Yang and Kim, 2006; Yang et al., 2011; Zhang et al., 2001). Hence, the emergence of several database models like probabilistic, possibilistic, fuzzy and evidential databases. These database models can store and query the imperfect data in order to provide reliable results. Any imperfect database model can be modeled in two equivalent representations: the first is compact and the second is a distribution of possible worlds.

This chapter presents briefly the general context related to this thesis dissertation. Indeed, it provides a typology of the imperfect information. Then, it introduces theories and database models of imperfection.

1.2 Typology of Imperfection

Imperfect data is so obvious in several real life applications (like medicine, genetics, industry, meteorology, etc). Types of imperfection can vary depending on the nature of fields, on the reliability of sources, etc. In fact, several types of imperfect data can be detected and they are partitioned into three major categories (Smets, 1996):

1. **Uncertainty**: is the lack of certainty or a state of a limited knowledge. It is a property related to the veracity of the information.
2. **Imprecision**: is the lack of precision or exactness. It is a property related to the content of the information.
3. **Inconsistency**: is the lack of coherence. It is a property related to the content of the information.

Example 1 *Suppose that we have four sources, each source gave an information as follows:*

- *John has at least two children and I am not sure about it.*

  $\Longrightarrow$ *This information is imprecise and uncertain.*

- *John has at least two children and I am sure about it.*

  $\Longrightarrow$ *This information is imprecise and certain.*

- *John has three children and I am not sure about it.*

  $\Longrightarrow$ *This information is precise and uncertain.*

- *John has three children and I am sure about it.*

  $\Longrightarrow$ *This information is precise and certain.*

1. **Uncertainty** can be:

   - **Objective**: based on random events and applies the chance concepts.

     Example 2 *Suppose having two same 6-sided dice. The chance of one dice being a particular number is* $\frac{1}{6}$ *. The chance of two dice being a same number is* $\frac{1}{6} * \frac{1}{6} = \frac{1}{36}$.

   - **Subjective**: based on experts' opinions about the truthfulness of a statement.

     For example, John is learning how to ride a bike, there are many chances that he falls.

2. **Imprecision** can be itself divided into two parts:

   2.1. **Imprecision without error**:

   - **Ambiguity**: is an indefinite information.
     As an example, the food is *hot.*
     $\Longrightarrow$ The food is *spicy* or the food's temperature is *high.*
   - **Vagueness**: is an unclear or not explicitly expressed information.
     For example, John is *young.*
     $\Longrightarrow$ Young represents a vague information.
   - **Incompleteness**: An information lacking some part.
     As an example, John does not have the last score of the game.

   2.2. **Imprecision with error**:

   - **Invalidity**: is a deficient information.
     For an example, John is *2000* years old.
   - **Bias**: a misrepresented information, i.e, the set of individuals does not represent the population.
     For example, 250 ml of the persons living on earth have severe obesity.

3. **Inconsistency** can be:

- **Conflict**: is a state of disagreement about one idea or information.

  For example, One doctor is sure that the disease of the patient is anemia and the second doctor is sure that the same patient does not have anemia.

- **confusion**:. is an unclear and incomprehensible information.

  For example, I am waiting for a bill of 150 euros and i get one with 300 euros.

Table 1.1 summarizes the types of imperfection.

Uncertainty	**Imprecision**	**Inconsistency**
objective subjective	ambiguity vagueness incompleteness invalidity bias	conflict confusion

Table 1.1: Recapitulation of imperfection types

Dealing with imperfect data requires the use of mathematical theories that can model and explicitly represent these imperfect data. Uncertainty and imprecision are modeled via these theories, except for inconsistency that can only be detected after a combination[1].

1.3 Theories of Imperfection

Theories of imperfection were introduced to represent and manage the imperfect data, depending on their types. The most known and classical theory is the probability theory (Laplace, 1812). It was introduced to handle the uncertain data. Then, other non-classical theories were proposed to offer different models like the possibility theory (Zadeh, 1978) that represents the uncertain and the imprecise data, the fuzzy sets theory (Zadeh, 1965) that represents the ambiguous data, the rough sets theory (Pawlak, 1982) that represents uncertain and vague data and the belief functions theory (Shafer, 1976; Dempster, 1967) that represents uncertainty, imprecision and ignorance. In this section we give a brief refresher about three well known theories: the probability theory, the possibility theory and the evidence theory.

[1] Combining two information coming from two different sources (two doctors for example) about one variable can cause the conflict or a confusion

1.3.1 The Theory of Probabilities

The probability theory has it roots in attempts to solve problems related to the games of chance in the sixteenth century. It is the oldest and the most acknowledged among theories of imperfection. The theory of probabilities was fully defined and considered in the nineteenth century by Pierre Laplace (Laplace, 1812).

Defintion 1 *Let Ω be a finite and countable set, called sample space, that relates to the set of all possible outcomes. Let x be an element $\in \Omega$, for each x a probability distribution is attached and that satisfies the following properties:*

$$p(x) \in [0;1], \forall x \in \Omega \tag{1.1}$$

$$\sum_{x \in \Omega} p(x) = 1 \tag{1.2}$$

An event is defined as any subset E of the sample space Ω with $E \subseteq \Omega$. The probability of event E is defined such as:

$$P(E) = \sum_{x \in E} p(x) \tag{1.3}$$

Properties

- $P(\emptyset) = 0$ and $P(\Omega) = 1$
- $\forall\ E_1, E_2 \subset \Omega$, if $E_1 \cap E_2 = \emptyset$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ (Additivity)
- $\forall\ E \subset \Omega$, $P(E) = 1 - P(\overline{E})$ (Duality)

1.3.2 The Possibility Theory

The possibility theory was introduced by (Zadeh, 1978) and developed by Dubois and Prade (Dubois and Prade, 1987). It handles uncertainty in a qualitative way based on the min/max algebra. This theory offers double representation/interpretation: the numerical representation and the ordinal representation.

Defintion 2 *Let Ω be the universe of discourse and x be a variable that takes its values from Ω. A possibility distribution π_x is attached to the variable x such that:*

$$\pi_x : \Omega \longmapsto L$$

L is a scale of a totally ordered plausibilities ([0,1], finite,...)

An event E is defined as a subset of Ω. Each event associates two measures, the possibility degree $\Pi(E)$ and the necessity degree $N(E)$.

The possibility and the necessity are dual measures such that:

$$N(E) = 1 - \Pi(\overline{E})$$

The possibility theory provides two interpretations as detailed in Figure 1.1:

- **Numerical interpretation**: When degrees of values reflect a specific sens.
  - $\Pi(E) = 1$
    $\Longrightarrow$ The event E is completely possible.
  - $\Pi(E) = 0$
    $\Longrightarrow$ The event E is completely impossible.
- **Ordinal interpretation**: When values reflect an ordering between the different states of the world.
  - $\Pi(E_1) > \Pi(E_2)$
    $\Longrightarrow$ E_1 is more possible/more preferred than E_2.

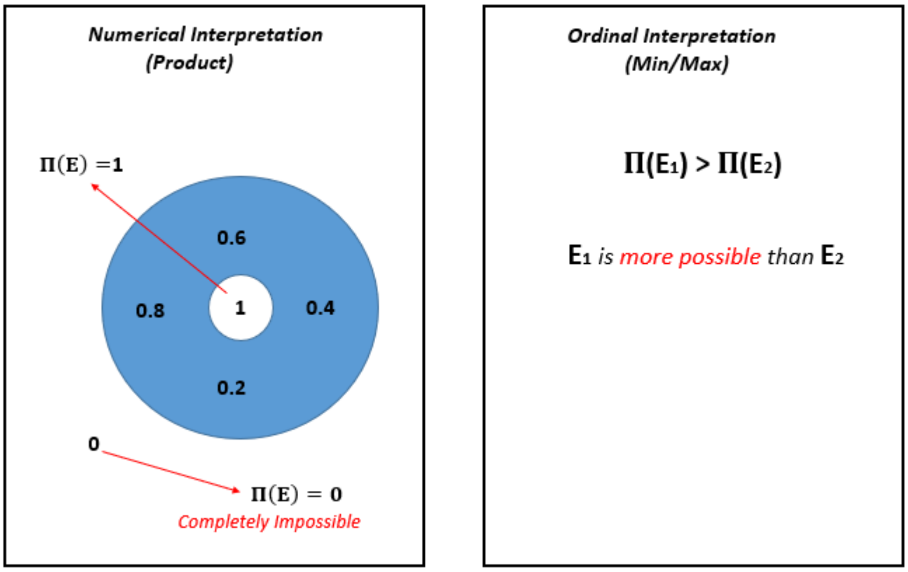


Figure 1.1: Numerical and Ordinal interpretations of the Possibility Theory

1.3.3 The Evidence Theory

The theory of evidence[2] also called *the theory of belief functions or the Dempster-Shafer theory*, was introduced by Dempster (Dempster, 1967) in the context of statistical inference, mathematically formalized by Shafer (Shafer, 1976) and later popularized by Smets (Smets, 1998a) with the transferable belief model (TBM). The theory of belief functions models and manages imperfect information through two levels: the credal and the pignistic levels. In the credal level, beliefs are quantified, modeled and can be aggregated. In the pignistic level, decisions are made considering the learned beliefs.

Defintion 3 *Let* Θ *be a frame of discernment (the domain) containing all entities of a given situation* $\Theta = \{\theta_1, \theta_2, .., \theta_n\}$ *and* 2^Θ *is the set of all subsets of* Θ. *A basic belief mass is the degree of belief of an hypothesis* A *from the set* 2^Θ *such that:*

$$\sum_{A \subseteq \Theta} m^\Theta(A) = 1$$

The mass m *is the truthfulness degree about hypothesis* A.

1.4 Imperfect Databases

Classical relational databases deal with certain data, contrary to imperfect databases that store sets of imperfect data. Indeed, several databases like probabilistic databases (Cavallo and Pittarelli, 1987), possiblistic databases (Bosc and Pivert, 2005), evidential databases (Bell et al., 1996; Choenni et al., 2006; Lee, 1992a; Lee, 1992b), etc, were introduced for the aim of storage, querying, datamining and reporting.

An imperfect database can be modeled as a distribution of candidate representative databases throughout treating the imperfection. Thus, two forms are studied in the literature:

- *The compact form* where attributes' values are distributions.
- *The possible worlds' form* where the database is a distribution of candidate databases.

When a result of a query processed over a compact database is the same when that query is applied over its possible worlds, we say that the result is reliable and that the model is a *strong representation system.*

[2] More details about this theory are explicitly presented in chapter 2

1.4.1 Probabilistic Databases

Probabilistic databases were introduced by (Cavallo and Pittarelli, 1987), then popularized by (Barbara et al., 1992). Later on, (Fuhr, 1990; Fuhr, 1992a; Fuhr, 1992b; Fuhr, 1993) defined a more elaborated model of the probabilistic databases.

1. A probabilistic database, on its *compact form*, is an imperfect database that includes N objects and D attributes where the imperfect values are expressed via the probability theory (Cavallo and Pittarelli, 1987; Barbara et al., 1992; Fuhr, 1990; Fuhr, 1992a; Fuhr, 1992b; Fuhr, 1993).

2. A probabilistic database, on its *non compact form*, is a set of *possible worlds* associated with degrees of probability. The probabilistic possible worlds represent the set of all possible states generated from this database, where each probabilistic possible world is a candidate to represent the compact probabilistic database (Abiteboul et al., 1995b; Suciu et al., 2011).

Figure 1.2 is an illustration of both forms of a probabilistic database.

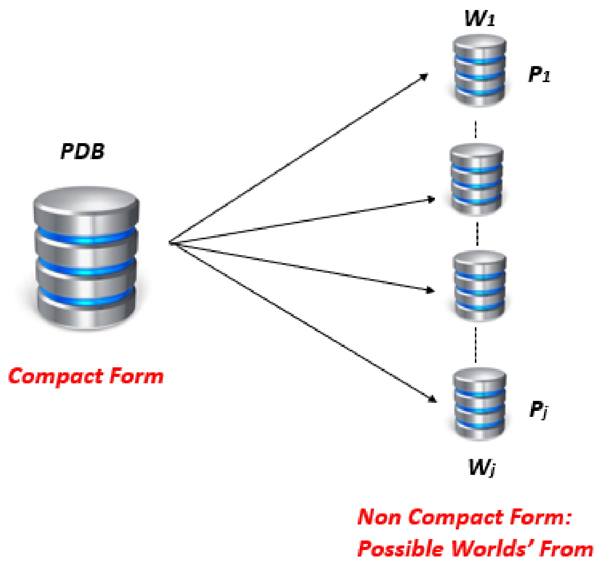


Figure 1.2: A Probabilistic Database

A probabilistic database PDB includes two levels of uncertainty:

- *The tuple level uncertainty*: a degree of probability P is associated to each tuple, $P \in [0, 1]$ where 0 represents a non existent tuple and 1 represents a certain one.

- *The attribute level uncertainty*: a degree of probability is assigned to hypotheses that reflect their veracity degree in the attribute.

Example 3 *As pieces of information can be imperfect in the medical domain, let us assume that a doctor gave his diagnoses about three patients as shown in the probabilistic database in Table 1.2. He expresses his beliefs about each hypothesis using the probability degrees. For instance, the tuple w.r.t $ID\#1$ exists in the PDB with a probability of 0.3. For the tuple $ID\#3$, the disease is a dataset of values, i.e., Asthma with a probability degree of 0.5 or Anemia with a probability degree of 0.5.*

ID	**Disease**	**P**
1	Flu	0.3
2	Anemia	0.2
3	Asthma 0.5 Anemia 0.5	0.5

Table 1.2: An example of a probabilistic table PDB

Possible Worlds	**P**
W_1={(1,Flu);(2,Anemia);(3,Asthma)}	= 0.3 * 0.2 * 0.25 = 0.015
W_2={(1,Flu);(2,Anemia);(3,Anemia)}	= 0.3 * 0.2 * 0.25 = 0.015
W_3={(1,Flu);(2,Anemia)}	=0.3 * 0.2 * (1 - 0.5) = 0.03
W_4={(1,Flu);(3,Asthma)}	= 0.3 * (1 - 0.2) * 0.25= 0.06
W_5={(1,Flu);(3,Anemia)}	= 0.3 * (1 - 0.2) * 0.25= 0.06
W_6={(2,Anemia);(3,Anemia)}	= (1 - 0.3) * 0.2 * 0.25= 0.035
W_7={(2,Anemia);(3,Asthma)}	= (1 - 0.3) * 0.2 * 0.25= 0.035
W_8={(1,Flu)}	= 0.3*(1-0.2)*(1-0.5) = 0.12
W_9={(2,Anemia)}	= (1-0.3)*0.2*(1-0.5)=0.07
W_{10}={(3,Anemia)}	= (1-0.3)*(1-0.2)*0.25=0.14
W_{11}={(3,Asthma)}	= (1-0.3)*(1-0.2)*0.25=0.14
W_{12}={∅}	= (1-0.3)*(1-0.2)*(1-0.5)= 0.28

Table 1.3: Possible worlds of the probabilistic table PDB

*In Table 1.3, we provide the different possible worlds that correspond to PDB of Table 1.2. For instance, the world W_1 that includes {(1,Flu);(2,Anemia);(3,Asthma)} is a possible candidate with a probability of 0.015 (=0.3 * 0.2 * 0.25).*

Applying a query Q over a probabilistic database returns a set of tuple-probability pairs $\{(t_1, p_1), (t_2, p_2), ...\}$.

To evaluate the querying methods of the probabilistic database model, three steps are required:

1. Querying the probabilistic compact form: It gives a compact result, which is a set of tuples and their probabilities responding to the condition of the query.
2. Querying the possible worlds' form of the probabilistic database: It means applying the query Q on each possible world. Each possible world gives one or more possible answers. These answers are the set of tuples and with their probabilities constitute the reply to Q.
3. Comparing results: It is about checking if the results derived from the compact form are equivalent or not to the results derived from the possible worlds' form. If these results are equivalent the database model is said a strong representation system.

Figure 1.3 summarizes the querying and the evaluation process of the probabilistic database model.

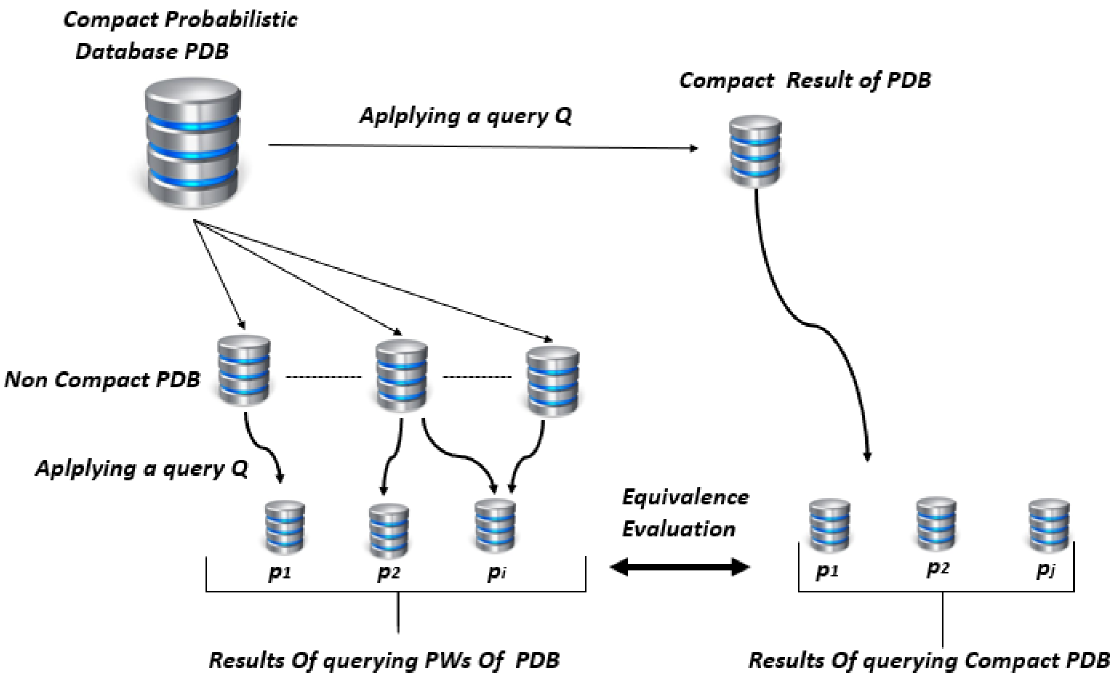


Figure 1.3: Querying process of the Probabilistic Databases (where PWs refers to possible worlds)

1.4.2 Possibilistic Databases

Possibilistic databases were introduced by Prade and Testmale (Prade and Testemale, 1984), then were popularized by Bosc and Pivert (Bosc and Pivert, 2005).

1. A possibilistic database, on its compact form, has N objects, D attributes in which imperfect values are expressed via the possibility theory (Bosc et al., 2003; Bosc and Pivert, 2005; Bosc and Pivert, 2010).

2. A possibilistic database, on its non compact form, defined as a set of possible worlds that associate degrees of possibilities. Each possibilistic world is a candidate to represent the real world described by means of the compact database (Bosc and Pivert, 2010).

Figure 1.4 is an illustration of both forms of a possibilistic database.

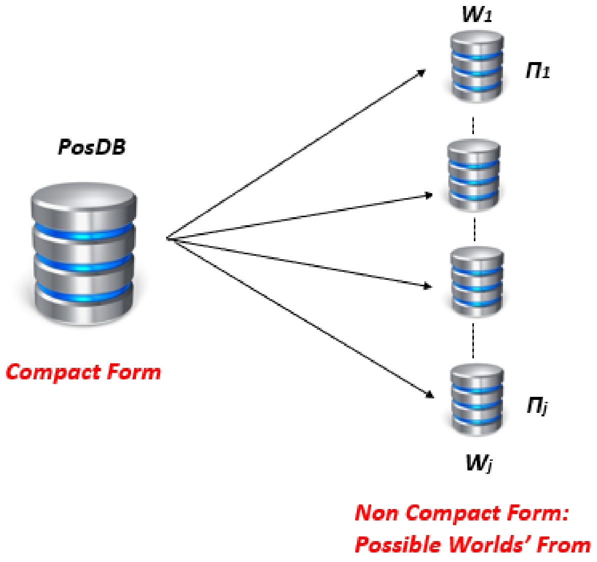


Figure 1.4: A Possibilistic Database

A possibilistic database $PosDB$ can include two levels of uncertainty:

- The tuple level uncertainty: a degree of possibility Π or necessity N is assigned to each tuple.

- The attribute level uncertainty: a degree of possibility is assigned to hypotheses in attributes. When the possibility of one hypothesis is equal to 1 then it is completely possible and when the degree is equal to 0 then the hypothesis is completely impossible.

Example 4 *Let us suppose now that some diagnoses given by a doctor are represented in the possibilistic database of Table 1.4. This table includes two attributes where attribute Disease contains imperfect data represented via the possibility theory. For example, the first patient can either have the Flu with a possibility degree* $\Pi(Flu) = 1$ *or Asthma with a possibility degree* $\Pi(Asthma) = 0.7$. *Possible worlds of Table 1.4 are shown in Table 1.5.*

ID	**Disease**
1	1/Flu + 0.7/Asthma
2	1/Cancer + 0.3/Anemia
3	Asthma
4	Anemia

Table 1.4: An example of a possibilistic database

W_1	W_2	W_3	W_4
$(1, 1/Flu\ )$	$(1, 1/Flu)$	$(1, 0.7/Asthma)$	$(1, 0.7/Asthma)$
$(2, 1/Cancer)$	$(2, 0.3/Anemia)$	$(2, 1/Cancer)$	$(2, 0.3/Anemia)$
$(3, 1/Asthma)$	$(3, 1/Asthma)$	$(3, 1/Asthma)$	$(3, 1/Asthma)$
$(4, 1/Anemia)$	$(4, 1/Anemia)$	$(4, 1/Anemia)$	$(4, 1/Anemia)$
min(1,1,1,1)=1	min(1,0.3,1,1)=0.3	min(0.7,1,1,1)=0.7	min(0.7,0.3,1,1)=0.3

Table 1.5: Possible worlds of the possibilistic database

In Table 1.5, we give the different possible worlds that correspond to PosDB of Table 1.4. For example, the possible world W_4 *that includes* $\{$*(1,0.7/Asthma), (2,0.3/Anemia), (3,1/Asthma), (4,1/Anemia)*$\}$ *has a possibility degree of 0.3 (=min(0.7, 0.3, 1, 1)).*

Applying a query Q over a possibilistic database returns a set of tuple-possibility pairs $\{(t_1, \Pi_1), (t_2, \Pi_2), ...\}$

To evaluate the querying methods of the possibilistic database model, three steps are needed:

- Querying the compact possibilistic database: It provides a compact result, which is a set of tuples and their degrees of possibility responding to query Q.
- Querying the possible worlds of the possibilistic database: Each possible world is interrogated with Q independently of other possible worlds. The set of the derived results from each world constitutes the set of answers, where each answer associates a possibility degree.

- Comparing results: Answers derived form the compact form and answers derived from the possible worlds' form are evaluated. Their equivalence means that the possibilistic database model is a strong representation system.

Figure 1.5 clarifies the querying and the evaluation process of the possibilistic database model.

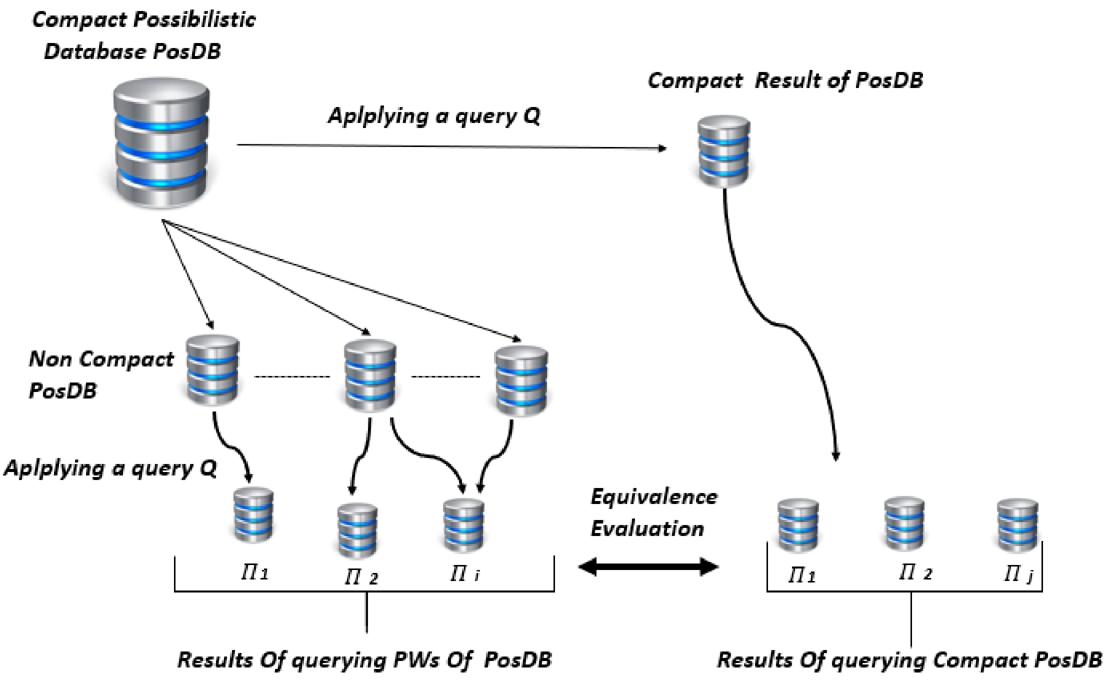


Figure 1.5: Querying process of the Possibilistic Databases

1.4.3 Evidential Databases

Evidential databases (EDBs) were introduced by Lee et al., (Lee, 1992b; Lee, 1992a; Bell et al., 1996). It is an imperfect database model that permits to represent certain and uncertain data using the theory of belief functions.

1. An evidential database, on its compact form, is a set of N tuples and D attributes where imperfect information is expressed by means of the evidence theory (Lee, 1992b; Lee, 1992a; Bell et al., 1996).

2. An evidential database, on its non compact form, is a set of possible worlds with

degrees of belief and plausibility[3] where each world is a candidate to represent the compact model (Bousnina et al., 2015).

Figure 1.6 is an illustration of both forms of an evidential database.

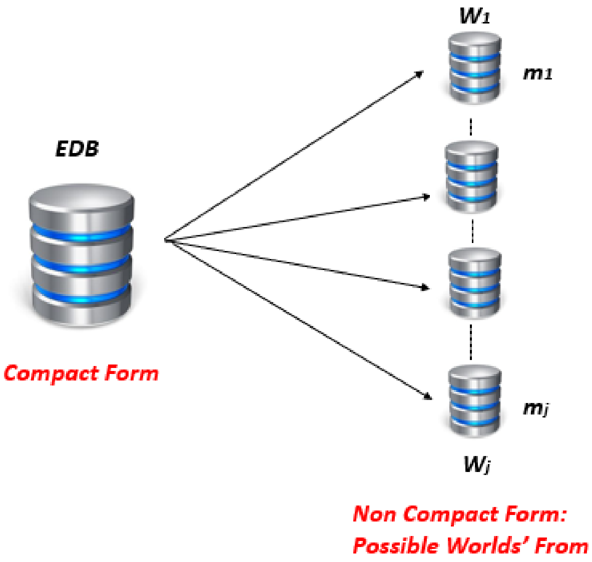


Figure 1.6: An Evidential Database

An evidential database has two levels of uncertainty:

- *The tuple level uncertainty*: an interval of minimum and maximum degrees of credibility are assigned to each tuple. This interval, named *confidence level*, reflects degrees of faith about the existence of the tuple.
- *The attribute level uncertainty*: a mass function is assigned to each hypothesis in attributes. It reflects the degree of truthfulness about the hypothesis.

Example 5 *We give an example of a doctor that expresses his medical diagnoses about two patients by giving multiple hypotheses and their assignments. Its diagnoses are modeled using the belief functions' theory and stored in the following compact evidential Table 1.6. The latter includes three attributes* ID, Disease, Symptom*. The domain of attribute* Disease *is* $\Theta_{DE} = \{Diabetes, Anemia, Stroke, Asthma\}$ *and the domain of attribute* Symptom $\Theta_{SY} = \{Fatigue, Nausea, Vertigo\}$*. For instance, the doctor believes that the second patient ID#2 has either the diabetes with a degree of 0.1 or a stroke with a degree of 0.9. The degree of belief about each hypothesis given by the doctor is called a mass (m). Possible worlds of EDB are detailed in Table 1.7.*

ID	Disease	Symptom
1	*Diabetes*	$Fatigue$ 0.4 $\{Fatigue, Nausea\}$ 0.6
2	*Diabetes* 0.1 *Stroke* 0.9	$Vertigo$

Table 1.6: An evidential table EDB

Possible Worlds	Confidence Level
W_1={(1,Diabetes,Fatigue);(2,Diabetes,Vertigo)}	[0.04;0,1]
W_2={(1,Diabetes,Fatigue);(2,Stroke,Vertigo)}	[0.36;0,9]
W_3={(1,Diabetes,Nausea);(2,Diabetes,Vertigo)}	[0;0,06]
W_4={(1,Diabetes,Nausea);(2,Stroke,Vertigo)}	[0;0,54]

Table 1.7: Possible worlds of the evidential table EDB

In Table 1.7 we provide the different possible worlds issued from Table 1.6. As an example, the possible world W_1 involves W_1={(1,Diabetes,Fatigue);(2,Diabetes, Vertigo)} with a confidence interval of [0.04;0,1][4].

Applying a query Q over an evidential database returns a set of tuple-interval pairs $\{(t_1, CL_1), (t_2, CL_2), ...\}$

To evaluate the querying methods of the evidential database model, three steps are needed:

1. Querying the compact evidential database: It provides a compact result, which is a set of tuples and their confidence levels responding to query Q. A confidence level is an interval of belief and plausibility.

2. Querying the possible worlds of the evidential database: Each queried possible world generates a possible answer. The set of answers involves a set of tuples and their confidence levels.

3. Comparing results: Answers derived from the compact form and answers derived from the possible worlds' form are compared. If they are equivalent the evidential database model is a strong representation system, otherwise it is not.

Figure 1.7 clarifies the querying and the evaluation process of the evidential database model.

[3] Notions of belief and plausibility are defined in Chapter 2.

[4] Details of confidence intervals' computations will be explained in Chapter 3.

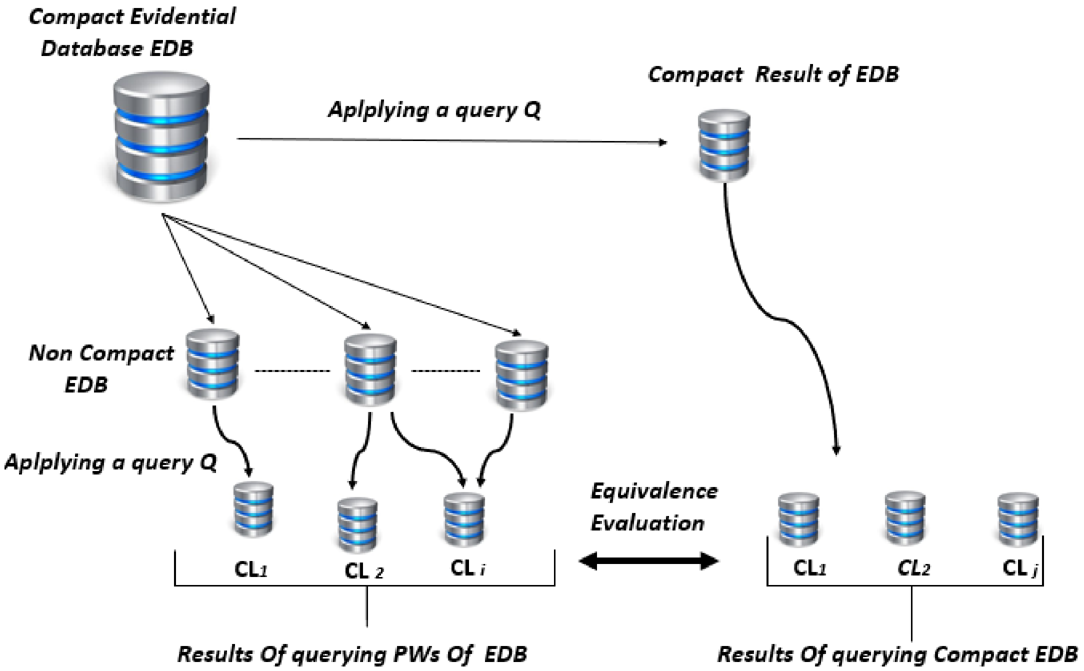


Figure 1.7: Querying process of the Evidential Databases

To the best of our knowledge, no previous work focused on representations of evidential databases and the querying evaluation process. The compact form is the only conceivable form in practice but modeling the non compact from remains fundamental to validate the querying methods of the compact representation. In this thesis, we investigate the evidential database representations issue.

1.5 Conclusion

Imperfection is widespread in several domains, hence the emergence of theories that represent and handle the incomplete data. These theories like the probability theory, the possibility theory, the belief functions theory, etc., came as powerful solutions to effectively manage multiple types of imperfection. Their appearance throughout the past years engendered the appearance of the database models of imperfection. These database models are based on these theories and they came as a natural step to store and query the imperfect data.

In this chapter, we presented a review about typologies of imperfection. Then, we briefly presented three theories of imperfection (probability, possibility and belief functions) and the database models that support these models. The evidence theory and the evidential databases represent the core of this dissertation, they are detailed in the coming chapters.

2
Theory of Evidence

Contents

2.1 Introduction . . . 30
2.2 Basic Concepts . . . 30
2.2.1 Frame of discernment . . . 30
2.2.2 Basic belief assignment . . . 31
2.2.3 Body of evidence . . . 32
2.2.4 Belief and plausibility functions . . . 32
2.2.5 Special belief functions . . . 38
2.3 Combination Rules . . . 42
2.3.1 Dempster's rule of combination . . . 43
2.3.2 Conjunctive rule of combination . . . 44
2.3.3 Disjunctive rule of combination . . . 45
2.4 Discounting . . . 46
2.5 Compatible Frames of Discernments . . . 47
2.5.1 Vacuous Extension . . . 47
2.5.2 Refinement and Coarsening . . . 49
2.6 Cognitive and Evidential Independences . . . 50
2.6.1 Cognitive Independence . . . 50
2.6.2 Evidential Independence . . . 51
2.7 Decision Making . . . 52
2.7.1 Pignistic Probability . . . 52
2.7.2 Maximum of Credibility . . . 53
2.7.3 Maximum of Plausibility . . . 54
2.8 Conclusion . . . 54

Summary

In this chapter, basic concepts of the evidence theory are presented. Other tools like combination rules, vacuous extension, cognitive and evidential independences are illustrated to manage imperfect data using this theory.

2.1 Introduction

The evidence theory is used in lots of domains like chemistry (Samet and Dao, 2015), sensor detection and data fusion (Ayoun and Smets, 2001), etc. In fact, this theory (Dempster, 1967; Shafer, 1976; Smets, 1998a) models and manages imperfect information through an explicit representation of uncertainty, imprecision and ignorance. It extends the probability theory by making a difference between equi-probability and ignorance. For example, an expert and a newbie will give their opinions about racing horses $\{h_1, h_2, h_3\}$. The expert thinks that all the three horses have the same chance to win the race. The newbie does not have any idea about racing and horses.

- In probability theory their opinions are modeled such that:
  - Expert: $P(h_1)$= 1/3; $P(h_2)$= 1/3; $P(h_3)$= 1/3
  - Newbie: $P(h_1)$= 1/3; $P(h_2)$= 1/3; $P(h_3)$= 1/3
- In evidence theory their opinions are modeled such that:
  - Expert: $m(h_1)$= 1/3; $m(h_2)$= 1/3; $m(h_3)$= 1/3 $\Rightarrow$ Equi-probability.
  - Newbie: $m(\{h_1, h_2, h_3\})$=1 $\Rightarrow$ Ignorance.

  Note that m can be interpreted as a subjective probability.

The belief functions theory manages the imperfect information via two levels. The first is the credal level where beliefs are quantified, represented and combined. The second is the pignistic level where decisions are made (See Figure 2.1). Indeed, the evidence theory provides the convenient tools to model, manage, aggregate and decide when having imperfect information.

This chapter is dedicated to the basic notions of the belief functions theory that will be used in the next chapters.

2.2 Basic Concepts

In this section, basic concepts of evidence theory as *the frame of discernment, the basic belief assignment, the body of evidence, belief and plausibility functions* are detailed.

2.2.1 Frame of discernment

In the theory of belief functions, a frame of discernment (or universe of discourse) is a set that contains all hypotheses of a given situation. Thus, $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ is a finite, non empty and exhaustive set of n elementary and mutually exclusive hypotheses related to a given problem.

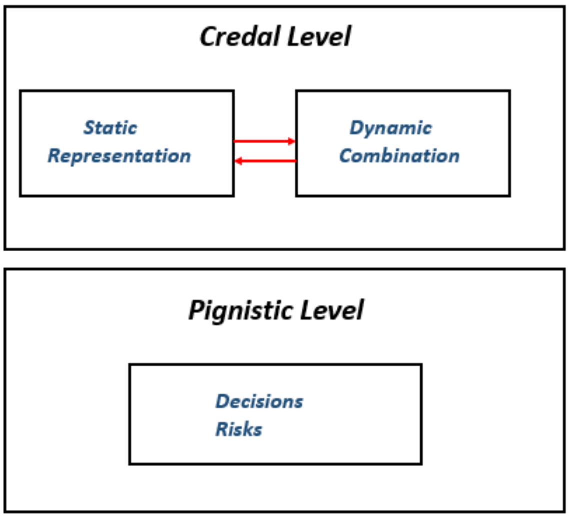


Figure 2.1: Credal and Pignistic Levels of the Evidence theory

The *power set* 2^{Θ} is the set of all subsets of Θ. It includes hypotheses of Θ and all disjunctions of Θ. It is defined as follows:

$$2^{\Theta} = \{A : A \subseteq \Theta\} = \{\varnothing, \theta_1, \theta_2, ..., \theta_n, \{\theta_1, \theta_2\}, .., \{\theta_1, \theta_2, ..., \theta_n\}\} \tag{2.1}$$

Each element of 2^{Θ} is a *proposition* (an *event*, an *alternative* or a *solution*).

Example 6 *Let us consider a medical problem where a doctor gives his propositions about the possible diseases of some patients. The possible diseases in our example are either Flu, Asthma or Anemia. Thus, the frame of discernment is defined as follows:*
$\Theta_{DE} = \{Flu,\ Asthma,\ Anemia\}$.

The different possible combinations for Θ_{DE} are defined in the following set:
$2^{\Theta_{DE}} = \{\emptyset,\ \{Flu\},\ \{Asthma\},\ \{Anemia\},\ \{Flu,\ Asthma\},\ \{Flu,\ Anemia\},$
$\{Asthma,\ Anemia\}, \{Flu,\ Asthma,\ Anemia\}\}$.

This application will be used through out this manuscript.

2.2.2 Basic belief assignment

A *basic belief assignment* (*bba*), noted m^{Θ}, 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:

$$\sum_{A \subseteq \Theta} m^{\Theta}(A) = 1 \tag{2.2}$$

The amount $m^{\Theta}(A)$ is called *basic belief mass* (bbm) and *mass* for short. It represents the degree of faith on the truth of hypothesis A. The mass $m^{\Theta}(A)$ is the degree of belief on A that is not distributed on its subsets.

Example 7 *The doctor believes that this patient suffers from Flu with a degree of belief equal to* 0.7 *or from Flu or Anemia with a degree of belief equal to* 0.3. *The following basic belief assignment illustrates the beliefs of the doctor:*

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

.

2.2.3 Body of evidence

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

$$m^{\Theta}(A) > 0 \tag{2.3}$$

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:

$$\varphi = \bigcup_{A \in F : m^{\Theta}(A) > 0} A \tag{2.4}$$

Example 8 *Let us consider the focal elements* $\{Flu\}$ *and* $\{Flu, Anemia\}$ *of* $m^{\Theta_{DE}}$ *such that:*

$F = \{\{Flu\}, \{Flu, Anemia\}\}$*: is the set of focal elements.*
The couple $(F, m^{\Theta_{DE}})$*: is the body of evidence.*
$\varphi^{DE} = \{Flu\} \cup \{Flu,\ Anemia\} = \{Flu,\ Anemia\}$*: is the core.*

2.2.4 Belief and plausibility functions

The belief and the plausibility are functions derived from the basic belief mass (m). They reflect degrees of faith about some hypotheses. The belief function (bel) and the plausibility (pl) are considered as different expressions of the same information.

Belief function

The *belief function* denoted *bel*, is the degree of faith exactly committed to hypothesis A (Shafer, 1976). It assigns to each subset A of Θ the sum of the masses of belief committed exactly to each subset of A by m^{Θ}. The belief function is considered as the *minimal* degree of belief given to A and it is defined as follows:

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

$$bel(A) = \sum_{B,A\subseteq\Theta:B\subseteq A} m^{\Theta}(B) \tag{2.5}$$

Properties

- $bel(\emptyset) = 0$ and $bel(\Theta) = 1$ if $m(\emptyset) = 0$
- $bel(A) + bel(\overline{A}) \leq 1$ (Sub additivity)
- $A \subseteq B \Rightarrow bel(A) \leq bel(B)$ (Monotonicity)
- $bel(A \cup B) \geq bel(A) + bel(B)$

Note that $\overline{A}$ is the complement of A.

Example 9 *The belief function (bel) corresponding to the basic belief assignment of Example 7 is as follows:*

$bel(\emptyset) = 0$
$bel(\{Flu\}) = 0.7$
$bel(\{Anemia\}) = 0$
$bel(\{Asthma\}) = 0$
$bel(\{Flu, Anemia\}) = 0.3 + 0.7 = 1$
$bel(\{Flu, Asthma\}) = 0.7 + 0 = 0.7$
$bel(\{Anemia, Asthma\}) = 0$
$bel(\Theta_{DE}) = bel(\{Flu, Anemia, Asthma\}) = 0 + 0.3 + 0.7 = 1$

Plausibility function

The *plausibility function* denoted *pl* is equal to the sum of masses relative to subsets of hypothesis B that do not contradict hypothesis A. It contains those parts of belief that are compatible with A. The plausibility function is considered as the *maximal* amount of belief given to hypothesis A and it is defined as follows:

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

$$pl(A) = \sum_{B,A \subseteq \Theta : A \cap B \neq \emptyset} m^{\Theta}(B) \tag{2.6}$$

Properties

- $pl(A) = 1 - bel(\overline{A})$ (Duality)
- $pl(\Theta) = 1 \quad \text{and } pl(\emptyset) = 0 \text{ if } m(\emptyset) = 0$
- $pl(A \cup B) \leq pl(A) + pl(B)$
- $bel(A) \leq pl(A)$
- $A \subseteq B \Rightarrow Pl(A) \leq Pl(B)$ (Monotonicity)
- $Pl(A) + Pl(A) \geq 1$ (Sub additivity)

Example 10 *The plausibility function corresponding to the basic belief assignment of Example 7 is represented as follows:*

$pl(\emptyset) = 0$
$pl(\{Flu\}) = 0.7 + 0.3 = 1$
$pl(\{Anemia\}) = 0.3$
$pl(\{Asthma\}) = 0$
$pl(\{Flu, Anemia\}) = 0.3 + 0.7 = 1$
$pl(\{Flu, Asthma\}) = 1$
$pl(\{Anemia, Asthma\}) = 0 + 0.3 = 0.3$
$pl(\Theta_{DE}) = 1$

Duality between bel and pl

The duality between the belief and the plausibility functions is one the properties proposed by (Shafer, 1976).

Let A and $\overline{A}$ be two independent[1] hypotheses such that :

$$bel(\overline{A}) = 1 - pl(A) \quad \text{since} \quad pl(A) = 1 - bel(\overline{A}) \tag{2.7}$$

$$pl(\overline{x}) = 1 - bel(x) \quad \text{since} \quad bel(x) = 1 - pl(\overline{x}) \tag{2.8}$$

[1] Two hypotheses are said to be independent when the occurrence of one does not affect the assignment of occurrence of the other.

Extended belief and plausibility functions

In the standard evidence theory, definitions of the belief and plausibility functions can not handle comparisons like ($=$, $\neq$, $<$, $>$,$\leq$, $\geq$). That is why the definition of belief *bel* and plausibility *pl* functions were extended in (Lee, 1992a; Lee, 1992b; Bell et al., 1996) to deal with comparisons between two independent basic belief assignments (*bbas*).

Defintion 4 *(Equality) Let x and y be two random independent variables and their mass functions, m_x^{Θ}, m_y^{Θ}: $2^{\Theta} \longrightarrow [0,1]$, $A, B \subseteq \Theta$ are their respective focal elements. The equality between x and y are defined in the following way (Bell et al., 1996):*

$$bel(x = y) = \sum_{|A|=1} m_x^{\Theta}(A) * m_y^{\Theta}(A) \tag{2.9}$$

$$pl(x = y) = \sum_{A \cap B \neq \emptyset} m_x^{\Theta}(A) * m_y^{\Theta}(B) \tag{2.10}$$

*Note that * is the product operator.*

Example 11 *Let us consider information about Diseases of two patients given by a doctor. The latter believes that the first patient is either diabetic or has a stroke. He thinks that the second patient has either Anemia or Diabetes or stroke. His degrees of belief are represented as follows in Table 2.1:*

Patient	**Disease**
x	Diabetes 0.1 Stroke 0.9
y	Anmeia 0.3 {Diabetes, Stroke} 0.7

Table 2.1: A medical diagnosis for two patients

We want to compute the belief (bel) and the plausibility pl values for the proposition that both patients have the same disease.

Thus, according to definition 4:

- $bel(x = y) = 0$
- $pl(x = y) = 0.1 * 0.7 + 0.9 * 0.7 = 0.7$

Defintion 5 *(Inequality) Let x and y be two random independent variables and their mass functions, m_x^{Θ}, m_y^{Θ}: $2^{\Theta} \longrightarrow [0,1]$, $A, B \subseteq \Theta$ are their respective focal elements. The inequality between x and y is computed such that (Bell et al., 1996):*

$$bel(x \neq y) = \sum_{A \cap B = \emptyset} m_x^{\Theta}(A) * m_y^{\Theta}(B) \tag{2.11}$$

$$pl(x \neq y) = \sum_{A \subseteq \Theta} m_x^{\Theta}(A) * \sum_{B \subseteq \Theta \text{ and } [A=B \text{ implies } |A|>1]} m_y^{\Theta}(B) \tag{2.12}$$

Example 12 *Let us consider the same diagnosis of Table 2.1. We want now to compute the belief (bel) and the plausibility (pl) of both patients for the proposition that they have different diseases.*

Thus, according to definition 5:

- $bel(x \neq y) = 0.1 * 0.3 + 0.9 * 0.3 = 0.3$
- $pl(x \neq y) = 0.1 * 0.3 + 0.9 * 0.3 + 0.1 * 0.7 + 0.9 * 0.7 = 1$

Defintion 6 *(Inferior Inequality) Let x and y be two random independent variables and their mass functions, m_x^{Θ}, m_y^{Θ}: $2^{\Theta} \longrightarrow [0,1]$, $A, B \subseteq \Theta$ are their respective focal elements. The inferior inequality can be measured as follows (Bell et al., 1996):*

$$bel(x < y) = \sum_{A \subseteq \Theta} m_x^{\Theta}(A) * \sum_{B \subseteq \Theta \wedge A <^{\forall} B} m_y^{\Theta}(B) \tag{2.13}$$

where $A <^{\forall} B$ means that $a < b$ for all $a \in A$, $b \in B$.

$$pl(x < y) = \sum_{A \subseteq \Theta} m_x^{\Theta}(A) * \sum_{B \subseteq \Theta \wedge A <^{\exists} B} m_y^{\Theta}(B) \tag{2.14}$$

where $A <^{\exists} B$ means for every $a \in A$ there exists $b \in B$ such that $a < b$.

Defintion 7 *(Inferior or Equal Inequality) Let x and y be two random independent variables and their mass functions, m_x^{Θ}, m_y^{Θ}: $2^{\Theta} \longrightarrow [0,1]$, $A, B \subseteq \Theta$ are their respective focal elements. $x \leq y$ is computed as follows (Bell et al., 1996):*

$$bel(x \leq y) = \sum_{A \subseteq \Theta} m_x^{\Theta}(A) * \sum_{B \subseteq \Theta \wedge A \leq^{\forall} B} m_y^{\Theta}(B) \tag{2.15}$$

where $A \leq^{\forall} B$ means that $a \leq b$ for all $a \in A$, $b \in B$.

$$pl(x \leq y) = \sum_{A \subseteq \Theta} m_x^{\Theta}(A) * \sum_{B \subseteq \Theta \wedge A \leq^{\exists} B} m_y^{\Theta}(B) \tag{2.16}$$

where $A \leq^{\exists} B$ means for every $a \in A$ there exists $b \in B$ such that $a \leq b$.

Example 13 *The doctor orders diseases from the less to the most serious. Indeed, he says that Stroke is more serious than Diabetes which is it self more serious than Anemia. The descendant order is represented such that:*

Anemia < Diabetes < Stroke.

We want now to compare values of attribute Disease for both patients in Table 2.1. We compute the bel and the pl for the proposition $(x < y)$ *and* $(x \leq y)$*.*

Thus, according to Definitions 6 and 7.

- $bel(x < y) = 0$
- $pl(x < y) = 0.1 * 0.7 = 0.07$
- $bel(x \leq y) = 0.1 * 0.7 = 0.07$
- $pl(x \leq y) = 0.1 * 0.7 + 0.9 * 0.7 = 0.7$

Defintion 8 *(Superior Inequality) Let* x *and* y *be two random independent variables and their mass functions,* m_x^Θ, m_y^Θ: $2^\Theta \longrightarrow [0,1]$, $A, B \subseteq \Theta$ *are their respective focal elements. The superior inequality is calculated such that (Bell et al., 1996):*

$$bel(x > y) = \sum_{A \subseteq \Theta} m_x^\Theta(A) * \sum_{B \subseteq \Theta \wedge A >^\forall B} m_y^\Theta(B) \tag{2.17}$$

where $A >^\forall B$ *means that* $a > b$ *for all* $a \in A$, $b \in B$*.*

$$pl(x > y) = \sum_{A \subseteq \Theta} m_x^\Theta(A) * \sum_{B \subseteq \Theta \wedge A >^\exists B} m_y^\Theta(B) \tag{2.18}$$

where $A >^\exists B$ *means for every* $a \in A$ *there exists* $b \in B$ *such that* $a > b$*.*

Defintion 9 *(Superior or Equal Inequality) Let* x *and* y *be two random independent variables and their mass functions,* m_x^Θ, m_y^Θ: $2^\Theta \longrightarrow [0,1]$, $A, B \subseteq \Theta$ *are their respective focal elements.* $x \geq y$ *is measured as follows (Bell et al., 1996):*

$$bel(x \geq y) = \sum_{A \subseteq \Theta} m_x^\Theta(A) * \sum_{B \subseteq \Theta \wedge A \geq^\forall B} m_y^\Theta(B) \tag{2.19}$$

where $A \geq^\forall B$ *means that* $a \geq b$ *for all* $a \in A$, $b \in B$*.*

$$pl(x \geq y) = \sum_{A \subseteq \Theta} m_x^\Theta(A) * \sum_{B \subseteq \Theta \wedge A \geq^\exists B} m_y^\Theta(B) \tag{2.20}$$

where $A \geq^\exists B$ *means for every* $a \in A$ *there exists* $b \in B$ *such that* $a \geq b$*.*

Example 14 *Let us continue with the same example of Table 2.1. We want now to calculate the bel and the pl of both patients for the proposition (x>y) and (x≥y) for the attribute Disease.*

Thus, according to Definitions 6 and 7.

- $bel(x > y) = 0.1 * 0.3 + 0.9 * 0.3 = 0.3$
- $pl(x > y) = 0.1 * 0.3 + 0.9 * 0.3 + 0.9 * 0.7 = 0.93$
- $bel(x \geq y) = 0.1 * 0.3 + 09 * 0.3 + 0.9 * 0.7 = 0.93$
- $pl(x \geq y) = 0.1 * 0.3 + 09 * 0.3 + 0.9 * 0.7 + 0.1 * 0.7 = 1$

2.2.5 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.

Vacuous belief function

A *vacuous belief function* is a special belief function (Shafer, 1976) where Θ is its unique focal element. In this case, the basic belief assignment *bba* quantifies the *total ignorance*. In other words it represents a bba with no information. It is defined such that:

$$m^{\Theta}(\Theta) = 1 \quad \text{and} \quad m^{\Theta}(A) = 0 \quad \text{where} \quad A \neq \Theta \tag{2.21}$$

Example 15 *The frame of discernment relative to attribute* Disease *is* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$.

$m^{\Theta_{DE}}(\{Flu, Asthma, Anemia\}) = 1$ *is called a vacuous basic belief assignment.*

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. It is defined such that:

$$\begin{cases} m^{\Theta}(A) = 1, \quad A \in \Theta & \text{and} \mid A \mid = 1 \\ m^{\Theta}(B) = 0 \quad \forall B \neq A \end{cases} \tag{2.22}$$

Example 16 *The frame of discernment of attribute* Disease *is* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$.

$m^{\Theta_{DE}}(\{Anemia\}) = 1$ *is called a certain basic belief assignment. This bba reflects the full certainty about information Anemia.*

Bayesian belief function

A bayesian belief function is a particular case of belief function. When all focal elements are singletons the *bba* is named bayesian and the distribution is said to be *probabilistic* (Shafer, 1976). It is defined such that:

$$\begin{cases} m^{\Theta}(A) \in]0,1], \quad \text{if } \mid A \mid = 1 \\ m^{\Theta}(A) = 0 \quad \text{otherwise} \end{cases} \tag{2.23}$$

Figure 2.2 is an illustration of a bayesian mass function.

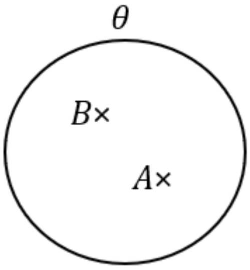


Figure 2.2: Bayesian mass function

Properties

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

Example 17 *Suppose the same frame of discernment of attribute Disease:* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$.

Let's consider the following masses:

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

Focal elements in this example are singletons. Thus, the distribution, in this case, is probabilistic and the bba is bayesian.

Consonant belief function

A consonant belief function is a function with nested focal elements such that:

$$A_1 \subset A_2 \subset ... \subset \Theta \tag{2.24}$$

Figure 2.3 is an illustration of a consonant mass function.

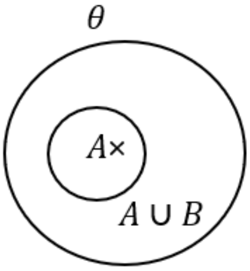


Figure 2.3: Consonant mass function

Properties

- $bel = (A \cap B) = min(bel(A), bel(B))$: Necessity measure.
- $pl(A \cup B) = max(pl(A), pl(B))$: Possibility measure.

Example 18 *Let's consider the same frame of discernment* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$. *We have the following masses:*

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

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

Simple support function

A simple support function is a mass that supports only one subset of Θ; it has at most one focal element different from Θ. This focal element is named the focus of the simple support function (Smets, 1995). It is defined such that:

$$m^{\Theta}(B) = \begin{cases} \omega & \text{if} \quad B = \Theta \\ 1-\omega & \text{if} \quad B = A, \quad \text{for some } A \subset \Theta \\ 0 & \text{otherwise} \end{cases} \tag{2.25}$$

Where:

A is the focus function.

ω is the degree of support of Θ, $\omega \in [0,1]$.

$1-\omega$ is the degree of support of the focus A.

Figure 2.4 is an illustration of a simple support mass function.

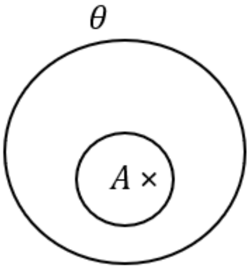


Figure 2.4: Simple support mass function

Example 19 *Suppose having the following bba:*
$m^{\Theta}(\{Anemia, Asthma\} = 0.5$
$m^{\Theta}(\{\Theta\}) = 0.5$
m^{Θ} *is a simple support function with* $\{Anemia, Asthma\}$ *is its focus.*

Dogmatic and non dogmatic functions

- *A dogmatic mass function* is defined such that (Smets, 1995):

$$m^{\Theta}(\Theta) = 0 \tag{2.26}$$

- *A non dogmatic mass function* is defined such that(Smets, 1995):

$$m^{\Theta}(\Theta) > 0 \tag{2.27}$$

Smets defined the notions of *the closed and the open worlds assumptions* (Smets, 1988). Thus, under the closed world assumption, all possible hypotheses are enumerated in Θ; where Θ is exhaustive. On the other side, when it is difficult to enumerate from the beginning all the hypotheses the open world assumption is considered. Under the open word assumption, the set of hypotheses are unknown; where Θ is not necessarily exhaustive. Note that the closed world assumption is the one considered by the Dempster-Shafer model (Dempster, 1967; Shafer, 1976).

Under the closed world assumption, the mass function is considered as a *normalized* basic belief function. Under the open world assumption, the mass function is a *non-normalized* one. The normalization can be executed such as:

$$\begin{cases} m^{*\Theta}(A) = \dfrac{m^{\Theta}(A)}{1 - m^{\Theta}(\emptyset)} & \forall A \subseteq \Theta \\ m^{*\Theta} = 0 \end{cases} \tag{2.28}$$

Information modeled and managed using the belief functions theory may come from different sources that can be reliable and/or unreliable. Sometimes it is necessary to merge them in order to make some decisions. Thus, the emergence of several combination rules under the evidential framework.

2.3 Combination Rules

In belief functions' theory, *Combination rules* appear as an interesting solution to get a more reliable information, specially in the presence of imperfect information (uncertain, imprecise, incomplete). The major interest of combining several sources is to have at the end one mass function that represents to the best all combined ones. Many combination rules are proposed in the framework of evidence theory. In this section, the most commonly used rules in literature are presented. Namely, the Dempster's rule (Dempster, 1967), the conjunctive and the disjunctive rules (Smets, 1993), Dubois and Prade combination rule (Dubois and Prade, 1988), Yager's rule (Yager, 1987).

In below, definitions of some basic notions related to combination rules are shown:

- *Independence* is the occurrence that one information does not affect the assignment of the occurrence of an other one.

- *Reliability* is the ability to provide a trustworthy information.

- *Normalization* is an adjustment of a measured values to either have a non negative ones or to asses them to a common scale.

2.3.1 Dempster's rule of combination

Suppose having M independent, distinct and reliable sources of information to combine. Each source S_j with $j \in [1, M]$ expresses its belief over a defined problem and gives the corresponding mass m_j^Θ (Dempster, 1967). The combination (called the *joint mass*) is calculated as follows:

$$m_\oplus^\Theta = m_1^\Theta \oplus .. \oplus m_M^\Theta \tag{2.29}$$

$\oplus$ is the orthogonal sum.

The Dempster's rule of combination is a normalized rule defined under the closed world assumption.

Defintion 10 *Let* m_1^Θ *and* m_2^Θ *be two independent mass functions, the* joint mass $m_{1\oplus 2}^\Theta$ *is computed such that:*

$$m_{1\oplus 2}^\Theta(A) = \begin{cases} \dfrac{\sum_{B\cap C=A} m_1^\Theta(B).m_2^\Theta(C)}{1-\sum_{B\cap C=\emptyset} m_1^\Theta(B).m_2^\Theta(C)} & \forall A \neq \emptyset \\ 0 & \forall A = \emptyset \end{cases} \tag{2.30}$$

The figure 2.5 shows how to combine M independent sources using the Dempster's rule.

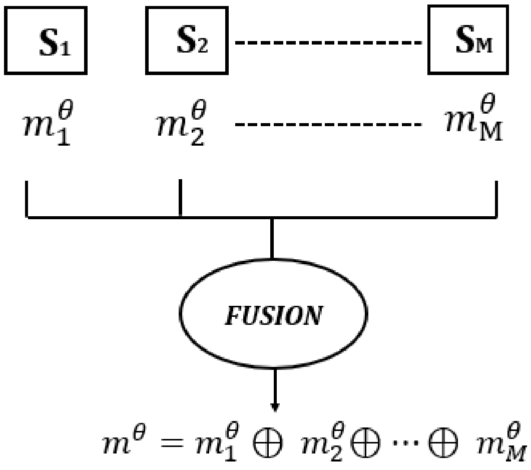

S1
S2
SM
m_1^θ
m_2^θ
m_M^θ
FUSION
$m^\theta = m_1^\theta \oplus m_2^\theta \oplus \cdots \oplus m_M^\theta$

Figure 2.5: Combination of M independent sources (Dempster, 1967)

Example 20 *Suppose having two different doctors* S_1 *and* S_2*, each one gives his diagnosis about the same patient. The combination of their diagnoses using the Dempster's* *rule of combination is shown in Table 20.*

$\oplus$	$m_2^{\Theta_{DE}}(\{Asthma\})$ =0.4	$m_2^{\Theta_{DE}}(\{Flu\})$ =0.5	$m_2^{\Theta_{DE}}(\{\Theta\})$ =0.1
$m_1^{\Theta_{DE}}(\{Asthma\})$ =0.5	$m_{1,2}^{\Theta_{DE}}(\{Asthma\})$ =0.2	$m_{1,2}^{\Theta_{DE}}(\emptyset)$ =0.25	$m_{1,2}^{\Theta_{DE}}(\{Asthma\})$ =0.05
$m_1^{\Theta_{DE}}(\{Asthma, Anemia\}) = 0.1$	$m_{1,2}^{\Theta_{DE}}(\{Asthma\})$ =0.04	$m_{1,2}^{\Theta_{DE}}(\emptyset)$ =0.05	$m_{1,2}^{\Theta_{DE}}(\{Asthma, Anemia\}) = 0.01$
$m_1^{\Theta_{DE}}(\{Flu\})$ =0.2	$m_{1,2}^{\Theta_{DE}}(\emptyset)$ =0.08	$m_{1,2}^{\Theta_{DE}}(\{Flu\})$ =0.1	$m\Theta_{DE1,2}(\{Flu\})$ =0.2
$m_1^{\Theta_{DE}}(\Theta)$ =0.2	$m_{1,2}^{\Theta_{DE}}(\{Asthma\})$ =0.08	$m_{1,2}^{\Theta_{DE}}(\{Flu\})$ =0.1	$m_{1,2}^{\Theta_{DE}}(\Theta)$ =0.02

Table 2.2: Dempster's Rule of combination over two different diagnoses of a same patient

- $m_{1\oplus 2}^{\Theta_{DE}}(\{Asthma\}) = \dfrac{0.2 + 0.05 + 0.04 + 0.08}{1 - (0.25 + 0.05 + 0.08)} = 0.56$
- $m_{1\oplus 2}^{\Theta_{DE}}(\{Asthma, Anemia\}) = \dfrac{0.01}{1 - (0.25 + 0.05 + 0.08)} = 0.05$
- $m_{1\oplus 2}^{\Theta_{DE}}(\{Flu\}) = \dfrac{0.1 + 0.02 + 0.1}{1 - (0.25 + 0.05 + 0.08)} = 0.355$
- $m_{1\oplus 2}^{\Theta_{DE}}(\Theta) = \dfrac{0.02}{1 - (0.25 + 0.05 + 0.08)} = 0.035$

2.3.2 Conjunctive rule of combination

The *conjunctive rule of combination* for two mass functions m_1^Θ and m_2^Θ defined on the same frame of discernment Θ was introduced by Smets (Smets, 1993) and it is defined as follows:

$$m_1^\Theta \circledcirc m_2^\Theta(C) = \sum_{A,B\subseteq\theta:A\cap B=C} m_1^\Theta(A).m_2^\Theta(B) \tag{2.31}$$

The conjunctive rule of combination merges *bbas* m_1^Θ and m_2^Θ provided by different, independent and reliable sources, the result is the joint *bba* induced from the combination of m_1^Θ and m_2^Θ. The conjunctive rule of combination is an unnormalized rule defined under the open world assumption.

Example 21 *Let us consider the following frame of discernment* $\Theta_{DE} = \{$*Flu, Asthma, Anemia*$\}$. *Sources* S_1 *and* S_2 *gave 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 2.3.

$\textcircled{$\cap$}$	$m_1^{\Theta_{DE}}(\{Flu, Anemia\})$ $= 0.7$	$m_1^{\Theta_{DE}}(\{Asthma, Anemia\})$ $= 0.3$
$m_2^{\Theta_{DE}}(\{Flu, Asthma\})$ $= 0.5$	$m_{1,2}^{\Theta_{DE}}(\{Flu\})$ $= 0.35$	$m_{1,2}^{\Theta_{DE}}(\{Asthma\})$ $= 0.15$
$m_2^{\Theta_{DE}}(\{Asthma, Anemia\})$ $= 0.5$	$m_{1,2}^{\Theta_{DE}}(\{Anemia\})$ $= 0.35$	$m_{1,2}^{\Theta_{DE}}(\{Asthma, Anemia\})$ $= 0.15$

Table 2.3: An application of the conjunctive rule of combination

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

2.3.3 Disjunctive rule of combination

The disjunctive rule of combination was also proposed by Smets (Smets, 1993). This combination rule relies on independent sources with at least one of them is reliable. It is based on the union of focal elements. Combining bbas m_1^{Θ} and m_2^{Θ} with the disjunctive rule leads to a combined *bba* which focal elements are the union of focal elements of m_1^{Θ} and m_2^{Θ}. Combining two bbas m_1^{Θ} and m_2^{Θ} defined on the same frame of discernment Θ is defined as follows:

$$m_1^{\Theta} \textcircled{$\cup$} m_2^{\Theta}(C) = \sum_{A,B \subseteq \theta : A \cup B = C} m_1^{\Theta}(A).m_2^{\Theta}(B). \tag{2.32}$$

Example 22 *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 2.4.*

$\ocup$	$m_1^{\Theta_{DE}}(\{Flu, Anemia\})$ $= 0.7$	$m_1^{\Theta_{DE}}(\{Asthma, Anemia\})$ $= 0.3$
$m_2^{\Theta_{DE}}(\{Flu, Asthma\})$ $= 0.5$	$m_{1,2}^{\Theta_{DE}}(\{Flu, Asthma, Anemia\}) = 0.35$	$m_{1,2}^{\Theta_{DE}}(\{Flu, Asthma, Anemia\}) = 0.15$
$m_2^{\Theta_{DE}}(\{Asthma, Anemia\})$ $= 0.5$	$m_{1,2}^{\Theta_{DE}}(\{Flu, Asthma, Anemia\}) = 0.35$	$m_{1,2}^{\Theta_{DE}}(\{Asthma, Anemia\})$ $= 0.15$

Table 2.4: An application of the disjunctive rule of combination

- $m_1^{\Theta_{DE}} \ocup\, m_2^{\Theta_{DE}}(\{Flu, Asthma, Anemia\})$ = *(0.5 * 0.7) + ((0.5 * 0.3) + (0.5 * 0.7) = 0.85*
- $m_1^{\Theta_{DE}} \ocup\, m_2^{\Theta_{DE}}(\{Asthma, Anemia\})$= *0.5 * 0.3 = 0.15*

Table 2.5 represents an overview of the use of the presented combination rules. Note that:

$\ocap$	$\oplus$	$\ocup$
Independent and reliable pieces of evidence non-normalized	Independent and reliable pieces of evidence Normalized	Independent and at least one of evidence pieces is reliable non-normalized

Table 2.5: An overview of combination rules properties

In case of having reliable and unreliable sources, where degrees of reliability can be measured, the discounting rule is used.

2.4 Discounting

A particular combination is the discounting that considers sources' reliabilities into their mass functions. It is a specific mechanism to the belief functions theory that discounts masses proportionally to their sources' reliabilities. However, sources' reliabilities need to be learned before the discounting.

The reliability factor α in [0, 1] characterizes the reliance of a source. Note that (i) $\alpha = 1$ represents a fully reliable source, (ii) $\alpha = 0$ represents an unreliable source. The discounting rate is $1 - \alpha$.

The discounted mass $m^{\Theta,\alpha}$ is computed as follows:

$$\begin{cases} m^{\Theta,\alpha}(A) = \alpha.m^{\Theta}(A) & \forall A \subset \Theta \\ m^{\Theta,\alpha}(\Theta) = \alpha.m^{\Theta}(\Theta) + (1-\alpha) \end{cases} \tag{2.33}$$

Example 23 *Let's consider this bba:*

- $m^{\Theta}_{DE}(\{Flu\}) = 0.4$
- $m^{\Theta}_{DE}(\{Anemia, Asthma\}) = 0.5$
- $m^{\Theta}_{DE}(\Theta_{DE}) = 0.1$

This bba is discounted with the reliability degree of the doctor which is $\alpha = 0.9$.

- $m^{\Theta_{DE},\alpha}(\{Flu\}) = 0.9 * 0.4 = 0.36$
- $m^{\Theta_{DE},\alpha}(\{Anemia, Asthma\}) = 0.9 * 0.5 = 0.45$
- $m^{\Theta_{DE},\alpha}(\Theta_{DE}) = 0.9 * 0.1 + (1 - 0.9) = 0.19$

Basic belief assignments can be combined only when they are defined on the same frame of discernment. Thus, when two basic belief assignments are defined on different frames of discernments, a compatible frame can be specified.

2.5 Compatible Frames of Discernments

Vacuous extension, coarsening and refinement (Shafer, 1976) are tools to define a relationship between compatible frames of discernment in order to specify beliefs on anyone of them.

2.5.1 Vacuous Extension

In some cases, we need to combine two *bbas* m_1^{Θ} and m_2^{Θ} 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* (Shafer, 1976) 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^{Θ} and m_2^{Θ}, 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^{Θ}, is defined as follows:

$$m^{\Theta_1 \uparrow \Theta_1 \times \Theta_2}(A) = \begin{cases} m^{\Theta_1}(B) & \text{where } A = B \times \Theta_2,\ B \subseteq \Theta_1 \\ 0 & \text{otherwise} \end{cases} \tag{2.34}$$

Figure 2.6 represents an illustration of the vacuous extension operator

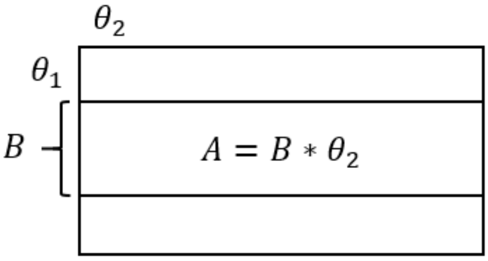


Figure 2.6: Vacuous Extension

Example 24 *Suppose we have the following frames of discernment:*

- *The set of Diseases* $\Theta_{DE} = \{Flu, Asthma, Anemia\}$
- *The set of Blood Types* $\Theta_{BT} = \{A, B, O\}$

The following Table 2.6 illustrates the doctor's diagnosis about two patients.

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

Table 2.6: Diagnoses of two patients

The vacuous extension of the evidential Table 2.6 is demonstrated in Table 2.7.

Note that Θ *denotes the joint frame of* Θ_{DE} *and* Θ_{BT} *and that* $\uparrow$ *denotes the vacuous extension.*

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 2.7: The vacuous extension of Table 2.6

2.5.2 Refinement and Coarsening

Let Θ and Ω be two different and compatible frames of discernment. The set Ω is a refinement of Θ provided by splitting hypotheses of Θ. The set Θ is called a coarsening of Ω obtained by gathering hypotheses of Ω. Figure 2.7 illustrates the relationship of coarsening and refinement between Θ and Ω.

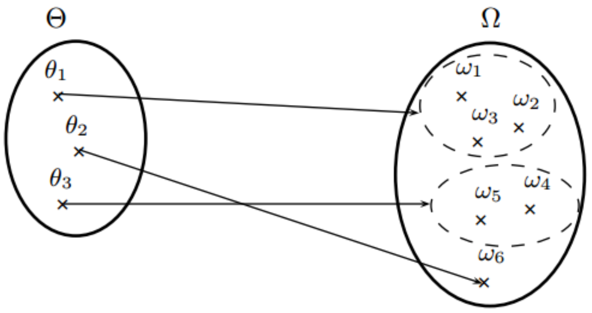


Figure 2.7: Coarsening and Refinement

Example 25 *Let's give an illustration with the same frame of discernment*
$\Theta = \{Flu, Asthma, Anemia\}$.
A refinement of Θ *can be such that:*
Ω=*{Flu type A, Flu type B, Flu type C, Asthma, Thalassaemia, Aplastic-anemia, Fanconi-anemia, Haemolytic-anemia}*

- $\omega(\{Flu\})$ =*{Flu type A, Flu type B, Flu type C}*.
- $\omega(\{Asthma\})$= *{Asthma}*.
- $\omega(\{Anemia\})$ =*{Thalassaemia, Aplastic-anemia, Fanconi-anemia, Haemolytic-anemia}*.

The set Θ *is a coarsening of* Ω.

2.6 Cognitive and Evidential Independences

The cognitive and the evidential independences are properties introduced by (Shafer, 1976). Then, other works based on the evidential independence were proposed (Chebbah et al., 2015).

2.6.1 Cognitive Independence

Defintion 11 *" Two frames of discernment may be called cognitively independent with respect to the evidence if new evidence that bears on only one of them will not change the degree of support for propositions discerned by the other"(Shafer, 1976).*

The cognitive independence is *the weak independence.* Two variables x and y are cognitively independent with respect to a mass function m^{Θ}[2] if new evidence that bears on only one of them does not change the degree of support for propositions discerned by the other one such that:

$$pl(x \wedge y) = pl(x) \times pl(y) \tag{2.35}$$

Example 26 *Let $\Theta_O = \{O, \overline{O}\}$, O for an obese person and $\overline{O}$ for a non obese person. The frame of discernment $\Theta_G = \{M, F\}$, M for a male and F for a female. The joint frame of discernment of Θ_O and Θ_G is Θ.*

Suppose a mass function m^{Θ} defined on the joint frame Θ such that:
$m^{\Theta}((M,O))$ = 0.26
$m^{\Theta}((F,O))$ = 0.16
$m^{\Theta}((M,O) \cup (F,O))$ = 0.58
The plausibilities $pl^{\Theta_O \uparrow \Theta}$ and $pl^{\Theta_G \uparrow \Theta}$ are computed in Table 2.8.

Θ_O	$pl^{\Theta_O \uparrow \Theta}$	Θ_G	$pl^{\Theta_G \uparrow \Theta}$
$\emptyset$	0	$\emptyset$	0
O	0.84	M	1
$\overline{O}$	0.74	F	0
$O \cup \overline{O}$	1	$M \cup F$	1

Table 2.8: $pl^{\Theta_O \uparrow \Theta}$ and $pl^{\Theta_G \uparrow \Theta}$

Variables "Obesity" and "Gender" are cognitively independent according to m^{Θ} when the following equalities are verified:

[2] $\Theta = \Theta_x \times \Theta_y$

$$\begin{cases} pl^{\Theta}((O,M)) = pl^{\Theta_O\uparrow\Theta}(O) \times pl^{\Theta_G\uparrow\Theta}(M) \\ pl^{\Theta}((O,F)) = pl^{\Theta_O\uparrow\Theta}(O) \times pl^{\Theta_G\uparrow\Theta}(F) \\ pl^{\Theta}((\overline{O},M)) = pl^{\Theta_O\uparrow\Theta}(\overline{O}) \times pl^{\Theta_G\uparrow\Theta}(M) \\ pl^{\Theta}((\overline{O},F)) = pl^{\Theta_O\uparrow\Theta}(\overline{O}) \times pl^{\Theta_G\uparrow\Theta}(F) \end{cases}$$

Thus,

$$\begin{cases} pl^{\Theta}((O,M)) = pl^{\Theta_O\uparrow\Theta}(O) \times pl^{\Theta_G\uparrow\Theta}(M) = 0.84 * 1 = 0.84 \\ pl^{\Theta}((O,F)) = pl^{\Theta_O\uparrow\Theta}(O) \times pl^{\Theta_G\uparrow\Theta}(F) = 0.84 * 0 = 0 \\ pl^{\Theta}((\overline{O},M)) = pl^{\Theta_O\uparrow\Theta}(\overline{O}) \times pl^{\Theta_G\uparrow\Theta}(M) = 0.74 * 1 = 0.74 \\ pl^{\Theta}((\overline{O},F)) = pl^{\Theta_O\uparrow\Theta}(\overline{O}) \times pl^{\Theta_G\uparrow\Theta}(F) = 0.74 * 0 = 0 \end{cases}$$

2.6.2 Evidential Independence

Defintion 12 *"Two frames of discernment are evidentially independent with respect to a support function if that support function could be obtained by combining evidence that bears on only one of them with evidence that bears on only the other"(Shafer, 1976).*

The evidential independence is *the strong independence*. Two variables x and y are evidentially independent with respect to a mass function m^{Θ} if m^{Θ} can be obtained by combining evidence that bears on only one of them with evidence that bears on only the other one such that:

$$\begin{cases} pl(x \wedge y) = pl(x) \times pl(y) \\ bel(x \wedge y) = bel(x) \times bel(y) \end{cases} \tag{2.36}$$

Example 27 *Variables "Obesity" and "Gender" are evidentially independent according to m^{Θ} when the following equalities are verified:*

$$\begin{cases} bel^{\Theta}((O,M)) = bel^{\Theta_O\uparrow\Theta}(O) \times bel^{\Theta_G\uparrow\Theta}(M) \\ bel^{\Theta}((O,F)) = bel^{\Theta_O\uparrow\Theta}(O) \times bel^{\Theta_G\uparrow\Theta}(F) \\ bel^{\Theta}((\overline{O},M)) = bel^{\Theta_O\uparrow\Theta}(\overline{O}) \times bel^{\Theta_G\uparrow\Theta}(M) \\ bel^{\Theta}((\overline{O},F)) = bel^{\Theta_O\uparrow\Theta}(\overline{O}) \times bel^{\Theta_G\uparrow\Theta}(F) \end{cases}$$

$$\begin{cases} pl^{\Theta}((O,M)) = pl^{\Theta_O\uparrow\Theta}(O) \times pl^{\Theta_G\uparrow\Theta}(M) \\ pl^{\Theta}((O,F)) = pl^{\Theta_O\uparrow\Theta}(O) \times pl^{\Theta_G\uparrow\Theta}(F) \\ pl^{\Theta}((\overline{O},M)) = pl^{\Theta_O\uparrow\Theta}(\overline{O}) \times pl^{\Theta_G\uparrow\Theta}(M) \\ pl^{\Theta}((\overline{O},F)) = pl^{\Theta_O\uparrow\Theta}(\overline{O}) \times pl^{\Theta_G\uparrow\Theta}(F) \end{cases}$$

The requirement on pl is already checked in Example 26. The beliefs $bel^{\Theta_O\uparrow\Theta}$ and $bel^{\Theta_G\uparrow\Theta}$ are computed in Table 2.9.

Θ_O	$bel^{Theta_O\uparrow\Theta}$	Θ_G	$bel^{\Theta_G\uparrow\Theta}$
$\emptyset$	0	$\emptyset$	0
O	0.26	M	1
$\overline{O}$	0.16	F	0
$O \cup \overline{O}$	1	$M \cup F$	1

Table 2.9: $bel^{\Theta_O\uparrow\Theta}$ and $bel^{\Theta_G\uparrow\Theta}$

$$\begin{cases} bel^{\Theta}((O,M)) = bel^{\Theta_O}(O) \times bel^{\Theta_G}(M) = 0.26 * 1 = 0.26 \\ bel^{\Theta}((O,F)) = bel^{\Theta_O}(O) \times bel^{\Theta_G}(F) = 0.26 * 0 = 0 \\ bel^{\Theta}((\overline{O},M)) = bel^{\Theta_O}(\overline{O}) \times bel^{\Theta_G}(M) = 0.16 * 1 = 0.16 \\ bel^{\Theta}((\overline{O},F)) = bel^{\Theta_O}(\overline{O}) \times bel^{\Theta_G}(F) = 0.16 * 0 = 0 \end{cases}$$

2.7 Decision Making

In the *credal level*, theory of belief functions permits to model and handle imperfect information. It also allows the combination of beliefs coming from different sources. In the *pignistic level*, evidence theory ensures the decision making by providing several solutions. This decision process can be not feasible because of the bbas' nature. Indeed, these bbas are modeled as singletons or as subsets. Hence, to facilitate the decision making process, the belief functions theory offers the following tools:

- Pignistic probability (Smets, 1998b; Smets and Kennes, 1994).
- Maximum of credibility (Janez, 1997; Basir and Yuan, 2007).
- Maximum of plausibility (Janez, 1997; Basir and Yuan, 2007).
- Distance based (Essaid et al., 2014).

2.7.1 Pignistic Probability

The pignistic probability, denoted $BetP$, was introduced by the transferable belief model (Smets, 1988; Smets and Kennes, 1994). This measure represents a compromise between the credibility function and the plausibility function. Thus, it transforms the beliefs to probability functions. The pignistic probability depends on choosing the most probable singleton hypothesis such that:

$$BetP(A) = \sum_{B \subseteq \Theta} \frac{|A \cap B| * m^{\Theta}(B)}{|B| * (1 - m^{\Theta}(\emptyset))} \ \forall A \subseteq \Theta \tag{2.37}$$

Example 28 *Suppose the following basic belief assignment where* Θ_{DE}*=*$\{$*Flu,Asthma,Anemia*$\}$*:*

$m(\{Flu\}) = 0.5$
$m(\{Asthma, Flu\}) = 0.2$
$m(\Theta) = 0.3$

The corresponding pignistic probability BetP to this bba is computed as follows:

- $BetP(\{Flu\}) = \frac{0.5}{1} + \frac{0.2}{2} + \frac{0.3}{3} = 0.7$
- $BetP(\{Asthma\}) = \frac{0.2}{2} + \frac{0.3}{3} = 0.2$
- $BetP(\{Anemia\}) = \frac{0.3}{3} = 0.1$

The hypothesis that maximizes the BetP function is $\{$*Flu*$\}$*. It is the most probable hypothesis.*

2.7.2 Maximum of Credibility

The maximum of credibility consists in selecting the most credible hypothesis. The latter has the maximal value of belief *bel*.

The maximum of credibility determines the best hypothesis with the least chance of being realized. It uses a pessimistic decision criterion (Janez, 1997) and it is defined such that:

$$argmax[bel(A)] \quad \forall A \subseteq \Theta \tag{2.38}$$

Example 29 *Let's compute the maximal credibility function of the same bba as the previous Example 28.*

- $bel(\{Flu\}) = 0.5$
- $bel(\{Asthma\} = 0$
- $bel(\{Anemia\} = 0$

The hypothesis that maximizes the bel function is $\{$*Flu*$\}$*. It is the most credible hypothesis.*

2.7.3 Maximum of Plausibility

The maximum of plausibility consists in selecting the most plausible hypothesis. The latter has the maximal value of plausibility *pl*.
The maximum of plausibility determines the best hypothesis with the most chance of being realized. It uses an optimistic decision criterion (Janez, 1997) and it is defined such that:

$$argmax[pl(A)] \quad \forall A \subseteq \Theta \tag{2.39}$$

Example 30 *Let's compute the maximal plausibility function of the same bba in Example 28.*

- $pl(\{Flu\}) = 1$
- $pl(\{Asthma\} = 0.5$
- $pl(\{Anemia\} = 0.3$

The hypothesis that maximizes the pl function is {Flu}. It is the most plausible hypothesis.

2.8 Conclusion

In this chapter, we presented the basic concepts of the belief functions theory. All notions like the frame of discernment, the belief function, the plausibility function, combination rules, vacuous extension, etc., were detailed with examples. We also presented the decision making tools as the pignistic probability, the maximum of credibility and the maximum of plausibility.

These tools will be used in the rest of our dissertation to model and handle evidential databases.

Part II

Evidential Database Models

3 Evidential Databases: The Compact Form

Contents

3.1 Introduction 58

3.2 Evidential Databases Model 58

3.3 Implementing Evidential Databases 59

3.3.1 Meta-model 60

3.3.2 Object-Relational Implementation 60

3.4 Querying Evidential Databases EDB 62

3.4.1 Evidential Relational Queries 63

3.4.2 Evidential Top-k Query 68

3.4.3 Evidential Skyline Query 79

3.5 Conclusion 83

Summary

This chapter is about the compact form of evidential databases (EDB) and it is divided into three major parts: first, we present the only existing model of evidential databases (Bell et al., 1996; Lee, 1992b; Lee, 1992a). Second, we present how we practically modeled and implemented this database model (Bousnina et al., 2016). Finally, we focus on querying the compact form of EDB. Thus, we recall the extended evidential relational queries (Bell et al., 1996). Then, we discuss two preferential queries; the evidential top-k query (Bousnina et al., 2017a) and the evidential skyline query (Elmi et al., 2014; Bousnina et al., 2017b).

3.1 Introduction

Relational databases store only certain data where each attribute contains singleton values with no additional information. Imperfect databases store sets of imperfect data like uncertain, imprecise, missing ones, etc, where each attribute can handle a set of values with other information about their existence in the table; hence, the divergence between the relational databases and the imperfect databases in terms of querying. Several models of imperfect databases and their querying methods were introduced such as probabilistic databases (Cavallo and Pittarelli, 1987), possiblistic databases (Bosc and Pivert, 2005), and evidential databases (Bell et al., 1996; Choenni et al., 2006; Lee, 1992a; Lee, 1992b).

In this chapter, our interest goes to the compact form of the evidential database (EDB). In fact, we present the most elaborated evidential database model (Bell et al., 1996; Lee, 1992a; Lee, 1992b). Then, we focus on the querying methods that can be applied over (EDB): The evidential relational queries (Bell et al., 1996), the evidential top-k queries (Bousnina et al., 2017a) and the evidential skyline queries (Bousnina et al., 2017b). We also present the Object-relational implementation of the evidential compact database EDB (Bousnina et al., 2016) that serves as a basis to practically evaluate the presented querying methods.

3.2 Evidential Databases Model

The most used and known evidential database model is the one introduced by (Lee, 1992b; Lee, 1992a) and extended later in (Bell et al., 1996). Thus, authors defined an evidential database model in a compact form semantics. It is formally defined as follow:

Defintion 13 *(Compact form)*

An Evidential Database, EDB, on its compact form, is a database with N tuples and D attributes, storing perfect and imperfect data. Imperfection is expressed in two levels:

- *The attribute level uncertainty, modeled via the evidential values. An evidential value, denoted V_{ta}, is the value of an attribute a for the tuple t. An evidential value is a bba, such that*

$$V_{ta} : 2^{\Theta_a} \rightarrow [0, 1] \tag{3.1}$$

$$\textit{with } m_{ta}^{\Theta}(\varnothing) = 0 \textit{ and } \sum_{B \subseteq \Theta_a} m_{ta}^{\Theta}(B) = 1 \tag{3.2}$$

The set of focal elements of the bba V_{ta} is noted F_{ta} such that:

$$F_{ta} = \{B \subseteq \Theta_a / m^{\Theta}_{ta}(B) > 0\} \tag{3.3}$$

The tuple level uncertainty, expressed through a particular attribute called confidence level. It is denoted CL and it stores confidence intervals given by experts about the existence of each tuple t in the evidential database. Each interval is a pair of belief and plausibility, such that:

$$CL = [bel; pl] \quad where \quad \Theta_{CL} = \{exist, \overline{exist}\} \quad CL \subseteq [0; 1] \quad bel \leq pl \tag{3.4}$$

Example 31 *Let us have an example of a doctor that expressed his medical diagnoses about three patients by giving multiple hypotheses and their assignments. His diagnoses are modeled using the belief functions' theory and stored in the following compact evidential Table 3.1. The latter includes four attributes* ID, Disease, Symptom *and the confidence level* CL*. The tuple level uncertainty is presented by the CL and the attribute level uncertainty is expressed by the mass functions in the Disease and Symptom attributes.*

ID	**Disease**	**Symptom**	**CL**
1	$Diabetes$	$Fatigue$ 0.4 $\{Fatigue, Nausea\}$ 0.6	[0.5 ; 1]
2	$Diabetes$ 0.1 $Stroke$ 0.9	$Vertigo$	[0.4 ; 0.8]
3	$Anemia$ 0.3 $\{Diabetes, Stroke\}$ 0.7	$Fainting$	[1 ; 1]

Table 3.1: A Medical Evidential Table

Since evidential databases store more complex data compared to relational databases, the relational structure seems to be not adequate to their storage. Hence, the need to introduce a model that fits the structure of this evidential data.

3.3 Implementing Evidential Databases

An evidential database (Bell et al., 1996) is a model that handles uncertainty, imprecision and ignorance. Storing these kind of data requires a specific model. At the best of our knowledge there is no efficient implementation relative to evidential databases for a querying purpose.

In our work (Bousnina et al., 2016), we introduce a meta-model and an object-relational implementation for EDBs that offer a scalable and flexible solution to manage evidential data.

3.3.1 Meta-model

In classical databases, tables store data in a matrix format; attributes in columns and objects in lines. In evidential databases, data are stored in object-relational format. As shown in the class diagram of the evidential meta-model of Figure 3.1. An evidential database is composed of tables where each table has N tuples and D attributes. These attributes are based on the structure of basic belief assignments, *bbas* . An attribute has a name and contains one *bba* for each tuple. A *bba* is composed of one or more focal elements. A focal element has a mass and contains hypotheses. Each hypothesis has a content. The belief and the plausibility functions compute respectively the minimal and the maximal degrees of believes about a set of hypotheses in a *bba*. They are defined as methods at the *bba* structure. Each evidential table stores N tuples. A tuple is identified by its ID and its confidence level, CL that quantifies the confidence level degree of the tuple's source.

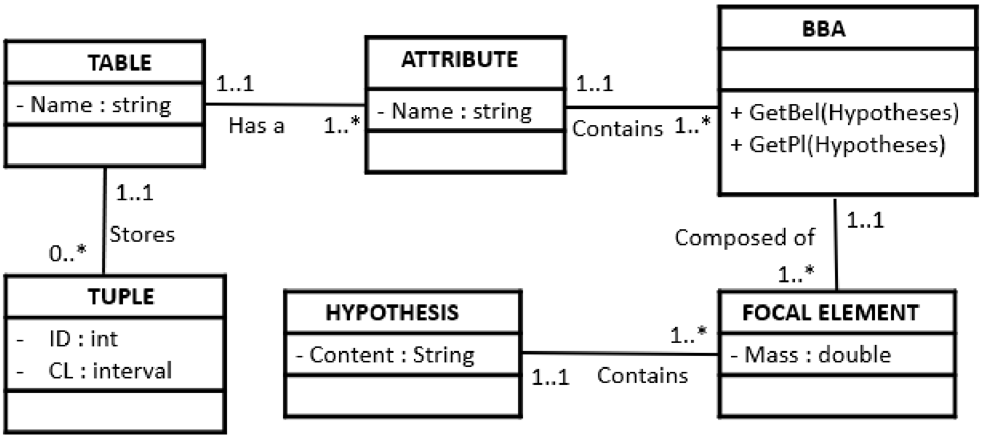


Figure 3.1: Meta-Model of Evidential Databases

3.3.2 Object-Relational Implementation

To put into practice the proposed meta-model, we relied on a commercial Object-Relational Database Management System (ORDBMAS), Oracle 10g. Its main asset is the Oriented-Object feature that facilitates the implementation of the complex structure of an evidential database, as designed in Figure 3.1. We define the *bba* type which is basically a collection of focal elements, whose type contains two compartments, a collection of hypotheses, and its mass value. As explained earlier, this modeling is impossible to implement in the relational framework, because of the first normal form constraint.

In addition to the previous advantage, a commercial ORDBMS offers an interesting I/O cost optimization. Indeed, Oracle uses a System Global Area (SGA) to keep in cache reused data (the database buffer cache) and reused queries (the shared pool). The database buffer cache contains previous extracted data. In case of querying some of (or these) data, the database server avoids disk access and returns the result directly from physical memory. In the other hand, queries execution plans whose computation is costly, are stored in the shared pool in the SGA. In case of executing an existing query in the pool, the system avoids syntactic analysis and execution plan computation which saves important CPU time. Added to that, the use of indexes in DBMSs that accelerates information extraction especially, in our case, when we use *Nested tables* vs. *Varying Arrays* to store collections. Indeed, a *varying arrays* attribute (in Oracle, the type is named varray) is physically stored in the same segment of the table. On the other hand, a *nested table* attribute is stored in a separate segment. Thus, data of that segment are indexable. For example, the attribute *symptom* in Table 3.1 is a collection of focal elements. The use of nested table type implies the possibility to create an index on that attribute which induces more efficiency when selecting tuples with a symptom criterion.

Moreover, we benefit from the use of SQL3 when employing the Object-Relational model in Oracle. Thus, if we want to create a *bba* type, which is mainly a collection of focal elements, we need to create first the type *hypos*; because the focal element structure contains a mass value, and also a collection of hypotheses (it can be one hypothesis or a set of hypotheses). The type *hypos* is a collection of a basic data type:

```
CREATE TYPE FOCALELEMENT AS OBJECT
(content HYPOS, mass NUMBER, MEMBER FUNCTION)
Includes (search FOCALELEMENT) RETURN NUMBER,
MEMBER FUNCTION Intersect (search FOCALELEMENT) RETURN NUMBER);
```

The *bba* type is defined as an object with one attribute; *content* and two methods; *bel* and *pl*. *Content* is defined as a collection of focal elements. *bel* and *pl* are methods that compute respectively belief and plausibility of comparison with another *bba*.

```
CREATE TYPE bba AS OBJECT (content FOCALELEMENT),
MEMBER FUNCTION bel_Comp (search bba, op CHAR(2)) RETURN NUMBER,
MEMBER FUNCTION pl_Comp (search bba, op CHAR(2)) RETURN NUMBER,
MEMBER FUNCTION bel(FOCALELEMENT F) RETURN NUMBER,
MEMBER FUNCTION pl(FOCALELEMENT F) RETURN NUMBER);
```

To integrate the tuple uncertainty level, we define the type CL that is an interval of numbers.

```
CREATE TYPE CL AS OBJECT (bel NUMBER, pl NUMBER);
```

Example 32 *To create the evidential Table of Example 3.1, we define the type on which is based the table, i.e., the type diagnosis:*

```
CREATE TYPE diagnosis AS OBJECT (ID NUMBER, Disease bba,
Symptom bba, CL CL);
```

And then we create the object table, called diagnoses, based on the type diagnosis:

```
CREATE TABLE diagnoses OF diagnosis (PRIMARY KEY(id),
NESTED TABLE Disease STORE AS
tab_diseases (NESTED TABLE content STORE AS hypos),
NESTED TABLE Symptom STORE AS
tab_symptoms (NESTED TABLE content STORE AS hypos);
```

The methods bel and pl are very important because we may select objects whose attributes' values (bbas) are compared as follows:

- *Symptom="Fatigue": The search criterion is a single value. Then, "Fatigue" is a bba with a single focal element, having a single hypothesis with a mass equal to one (a certain bba).*
- *Symptom= "Fatigue" OR "Nausea": It is a bba with one focal element that is {"Fatigue","Nausea"} having a mass equal to one.*
- *e.Symptom=d.Symptom: It consists in comparing two bbas. For example, comparing symptoms of patient (ID=1) and patient (ID=2).*

Querying is a form of questioning the stored data. Depending on the format of the data's storage, several families of queries can be applied like relational queries, ranking queries, preferential queries, etc.

3.4 Querying Evidential Databases EDB

Processing queries over evidential data is very challenging because of the complex structure of their modeling. Thus, each hypothesis can be uncertain and/or imprecise with two levels of assignment: at the attribute with masses and at the tuple with confidence intervals. Specific queries to deal with evidential data were introduced like the evidential relational queries (Bell et al., 1996), evidential top-k queries (Bousnina et al., 2017a; Bousnina et al., 2018b) and evidential Skyline queries (Elmi et al., 2014; Elmi et al., 2015; Bousnina et al., 2017b).

3.4.1 Evidential Relational Queries

The relational operators (selection, projection, Cartesian product, union and difference) were introduced by (Codd, 1970) based on the relational algebra (Kamel, 1954) and set theory (Moore, 1932). These operators represent the fundamental operators to query any database. Hence, authors (Lee, 1992b; Lee, 1992a; Bell et al., 1996) extended the classical relational operators to be suitable with the complex structure of an evidential database.

Extended Select Operator

The relational select operator consists in extracting some tuples from a given table whose attributes satisfy some conditions defined via the query. In the evidential database model as presented in (Lee, 1992a; Bell et al., 1996), the evidential select operator behaves almost the same as the classical ones except for the computation of confidence levels.

A new CL_t defined on $\Theta_a \ltimes \Theta_{CL}$ is computed such that:

$$CL_t = [bel_a^{\Theta_a}(a = h) \times bel_t^{\theta_1 CL}; pl_a^{\Theta_a}(a = h) \times pl_t^{\theta_1 CL}] \tag{3.5}$$

Where bel_a and pl_a are respectively the belief and the plausibility of attribute a with an hypothesis h responding to the condition of the query; bel_t and pl_t are respectively the belief and the plausibility of the tuple.

Example 33 *Let's apply the following select query over Table 3.1.*

Q_1: `SELECT ID FROM EDB WHERE Disease = ‘‘Stoke’’`

Tuples that do not satisfy the condition are eliminated. The returned relation includes the two tuples that verify the criterion. The confidence level of each tuple in the result is computed using Equation (3.5) *as shown in Table 3.2.*

Extended Project Operator

The relational project operator consists in taking as an input a relation and then giving as an output one attribute or more from attributes of the same relation. The evidential project operator as presented in (Bell et al., 1996; Choenni et al., 2006; Lee, 1992a) is

ID	CL
2	[0.9*0.4 ; 0.9*0.9] = [0.36 ; 0.81]
3	[0*1; 0.7*1] = [0 ; 0.7]

Table 3.2: Result of query Q_1

quite similar to classical one. Thus, it gives as a result the attribute values associated with their masses.

Example 34 *Let's apply the following project query over same table 3.1.*

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

Attributes that do not satisfy the query are eliminated. The result is presented in Table 3.3.

ID	*Disease*
1	Diabetes 1
2	Diabetes 0.1 Stroke 0.9
3	Anemia 0.3 {Diabetes, Stroke} 0.7

Table 3.3: Result of query Q_2

Extended Cartesian Product

The relational Cartesian product of two relations, denoted ×, consists in making the product of every tuple of the first relation with every tuple of the second relation. The evidential Cartesian product is defined in the same way except for the value of CL. It is computed using the *conjunctive combination* between two tuples (Bell et al., 1996; Lee, 1992a).

Defintion 14 *Let $CL_1[bel_1; pl_1]$ and $CL_2[bel_2; pl_2]$ be two evidential confidence levels, the conjunctive combination is defined such that:*

$$[bel_1; pl_1] \wedge [bel_2; pl_2]) = [bel_1 * bel_2; pl_1 * pl_2] \tag{3.6}$$

Example 35 *Let Table 3.4 and Table 3.5 be two evidential tables. The Cartesian product of $EDB_1 \times EDB_2$ is presented in Table 3.6.*

ID	$Disease$	$Symptom$	CL
1	$Diabetes$	$Nausea$	$[0.1; 0.5]$
2	$Anemia$	$Vertigo$	$[0; 0.7]$

Table 3.4: EDB_1

$Disease$	$Blood\ Type$	CL
$Stroke$	A	$[0.8; 1]$
$Anemia$	O	$[0.5; 0.9]$

Table 3.5: EDB_2

ID	$Disease$	$Disease$	$Symptom$	$BloodType$	CL
1	$Diabetes$	$Stroke$	$Nausea$	A	$[0.08; 0.5]$
1	$Diabetes$	$Anemia$	$Nausea$	O	$[0.05; 0.45]$
2	$Anemia$	$Stroke$	$Vertigo$	A	$[0; 0.7]$
2	$Anemia$	$Anemia$	$Vertigo$	O	$[0; 0.63]$

Table 3.6: Cartesian Product: $EDB_1 \times EDB_2$

Extended Join

The relational join operation, denoted $\bowtie$, 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 evidential context, the extended join behaves the same except for the confidence levels values which are calculated using the conjunctive combination of equation (3.6) (Bell et al., 1996; Lee, 1992a; Lee, 1992b).

Example 36 *We apply the extended join operator over Tables 3.4 and 3.5. The result of the evidential join is shown in Table 3.7.*

ID	$Disease$	$Disease$	$Symptom$	$BloodType$	CL
2	$Anemia$	$Anemia$	$Vertigo$	O	$[0; 0.63]$

Table 3.7: Join: $EDB_1 \bowtie EDB_2$

Experiments

The Object-Relational implementation of the compact evidential database model (Bousnina et al., 2016) allowed us to evaluate some extended relational operators and to analyze the results. Indeed, we implemented the evidential select and project operators. These operators can make use of any comparison operator $\{<,>,\leq,\geq\}$ to define a criterion on a *bba* or to define a *bel* and a *pl* threshold.

Example 37 *As an example, the query in below is about selecting patients whose diseases are equal to "Anemia" with belief grater that 0.2. The searched symptom is processed by the bel function as a bba with a single focal element whose mass is equal to one.*

```
Q: SELECT d.* FROM disease d
 WHERE d.bel(bba(FOCALELEMENT('Anemia',1)), '=') > 0.2;
```

We evaluated the proposed Object–Relational implementation from a performance point of view. We used a windows 8 desktop with a 2.67 GHZ CPU and 8GB RAM. We also used SQL3 and PL/SQL for implementation on Oracle 10g server. The size of the System Global Area (SGA) in Oracle server is set to 1GB.

Data sets The used data sets are synthetic with the following parameters (1) N the size of the database, (2) $\%IR$ the imperfection rate of data, i.e., number of imperfect tuples over N, (3) nfe the maximum number of focal elements per *bba*, (4) sfe the maximum size of each focal element (5) D number of attributes and (6) d_i size of attribute domain.

To generate a synthetic evidential database, the used algorithm uses a main procedure that generates a synthetic *bba*. This procedure operates as follows: it computes randomly a fixed number of focal elements in the interval $[1, nfe]$. Then, for each focal element, it generates randomly a number in the interval $[1, sfe]$; that's the size of current focal element. Each hypothesis in the focal element is randomly generated in the interval $[1, d]$, d_i being the cardinality of our attribute domain. Masses of focal elements are generated in the interval $[0, 1]$. We used the random function of JAVA which is based on the uniform law.

We managed several constraints like uniqueness of hypotheses into one focal element, uniqueness of focal elements in a *bba*, and normalization of a *bba* (sum of focal elements' masses must be equal to one). Then, we generated for each tuple one *bba* per attribute. We repeat this operation for $\%IR$ of N. Remaining tuples are perfect and contain in every attribute one *bba* with one singleton focal element whose mass is equal to one.

Evaluation The experiments showed interesting results from a performance point of a view although it has some limits when some parameters of the database reach some thresholds.

Default values used in experiments are fixed to $N = 1000$, $nfe = 4$, $sfe = 2$, $d_i = 10$ and $\%IR = 70$. In the first experiments, we varied the parameter N from 1000 tuples to 500000. The system crashed when we set $N = 500000$. For $N = 100000$ the system answers our test query in approximately 17 seconds. Experiments produced very acceptable time of execution until the limit of 60000 tuples (about 6 seconds at the worst). Figure 3.2(a) shows the experiment results.

Then, we varied the number of focal elements. Until $nfe = 50$, the system answers with acceptable execution time, without exceeding 4 seconds. We judge the value of 50 as appropriate, because an expert (the doctor in our example) does not give a such great number of believable hypotheses (in our case diagnoses). Results of this experiment are shown in Figure 3.2(b).

For focal elements' sizes, we varied the parameter sfe from 2 to 10. Figure 3.2(c) shows the results. Execution time did not exceed 0.3 seconds. This acceptable performance is due to *bel_Comp* (respectively *pl_Comp*) PL/SQL implementation. First of all, PL/SQL is integrated in Oracle (such that every transactional programming language), which reduces I/O costs. Second, scanning focal elements for belief/plausibility computations' amounts to scanning Oracle nested tables. Searching these structures is optimized by Oracle which offers for them sequential and direct access.

Each attribute is characterized by its domain cardinality. It refers to the frame of discernment size. We varied this parameter from 10 to 1000, the results are presented in Figure 3.2(d). The performance is again very acceptable although we reached a high size of cardinality (1000). Note that complexity of a generated *bba* also depends on number and size of focal elements that are controlled by parameters nfe and sfe. If we do not control these parameters, we can reach 21000 focal elements, with sizes that could reach 1000. This situation is not realistic, because expressing a so huge number of focal elements in one *bba* is impossible.

To mimic real imperfect databases, our solution should process objects that are either perfect or imperfect. To show the impact of imperfect objects on our solution performance, we varied the rate of imperfect objects in the database from 20% to 100%. It's logical that performance decreases in case of high values of the imperfection rate. Processing evidential *bbas* v.s. certain *bbas* involves more belief/plausibility computations. However in general, the performance was acceptable (it didn't exceed 0.5 seconds) even when the rate reached 100%.

An important feature of commercial databases consists in using memory caches to speed-up queries answering. In general, two memory caches are used in most of commercial solutions: (i) the database buffer cache, for keeping previous extracted data; (ii) the shared pool for keeping syntactic and execution information of previous queries. To evaluate the contribution of caching data and queries in our solution, we compared first execution of the test query (without cache), with next executions(with cache). Naturally, the difference is clear and expected. Table 3.8 shows the result of this experimentation for different values of the parameter n (size of the database).

Database size	**First execution time (s)**	**Next executions' times (s)**
1000	0.2	0.03
5000	0.8	0.04
10 000	4.1	0.06
50 000	5.8	0.12
70 000	6.4	0.16
100 000	17	1.2

Table 3.8: Contribution of caches to queries re-execution

3.4.2 Evidential Top-k Query

Top-k queries, also known as *Ranking queries*, were introduced in the multimedia systems by Fagin (Fagin, 1996; Fagin, 1998).They represent a powerful tool to order queries' results and give only the most interesting answers. Generally, top-k queries rank the results using a defined *score function* where only the k ($k \geq 1$) most important ones are returned; i.e, only answers with the highest scores are returned.

Top-k queries are needed in real worlds applications: for example movies can be ordered by the most watched ones, music can be ranked by the most listened songs, researchers can be ranked by their H-index, athletes by their race time, etc.

Several top-k processing techniques exist in the literature. In the uncertain data context, they can be classified into three categories (Ilyas et al., 2008):

- *Exact methods over certain data*, where top-k queries and data are deterministic. The majority of top-k processing techniques are based on exact methods and certain data.

- *Approximate methods over certain data*, where processing top-k queries over certain data produces approximate results (Amato et al., ; Theobald et al., 2005).

- *Methods over uncertain data*, where top-k processing techniques deal with imperfect data. The top-k queries are based on different uncertainty models. At the best of our knowledge, only top-k queries' approaches that deal with probabilities exist in the literature (Re et al., 2007; Soliman et al., 2007) but there is no work that deals with other types of imperfect data.

Similarly to the probabilistic top-k queries (Soliman et al., 2007), *Evidential Top-k queries* should return the k answers that respond to an evidential query with the highest scores based on a scoring function that takes into consideration the degrees of imperfection in the database. Indeed, we introduced a new type of uncertain top-k queries; the *Evidential Top-k Queries* (Bousnina et al., 2017a) that we apply over an evidential relation. Thus, we introduced as a first step a new scoring function for evidential data that returns an interval bounded by a belief and a plausibility. To rank the evidential scores, we relied on the method of comparison of (Wang et al., 2005). We also presented a new imperfect top-k semantics specific to the evidential scenario.

Formalism

Processing queries over evidential databases gives answers, each one quantified with a degree of confidence in a form of interval. That degree reflects the lower and the upper bounds of trust in that response which is calculated from the database. In this case, one can not rank such answers, so the user is not able to choose the most interesting ones from the set of results according to some defined criteria. In order to provide the decision maker with a method to select the k best answers to this query, we need to introduce a new top-k approach specific for evidential databases.

Table 3.9 presents an example of an evidential table that stores some users' appreciations about books: b_1, b_2, b_3, b_4. It is a relation with three attributes: The first one is ID that represents the identifier of a specific reader. The second attribute is $BookRate$ where the reader expresses its preferred books using the evidence theory[1]. The uncertainty here is of the attribute level nature. The third attribute is CL, it stores the interval of confidence about the user's appreciations (it deals with uncertainty at the tuple level).

ID	BookRate	CL
1	b_1 0.3 $\{b_2, b_3\}$ 0.7	[0.5;1]
2	b_2 0.5 b_4 0.5	[0.3;0.8]
3	$\{b_1, b_2, b_3\}$ 1	[1;1]
4	b_3 1	[0.5;0.9]

Table 3.9: Books Appreciations' Table: BAT

[1] The literature is abundant in term of methods of preference elicitation using the evidence theory. We cite two main works (Ben Yaghlane et al., 2008; Ennaceur et al., 2014)

The new formalism of ranking evidential results is based on several steps. First, we process a query Q over an evidential database. Then, for each generated response an *Evidential Score* is computed. That score is an interval of belief and plausibility, defined as follows:

Defintion 15 *(Evidential Score) Let R_i be a response generated from processing a query Q over an evidential database EDB, $S(R_i)$ is the score function of R_i and $bel(R_i)$ and $pl(R_i)$ are respectively its belief and plausibility in the table.*

$$S(R_i) = [bel(R_i); pl(R_i)] \tag{3.7}$$

$$Where \qquad bel(R_i) = \frac{\sum_{l=1}^{N} bel_l(R_i) * bel_l}{N}$$

$$pl(R_i) = \frac{\sum_{l=1}^{N} pl_l(R_i) * pl_l}{N}$$

Note that bel_l and pl_l are respectively the belief and the plausibility of the tuple l, defined on $\Theta_{CL} = \{exist, \overline{exist}\}$.

The belief of an answer, $bel(R_i)$, is a disjunction of the response's beliefs in each object of the database. The belief of a response in one object l is the product of its belief in the attribute and the belief of that object. Same for the plausibility of an answer, $pl(R_i)$. It is the disjunction of the response's plausibilities in each object of the database where the plausibility of a response in one object t is the product of its plausibility in the attribute and the plausibility of that object (Bell et al., 1996; Lee, 1992a).

Example 38 *Let's process the query Q over the evidential Table 3.9 in order to get the most two appreciated books (k=2) by all readers in the table.*

```
Q2: SELECT BookRate FROM BAT ORDER BY S(BookRate) LIMIT 2;
```

The score of each item in the relation that may be a response to the query Q is computed as follows:

- *The first possible response is book b_1, it appears in tuples l_1 and l_3. Therefore:*

$$\begin{cases} bel(b_1) = \dfrac{(0.3 * 0.5) + (0 * 0.3) + (0 * 1) + (0 * 0.5)}{4} \\ pl(b_1) = \dfrac{(0.3 * 1) + (0 * 0.8) + (1 * 1) + (0 * 0.9)}{4} \end{cases}$$

Thus:
$S(b_1) = [bel(b_1); pl(b_1)] = [0.0375; 0.325]$

- *The second possible response is book b_2, it appears in tuples l_1, l_2 and l_3. Therefore:*

$$\begin{cases} bel(b_2) = \dfrac{(0*0.5)+(0.5*0.3)+(0*1)+(0*0.5)}{4} \\ pl(b_2) = \dfrac{(0.7*1)+(0.5*0.8)+(1*1)+(0*0.9)}{4} \end{cases}$$

Thus:
$S(b_2) = [bel(b_2); pl(b_2)] = [0.0375; 0.525]$

- *The third possible response is book b_3, it appears in objects l_1, l_3 and l_4. Therefore:*

$$\begin{cases} bel(b_3) = \dfrac{(0*0.5)+(0*0.3)+(0*1)+(1*0.5)}{4} \\ pl(b_3) = \dfrac{(0.7*1)+(0*0.8)+(1*1)+(1*0.9)}{4} \end{cases}$$

Thus:
$S(b_3) = [bel(b_3); pl(b_3)] = [0.125; 0.65]$

- *The final possible answer is book b_4, it appears only in object l_2. Therefore:*

$$\begin{cases} bel(b_4) = \dfrac{(0*0.5)+(0.5*0.3)+(0*1)+(0*0.5)}{4} \\ pl(b_4) = \dfrac{(0*1)+(0.5*0.8)+(0*1)+(0*0.9)}{4} \end{cases}$$

Thus:
$S(b_4) = [bel(b_4); pl(b_4)] = [0.0375; 0.1]$

The computed evidential scores are shown in Table 3.10.

Item	*EvidentialScore*
b_1	R_1= [0.0375 ; 0.325]
b_2	R_2= [0.0375 ; 0.525]
b_3	R_3= [0.125 ; 0.65]
b_4	R_4= [0.0375 ; 0.1]

Table 3.10: Evidential Score per Item

Classical Top-k queries are based on a defined score function that produces precise values. However, the evidential top-k queries are based on a score function that produces intervals bounded by belief and plausibility values. In (Wang et al., 2005), authors introduced an approach of ranking intervals based on preference degrees. Their method is the one that we adopted to rank scores previously generated.

Defintion 16 *(Preference Degree) Let $S(R_i)$=[$bel_i; pl_i$] and $S(R_j)$=[$bel_j; pl_j$] be two evidential scores. Each one is an interval composed of a degree of belief and a degree of plausibility. The degree of one interval to be greater than the other one is called a* degree of preference *and denoted P.*

The degree of preference that $S(R_i) > S(R_j)$ is defined such that:

$$P(S(R_i) > S(R_j)) = \frac{max(0, pl_i - bel_j) - max(0, bel_i - pl_j)}{(pl_i - bel_i) + (pl_j - bel_j)} \tag{3.8}$$

The degree of preference that $S(R_i) < S(R_j)$ is defined such that:

$$P(S(R_i) < S(R_j)) = \frac{max(0, pl_j - bel_i) - max(0, bel_j - pl_i)}{(pl_i - bel_i) + (pl_j - bel_j)} \tag{3.9}$$

The different cases of comparing intervals $S(R_i)$ and $S(R_j)$ are as follows :

- *If $P(S(R_i) > S(R_j)) > P(S(R_j) > S(R_i))$, then $S(R_i)$ is said to be superior to $S(R_j)$, denoted by $S(R_i) \succ S(R_j)$.*

- *If $P(S(R_i) > S(R_j)) = P(S(R_j) > S(R_i)) = 0.5$, then $S(R_i)$ is said to be indifferent to $S(R_j)$, denoted by $S(R_i) \sim S(R_j)$.*

- *If $P(S(R_j) > S(R_i)) > P(S(R_i) > S(R_j))$, then $S(R_i)$ is said to be inferior to $S(R_j)$, denoted by $S(R_i) \prec S(R_j)$.*

Theorem 3.1 *Let $S(R_i)$=[$bel_i; pl_i$] and $S(R_j)$=[$bel_j; pl_j$] be two evidential scores such that:*

- *Case 1: if $S(R_i) = S(R_j)$ then*

  $P(S(R_i)) > P(S(R_j)) = P(S(R_i)) < P(S(R_j)) = 0.5.$

- *Case 2: if $bel_i \geq pl_j$ then $P(S(R_i) > S(R_j)) = 1$.*

- *Case 3: if $bel_i \geq bel_j$ and $pl_i \geq pl_j$ then*

  $P(S(R_i) > S(R_j)) \geq 0.5$ *and* $P(S(R_j) > S(R_i)) \leq$ *0.5.*

In order to detect the dominant interval between the score of relation R_i denoted $S(R_i)$ and the score of relation R_j denoted $S(R_j)$, we need to compute the degree of preference that $S(R_i) > S(R_j)$ and the degree of preference that $S(R_i) < S(R_j)$. The complexity of this computation can be reduced thanks to the *complementarity* of $P(S(R_i) > S(R_j))$ and $P(S(R_i) < S(R_j))$.

The complementarity is only feasible when :

$$\begin{cases} S(R_i) \neq S(R_j) \\ bel_i < pl_j \end{cases} \tag{3.10}$$

Proof (Complementarity)

$$P(S(R_i) < S(R_j)) = \frac{max(0, pl_j - bel_i) - max(0, bel_j - pl_i)}{(pl_i - bel_i) + (pl_j - bel_j)}$$

$$P(S(R_j) < S(R_i)) = \frac{max(0, pl_i - bel_j) - max(0, bel_i - pl_j)}{(pl_i - bel_i) + (pl_j - bel_j)}$$

$P(S(R_i) < S(R_j)) + P(S(R_j) < S(R_i))$

$$\begin{aligned} &= \frac{max(0, pl_j - bel_i) - max(0, bel_j - pl_i)}{(pl_i - bel_i) + (pl_j - bel_j)} \\ &+ \frac{max(0, pl_i - bel_j) - max(0, bel_i - pl_j)}{(pl_i - bel_i) + (pl_j - bel_j)} \\ &= \frac{max(0, pl_j - bel_i) - 0 + max(0, pl_i - bel_j) - 0}{(pl_i - bel_i) + (pl_j - bel_j)} \\ &= \frac{pl_j - bel_i + pl_i - bel_j}{pl_i - bel_i + pl_j - bel_j} = 1 \end{aligned}$$

$P(S(R_i) < S(R_j)) + P(S(R_j) < S(R_i)) = 1$

□

Defintion 17 *(Optimized Preference Degree)*

Let $S(R_i)$=[$bel_i; pl_i$] and $S(R_j)$=[$bel_j; pl_j$] be two evidential scores. Every interval is composed of degrees of belief (bel) and plausibility (pl) and P is the calculated preference degree.

$$P(S(R_i) > S(R_j)) = \frac{max(0, pl_i - bel_j) - max(0, bel_i - pl_j)}{(pl_i - bel_i) + (pl_j - bel_j)} = \lambda \tag{3.11}$$

The different cases of comparing intervals $S(R_i)$ and $S(R_j)$ are as follows:

- *If $\lambda >$ 0.5 then $S(R_i) \succ S(R_j)$.*
- *If $\lambda =$ 0.5 , then $S(R_i) \sim S(R_j)$.*
- *If $\lambda <$ 0.5 then $S(R_i) \prec S(R_j)$.*

Figure 3.3 summarizes the different cases of evidential scores intervals. It indicates also which property to use for each case.

The transitivity property is necessary to achieve a complete ranking order for scores. In (Wang et al., 2005), authors proved that preference relations are *transitive.*

Property 1 *(Transitivity)*

Let $S(R_i) = [bel_i; pl_i]$, $S(R_j) = [bel_j; pl_j]$ and $S(R_k) = [bel_k; pl_k]$ be three intervals. If $S(R_i) \succ S(R_j)$ and $S(R_j) \succ S(R_k)$ then $S(R_i) \succ S(R_k)$.

Previous definitions provide a total ranking of answers that respond to the proposed top-k query. But how to interpret any evidential answer ?

The top-k queries in deterministic databases are semantically clear. However, the interpretation of top-k queries in imperfect databases is challenging. In (Soliman et al., 2007), authors introduced new semantics relative to probabilistic top-k queries. They defined them as *the most probable query answers.* Their work is based on the possible worlds' model and they proposed interpretations like: (i) *The top-k tuples in the most probable world.* (ii) *The most probable top-k tuples that belong to a valid possible world.*

The interpretations of probabilistic top-k queries can not be considered for evidential top-k queries. Thus, a new specific semantics for *Evidential Top-k Queries* is defined as follows:

Defintion 18 *(E-Top-k)*

Let EDB be an evidential database with N objects and A attributes; CL is an attribute where the intervals associated to objects reflects the degrees of confidence about these objects. Let $S(R_i)$ be a score function that maximizes both CL and the interval of belief on each result. Responses R are ordered according to the computed scores.

An E-top$k$ *returns* the k most credible answers from the set of answers *such that:*

$$S(R_i) = Argmax_{R_i \in R}([bel(R_i); pl(R_i)]) \tag{3.12}$$

Example 39 *Let's carry on with the same example of Table 3.9. We want to give a total ranking of the resulting evidential scores and we want to deduce the top-2 answers and their semantics.*

(i) *Since* $bel_{b_1} = bel_{b_2} = bel_{b_4}$ *and* $pl_{b_2} > pl_{b_1} > pl_{b_4}$

then $b_2 \succ b_1 \succ b_4$

(ii) *Since* $bel_{b_3} > bel_{b_2}$ *and* $pl_{b_3} > pl_{b_2}$

then $b_3 \succ b_2$

The final ranking deduced from (i) and (ii) is: $b_3 \succ b_2 \succ b_1 \succ b_4$.
The Top-2 appreciated books are :

- b_3 *with a confidence level [0.125 ; 0.65]*
- b_2 *with a confidence level [0.0375 ; 0.525]*

Books b_3 and b_2 are the most appreciated credible answers from the set of results.

Algorithms

The Object-Relational implementation of Evidential top-k queries (Bousnina et al., 2017a) involves two methods to handle the evidential scores (evidential intervals):

- The first method is naive, it consists in computing the preference degrees through three steps each time: (a) it computes the preference degree that the first interval is superior to the second one and then (b) it computes the preference degree that the first interval is inferior to the second one, (c) finally, it compares results and gives the partial ranking. This algorithm is presented in Table 3.11.
- The second method is an optimization of the first one. Indeed, it consists in computing only in one step the preference degree and then deduces the partial order between two intervals. This algorithm is detailed in Table 3.12.

After that, the final order is treated using a sorting algorithm, that ranks all evidential intervals and provides the k most interesting ones. The presented implementations offer two methods of evidential intervals' ranking. Both algorithms use the object-oriented paradigm for its programming benefits.

Naive Method

```
Initialization
Tuple a, b ;
begin
  if (a.bel=b.bel and a.pl=b.pl)
    return 0;
  if (a.pl<b.bel)
    return -1;
  if (b.pl<a.bel)
    return 1;
  if (a.bel>b.bel and a.pl>b.pl)
    return 1;
  if (b.bel>a.bel and b.pl>a.pl)
    return -1;
  if (score(a,b)>score(b,a))
    return 1;
  else return -1;
end
```

Naive Evidential Top-k Algorithm

```
Initialization
Integer m;
ArrayList Table;
begin
  for (int i ⟵ 0; i<Table.size()-1; i++)
  { m⟵i;
    for (int j⟵ i+1; j<Table.size(); j++)
    {
      if (NaiveMethod(Table.get(j), Table.get(m))=1)
      { m⟵j; }
    }
    if (NaiveMethod(Table.get(m) ,Table.get(i))=1)
    { Tuple c ⟵ Table.get(i);
      Table.set(i,Table.get(m));
      Table.set(m,c); }
  }
end
```

Table 3.11: ETop-k Naive Algorithm

Experiments

The Object-Relational implementation of evidential top-k queries (Bousnina et al., 2017a) allowed us to evaluate this new operator based on the Object-Relational implementation of the compact evidential database (Bousnina et al., 2016). Indeed, we implemented the evidential top-k operator using two methods of partial ranking scores (a naive and an optimized ones).

ETop-k Method	**Optimized ETop-k Algorithm**
Initialization	***Initialization***
Tuple a, b ;	Integer m;
begin	ArrayList Table;
if (a.bel=b.bel ***and*** a.pl=b.pl)	***begin***
return 0;	***for***(int i⟵0; i<Table.size()-1; i++)
if (a.pl<b.bel)	{
return -1;	***for*** (int j⟵i+1; j<Table.size(); j++)
if (b.pl<a.bel)	{
return 1;	***if***(*EtopKMethod*(Table.get(j), Table.get(m))=1)
if (a.bel>b.bel ***and*** a.pl>b.pl)	{m⟵j;}
return 1;	}
if (b.bel>a.bel ***and*** b.pl>a.pl)	***if***(*EtopKMethod*(Table.get(m) ,Table.get(i))=1)
return -1;	{ Tuple c ⟵ Table.get(i);
if (score(a,b)>0.5)	Table.set(i,Table.get(m));
return 1;	Table.set(m,c); }
else return -1;	}
end	***end***

Table 3.12: ETop-k Optimized Algorithm

Data sets We evaluated both algorithms from a performance point of view. We used a windows 10 operating system with 2.10 GHz CPU and 4 GB RAM. We also used Java programming language and NetBeans platform. We used synthetic data sets with the following parameters (a) N the size of the database, (b) S the evidential score which is an interval of belief and plausibility $[bel; pl]$ with bel, pl $\in$ [0;1] and bel $\leq$ pl[2]. To generate a synthetic evidential database, the used algorithm uses a procedure that generates a synthetic S. Indeed, the procedure computes randomly a fixed number of evidential scores in the interval $[0, 1]$. Then, we process the algorithms (naive or optimized) in order to rank the intervals (scores). Finally, a sorting function is used to provide the final complete ranking of all intervals. Note that each interval is associated to one unique item in the evidential database. In our example, the item is a book.

Evaluation Experiments showed interesting results from a performance point of view. In fact, we varied the database size parameter (N) from 10 to 3000. The execution time did not exceed 4 minutes and 50 seconds for both algorithms. Results of the impact of the database size for both methods is presented in Table 3.13. Both algorithms showed interesting results. Moreover, $OptETopK$ gave better ones as shown in Figure 3.4. For example, $OptETopK$ ranked 1500 tuples in 69 seconds against 60 seconds for

[2] bel and pl are the two functions defined in the object-relational implementation of evidential databases in (Bousnina et al., 2016)

$NaiETopKNote$. Note that complexity depends also on the intervals' nature generated randomly as detailed theoretically above.

Tuples Number (N)	**Execution Time (s)**	
	NaiETopK Method	**OptETopK Method**
10	1	0
50	2	0
100	2	0
200	3	0
300	4	1
500	8	5
800	19	13
1000	33	28
1500	69	60
2000	125	121
3000	279	277

Table 3.13: Impact of the database size for methods: NaiTopK and OptTopK

3.4.3 Evidential Skyline Query

The skyline operator (Börzsönyi et al., 2001) filters answers from a database to give only those tuples that are not worse than any other. For example, a user wants to select the cheapest and the closest hotels to the beach. The skyline operator in this case selects those hotels that are not worse than any others in price and distance. The evidential skyline operator (Elmi et al., 2014), denoted e-sky, is based on the belief and plausibility dominance in order to give answers that are not worse than any other.

This evidential skyline operator was applied over some reviews collected from a real platform: the $TripAdvisor$; $TripAdvisor$ is one of the most well known crowd-sourcing platforms where travelers express their opinions about hotels they visited through an evaluation form. Since the $TripAdvisor$ platform do not answer a multi-criteria request. Then, we proposed to use these collected reviews to answer users' queries about the best hotels regarding some criteria like distance, price, etc.

$TripAdvisor$ provides a reviewing form for travelers in order to evaluate hotels according to several criteria. A response about one criteria for a specific hotel can be in {-1;1;2;3;4;5}. If the response is in {1;2;3;4;5}, it is precise and certain. Thus, it induces a precise and certain belief function. If the response is -1, then it reflects the total ignorance. These reviews need to be transformed and then stored in the real $TripAdvisor$ evidential database, to be later queried with e-sky. Four major steps are made:

- *Construction of mass functions*: A reviewer response is translated into a *bba*, in the context of belief functions theory.
- *Reliability estimation and discounting*: Reviews are discounted based on the reliability of each reviewer.
- *Combination of reviews*: All responses to the same review (same hotel and same criteria) are combined in order to provide one *bba* that summarizes all the reviewers' evaluations.
- *Evidential skyline query*: The e-sky is applied over the obtained $TripAdvisor$ evidential database.

Construction of Mass Functions

Belief functions theory allows the construction of basic belief assignments (*bbas*) from the set of hypotheses. The mass of an hypothesis A as modeled in Equation (2.2) and denoted, $m(A)$, is interpreted as the degree of support given by an expert and that reflects his belief on that response A. This mass can not be divided on subsets of A.

Reviewers	*Hotels*	*Price*	*Place*	*Service*	*Score*
Rev_1	h_1	3	4	3	1510
Rev_2	h_1	-1	4	2	22800
Rev_3	h_2	4	-1	5	400
Rev_4	h_2	3	5	-1	8140

Table 3.14: Reviews about Hotels

In *TripAdvisor* platform, each traveler chooses one rate from 1 to 5. If he does not provide a rate, his response is interpreted as -1. From the evidence theory point of view, the frame of discernment is $\Theta = \{1, 2, 3, 4, 5\}$ and -1 is interpreted as total ignorance, $m^{\Theta}(\Theta) = 1$. Each non empty response is interpreted as certain and precise belief functions over Θ.

Example 40 *The first reviewer Rev_1 gives a rate 3 for the service of hotel h_1. Its response is interpreted as $m^{\Theta}(3) = 1$.*

Table 3.15 is an interpretation of Table 3.14, in the context of belief functions' theory for criteria: Price, Place and Service.

Reviewers	*Hotels*	*Price*	*Place*	*Service*	*Score*
Rev_1	h_1	$m^{\Theta}(3) = 1$	$m^{\Theta}(4) = 1$	$m^{\Theta}(3) = 1$	0.136
Rev_2	h_1	$m^{\Theta}(\Theta) = 1$	$m^{\Theta}(4) = 1$	$m^{\Theta}(2) = 1$	0.99
Rev_3	h_2	$m^{\Theta}(4) = 1$	$m^{\Theta}(\Theta) = 1$	$m^{\Theta}(5) = 1$	0.036
Rev_4	h_2	$m^{\Theta}(3) = 1$	$m^{\Theta}(5) = 1$	$m^{\Theta}(\Theta) = 1$	0.733

Table 3.15: Construction of mass functions

The obtained mass functions need to be combined in order to have only one *bba* for each object. However, the reviews must be discounted by the estimated reliabilities before their combination.

Reliability Estimation and Discounting

The conflict may appear according to different cases (Martin, 2019). in our case, the conflict reflects the unreliability of at least one of the experts' opinions. A user's estimated reliability is used to weaken its given opinions modeled through the basic belief assignments (*bbas*).

The *TripAdvisor* platform gives to each reviewer a number of points depending to its contributions. These points are accumulated when the traveler (reviewer) gives an opinion about a hotel that he visited. Figure 3.5(a) shows how the *TripAdvisor* rewards reviewers that add photos, videos, helpful reviews, etc. Added to that, *TripAdvisor* divides its reviewers into 6 levels, shown in Figure 3.5(b): the first level is assigned to travelers having 300 to 2499 points and the final and the sixth level is affected to travelers with points starting from 10.000. Method of rewarding travelers as illustrated in Figure 3.5 is fixed by the *TripAdvisor* platform.

We proposed to estimate the reliability of each reviewer based on points and levels given by the *TripAdvisor* platform. Thus, two methods are proposed: the first is to calculate a reliability for each reviewer having points from 300 to 10.000 relatively to the sixth level, as shown in Equation (3.13), and the second is to compute the reliability score for reviewers having more than 10.000 point (i.e, travelers that acquire the last level and accumulating more points), as shown in Equation (3.14).

The maximal score is fixed to 0.9 for the 10.000 points. Based on that, a reliability is computed for reviewers having points under 10.000, such that:

$$Score = (points * 0.9)/10.000 \tag{3.13}$$

When the number of points accumulated by a reviewer are greater that 10000, the reliability is computed such that:

$$Score = 1 - (1/points) \tag{3.14}$$

Figure 3.6 shows the reviewers' reliabilities according to accumulated points.

Example 41 *the first reviewer R_1 in Table 3.14 has accumulated* 1510 *points and since his number of points is lower than* 10000 *then his reliability score is computed using method (i): $Score_{R_1}$ = 1510 * 0.9 / 10000 =* 0.136*. The second reviewer R_2 has accumulated more points than* 10000 *then his reliability score is computed using method (ii): $Score_{R_2}$ = 1 - (1 /22800) =* 0.99*. Estimated reliabilities for all reviewers are shown in Table 3.15.*

The reliability estimated for each reviewer is used to discount the basic belief assignments that reflect their reviews about hotels using Equation (2.33).

Example 42 *The reviewer R_1, the reliability degree is α = 0.136. Thus:*
$m^{\alpha}_{Price}(3) = 0.136 * 1 = 0.136$
$m^{\alpha}_{Price}(\Theta) = 0.136 * 0 + (1 - 0.136) = 0.864$

Results of discounted mass functions are shown in Table 3.16.

Once the reviews, modeled as *bba*s, are discounted, they are combined.

Reviewers	*Hotels*	*Price*	*Place*	*Service*
Rev$_1$	h_1	$m^\Theta(3) = 0.136$ $m^\Theta(\Theta) = 0.864$	$m^\Theta(4) = 0.136$ $m^\Theta(\Theta) = 0.864$	$m^\Theta(3) = 0.136$ $m^\Theta(\Theta) = 0.864$
Rev$_2$	h_1	$m^\Theta(\Theta) = 1$	$m^\Theta(4) = 0.99$ $m^\Theta(\Theta) = 0.01$	$m^\Theta(2) = 0.99$ $m^\Theta(\Theta) = 0.01$
Rev$_3$	h_2	$m^\Theta(4) = 0.036$ $m^\Theta(\Theta) = 0.964$	$m^\Theta(\Theta) = 1$	$m^\Theta(5) = 0.036$ $m^\Theta(\Theta) = 0.964$
Rev$_4$	h_2	$m^\Theta(3) = 0.733$ $m^\Theta(\Theta) = 0.267$	$m^\Theta(5) = 0.733$ $m^\Theta(\Theta) = 0.267$	$m^\Theta(\Theta) = 1$

Table 3.16: Discounting of mass functions

Combination of Reviews

In theory of belief functions, combination rules aggregate data from different sources to get one mass function that reflects all sources' opinions.

Example 43 *Reviews about hotel h_2 for attribute Price are combined as shown in Table 3.17. Note that m_3^Θ is the mass function given by reviewer 3 and that m_4^Θ is the mass function given by reviewer 4.*

Price $\setminus h_2$	$m_3^\Theta(\Theta) = 0.964$	$m_3^\Theta(4) = 0.036$
$m_4^\Theta(\Theta) = 0.267$	$m^\Theta(\Theta) = 0.26$	$m^\Theta(4) = 0.01$
$m_4^\Theta(3) = 0.733$	$m^\Theta(3) = 0.7$	$m^\Theta(\emptyset) = 0.03$

Table 3.17: Combination of *bba*s about the Price of h_2

The joint mass of reviewers Rev$_3$ *and Rev*$_4$, $m_{3\oplus4}^\Theta$ *about the price of hotel h_2 is:*
(i) $m_{3\oplus4}^\Theta(3) = 1/(1-0.03)$ ** 0.7 = 0.72*
(ii) $m_{3\oplus4}^\Theta(4) = 1/(1-0.03)$ ** 0.03 = 0.012;*
(iii) $m_{3\oplus4}^\Theta(\Theta) = 1/(1-0.03)$ ** 0.26 = 0.268.*

Similarly, we combine all bbas for each attribute for the different hotels. The obtained evidential database EDB is in Table 3.18.

The obtained database is evidential with either precise *bba*s, or partial ignorance *bba*s. This *EDB* is then queried with preference conditions using the skyline operator. Preference conditions may deal either with one attribute like Price, Place or Service or with a combination of these attributes leading to the multi criteria filtering.

$Hotels$	$Price$	$Place$	$Service$
h_1	$m^{\Theta}(3) = 0.136$ $m^{\Theta}(\Theta) = 0.864$	$m^{\Theta}(4) = 0.9814$ $m^{\Theta}(\Theta) = 0.0086$	$m^{\Theta}(2) = 0.98$ $m^{\Theta}(3) = 0.01$ $m^{\Theta}(\Theta) = 0.01$
h_2	$m^{\Theta}(3)= 0.72$ $m^{\Theta}(4) = 0.012$ $m^{\Theta}(\Theta) = 0.268$	$m^{\Theta}(4) = 0.992$ $m^{\Theta}(\Theta) = 0.008$	$m^{\Theta}(5) = 0.036$ $m^{\Theta}(\Theta) = 0.964$

Table 3.18: The obtained Evidential Database from *TripAdvisor*

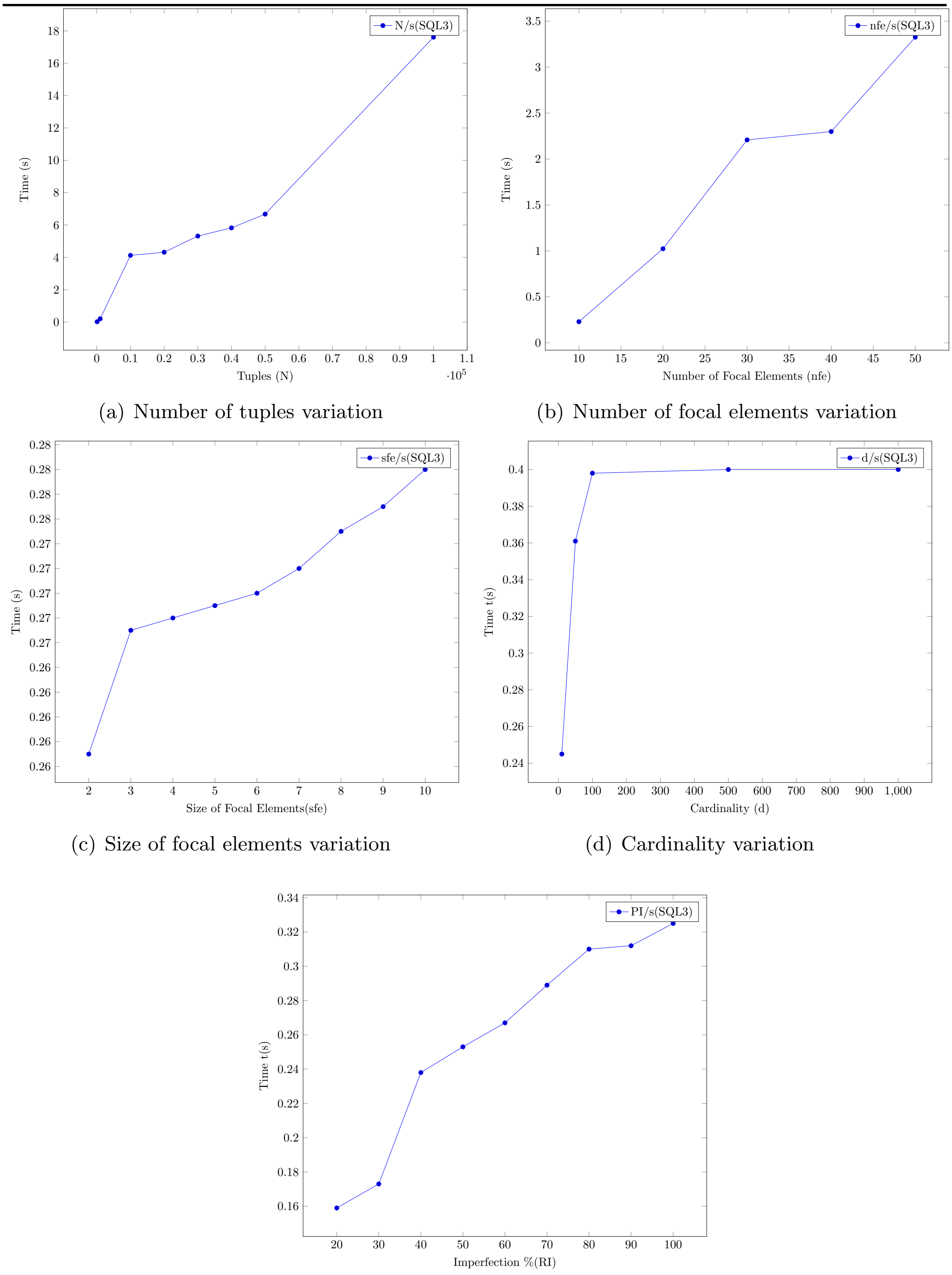
Evidential Skyline Operator

The evidential skyline operator (Elmi et al., 2014) is based on two methods: the b-dominance and the p-dominance; i.e, the skyline is computed using the dominance of beliefs and plausibilities. This formalism was applied over the extracted data from the *TripAdvisor* platform. As a consequence, the evidential skyline query in the *TripAdvisor* context (Bousnina et al., 2017b) ameliorates the skyline performance compared to the basic evidential skyline (Elmi et al., 2014). This amelioration is due to the simple modeling of reviews in the evidential table which reduces the computations' cost.

3.5 Conclusion

In this chapter, we presented the most known and used evidential database model on its compact form (Lee, 1992b; Lee, 1992a; Bell et al., 1996). In fact, we presented the Object-Relational implementation of this model using Java and SQL3 (Bousnina et al., 2016). The proposed implementation was used to apply queries like extended relational operators (Bell et al., 1996), evidential skyline and evidential Top-k (Bousnina et al., 2017b). The latter is a ranking query that provides the k best credible answers via an interval scores.

To evaluate the compact evidential model (EDB), we need to model it through its possible worlds' representation. The non compact form is fundamental to evaluate the querying methods of (EDB).



(a) Number of tuples variation

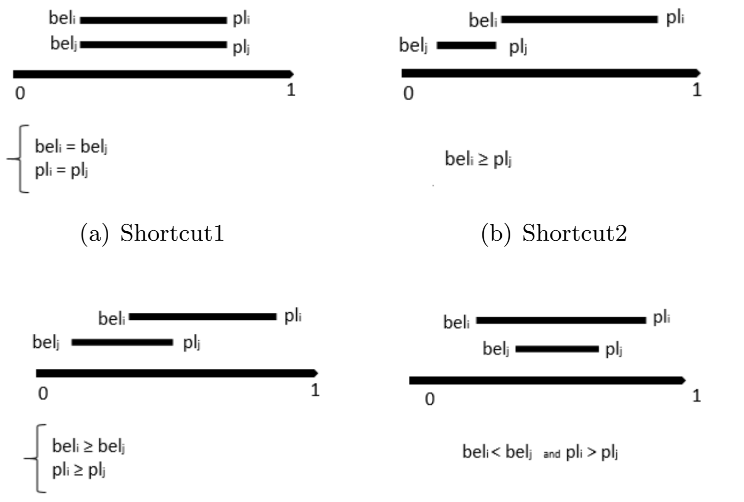
(b) Number of focal elements variation

(c) Size of focal elements variation

(d) Cardinality variation

(e) Imperfection rate variation

Figure 3.2: Experimentation Results



(a) Shortcut1

(b) Shortcut2

(c) Shortcut3

(d) Evidential Score / Complementarity

Figure 3.3: Comparison of Evidential Scores

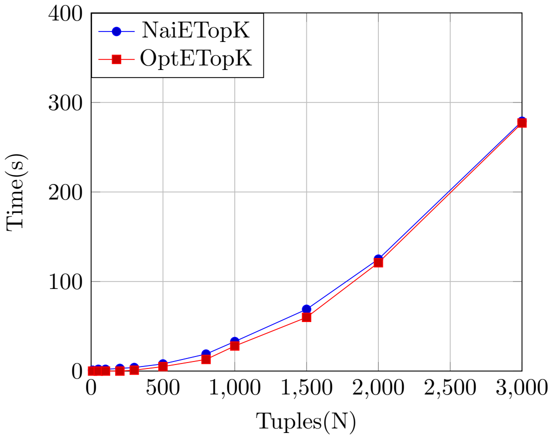
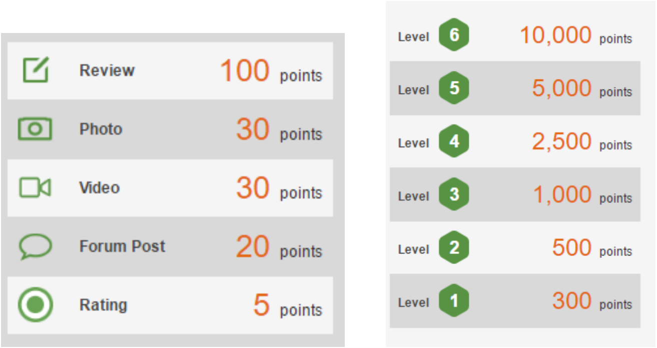


Figure 3.4: Comparison of performance of NaiETopK and OptETopK



Review	100 points
Photo	30 points
Video	30 points
Forum Post	20 points
Rating	5 points

(a) Point

Level 6	10,000 points
Level 5	5,000 points
Level 4	2,500 points
Level 3	1,000 points
Level 2	500 points
Level 1	300 points

(b) Levels

Figure 3.5: Computation of points in *TripAdvisor* and their corresponding levels

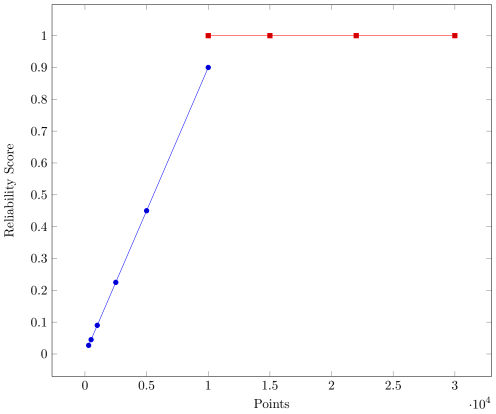


Figure 3.6: Estimated Reviewers' Reliabilities

4
Evidential Databases: The Possible Worlds' Form

Contents

4.1 Introduction 88

4.2 Representation Systems 88

4.3 Evidential Database as Possible Worlds 90

4.3.1 Imprecise Possible Worlds 92

4.3.2 Uncertain Possible Worlds 93

4.3.3 Possible Worlds 98

4.4 Querying Possible Worlds 106

4.5 Generalizing Probabilistic Databases 109

4.6 EDB: What Kind of Representation System? 110

4.7 Conclusion 114

Summary

This chapter is about modeling and querying evidential databases as possible worlds. First, we define representation systems. After that, we present the evidential databases (EDB) on their non compact form, i.e, the possible worlds and their querying process. Finally, we evaluate querying methods applied over this database model in order to determine what kind of representation system it is.

4.1 Introduction

The compact form is the only conceivable representation in practice. In fact, implementing and querying the possible worlds is expensive in terms of memory and CPU. However, the possible worlds' form remains fundamental; first, it defines clear semantics for the imperfect database, and second, it is used to validate querying methods over the compact form. Although, a model is considered as a strong representation system when querying the compact form is equivalent to querying the possible worlds of the imperfect database.

Several researches (Bell et al., 1996; Choenni et al., 2006; Lee, 1992a; Lee, 1992b) focused on modeling, querying and mining evidential databases on their compact forms. In (Bousnina et al., 2015), we proposed a method for modeling and querying an evidential database on its possible worlds form treating the attribute level uncertainty (ALU). Then, we generalized the evidential possible worlds model by considering the tuple level uncertainty (TLU) in addition to the attribute level uncertainty.

Table 4.1 is an example of an evidential table that stores blood types of two patients. The blood types' domain is $\Theta_{BT} = \Theta_{BTF} = \{A, B, O\}$. This table contains four attributes where *Blood Type*, *BloodTypeFather* involve imperfect information modeled thanks to the mass functions (to express the ALU) and the confidence level CL (to express the TLU).

ID	**BloodType**	**BloodTypeFather**	**CL**
1	A 0.5 B 0.5	A	[0.5;1]
2	A 0.3 $\{A, B\}$ 0.7	B 0.8 $\{A, B\}$ 0.2	[0.3;0.8]

Table 4.1: Medical Evidential Table about Blood Types

4.2 Representation Systems

The concept of *representation systems* was introduced by Imielinski and Lipski (Imielinski and Lipski, 1984). It constitutes the *way* of modeling an imperfect database from its compact form to its possible worlds' form and it is formally defined such that:

Defintion 19 *A Representation System is a set of instances and a function Rep that associates to each imperfect database DB the set of instances Rep(DB).*

Example 44 *Figure 4.1 illustrates an incomplete database denoted DB as well as $Rep(DB)$; its possible worlds representation.*

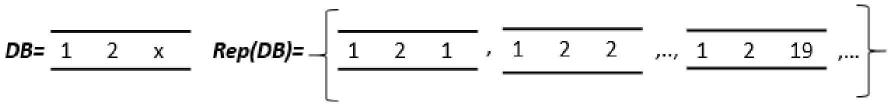


Figure 4.1: An imperfect database DB and its possible worlds Rep(DB)

A representation system (Abiteboul et al., 1995b) should be able to represent any imperfect database. This property is called *completeness* of the RS. In addition, it should be able to represent any query answers under the possible worlds' form. This property is called *closure* and is a consequence of the first one (the opposite is not true (Suciu et al., 2011)). More formally, completeness and closure are defined as follow:

Defintion 20 *(Completeness) A representation system is said to be* ***complete*** *if it can represent any imperfect database (Sarma et al., 2006).*

Defintion 21 *(Closure) A representation system is said to be* ***closed*** *under a query language if for any query Q and any database DB there is a database DB' that represents $Q(Rep(DB))$ (Sarma et al., 2006).*

A *Strong representation system*, denoted SRS, is closed when answers generated from the compact form and answers generated from the possible worlds form are *equivalent* (Abiteboul et al., 1995b). Figure 4.2 shows in details the steps towards the validation of any imperfect database representation system.

Defintion 22 *(Strong Representation System)*

An imperfect database DB is said to be a strong representation system if:

$$Rep(Q(DB)) \equiv Q(Rep(DB)) \tag{4.1}$$

Where $Rep(DB) = \{W_1..W_j\}$ is the set of possible worlds generated from DB and $\{W'_1..W'_k\}$ are possible worlds generated from $Q(DB)$.

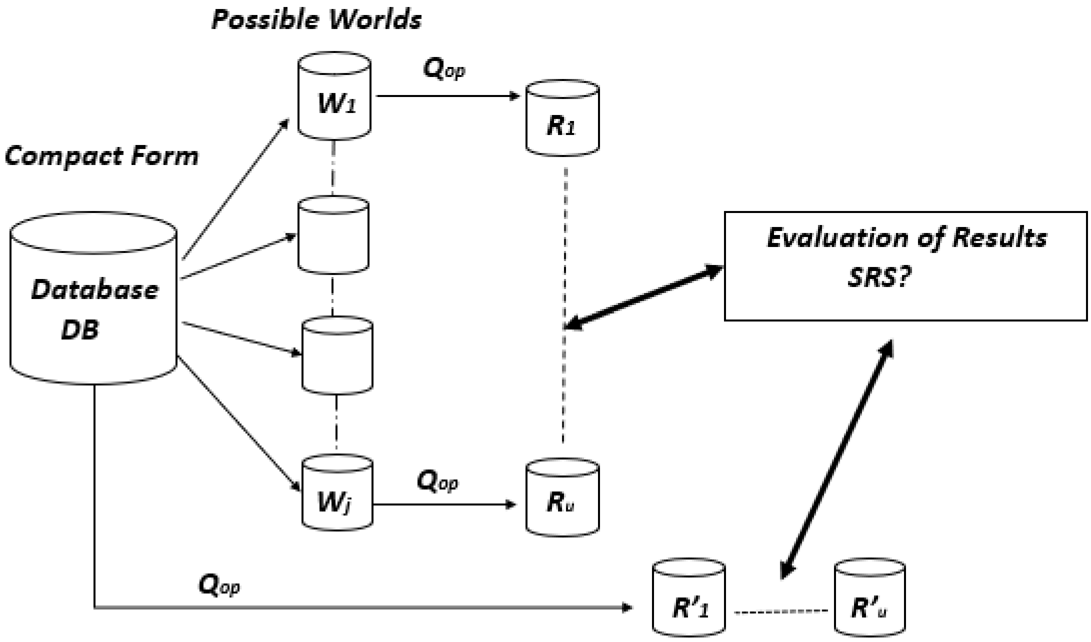


Figure 4.2: Querying formalism towards a *SRS*

4.3 Evidential Database as Possible Worlds

The definition of the possible worlds form is motivated by two reasons:

1. Semantic interpretation: Most of uncertainty theories represent imperfect information through a distribution of hypotheses where each one matches the solution with a degree of uncertainty. In the same way, the possible worlds form models an imperfect database with some degree of uncertainty. Thus, it is considered as a clearer semantics than the compact form. It represents an imperfect database as a distribution of candidate databases. Each candidate is a consistent database that can match to the real database with a certain degree of credibility.

2. Use for proofs: Querying an imperfect database should be equivalent to querying its possible worlds as stated in (Imielinski and Lipski, 1984) and well presented in (Abiteboul et al., 1995a). A querying model over imperfect databases is considered as a strong representation system when it is equivalent to querying their possible worlds.

In the remaining of this section, we explain how to produce possible worlds from a compact evidential database using the Dempster-Shafer tools. A summarizing schema in Figure 4.3; and a recapitulate example of Figure 4.4 are provided at the end of this section to illustrate the overall model.

The evidential non compact form is obtained by expanding the compact form into different states where each state is representative of the evidential database (compact form) with a degree of support. Thus, an evidential database EDB on its non compact form, is a set of evidential objects.

Defintion 23 *An evidential object is a basic belief assignment computed via the combination of all object's evidential values. Its frame of discernment Θ is the joint frame of all attributes' domains. The mass function m_t^{Θ} relative to the tuple t is the conjunctive combination of all evidential values m_{ta}^{Θ} issued from the vacuous extension to the joint frame. An evidential object is defined such that:*

$$m_t^{\Theta} = \circledcirc_{a\in[1,D]} m_{ta}^{\Theta_a \uparrow \Theta} \quad \textit{with } \Theta = \bigotimes_{a\in[1,D]} \Theta_a \tag{4.2}$$

Note that m_{ta} is the mass function of an attribute a for a tuple t, with D is the number of attributes.

Example 45 *Obtaining the evidential objects of Table 4.1 is based on two steps* [1]*:*

- *First, evidential values are extended to the joint frame Θ (using the vacuous extension given in Section 2.5.1) as shown in Table 4.2.*
- *Then, the extended evidential values are combined (using the conjunctive rule of combination described in Section 2.3.2) as detailed in Table 4.3.*

$\Theta_{ID} \uparrow \Theta$	$\Theta_{BT} \uparrow \Theta$	$\Theta_{BTF} \uparrow \Theta$	CL
$1 \times \Theta_{BT} \times \Theta_{BTF}$	$A \times \Theta_{ID} \times \Theta_{BTF}$ $B \times \Theta_{ID} \times \Theta_{BTF}$	$A \times \Theta_{ID} \times \Theta_{BT}$	[0.5;1]
$2 \times \Theta_{BT} \times \Theta_{BTF}$	$A \times \Theta_{ID} \times \Theta_{BTF}$ $\{A, B\} \times \Theta_{ID} \times \Theta_{BTF}$	$B \times \Theta_{ID} \times \Theta_{BT}$ $\{A, B\} \times \Theta_{ID} \times \Theta_{BT}$	[0.3;0.8]

Table 4.2: Vacuous extension of the focal elements of the evidential Table 4.1

[1] Note that Θ_{ID}, Θ_{BT} and Θ_{BTF} are respectively the frames of discernment of the attributes ID, $BloodType$ and $Father'sBloodType$

$\textcircled{\cap}$	CL
$(1, A, A)$ $0.5 = 1 \times 0.5 \times 1$ $(1, B, A)$ $0.5 = 1 \times 0.5 \times 1$	[0.5;1]
$(2, A, B)$ $0.24 = 1 \times 0.3 \times 0.8$ $(2, \{A, B\}, B)$ $0.56 = 1 \times 0.7 \times 0.8$ $(2, A, \{A, B\})$ $0.06 = 1 \times 0.3 \times 0.2$ $(2, \{A, B\}, \{A, B\})$ $0.14 = 1 \times 0.7 \times 0.2$	[0.3;0.8]

Table 4.3: Combination of the extended *bbas* using the conjunctive rule of combination

The expansion of the compact form leads to two intermediate results before getting the possible worlds form (non compact form), the first is the generation of the *imprecise worlds* and the second is the generation of *uncertain worlds.*

4.3.1 Imprecise Possible Worlds

The imprecise possible worlds form (Bousnina et al., 2015) is defined as follows:

Defintion 24 *The non-compact form of an EDB is a finite set of* imprecise possible worlds *such that $EDB = \{IW_1, IW_2, ..., IW_i\}$. Each imprecise possible world has N objects, where each object involves one focal element per attribute (Bousnina et al., 2015).*

The number of all imprecise possible worlds, I, is induced from sizes of sets $\mathcal{F}_{ta}$ ($\forall t \in [1; N]$ and $\forall a \in [1; D]$, where N is the size of tuples and D is the size of attributes) as follows:

$$I = \prod_{t=1}^{N} \prod_{a=1}^{D} | \mathcal{F}_{ta} | \tag{4.3}$$

Example 46 *The number of imprecise possible worlds generated from Table 4.1 is $I = 8$. Table 4.4 is an example of one imprecise world generated from Table 4.1.*

ID	$BloodType$	$BloodTypeFather$
1	A	A
2	A	{A,B}

Table 4.4: An Imprecise World

These worlds are qualified by *imprecise* because they include imprecise values as attributes' values. For example, $\{A, B\}$ is an imprecise value of the attribute

$BloodTypeFather$ for the second tuple in Table 4.4.

An evidential database EDB is the union of its evidential objects, similarly an evidential database on its non compact form is the disjunctive combination of its evidential objects (Bousnina et al., 2015).

Defintion 25 *A non-compact evidential database is presented via a basic belief assignment derived from the disjunctive combination of the set of its evidential objects. That mass is defined such that:*

$$m^{\Theta} = \circledcirc_{t\in[1,N]} m_t^{\Theta} \tag{4.4}$$

Example 47 *The disjunctive combination of all evidential objects is presented in Table 4.5. The bba* $m^{\Theta_{IW}}$ *is defined on* Θ_{IW}. $\Theta_{IW} = \{IW_1, IW_2, IW_3, IW_4, IW_5, IW_6, IW_7, IW_8\}$ *is the set of all imprecise possible worlds , denoted* IW_i.

$m^{\Theta_{IW}}(\{IW_1\}) = 0.12$
$m^{\Theta_{IW}}(\{IW_2\}) = 0.07$
$m^{\Theta_{IW}}(\{IW_3\}) = 0.12$
$m^{\Theta_{IW}}(\{IW_4\}) = 0.07$
$m^{\Theta_{IW}}(\{IW_5\}) = 0.03$
$m^{\Theta_{IW}}(\{IW_6\}) = 0.28$
$m^{\Theta_{IW}}(\{IW_7\}) = 0.03$
$m^{\Theta_{IW}}(\{IW_8\}) = 0.28$

The bba $m^{\Theta_{IW}}$ *reflects the degree of belief on which imprecise possible world represents the compact* EDB *in its imprecise state.*

The confidence level CLs *of the tuples in each imprecise world* IW_i *are inherited from objects' CL of the compact from (see Table 4.1). For example, The first tuple of* IW_1 *inherits its* CLs *from the first tuple of Table 4.1.*

Imprecise possible worlds are the first intermediate non compact form. They represent the compact form into several states but include imprecision. Uncertain possible worlds are the second intermediate non compact form after imprecise worlds. They treat the imprecision of the previous form but include uncertainty.

4.3.2 Uncertain Possible Worlds

Each imprecise possible world has different states called *Uncertain Possible Worlds* UW_j. In fact, each imprecise possible world can be expanded into several uncertain possible worlds by splitting imprecise values (attributes' values) into several possible precise

IW_1	
$EvidentialObjects$	CL
$(1, A, A)$ $(2, A, B)$	[0.5;1] [0.3;0.8]

IW_2	
$EvidentialObjects$	CL
$(1, A, A)$ $(2, \{A, B\}, \{A, B\})$	[0.5;1] [0.3;0.8]

IW_3	
$EvidentialObjects$	CL
$(1, B, A)$ $(2, A, B)$	[0.5;1] [0.3;0.8]

IW_4	
$EvidentialObjects$	CL
$(1, B, A)$ $(2, \{A, B\}, \{A, B\})$	[0.5;1] [0.3;0.8]

IW_5	
$EvidentialObjects$	CL
$(1, A, A)$ $(2, A, \{A, B\})$	[0.5;1] [0.3;0.8]

IW_6	
$EvidentialObjects$	CL
$(1, A, A)$ $(2, \{A, B\}, B)$	[0.5;1] [0.3;0.8]

IW_7	
$EvidentialObjects$	CL
$(1, B, A)$ $(2, A, \{A, B\})$	[0.5;1] [0.3;0.8]

IW_8	
$EvidentialObjects$	CL
$(1, B, A)$ $(2, \{A, B\}, B)$	[0.5;1] [0.3;0.8]

Table 4.5: Combination of the evidential objects of EDB using the disjunctive rule of combination

values. Worlds are still uncertain because of tuples' confidence levels. These confidence levels measure the credibility and the plausibility about the existence of tuples. Note that the attribute level uncertainty is treated at this step (the step of generating uncertain possible worlds). For example IW_5 has two uncertain states {(1,A,A);(2,A,A)} and {(1,A,A);(2,A,B)}.

The number of uncertain possible worlds, J, is computed from sizes of cores $\varphi_{m_{ta}^{\Theta_a}}$ as follows:

$$J = \prod_{t=1}^{N} \prod_{a=1}^{D} |\varphi_{m_{ta}^{\Theta_a}}| \tag{4.5}$$

The set Θ_{IW} is the set of imprecise possible worlds IW_i. Each IW_i leads to one or more uncertain possible worlds. The set Θ_{UW} enumerates all uncertain possible worlds. Therefore, Θ_{UW} is a *refinement* of Θ_{IW} and Θ_{IW} is a *coarsening* of Θ_{UW}.

Thus, the mass function $m^{\Theta_{UW}}$ on which uncertain possible world is the best candidate to the evidential table, is deduced from $m^{\Theta_{IW}}$.

Example 48 *Table 4.6 presents the set of uncertain possible worlds derived from the imprecise worlds of Table 4.5 with the basic belief assignment $m^{\Theta_{UW}}$ derived from $m^{\Theta_{IW}}$. Note that $\Theta_{IW} = \{IW_1, \ldots, IW_8\}$ is the set of all imprecise possible worlds, and that $\Theta_{UW} = \{UW_1, \ldots, UW_8\}$ represents the set of all uncertain possible worlds,* *such that* Θ_{UW} *is a refinement of* Θ_{IW} *and:*

$IW_1 \rightarrow \{UW_1\}$
$IW_2 \rightarrow \{UW_1, UW_3, UW_4, UW_5\}$
$IW_3 \rightarrow \{UW_2\}$
$IW_4 \rightarrow \{UW_2, UW_6, UW_7, UW_8\}$
$IW_5 \rightarrow \{UW_1, UW_3\}$
$IW_6 \rightarrow \{UW_1, UW_5\}$
$IW_7 \rightarrow \{UW_2, UW_6\}$
$IW_8 \rightarrow \{UW_2, UW_7\}$

UW_1		UW_2	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, A, A)$	[0.5;1]	$(1, B, A)$	[0.5;1]
$(2, A, B)$	[0.3;0.8]	$(2, A, B)$	[0.3;0.8]
UW_3		UW_4	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, A, A)$	[0.5;1]	$(1, A, A)$	[0.5;1]
$(2, A, A)$	[0.3;0.8]	$(2, B, A)$	[0.3;0.8]
UW_5		UW_6	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, A, A)$	[0.5;1]	$(1, B, A)$	[0.5;1]
$(2, B, B)$	[0.3;0.8]	$(2, A, A)$	[0.3;0.8]
UW_7		UW_8	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, B, A)$	[0.5;1]	$(1, B, A)$	[0.5;1]
$(2, B, B)$	[0.3;0.8]	$(2, B, A)$	[0.3;0.8]

Table 4.6: Uncertain Possible worlds of EDB

$m^{\Theta_{UW}}(\{UW_1\}) = m^{\Theta_{IW}}(\{IW_1\}) = \mathit{0.12}$

$m^{\Theta_{UW}}(\{UW_1, UW_3, UW_4, UW_5\}) = m^{\Theta_{IW}}(\{IW_2\}) = \mathit{0.07}$

$m^{\Theta_{UW}}(\{UW_2\}) = m^{\Theta_{IW}}(\{IW_3\}) = \mathit{0.12}$

$m^{\Theta_{UW}}(\{UW_2, UW_6, UW_7, UW_8\}) = m^{\Theta_{IW}}(\{IW_4\}) = \mathit{0.07}$

$m^{\Theta_{UW}}(\{UW_1, UW_3\}) = m^{\Theta_{IW}}(\{IW_5\}) = \mathit{0.03}$

$m^{\Theta_{UW}}(\{UW_1, UW_5\}) = m^{\Theta_{IW}}(\{IW_6\}) = 0.28$

$m^{\Theta_{UW}}(\{UW_2, UW_6\}) = m^{\Theta_{IW}}(\{IW_7\}) = 0.03$

$m^{\Theta_{UW}}(\{UW_2, UW_7\}) = m^{\Theta_{IW}}(\{IW_8\}) = 0.28$

It is obvious that a same uncertain possible world can be derived from various imprecise worlds. For example, UW_1 is derived from IW_1 and IW_5. Uncertain possible worlds inherit their degrees of belief from their imprecise possible worlds.

CLs per uncertain world UW need to be taken into consideration. In fact, they are computed based on their mass functions as presented above.

$UW_1 \in \{IW_1, IW_2, IW_5, IW_6\}$ then its confidence level is UW_1 [0.12;0.5] and it is computed as follows:

- bel($\{UW_1\}$)= 0.12

- pl($\{UW_1\}$) = 0.12 + 0.07 + 0.03 + 0.28 = 0.5

$UW_2 \in \{IW_3, IW_4, IW_7, IW_8\}$ then UW_2 [0.12;0.5]

$UW_3 \in \{IW_2, IW_5\}$ then UW_3 [0;0.1]

$UW_4 \in \{IW_2\}$ then UW_4 [0;0.07]

$UW_5 \in \{IW_2, IW_6\}$ then UW_5 [0;0.35]

$UW_6 \in \{IW_4, IW_7\}$ then UW_6 [0;0.1]

$UW_7 \in \{IW_4, IW_8\}$ then UW_7 [0;0.35]

$UW_8 \in \{IW_4\}$ then UW_8 [0;0.07]

UW_1 **[0.12;0.5]**		UW_2 **[0.12;0.5]**	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, A, A)$	[0.5;1]	$(1, B, A)$	[0.5;1]
$(2, A, B)$	[0.3;0.8]	$(2, A, B)$	[0.3;0.8]
UW_3 **[0;0.1]**		UW_4 **[0;0.07]**	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, A, A)$	[0.5;1]	$(1, A, A)$	[0.5;1]
$(2, A, A)$	[0.3;0.8]	$(2, B, A)$	[0.3;0.8]
UW_5 **[0;0.35]**		UW_6 **[0;0.1]**	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, A, A)$	[0.5;1]	$(1, B, A)$	[0.5;1]
$(2, B, B)$	[0.3;0.8]	$(2, A, A)$	[0.3;0.8]
UW_7 **[0;0.35]**		UW_8 **[0;0,07]**	
$EvidentialObjects$	CL	$EvidentialObjects$	CL
$(1, B, A)$	[0.5;1]	$(1, B, A)$	[0.5;1]
$(2, B, B)$	[0.3;0.8]	$(2, B, A)$	[0.3;0.8]

Table 4.7: Uncertain Possible worlds of EDB

These confidence levels at uncertain worlds' level (see Table 4.7) are considered when computing the confidence levels at the tuple level as shown in Table 4.8.

UW_1	CL	UW_2	CL
$(1, A, A)$	[0.06;0.5]	$(1, B, A)$	[0.06;0.5]
$(2, A, B)$	[0.036;0.4]	$(2, A, B)$	[0.036;0.5]
UW_3	CL	UW_4	CL
$(1, A, A)$	[0;0.1]	$(1, A, A)$	[0;0.07]
$(2, A, A)$	[0;0.08]	$(2, B, A)$	[0;0.056]
UW_5	CL	UW_6	CL
$(1, A, A)$	[0;0.35]	$(1, B, A)$	[0;0.1]
$(2, B, B)$	[0;0.28]	$(2, A, A)$	[0;0.08]
UW_7	CL	UW_8	CL
$(1, B, A)$	[0;0.35]	$(1, B, A)$	[0;0.07]
$(2, B, B)$	[0;0.28]	$(2, B, A)$	[0;0.056]

Table 4.8: Uncertain Possible worlds of EDB

Each uncertain possible world UW_j can be expanded itself into different states called *Possible Worlds*, denoted W_{jp}. In possible worlds, tuple uncertainty level is treated in addition to the attribute uncertainty level (already handled in uncertain possible worlds).

4.3.3 Possible Worlds

Generating possible worlds from uncertain possible worlds induces the treatment of the tuple level uncertainty.

Each possible world W_{jp} is a candidate to represent the evidential relation where j $(1 \leq j \leq J)$ and p $(1 \leq p \leq P)$.

The number of possible worlds P is induced from possibilities about the existence of tuples in uncertain worlds UW_j such that:

$$P = 2^N \tag{4.6}$$

Thus, an uncertain world UW_j is expanded into 2^N possible worlds where each W_{jp} is a combination of tuples belonging to UW_j.

Defintion 26 *A possible world W_{jp} is a precise subset from an uncertain possible world UW_j. Each uncertain possible world can be divided into one or more precise possible worlds by handling uncertainty due to the confidence level. Each uncertain possible world generates $N+2$ possible worlds. The confidence level is bounded by the bel and the pl of the existence of each object in the table. Therefore, Θ_W is a refinement of Θ_{UW} and Θ_{UW} is a coarsening of Θ_W.*

Example 49 *Four possible worlds are carried out from UW_1 :*

- $\{t_1, t_2\}$ *: Both tuples 1 and 2 exist simultaneously.*
- $\{t_1\}$ *: Only tuple 1 exists.*
- $\{t_2\}$ *: Only tuple 2 exists.*
- $\{\varnothing\}$ *: None of the tuples exist.*

Thus:
$W_{11} = \{(1, A, A); (2, A, B)\}$ $W_{13} = \{(2, A, B)\}$

$W_{12} = \{(1, A, A)\}$ $W_{14} = \{\varnothing\}$

Defintion 27 *The number of possible worlds, denoted P, varies between J and $J * 2^N$ with J is the number of uncertain possible worlds and N is the number of objects of EDB.*

- $P = J$ *: if all confidence levels are equal to [1;1]; in this case, the tuple level uncertainty is not managed (Bousnina et al., 2015).*
- $P = J \times 2^N$ *: if all confidence levels are different from [0;0] and [1;1].*

- $P \in \left]J, J \times 2^N\right[$ *: if some confidence levels are either [0;0] or [1;1].*

Example 50 *Table 4.9 includes the* $32 = 8 * 4$ *possible worlds derived from the uncertain worlds:*

UW_1	UW_2
$W_{11} = \{(1, A, A); (2, A, B)\}$	$W_{21} = \{(1, B, A); (2, A, B)\}$
$W_{12} = \{(1, A, A)\}$	$W_{22} = \{(1, B, A)\}$
$W_{13} = \{(2, A, B)\}$	$W_{23} = \{(2, A, B)\}$
$W_{14} = \{\varnothing\}$	$W_{24} = \{\varnothing\}$
UW_3	UW_4
$W_{31} = \{(1, A, A); (2, A, A)\}$	$W_{41} = \{(1, A, A); (2, B, A)\}$
$W_{32} = \{(1, A, A)\}$	$W_{42} = \{(1, A, A)\}$
$W_{33} = \{(2, A, A)\}$	$W_{43} = \{(2, B, A)\}$
$W_{34} = \{\varnothing\}$	$W_{44} = \{\varnothing\}$
UW_5	UW_6
$W_{51} = \{(1, A, A); (2, B, B)\}$	$W_{61} = \{(1, B, A); (2, A, A)\}$
$W_{52} = \{(1, A, A)\}$	$W_{62} = \{(1, B, A)\}$
$W_{53} = \{(2, B, B)\}$	$W_{63} = \{(2, A, A))\}$
$W_{54} = \{\varnothing\}$	$W_{64} = \{\varnothing\}$
UW_7	UW_8
$W_{71} = \{(1, B, A); (2, B, B)\}$	$W_{81} = \{(1, B, A); (2, B, A)\}$
$W_{72} = \{(1, B, A)\}$	$W_{82} = \{(1, B, A)\}$
$W_{73} = \{(2, B, B)\}$	$W_{83} = \{(2, B, A)\}$
$W_{74} = \{\varnothing\}$	$W_{84} = \{\varnothing\}$

Table 4.9: Possible Worlds

Defintion 28 *Let W be a possible world generated from an uncertain world UW. The confidence level of W is computed as follows:*

$$CL = [bel_W^{\Theta_W}; pl_W^{\Theta_W}] \tag{4.7}$$

$$\begin{aligned} bel_W^{\Theta_W} &= \prod bel(t_i) * \prod (1 - pl(t_j)) \\ pl_W^{\Theta_W} &= \prod pl(t_i) * \prod (1 - bel(t_j)) \end{aligned} \tag{4.8}$$

$$\textit{where } t_i, t_j \in UW \textit{ and } t_i \in W; t_j \notin W$$

Example 51 *Confidence levels of possible worlds are computed using Definition 28 based on equations* (2.8) (2.7), (2.36) *such that:*

$\{t_1, t_2\}$:

- $bel(t_1 \wedge t_2) = bel_1 * bel_2$
- $pl(t_1 \wedge t_2) = pl_1 * pl_2$

$\{t_1\}$:

- $bel(t_1 \wedge \overline{t_2}) = bel_1 * (1 - pl_2)$
- $pl(t_1 \wedge \overline{t_2}) = pl_1 * (1 - bel_2)$

$\{t_2\}$:

- $bel(\overline{t_1} \wedge t_2) = (1 - pl_1) * bel_2$
- $pl(\overline{t_1} \wedge t_2) = (1 - bel_1) * pl_2$

$\{\varnothing\}$:

- $bel(\overline{t_1} \wedge \overline{t_2}) = (1 - pl_1) * (1 - pl_2)$
- $pl(\overline{t_1} \wedge \overline{t_2}) = (1 - bel_1) * (1 - bel_2)$

Thus, confidence levels of possible worlds: W_{11}, W_{12}, W_{13}, W_{14} generated from UW_1 are computed as follows:

W_{11}:

- $bel(t_1 \wedge t_2) = bel_1 * bel_2 = 0.06 * 0.036 = 0.0021$
- $pl(t_1 \wedge t_2) = pl_1 * pl_2 = 0.5 * 0.4 = 0.2$

$CL_{11} = [0.0021; 0.2]$

W_{12}:

- $bel(t_1 \wedge \overline{t_2}) = bel_1 * (1 - pl_2) = 0.06 * (1\text{-}0.4) = 0.048$
- $pl(t_1 \wedge \overline{t_2}) = pl_1 * (1 - bel_2) = 0.5 * (1\text{-}0.036)= 0.482$

$CL_{12} = [0.048; 0.482]$

W_{13}:

- $bel(\overline{t_1} \wedge t_2) = (1 - pl_1) * bel_2 = 0.036 * (1\text{-}0.5) = 0.018$
- $pl(\overline{t_1} \wedge t_2) = (1 - bel_1) * pl_2 = 0.4 * (1\text{-}0.06) = 0.376$

$CL_{13} = [0.018; 0.376]$

W_{14}*:*

- $bel(\overline{t_1} \wedge \overline{t_2}) = (1 - pl_1) * (1 - pl_2)$ *= (1-0.5) * (1-0.2) = 0.4*
- $pl(\overline{t_1} \wedge \overline{t_2}) = (1 - bel_1) * (1 - bel_2)$ *= (1-0.06)*(1-0.036) = 0.9*

$CL_{14} = [0.4; 0.9]$

All possible worlds and their confidence levels of Table 4.1 are presented in Table 4.10.

UW_1	CL	UW_2	CL
$W_{11} = \{(1, A, A); (2, A, B)\}$	$CL_{11} = [0.0021; 0.2]$	$W_{21} = \{(1, B, A); (2, A, B)\}$	$CL_{21} = [0.0021; 0.2]$
$W_{12} = \{(1, A, A)\}$	$CL_{12} = [0.048; 0.482]$	$W_{22} = \{(1, B, A)\}$	$CL_{22} = [0.048; 0.482]$
$W_{13} = \{(2, A, B)\}$	$CL_{13} = [0.018; 0.376]$	$W_{23} = \{(2, A, B)\}$	$CL_{23} = [0.018; 0.376]$
$W_{14} = \{\varnothing\}$	$CL_{14} = [0.4; 0.9]$	$W_{24} = \{\varnothing\}$	$CL_{24} = [0.4; 0.9]$
UW_3	CL	UW_4	CL
$W_{31} = \{(1, A, A); (2, A, A)\}$	$CL_{31} = [0; 0.008]$	$W_{41} = \{(1, A, A); (2, B, A)\}$	$CL_{41} = [0; 0.004]$
$W_{32} = \{(1, A, A)\}$	$CL_{32} = [0; 0.1]$	$W_{42} = \{(1, A, A)\}$	$CL_{42} = [0; 0.07]$
$W_{33} = \{(2, A, A)\}$	$CL_{33} = [0; 0.8]$	$W_{43} = \{(2, B, A)\}$	$CL_{43} = [0; 0.056]$
$W_{34} = \{\varnothing\}$	$CL_{34} = [0.83; 1]$	$W_{44} = \{\varnothing\}$	$CL_{44} = [0.88; 1]$
UW_5	CL	UW_6	CL
$W_{51} = \{(1, A, A); (2, B, B)\}$	$CL_{51} = [0; 0.098]$	$W_{61} = \{(1, B, A); (2, A, A)\}$	$CL_{61} = [0; 0.008]$
$W_{52} = \{(1, A, A)\}$	$CL_{52} = [0; 0.35]$	$W_{62} = \{(1, B, A)\}$	$CL_{62} = [0; 0.1]$
$W_{53} = \{(2, B, B)\}$	$CL_{53} = [0; 0.28]$	$W_{63} = \{(2, A, A))\}$	$CL_{63} = [0; 0.08]$
$W_{54} = \{\varnothing\}$	$CL_{54} = [0.47; 1]$	$W_{64} = \{\varnothing\}$	$CL_{64} = [0.83; 1]$
UW_7	CL	UW_8	CL
$W_{71} = \{(1, B, A); (2, B, B)\}$	$CL_{71} = [0; 0.098]$	$W_{81} = \{(1, B, A); (2, B, A)\}$	$CL_{81} = [0; 0.004]$
$W_{72} = \{(1, B, A)\}$	$CL_{72} = [0; 0.35]$	$W_{82} = \{(1, B, A)\}$	$CL_{82} = [0; 0.07]$
$W_{73} = \{(2, B, B)\}$	$CL_{73} = [0; 0.28]$	$W_{83} = \{(2, B, A)\}$	$CL_{83} = [0; 0.056]$
$W_{74} = \{\varnothing\}$	$CL_{84} = [0.47; 1]$	$W_{84} = \{\varnothing\}$	$CL_{84} = [0.88; 1]$

Table 4.10: Possible Worlds with their CLs

Generating the possible worlds of an evidential database is a complex process as shown above. Figure 4.3 summarizes this process. Thus, it illustrates the obtained intermediate non-compact forms:

- Imprecise Possible Worlds (IW).
- Uncertain Possible Worlds (UW).
- Possible Worlds (W).

It also details the used Dempster-Shafer theory tools and concepts:

- Conjunctive Rule of Combination $\textcircled{\cap}$ and Disjunctive Rule of Combination $\textcircled{\cup}$.
- Splitting focal elements for the attribute level uncertainty treatment.
- Evidence Independence for the tuple level uncertainty treatment.

Figure 4.4 illustrates, through an example, the overall process:

- Starting from the EDB in the left, we obtain the imprecise worlds by using the conjunctive rule of combination on evidential values of the same object (here we have only one attribute), and then by applying the disjunctive rule of combination on the obtained evidential objects.
- The result is a bba of imprecise worlds whose masses are mentioned between parentheses in the first lines. Then, we consider only the imprecise world IW_2. We split the composite focal elements and we obtain two uncertain worlds UW_1 and UW_2.
- At this stage, each world is precise, but still uncertain because of the CL of each object. Then, we show how the world UW_2 is expanded into four possible worlds which is the number of possible combinations of coexistence of two uncertain objects. The confidence levels of the obtained possible worlds are computed using the notion of evidential independence. Note that obtained confidence levels are mentioned in the first worlds' lines.

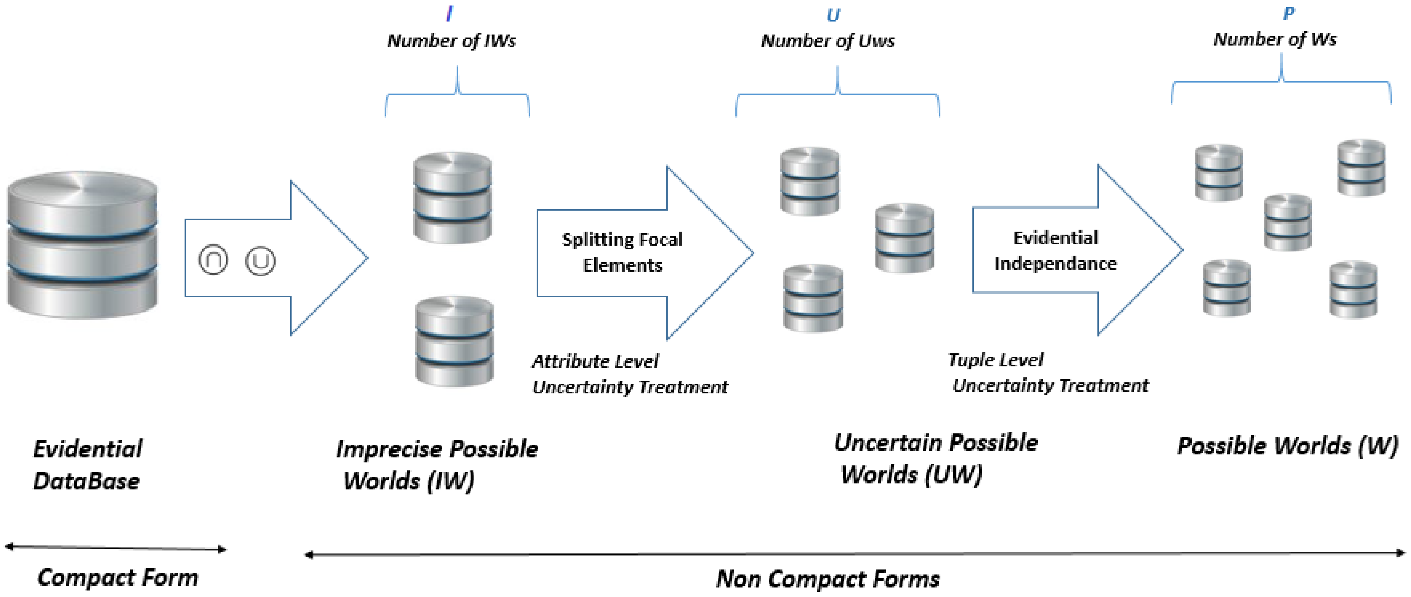
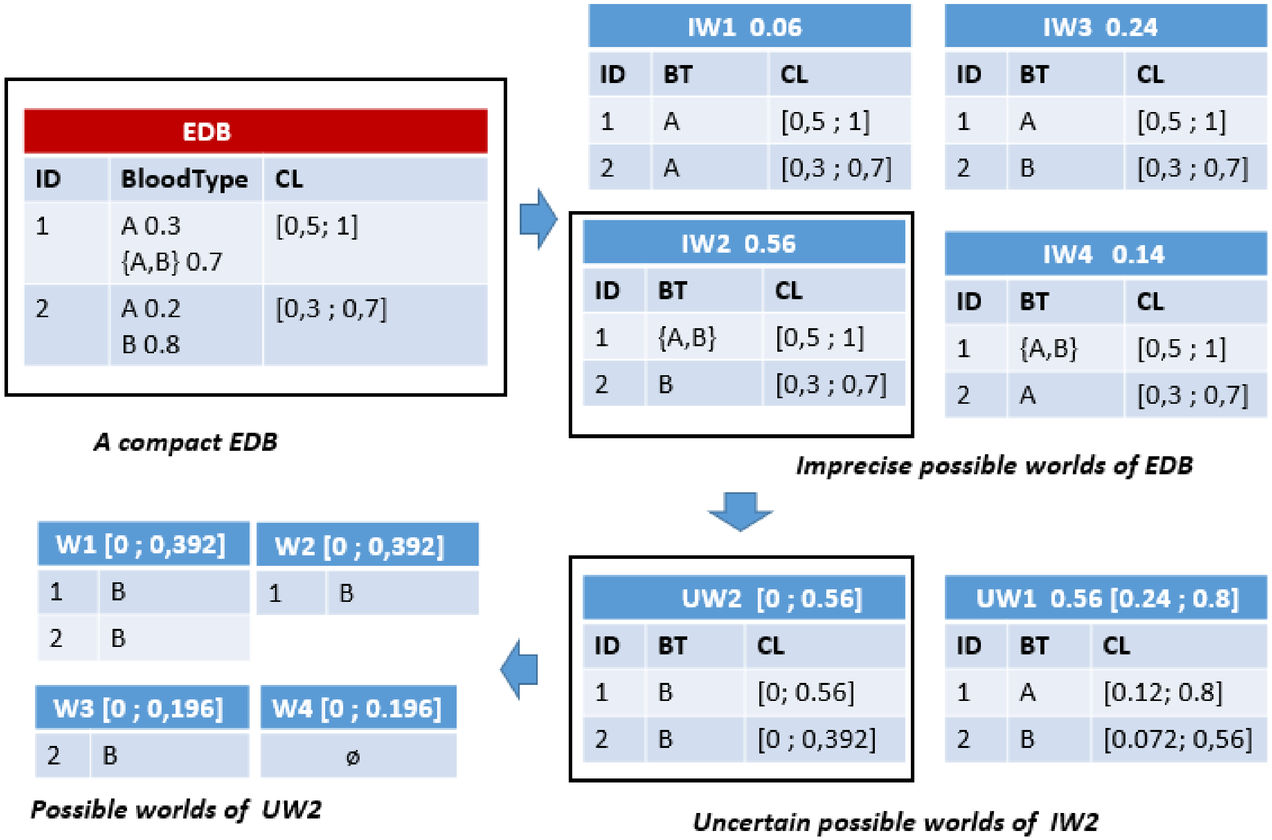


Figure 4.3: Process of producing the possible worlds form



EDB

ID	BloodType	CL
1	A 0.3 {A,B} 0.7	[0,5; 1]
2	A 0.2 B 0.8	[0,3 ; 0,7]

A compact EDB

IW1 0.06

ID	BT	CL
1	A	[0,5 ; 1]
2	A	[0,3 ; 0,7]

IW3 0.24

ID	BT	CL
1	A	[0,5 ; 1]
2	B	[0,3 ; 0,7]

IW2 0.56

ID	BT	CL
1	{A,B}	[0,5 ; 1]
2	B	[0,3 ; 0,7]

IW4 0.14

ID	BT	CL
1	{A,B}	[0,5 ; 1]
2	A	[0,3 ; 0,7]

Imprecise possible worlds of EDB

UW2 [0 ; 0.56]

ID	BT	CL
1	B	[0; 0.56]
2	B	[0 ; 0,392]

UW1 0.56 [0.24 ; 0.8]

ID	BT	CL
1	A	[0.12; 0.8]
2	B	[0.072; 0,56]

Uncertain possible worlds of IW2

W1 [0 ; 0,392]

1	B
2	B

W2 [0 ; 0,392]

1	B

W3 [0 ; 0,196]

2	B

W4 [0 ; 0.196]

ø

Possible worlds of UW2

Figure 4.4: Illustrative example for generating possible worlds of an EDB

4.4 Querying Possible Worlds

Querying possible worlds is an essential step after the modeling in order to evaluate the representation system in question. Indeed, we introduced in (Bousnina et al., 2015) how to query possible worlds' form.

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

$$R_u = Q(W)$$

Example 52 *We evaluate the following query Q over the possible worlds' form of Table 4.10 possible worlds form.*

```
Q: SELECT * FROM BT WHERE (BloodType==BloodTypeFather);
```

Eight possible answers responded to query Q. Responses and their confidence levels are presented as follows:

$R_1 = Q(W_{11}) = \{(1, A, A)\}; CL_{R_1}$ = *[0.0021 ; 0.2]*

$R_1 = Q(W_{12}) = \{(1, A, A)\}; CL_{R_1}$ = *[0.048 ; 0.482]*

$R_2 = Q(W_{13}) = \{\emptyset\}; CL_{R_2}$ = *[0.018 ; 0.376]*

$R_2 = Q(W_{14}) = \{\emptyset\}; CL_{R_2}$ = *[0.4; 0.9]*

$R_2 = Q(W_{21}) = \{\emptyset\}; CL_{R_2}$ = *[0.0021; 0.2]*

$R_2 = Q(W_{22}) = \{\emptyset\}; CL_{R_2}$ = *[0.048; 0.482]*

$R_2 = Q(W_{23}) = \{\emptyset\}; CL_{R_2}$ = *[0.018; 0.376]*

$R_2 = Q(W_{24}) = \{\emptyset\}; CL_{R_2}$ = *[0.4; 0.9]*

$R_3 = Q(W_{31}) = \{(1, A, A); (2, A, A)\}; CL_{R_3}$ = *[0 ; 0.008]*

$R_1 = Q(W_{32}) = \{(1, A, A)\}; CL_{R_1}$ = *[0; 0.1]*

$R_4 = Q(W_{33}) = \{(2, A, A)\}; CL_{R_4} =$ *[0 ; 0.08]*

$R_2 = Q(W_{34}) = \{\emptyset\}; CL_{R_2} =$ *[0.83; 1]*

$R_2 = Q(W_{41}) = \{\emptyset\}; CL_{R_2} =$ *[0; 0.004]*

$R_1 = Q(W_{42}) = \{(1, A, A)\}; CL_{R_1} =$ *[0; 0.07]*

$R_2 = Q(W_{43}) = \{\emptyset\}; CL_{R_2} =$ *[0; 0.056]*

$R_2 = Q(W_{44}) = \{\emptyset\}; CL_{R_2} =$ *[0.88; 1]*

$R_5 = Q(W_{51}) = \{(1, A, A); (2, B, B)\}; CL_{R_5} =$ *[0; 0.098]*

$R_1 = Q(W_{52}) = \{(1, A, A)\}; CL_{R_1} =$ *[0; 0.35]*

$R_6 = Q(W_{53}) = \{(2, B, B)\}; CL_{R_6} =$ *[0 ; 0.28]*

$R_4 = Q(W_{63}) = \{(2, A, A)\}; CL_{R_4} =$ *[0 ; 0.08]*

$R_2 = Q(W_{54}) = \{\emptyset\}; CL_{R_2} =$ *[0.47; 1]*

$R_4 = Q(W_{61}) = \{(2, A, A)\}; CL_{R_4} =$ *[0 ; 0.008]*

$R_2 = Q(W_{62}) = \{\emptyset\}; CL_{R_2} =$ *[0; 0.1]*

$R_4 = Q(W_{63}) = \{(2, A, A)\}; CL_{R_4} =$ *[0 ; 0.08]*

$R_2 = Q(W_{64}) = \{\emptyset\}; CL_{R_2} =$ *[0.83; 1]*

$R_6 = Q(W_{71}) = \{(2, B, B)\}; CL_{R_6} =$ *[0; 0.098]*

$R_2 = Q(W_{72}) = \{\emptyset\}; CL_{R_2} =$ *[0; 0.35]*

$R_6 = Q(W_{73}) = \{(2, B, B)\}; CL_{R_6} =$ *[0 ; 0.28]*

$R_2 = Q(W_{74}) = \{\emptyset\}; CL_{R_2} =$ *[0.47; 1]*

$R_2 = Q(W_{81}) = \{\emptyset\}; CL_{R_2} =$ *[0; 0.004]*

$R_2 = Q(W_{82}) = \{\emptyset\}; CL_{R_2}$ = *[0; 0.07]*

$R_2 = Q(W_{83}) = \{\emptyset\}; CL_{R_2}$ = *[0; 0.056]*

$R_2 = Q(W_{84}) = \{\emptyset\}; CL_{R_2}$ = *[0.88; 1]*

Redundancy comes from different beliefs for identical tuples. In relational databases, redundancy is considered automatically using the property of sets of the table. In evidential databases, redundancy is handled in a different way such that (Hau and Kashyap, 1990; Bell et al., 1996; Lee, 1992b):

Defintion 29 *Let DB be an evidential database and t_1 and t_2 be two identical tuples with $t_1.CL_1 \neq t_2.CL_2$ and CL_1 and CL_2 are $\in$ [0;1], then t_1 and t_2 are called redundant tuples.*

Defintion 30 *Let E_1 and E_2 be two independent events. Each event has a confidence level of belief and plausibility where $CL_1 = [bel_1; pl_1]$ and $CL_2 = [bel_2; pl_2]$. The disjunction of the events E_1 and E_2 are defined such that:*

$$[bel(E_1 \vee E_2); pl(E_1 \vee E_2)] = [1 - (1 - bel_1)(1 - bel_2); 1 - (1 - pl_1)(1 - pl_2)] \quad (4.9)$$

Note that in the case where both confidence levels are equal to [1;1], the redundancy is removed automatically via the set property.

Example 53 *Let us carry on with the same example and compute the CLs for each answer by treating the redundancy:*

- *R_1={(1,A,A)}; CL_{R_1}= [(1-(1-0.0021)*(1-0.048)*(1-0)*(1-0)*(1-0)*(1-0));(1-(1-0.2)*(1-0.482)*(1-0.1)*(1-0.35)*(1-0.004)*(1-0.07)] = [0.049 ; 0.775]*
- *R_2={$\emptyset$}; CL_{R_2}= [(1-(1-0.18)*(1-0.4)*(1-0.0021)*(1-0.048)*(1-0.018)*(1-0.4)*(1-0.83)*(1-0)*(1-0.56)*(1-0.47)*(1-0)*(1-0.83)*(1-0)*(1-0.47)*(1-0)*(1-0)*(1-0)*(1-0.88));(1-(1-0.376)*(1-0.9)*(1-0.2)*(1-0.482)*(1-0.376)*(1-0.9)*(1-1)*(1-0.88)*(1-1)*(1-1)*(1-0.1)*(1-1)*(1-0.35)*(1-1)*(1-0.004)*(1-0.7)*(1-0.056)*(1-1)] = [0.99 ; 1]*
- *R_3={(1,A,A),(2,A,A)}; CL_{R_3}= [0 ; 0.008]*
- *R_4={(2,A,A)}; CL_{R_4}= [(1-(1-0)*(1-0)*(1-0));(1-(1-0.8)*(1-0.008)*(1-0.08))] = [0 ; 0.81]*
- *R_5={(1,A,A),(2,B,B)}; CL_{R_5}= [0 ; 0.098]*

- R_6={(2,B,B)}; CL_{R_6}= [(1-(1-0)*(1-0)*(1-0));(1-(1-0.28)*(1-0.28)*(1-0.098))] = [0;0.53]

4.5 Generalizing Probabilistic Databases

Evidential database model as introduced in the pioneer work (Lee, 1992b) is an extension of the relational model where two levels of uncertainties are considered; the tuple-level and the attribute-level, managed through the evidence theory. The main advantage of this kind of imperfect database resides in its theoretical basis; the evidence theory. This framework manages several kinds of imperfection, including those handled by probability and possibility theories. Therefore, in this section, we discuss the particular cases where evidential values are exclusively (1) probabilistic (*bbas* with singleton focal elements) and (2) possibilistic (*bbas* with nested focal elements).

In the first case, the database is typically probabilistic. Attributes store probability distributions (which is a special *bba*) and tuples are characterized by a *maybe* probability[2] which is the belief bound of the confidence level. If we perform the process of generating possible worlds (shown in Figure 4.3), we will start, for each tuple, by combining the distributions of its attributes. The conjunctive rule of combination is based on the cross product, which is exactly the same technique used for the probabilistic model[3]. We obtain a normalized distribution for each original tuple. These distributions are again combined, using the disjunctive rule of combination, that is also based on multiplying masses of focal elements (tuples in our case). We obtain a probabilistic distribution of possible worlds. The resulted possible worlds could not be *imprecise* because the input distributions do not include composite focal elements. At this stage, we treated the attribute level uncertainty. Then, each tuple t in each world has a maybe probability, which corresponds to the tuple belief in the evidential model. When a tuple t has a probability $p_t < 1$, we have two possible scenarios; t exists with a probability p_t and t does not exist with the probability $1 - p_t$. In our evidential model, t has a belief (probability) $bel(t)$. Thus t exists with the belief: $bel(t)$, and does not exist with the belief $1 - pl(t)$ where $pl(t)$ is the plausibility of t. However, when an evidential scenario is particularly probabilistic, belief and plausibility have exactly the same value (Shafer, 1976). Thus, we fall into the same model, and the overall evidential process to generate possible worlds matches the probabilistic process.

This observation is not surprising because of two statements. First, conjunctive and disjunctive rules combine *bbas* by multiplying masses, which matches the probability combination method. Thus, the attribute uncertainty level is handled in the same way.

[2] Term introduced in (Agrawal et al., 2006) and refers to the probability of a tuple's existence

[3] The process of generating possible worlds in the case of probabilistic database is described in (Suciu et al., 2011); an excellent survey about probabilistic databases.

Second, belief and plausibility measures provide the same value in the probabilistic case, matching the probability value. Therefore, the tuple uncertainty level is also performed in the same way.

In case of possibilistic data (i.e. when *bbas* have consonant focal elements), we obtain the same possible worlds. Indeed, both models[4] use the cross product operator to generate possible tuples and then worlds (for processing the attribute uncertainty level). They also handle in the same way the problem of tuples' existences, and generate for each tuple the two possibilities of existence and non-existence. Therefore, the two models generate the same possible worlds. However, obtained distributions produce different uncertainty measures. This is inherent from the theories on which the models are based. Indeed, if combination rules in evidence theory uses the multiplication operator, in possibility theory (Zadeh, 1965), combination is based on the minimum/maximum operators.e singletons, e.g. the first tuple of attribute BT in Table 4.1. Possibilistic data are also a special case for evidential data when focal elements are nested, e.g. the second tuple of attribute BT in Table 4.1. As a consequence, semantics relative to probabilistic and possibilistic worlds can be generalized with evidential possible worlds (Agrawal and Widom, 2010).

4.6 EDB: What Kind of Representation System?

As explained earlier, an evidential database EDB (Lee, 1992b; Lee, 1992a; Bell et al., 1996) has two representations: (i) the compact form and (ii) the possible worlds' form.

(i) Querying the compact form provides a compact answer Q(EDB). This evidential compact answer is expanded into several possible worlds that give themselves the possible answers $\{R_1, \dots, R_u\}$ and their confidence levels $\{CL_1, \dots, CL_u\}$.

(ii) Querying the possible worlds' form provides the possible answers $\{R'_1, \dots, R_s\}$ and their confidence levels $\{CL'_1, \dots, CL'_s\}$.

The derived results from (i) and (ii) are compared in order to check weather the evidential database is a strong representation system (SRS).

$$\begin{cases} \text{If } Q(Rep(EDB)) \equiv Rep(Q(EDB)) & \text{then EDB is said a SRS} \\ \text{If } Q(Rep(EDB)) \neq Rep(Q(EDB)) & \text{then EDB is not a SRS} \end{cases}$$

Figure 4.5 shows the process of querying the compact EDB and its possible worlds as detailed above.

[4] Please refer to (Bosc and Pivert, 2010) for an excellent state of the art about possibilistic database models.

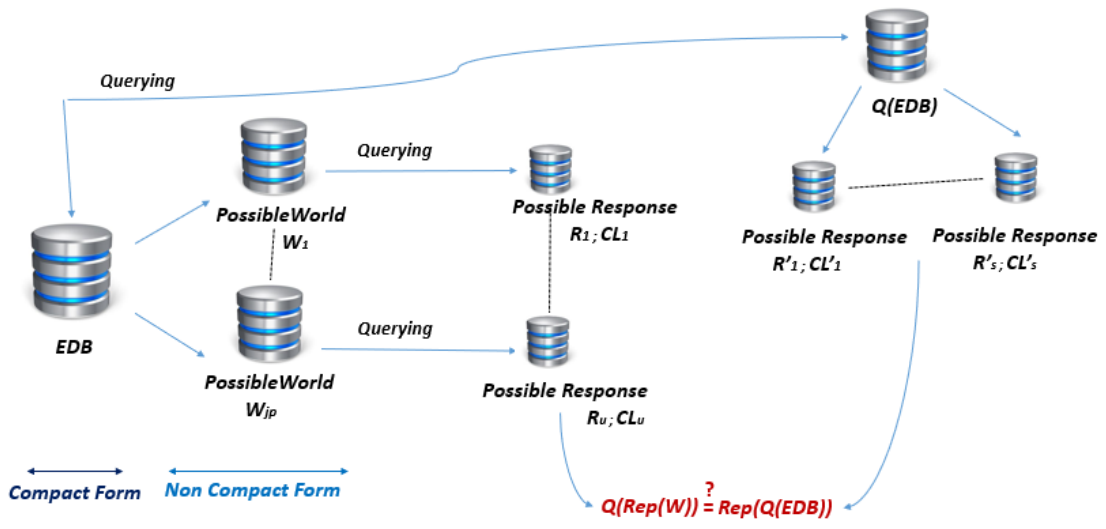


Figure 4.5: The process of querying both forms of EDB

Example 54 *We take an evidential Table 4.11 about medical diagnosis that includes three attributes: ID, Disease which contains hypotheses and their masses given by the doctor about his patients and CL that reflect the tuple level uncertainty. In this example, we choose a certain CL [1;1] to simplify computations. Our aim is to check the non-equivalence between the compact form and the possible worlds' form.*

ID	Disease	CL
1	$\{Cancer, Anemia\}$ 0.2 $Anemia$ 0.8	[1 ; 1]
2	$Anemia$ 1	[1 ; 1]

Table 4.11: An Evidential Table EDB

- ***Step 1:*** *Modeling EDB from the compact form to the possible worlds' form.*
  - *Table 4.11 can be expanded into two imprecise possible worlds $\{IW_1, IW_2\}$ as shown in Table 4.12.*

    The computed mass functions of IW_1 and IW_2 are:

    $m^{\Theta_{IW}}(\{IW_1\}) = 0.2 * 1 = 0.2$

IW_1			IW_2		
ID	**Disease**	**CL**	**ID**	**Disease**	**CL**
1	$\{Cancer, Anemia\}$	[1;1]	1	*Anemia*	[1;1]
2	*Anemia*	[1;1]	2	*Anemia*	[1;1]

Table 4.12: Imprecise Worlds of EDB

$m^{\Theta_{IW}}(\{IW_2\}) = 0.8 * 1 = 0.8$

The confidence levels, in this example, have no impact because tuples are certain ones.

- *Table 4.12 itself is exploded into two possible worlds $\{W_1, W_2\}$ as shown in Table 4.13.*

W_1		W_2	
ID	**Disease**	**ID**	**Disease**
1	*Cancer*	1	*Anemia*
2	*Anemia*	2	*Anemia*

Table 4.13: Possible Worlds of EDB

Their masses are inherited from masses of their imprecise possible worlds.

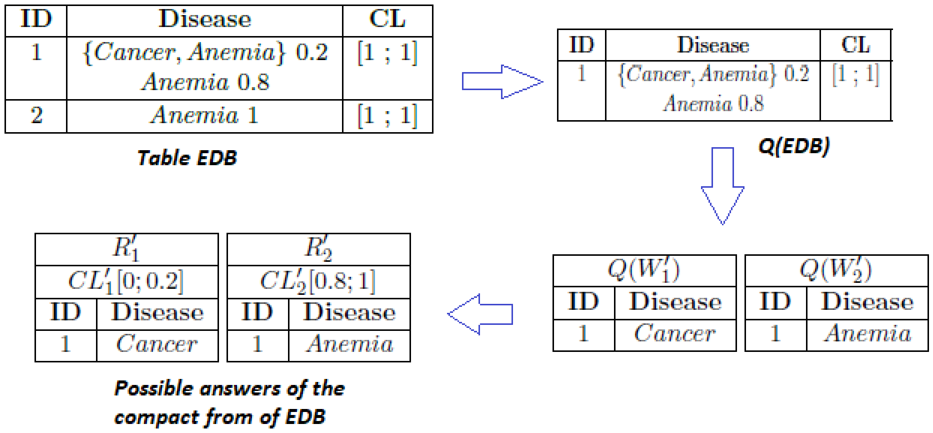
$m(\{W_1, W_2\}) = m(\{IW_1\}) = 0.2$
$m(\{W_2\}) = m(\{IW_2\}) = 0.8$

- ***Step 2:*** *Querying EDB from the compact form to the possible worlds' form.*

  We suppose the following query:

  Q : *SELECT * FROM* EDB *WHERE* $< Disease = Cancer >$

  - *We apply query Q over the compact Table 4.11. The result is the compact answer $Q(EDB)$ that contains tuples that only respond to query Q. This relation is imperfect, i.e, it includes uncertainty and imprecision. Indeed, it is modeled on its non compact form. In fact, the relation $Q(EDB)$ is expanded into possible worlds whose provide the possible answers $\{R'_1, R'_2\}$ and their confidence levels $\{CL'_1, CL'_2\}$. These steps are detailed in Figure 4.6.*



ID	Disease	CL
1	$\{Cancer, Anemia\}$ 0.2 $Anemia$ 0.8	[1 ; 1]
2	$Anemia$ 1	[1 ; 1]

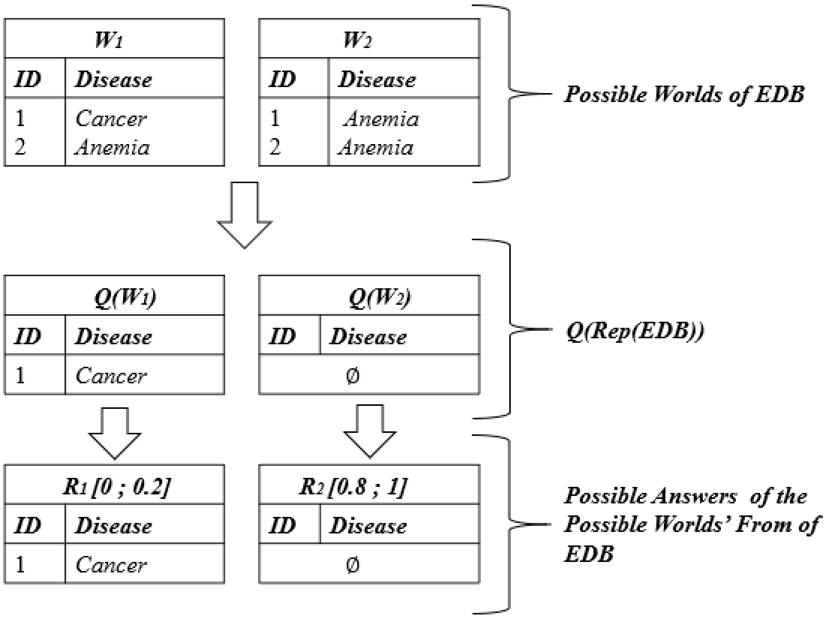
ID	Disease	CL
1	$\{Cancer, Anemia\}$ 0.2 $Anemia$ 0.8	[1 ; 1]

$Q(W'_1)$		$Q(W'_2)$	
ID	Disease	ID	Disease
1	$Cancer$	1	$Anemia$

R'_1		R'_2	
$CL'_1[0; 0.2]$		$CL'_2[0.8; 1]$	
ID	Disease	ID	Disease
1	$Cancer$	1	$Anemia$

Figure 4.6: The querying process the compact form of EDB

- *We apply now the same query Q over the possible worlds in Table 4.13 of the evidential table. The result is a set of answers $Q(Rep(EDB))$. These answers $\{R_1, R_2\}$ are quantified by confidence levels $\{CL_1, CL_2\}$. These steps are detailed in Figure 4.7.*



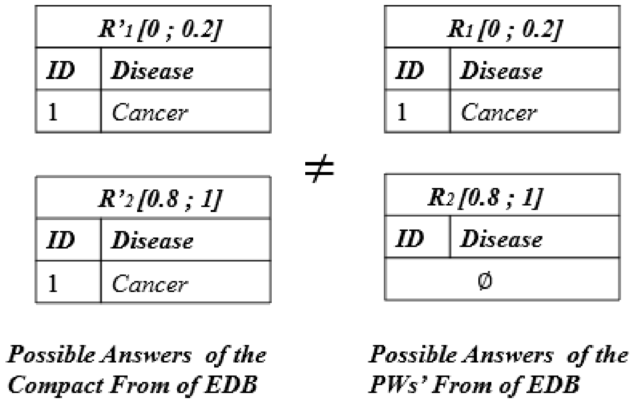
W1		*W2*	
ID	*Disease*	*ID*	*Disease*
1	*Cancer*	1	*Anemia*
2	*Anemia*	2	*Anemia*

Q(W1)		*Q(W2)*	
ID	*Disease*	*ID*	*Disease*
1	*Cancer*		Ø

R1 [0 ; 0.2]		*R2 [0.8 ; 1]*	
ID	*Disease*	*ID*	*Disease*
1	*Cancer*		Ø

Figure 4.7: The querying process of the non compact form of EDB

- ***Step 3:*** *Checking the equivalence between results issued from the compact form and the possible worlds' form of EDB.*



R'_1 [0 ; 0.2]	
ID	Disease
1	Cancer

R'_2 [0.8 ; 1]	
ID	Disease
1	Cancer

≠

R_1 [0 ; 0.2]	
ID	Disease
1	Cancer

R_2 [0.8 ; 1]	
ID	Disease
∅	

Possible Answers of the Compact From of EDB

Possible Answers of the PWs' From of EDB

Figure 4.8: Checking the equivalence between possible answers issued from both forms of EDB

As shown in Figure 4.8, answers issued from the compact form of EDB are different form those issued from its possible worlds' form. Thus, while the compact representation gave the answers:

$R'_1 = \{1, Cancer\}$ *with* $CL'_1 = [0; 0.2]$

$R'_2 = \{1, Anemia\}$ *with* $CL'_2 = [0.8; 1]$

The possible worlds' form gave the answers:

$R_1 = \{1, Cancer\}$ *with* $CL_1 = [0; 0.2]$

$R_2 = \emptyset$ *with* $CL_2 = [0.8; 1]$

Results are not the same which is due to structure of an evidential database that returns the hole tuple when one of its elements responds to a given query.

This counter example proves that the evidential database model as presented in the pioneer work (Lee, 1992b; Lee, 1992a; Bell et al., 1996) is not a strong representation system.

4.7 Conclusion

Representation systems are defined as the way of modeling an imperfect database from its compact form to its possible worlds' form. In this chapter, we presented how to model the compact evidential database (Lee, 1992b; Lee, 1992a; Bell et al., 1996) into its possible worlds' form by treating the tuple level and the attribute level uncertainties. This process provided two intermediate non compact forms: the imprecise possible worlds and the uncertain possible worlds. Moreover, we showed that querying this database model through its two equivalent forms is not a strong representation system. Hence, the need to introduce an evidential database model that revises the weaknesses of the EDB model towards a strong representation system under the relational algebra.

5
Evidential Conditional databases

Contents

5.1 Introduction 118

5.2 Conditional Tables 118

5.3 Evidential Conditional Tables 120

5.3.1 Modeling ec-tables: Compact form 120

5.3.2 Modeling ec-tables: Possible Worlds' Form 121

5.3.3 Querying ec-tables: Possible worlds form 124

5.3.4 Querying ec-tables: Compact form 126

5.3.5 Checking the equivalence 128

5.4 Discussion 128

5.5 Conclusion 140

Summary

This chapter is about modeling and querying evidential conditional databases, called ec-tables. In fact, we use the strengths of classical conditional databases and evidence theory to introduce a new evidential database model called evidential conditional databases (ECD). First, we represent this model into its both representations: the compact and the possible worlds; then, we present how to interrogate it using the relational select-project operators. Indeed, the querying methods are used to check the equivalence between both forms. Finally, we discuss the evidential conditional databases' specificities in order to make the ECD model a strong representation system for evidential databases under select-project operators.

5.1 Introduction

As shown in the previous chapter, an evidential database can be modeled through a distribution of candidate databases called the possible worlds. Defining possible worlds is a fundamental step towards the evaluation of the querying methods applied over an evidential database model. We proved with a counter example that the evidential database model (EDB) (Bell et al., 1996; Lee, 1992a; Lee, 1992b) is not a strong representation system, i.e, querying the compact from is not equivalent to querying the possible worlds' form. Therefore, the need to introduce a new model that can treat the weaknesses of the EDB model.

In this chapter, we introduce the *evidential conditional databases*, named *ec-tables* and its assets. In fact, we use the strengths of the conditional databases and theory of belief functions to elaborate the ec-tables model that represents an advanced step towards strong representation systems under relational select-project operators. We also discuss the specificities of ec-tables and how to translate a table from the classical EDB model to the evidential conditional database model ECD.

5.2 Conditional Tables

Conditional databases were introduced to condition assumptions that can not coexist at the same time. For example, in the medical domain, we can find two diseases that can not appear together or two medicines that can not be used at the same time. In addition to that, the exclusivity of elements in the frame of discernment is based on the fact that the existence of one hypothesis implies the non existence of other hypotheses. Indeed, an evidential conditional database model relays on these two observations.

Several representations systems that cope with incomplete and probabilistic databases, were introduced in the literature: first, the *Codd tables* (Codd, 1972) appeared as relations that model incomplete data with null values. Later on, the *v-tables* extended Codd tables (Imielinski and Lipski, 1984). Then, other representation systems like *e-tables*, *g-tables*, *maybe tables*, *or-set-?-tables*, *c-tables*, (Imielinski and Lipski, 1984; Abiteboul et al., 1995b; Sarma et al., 2006) were introduced where each representation system handles imperfect data in a different context.

- **Codd tables**: Relations are annotated with constants and null variables (Codd, 1972). In the Codd table in Table 5.1, the name and the disease fields of lines 2 and 3 are unknown. Null values @ represent here the incompleteness of the data.

- **v-tables**: Relations involve constants and also variables that represent incomplete data (Imielinski and Lipski, 1984). Note here that variables define a correlation although they are unknown. For example, in Table 5.1, the v-table includes two unknown names for lines 2 and 3 that are *different*. Such information is not representable in Codd tables, because values are null without any additional information.

- **or-set-?-tables**: Relations include attributes on finite domains. Each attribute may be imprecise; i.e., it involves several values where only one is true. The "?" denotes the uncertainty that the tuple exists (Sarma et al., 2006). Table 5.1 illustrates an or-set-?-table where name and disease of the first and the second tuples, respectively, are imprecise. Also note that the second tuple may not exist at all.

An illustrative example is presented in Table 5.1 to clarify these representation systems of imperfect databases.

Codd table		**v-table**	
Name	**Disease**	**Name**	**Disease**
Smith	Anemia	Smith	Anemia
@	Asthma	x	Asthma
Brown	@	y	Asthma

or-set-?-table		
Name	**Disease**	
Smith	$< Anemia, Asthma >$	?
$< Smith, Brown >$	Asthma	
Brown	Asthma	

Table 5.1: Examples of Representation Systems of Imperfect Databases

- **c-tables**: It is a relation where the existence of each tuple is conditioned by a propositional formula over random variables called condition (Imielinski and Lipski, 1984). A conditional database is defined in (Suciu et al., 2011) as follows:

Defintion 31 *(Conditional Database CD)*

Named c-tables for short, $CD = \{R_1, ..., R_K, \Phi\}$ where $\{R_1, .., R_k\}$ is a relational database instance and Φ assigns a propositional formula Φ_t for each tuple t in each relation $R_1, .., R_k$. Given a valuation ∂ of variables in Φ, the world associated with ∂ is $W^\partial = \{R_1^\partial, .., R_k^\partial\}$ where $R_i^\partial = \{t | t \in R_i, \Phi_t[\partial] = true\}$ for each $i \in \{i, .., k\}$.

Example 55 *Table 5.2 is an illustration of a c-table that includes four attributes: the ID, the social security number SSN, the $Name$ and the $Condition$ over random variables for each tuple. The attribute $Condition$ involves conditions on random variables reflecting the existence of the labeled tuple and its relation with the other tuples. For example, the first $Name$ is 'Smith' that can not exist twice with two different social* *security numbers. Therefore, conditions on a random variable X is added to the first two tuples such that only one of them holds. Thus, if $X = 1$, the SSN of Smith is* 185*. In this case the second tuple does not hold. Otherwise, if $X \neq 1$, the first tuple does not hold and the SSN of Smith is* 785*. This table can be expanded into two possible worlds as detailed in Table 5.3.*

ID	**SSN**	**Name**	**Condition**
1	185	Smith	X=1
1	785	Smith	X≠1
2	186	Brown	Y=1

Table 5.2: Example of a c-table

W_1				W_2			
ID	**SSN**	**Name**	**Condition**	**ID**	**SSN**	**Name**	**Condition**
1	185	Smith	X=1	1	785	Smith	X≠1
2	186	Brown	Y=1	2	186	Brown	Y=1

Table 5.3: Possible worlds of the c-table of Table 5.2

(Suciu et al., 2011) have shown that the c-tables are strong representation systems. Thus, querying the compact c-table is equivalent to querying its possible worlds. In the general case, to prove that a representation system is strong, the equivalence between querying the compact form and the possible worlds' form should be demonstrated. There is no work that interested to combine the strengths of c-tables and the belief functions theory to build a strong representation database model.

5.3 Evidential Conditional Tables

In conditional tables (c-tables), a propositional formula is assigned to each tuple. This mechanism is very interesting because it defines the co-existence of one relation's tuples. For evidential conditional tables (ec-tables), this property is kept and a confidence level is assigned to each tuple in order to define the belief on tuples' existence. An evidential conditional database (ECD) has two equivalent forms: (i) the compact form and (ii) the possible worlds' form.

5.3.1 Modeling ec-tables: Compact form

Formally, evidential conditional tables are defined as follow:

Defintion 32 *(Evidential Conditional Tables ECD)*

An ***Evidential Conditional Database*****,** ***ec-tables*** *for short, on its compact form has N tuples and D attributes. An ec-table, is an evidential table with conditions over tuples. A confidence level, denoted* CL *is an attribute that contains degrees of confidence about the existence of each tuple. It is a pair of belief bel and plausibility pl such that:* $\{[bel; pl] \in [0;1] | bel \leq pl\}$

ID	**Disease**	**Condition**	**CL**
1	*Anemia*	x=1	[0.2; 0.8]
1	*Asthma*	x≠1	[0.5 ; 0.7]
2	*Anemia*	y=1	[0.5 ; 0.5]
2	*Cancer*	y≠1	[0.3 ; 0.9]

Table 5.4: A medical ec-table

Table 5.4 is an ec-table where a doctor expresses his diagnoses about several patients. In the first tuple, the doctor reveals that the first patient has either *Anemia* with a confidence level [0.2 ; 0.8] or *Asthma* with a confidence level [0.5 ; 0.7]. The attribute condition here indicates that when the first hypothesis is true (Disease=‘Anemia’), the second hypothesis (Disease=‘Asthma’) is false and vis versa. Confidence levels reflect the minimal and maximal beliefs about each tuple.

5.3.2 Modeling ec-tables: Possible Worlds’ Form

Most of theories that deal with imperfection represent the imperfect information as a distribution of hypotheses. Each one of them performs as a candidate to the solution. The main objective is to query the compact model and to get back a correct and reliable answers. Even though the only feasible model in practice is the compact one, the possible worlds’ form is a fundamental step to validate the querying methods. Thus, possible worlds provide shaper semantics when querying the imperfect database. In fact, generating possible worlds is a way to model the compact database into several states by treating its uncertainties.

For the ECD case, generating the possible worlds; i.e, the non compact form, goes through an intermediate representation where imprecision is treated.

Defintion 33 *(Non compact form of ECD)*

A Conditional Evidential Database, ECD for short, on its non compact form is a finite set of uncertain possible worlds such that:

$ECD = \{UW_1, UW_2, ..., UW_i\}$

Each uncertain possible world includes N tuples where each tuple contains one singleton focal element per attribute.

Example 56 *In Table 5.5, four uncertain worlds UW_1, UW_2, UW_3 and UW_4 are generated from Table 5.4.*

UW_1				UW_2			
1	*Anemia*	x=1	[0.2 ; 0.8]	1	*Anemia*	x=1	[0.2 ; 0.8]
2	*Anemia*	y=1	[0.5 ; 0.5]	2	*Cancer*	y≠1	[0.3 ; 0.9]
UW_3				UW_4			
1	*Asthma*	x≠1	[0.5 ; 0.7]	1	*Asthma*	x≠1	[0.5 ; 0.7]
2	*Anemia*	y=1	[0.5 ; 0.5]	2	*Cancer*	y≠1	[0.3 ; 0.9]

Table 5.5: Uncertain possible worlds of a medical ec-table

Then, each UW_i generates itself one or more possible worlds. Each possible world W_j is a combination of the uncertain world's tuples. Generating possible worlds is due to handling confidence levels. Thus, each possible world is a subset of tuples that can co-exist according to their defined propositional formulas, using equations (2.8) (2.7), (2.36).

Defintion 34 *(Possible World)*

A possible world, denoted W_j, is a subset of an uncertain world UW_i. Each generated possible world W_j has a confidence level CL_j that represents the bel_{W_j} and the pl_{W_j} of the existence of each tuple in the conditional table and it is defined such that:

$$CL = [bel_{W_j}; pl_{W_j}] \tag{5.1}$$

$$\begin{aligned} bel_{W_j} &= \prod bel(t_v) * \prod (1 - pl(t_z)) \\ pl_{W_j} &= \prod pl(t_v) * \prod (1 - bel(t_z)) \end{aligned} \tag{5.2}$$

$$\text{where } t_v, t_z \in UW_i \text{ and } t_v \in W_j; t_v \notin W_j$$

Note that equation (5.2) *is based on equations* (2.8),(2.7) *and* (2.36).

Suppose we have uncertain worlds with two tuples, the generated possible worlds and their corresponding confidence levels are computed based on tuples' existence as follows:

$\{t_1, t_2\}$:

- $bel(t_1 \wedge t_2) = bel_1 * bel_2$
- $pl(t_1 \wedge t_2) = pl_1 * pl_2$

$\{t_1\}$:

- $bel(t_1 \wedge \overline{t_2}) = bel_1 * (1 - pl_2)$
- $pl(t_1 \wedge \overline{t_2}) = pl_1 * (1 - bel_2)$

$\{t_2\}$:

- $bel(\overline{t_1} \wedge t_2) = (1 - pl_1) * bel_2$
- $pl(\overline{t_1} \wedge t_2) = (1 - bel_1) * pl_2$

$\{\varnothing\}$:

- $bel(\overline{t_1} \wedge \overline{t_2}) = (1 - pl_1) * (1 - pl_2)$
- $pl(\overline{t_1} \wedge \overline{t_2}) = (1 - bel_1) * (1 - bel_2)$

Example 57 *The number of possible worlds generated from Table 5.5 is $P = 16$. Confidence levels are computed using Definition 34. Possible worlds are shown in Table 5.6.*

UW_1	
W_{11}= {(1,Anemia),(2,Anemia)}	[0.1;0.4]
W_{12}={(1,Anemia)}	[0.1;0.4]
W_{13}= {(2,Anemia)}	[0.2;0.4]
W_{14}= {$\emptyset$}	[0.1;0.4]

UW_2	
W_{21}= {(1,Anemia),(2,Cancer)}	[0.06;0.72]
W_{22}={(1,Anemia)}	[0.02;0.56]
W_{23}= {(2,Cancer)}	[0.06;0.72]
W_{24}= {$\emptyset$}	[0.02;0.56]

UW_3	
W_{31}= {(1,Asthma),(2,Anemia)}	[0.25;0.35]
W_{32}={(1,Asthma)}	[0.25;0.35]
W_{33}= {(2,Anemia)}	[0.15;0.25]
W_{34}= {$\emptyset$}	[0.15;0.25]

UW_4	
W_{41}= {(1,Asthma),(2,Cancer)}	[0.15;0.63]
W_{42}={(1,Asthma)}	[0.05;0.49]
W_{43}= {(2,Cancer)}	[0.09;0.45]
W_{44}= {$\emptyset$}	[0.03;0.35]

Table 5.6: Possible worlds of ec-table

Since the ECD is an imperfect database model, querying its compact form using the relational operators requires their evaluation through querying also the possible worlds' form. Figure 5.1 illustrates the querying and the evaluation process for the ECD model under its both forms: the compact and the possible worlds.

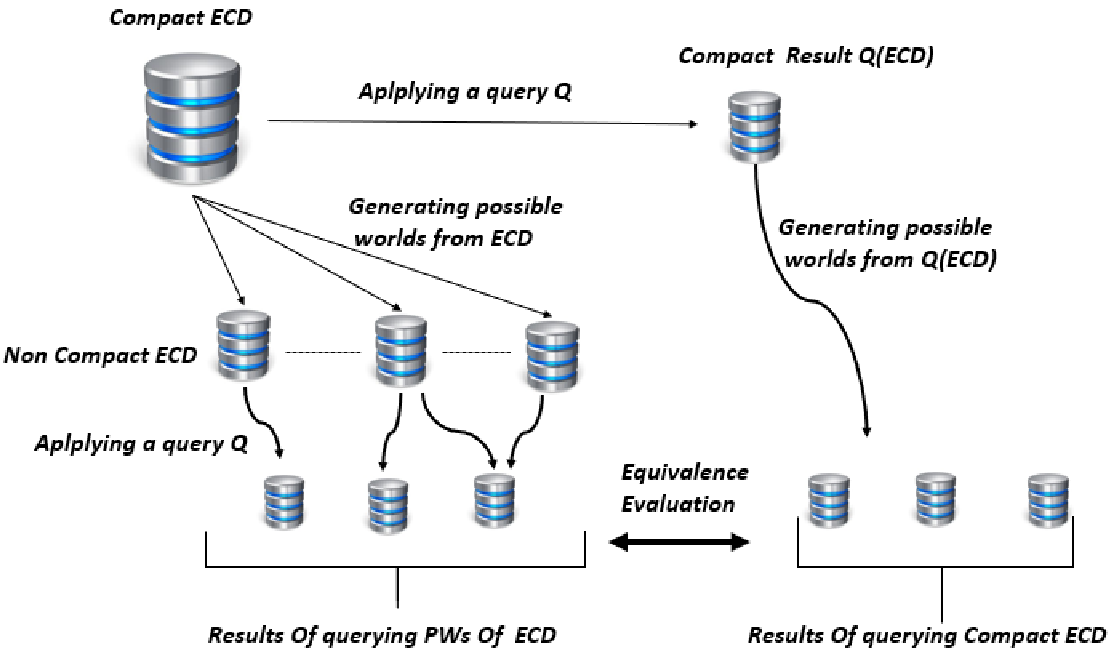

Compact ECD
Aplplying a query Q
Compact Result Q(ECD)
Generating possible worlds from ECD
Generating possible worlds from Q(ECD)
Non Compact ECD
Aplplying a query Q
Equivalence Evaluation
Results Of querying PWs Of ECD
Results Of querying Compact ECD

Figure 5.1: ECD Model

5.3.3 Querying ec-tables: Possible worlds form

Querying possible worlds of an ec-table denoted $Q(W_j)$ gives possible answers $\{R_1, ..., R_u\}$.

Example 58 *We process a query Q over possible worlds presented in Table 5.6.*

Q : *SELECT * FROM* ECD *WHERE* $< Disease = Cancer >$

$Q(W_{11}) = \{\emptyset\}$ *with [0.1 ; 0.4]* $= R_1$
$Q(W_{12}) = \{\emptyset\}$ *with [0.1 ; 0.4]* $= R_1$
$Q(W_{13}) = \{\emptyset\}$ *with [0.2 ; 0.4]* $= R_1$
$Q(W_{14}) = \{\emptyset\}$ *with [0.1 ; 0.4]* $= R_1$
$Q(W_{21}) = \{(2, Cancer)\}$ *with [0.06 ; 0.72]* $= R_2$
$Q(W_{22}) = \{\emptyset\}$ *with [0.02 ; 0.56]* $= R_1$
$Q(W_{23}) = \{(2, Cancer)\}$ *with [0.06 ; 0.72]*$= R_2$
$Q(W_{24}) = \{\emptyset\}$ *with [0.02 ; 0.56]* $= R_1$
$Q(W_{31}) = \{\emptyset\}$ *with [0.25 ; 0.35]* $= R_1$
$Q(W_{32}) = \{\emptyset\}$ *with [0.25 ; 0.35]* $= R_1$
$Q(W_{33}) = \{\emptyset\}$ *with [0.15 ; 0.25]* $= R_1$

Q(W_{34}) = $\{\emptyset\}$ with [0.15 ; 0.25] = R_1
Q(W_{41}) = $\{(2, Cancer)\}$with [0.15 ; 0.63] = R_2
Q(W_{42}) = $\{(2, Cancer)\}$ with [0.09 ; 0.45] = R_2
Q(W_{43}) = $\{\emptyset\}$ with[0.05 ; 0.49] = R_1
Q(W_{44}) = $\{\emptyset\}$ with [0.03 ; 0.35] = R_1

Applying the query Q over possible worlds may provide redundant tuples with different confidence levels. This is due to the combination of the tuples when generating the possible worlds; i.e. the CL of a same tuple can be taken into consideration in several possible worlds. The redundancy is treated using Definitions 29, 30.

Example 59 *If we carry on with the same example, two possible answers $\{R_1, R_2\}$ were generated. Each response with the two ones has different confidence levels.*

- *The first response $R_1 = \{\emptyset\}$ has 12 different CLs coming from W_{11}, W_{12}, W_{13}, W_{14}, W_{22}, W_{24}, W_{31}, W_{32}, W_{33}, W_{34}, W_{42} and W_{44}.*
- *The second response $R_2 = \{(2, Cancer)\}$ has 4 different CLs coming from W_{21}, W_{23}, W_{41} and W_{43}.*

Redundant answers are treated using Definition 30 such that:

- *R_1= $\{\emptyset\}$ [1-(1-0.1)*(1-0.1)*(1-0.2)*(1-0.1)*(1-0.02)*(1-0.02)*(1-0.25)*(1-0.25)*(1-0.15)*(1-0.15)*(1-0.05)*(1-0.03) ; 1-(1-0.4)*(1-0.4)*(1-0.4)*(1-0.4)*(1-0.56)*(1-0.56)*(1-0.35)*(1-0.35)*(1-0.25)*(1-0.25)*(1-0.49)*(1-0.35)] = [0.79; 0.998]*

- *R_2= $\{2, Cancer\}$ = [1-(1-0.06)*(1-0.06)*(1-0.15)*(1-0.09) ; 1-(1-0.72)*(1-0.72)*(1-0.63)*(1-0.45)] = [0.31;0.98]*

Thus, the final CLs for answers derived from the possible worlds' form when applying the query Q are then:

R_1= $\{\emptyset\}$ with CL_{R_1}=[0.79 ; 0.998]
R_2= $\{2, Cancer\}$ with CL_{R_2}=[0.31;0.98]

After querying the possible worlds of an ECD, we query the compact form in order to evaluate the equivalence between results in a further step.

5.3.4 Querying ec-tables: Compact form

Querying the compact form ECD gives a relation $Q(ECD)$ that responds to query Q. This result needs to be expanded into several possible states, called possible answers $\{R'_1, ..., R'_s\}$.

Example 60 *We process the same query Q over the compact form (Table 5.4), the result $(Q(ECD))$ is presented in Table 5.7.*

$Q(ECD)$		
2	*Cancer*	[0.3 ; 0.9]

Table 5.7: Compact result of query Q: $Q(ECD)$

The compact result itself is expanded into possible worlds as shown in Table 5.8. The first possible world W'_1 reveals the existence of the tuple result and the possible worlds W'_2 reveals the non existence of the tuple result.

W'_1		W'_2
2	Cancer	$\emptyset$

Table 5.8: The possible worlds of the compact result of query Q

The possible answer of query Q over the compact form are:
$R'_1 = \{2, Cancer\}$
$R'_2 = \{\emptyset\}$

- *The first response R'_1=\{(2,cancer)\} is coming from the following cases:*

  1. *When the first tuple with the condition (x=1) and the second tuple with the condition ($y \neq 1$) exist together. Hence, the first tuple that store the value $<$ Anemia $>$, does not appear in the set of results. The computed CL for this response $< \emptyset, (2, cancer) >$ is [0.06 ; 0.72].*
  2. *When the first tuple with condition (x=1) does not exist but the second tuple with the condition ($y \neq 1$) exists. Its computed CL is [0.06 ; 0.72].*
  3. *When the second tuple with condition ($y \neq 1$) exists with the first tuple with condition ($x \neq 1$). The latter tuple does not respond to the query Q. The computed CL for this response $< \emptyset, (2, cancer) >$ is [0.15 ; 0.63].*
  4. *When the first tuple with condition ($x \neq 1$) does not exist but the second tuple with ($y \neq 1$) exists. Its computed CL is [0.15 ; 0.63].*

- *The second response $R'_2 = \{\emptyset\}$ is also coming from 12 cases, are the following:*

  1. *Tuples with (x=1) and (y=1) exist together. Its CL is [0.1 ; 0.4].*
  2. *Tuple with (x=1) exists and tuple with (y=1) does not exist. Its CL is [0.1 ; 0.4].*
  3. *Tuple with (x=1) does not exist and tuple with (y=1) exists. Its CL is [0.2 ; 0.4].*
  4. *Both tuples with (x=1) and (y=1) do not exist. Its CL is [0.1 ; 0.4].*
  5. *Tuple with (x=1) exists and tuple with ($y \neq 1$) does not exist. Its CL is [0.02 ; 0.56].*
  6. *Both tuples with (x=1) and ($y \neq 1$) do not exist. Its CL is [0.02 ; 0.56].*
  7. *Tuples with ($x \neq 1$) and (y=1) exist together. Its CL is [0.25 ; 0.35].*
  8. *Tuple with (x≠1) exists and tuple with ($y = 1$) does not exist. Its CL is [0.25; 0.35].*
  9. *Tuple with (y=1) exists and tuple with ($x \neq 1$) does not exist. Its CL is [0.15; 0.25].*
  10. *None of tuples with conditions ($x \neq 1$) and ($y = 1$) exist. Its CL is [0.15 ; 0.25].*
  11. *Tuple with ($x \neq 1$) exists and tuple with ($y \neq 1$) does not exist. Its CL is [0.05 ; 0.49].*
  12. *None of tuples with conditions ($x \neq 1$) and ($y \neq 1$) exist. Its CL is [0.03 . 0.35].*

Answers R'_1 and R'_2 are redundant. Redundancy is treated using Definition 30. Their CLs are combined as follows:

- R'_1= $\{2, Cancer\}$ *= [1-(1-0.06)*(1-0.06)*(1-0.15)*(1-0.09) ; 1-(1-0.72)*(1-0.72)*(1-0.63)*(1-0.45)] = [0.31;0.98]*

- R'_2= $\{\emptyset\}$ *[1-(1-0.1)*(1-0.1)*(1-0.2)*(1-0.1)*(1-0.02)*(1-0.02)*(1-0.25)*(1-0.25)*(1-0.15)*(1-0.15)*(1-0.05)*(1-0.03) ; 1-(1-0.4)*(1-0.4)*(1-0.4)*(1-0.4)*(1-0.56)*(1-0.56)*(1-0.35)*(1-0.35)*(1-0.25)*(1-0.25)*(1-0.49)*(1-0.35)] = [0.79 ; 0.998]*

Thus, the final CLs for answers derived from the compact form when applying the query Q are:

R'_1= $\{2, Cancer\}$ *[0.31;0.98]*
R'_2= $\{\emptyset\}$ *[0.79 ; 0.998]*

5.3.5 Checking the equivalence

An ec-table is said to be SRS if querying its compact form is equivalent to querying its possible worlds. Note that according to the query of Example 58, we obtained exactly the same results for the possible worlds form and for the compact form, which are {(2,Cancer)}, $\{\emptyset\}$. This result was not surprising, since c-tables were proved to be a strong representation system (Imielinski and Lipski, 1984). In fact, c-tables are considered as a particular case of ec-tables where the CL is [1 ; 1]. Considering a CL $\subseteq$ [1 ; 1] has not an effect on content of the possible worlds but on their quantifications, i.e, their computed CLs.

Example 61 *For our example, the equivalence under the select-project query is valid. Indeed, possible answers are equivalent:* $\{R_1, R_2\}$=$\{R'_1, R'_2\}$.

Formally, according to Definition 22, ec-tables are strong representation system if $Rep(Q(ECD)) \equiv Q(Rep(ECD))$. $Rep(ECD) = W_{1..j}$ in a set of couples (W_j, CL_j); W_j is a possible world and CL_j is its associated confidence level. Thus, $Q(W_{1..j})$ is the application of Q on each world in $(Rep(ECD))$, and the result is the set of tuples satisfying Q in each world denoted R_u. On the other hand, $Q(ECD)$ provides the tuples in the ec-table that satisfy the query Q. It produces couples (t_j, CL_j) where t_j is a tuple satisfying Q and CL_j its confidence level. This process is exactly the same as in the regular c-tables, proved to be SRS in (Imielinski and Lipski, 1984) under the relational algebra. Now to quantify belief and plausibility of each world, either generated from $Q(Rep(ECD))$ or from $Rep(Q(ECD))$, we use the same tool, i.e., the evidential and cognitive independence formulas, based on the same confidence levels associated to the ec-table tuples.

5.4 Discussion

Evidential databases as defined in (Bell et al., 1996) can handle several types of imperfection like uncertainty, imprecision and ignorance. However, this model was proven not to be a strong representation system (cf. section 4.6). Transforming the evidential databases into ec-tables is a way to remodel this representation towards a strong representation system. In Example 62, below, we show how this transformation may be performed.

Example 62 *Table 5.9 represents an evidential database EDB as introduced in (Bell et al., 1996), with three attributes: ID, Disease and CL; where experts expressed their beliefs about some patients' diseases. Table 5.10 constitutes an ec-table with four attributes: ID, Disease, CL and Condition. Thus, it represents a translation of Table 5.9. Indeed, hypotheses are transformed to singletons and their confidence levels are a*

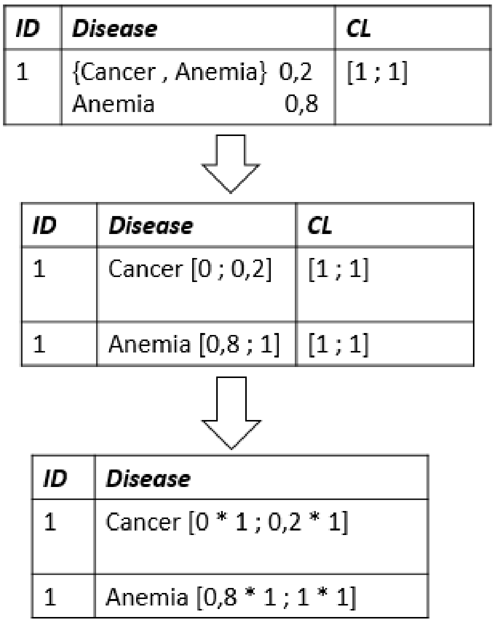
combination of masses and CLs. First, the belief and the plausibility of each hypothesis are deduced from its mass functions. Then, a confidence level is associated to this hypothesis. This CL is the product of bel and pl deduced from the masses (at the attribute level) and bel and pl deduced from the tuple's existence (at the tuple level).
For our example, the first tuple in Table 5.9, includes two hypotheses $< Anemia; Cancer >$. *This tuple is interpreted in ec-table of Figure 5.2. Thus, hypothesis* $< Cancer >$ *of the first tuple associates a computed CL. This CL is the product of bel and pl of* $< Cancer >$ *coming from the attribute level, and bel and pl of* $< Cancer >$ *coming from the tuple level. Indeed, the CL of hypothesis* $< Cancer >$ *is [0*1 ; 0.2*1] = [0 ; 0.2]. The CL of* $< Anemia >$ *in the first tuple is computed the same way; it is a combination of its CL coming the attribute level [0.8 ; (0.8+0.2)] and its CL coming the tuple level [1 ; 1], which is [0.8*1 ; 1*1] = [0.8 ; 1]. The rest of hypothesis in the table are considered the same way. Results of computed CLs are shown in Table 5.10.*

ID	Disease	CL
1	$\{Cancer, Anemia\}$ 0.2 $Anemia$ 0.8	[1 ; 1]
2	$Anemia$ 1	[0.6 ; 0.9]
3	$Asthma$ 0.5 $Anemia$ 0.5	[0.7 ; 1]

Table 5.9: An Evidential table EDB

ID	Disease	CL	Condition
1	$Anemia$	[0.8 ; 1]	$x = 1$
1	$Cancer$	[0 ; 0.2]	$x \neq 1$
2	$Anemia$	[0.6 ; 0.9]	$y = 1$
3	$Asthma$	[0.35 ; 0.5]	$z = 1$
3	$Anemia$	[0.35 ; 0.5]	$z \neq 1$

Table 5.10: Obtained Conditional Evidential table



ID	Disease	CL
1	{Cancer , Anemia} 0,2 Anemia 0,8	[1 ; 1]

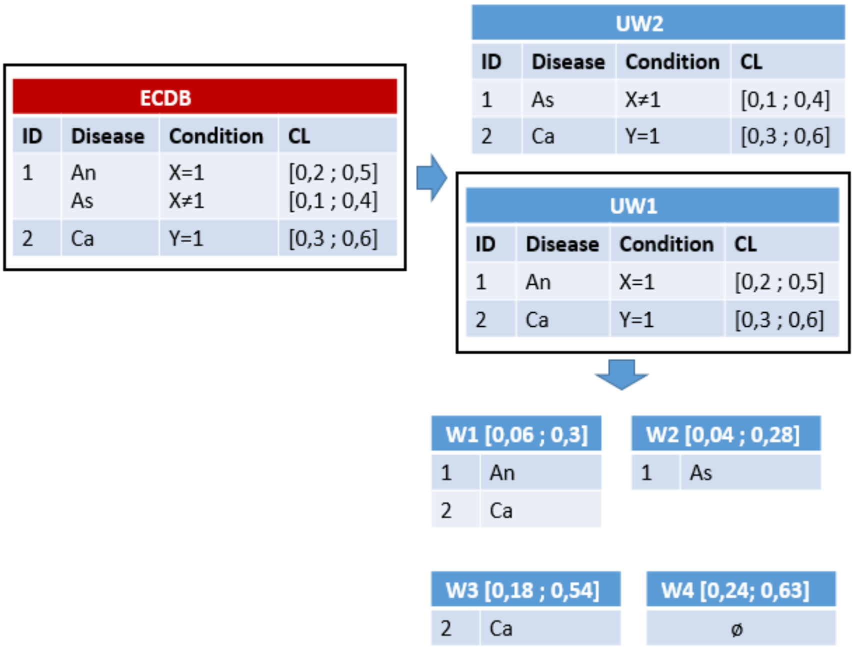
ID	Disease	CL
1	Cancer [0 ; 0,2]	[1 ; 1]
1	Anemia [0,8 ; 1]	[1 ; 1]

ID	Disease
1	Cancer [0 * 1 ; 0,2 * 1]
1	Anemia [0,8 * 1 ; 1 * 1]

Figure 5.2: The first tuple's interpretation from EDB to ECD

In addition, evidential conditional tables may be seen as a generalization of the probabilistic conditional tables, when considering only the belief degree of the confidence level. In this case, the belief degree is considered equivalent to the probability degree of each tuple. Moreover, when the confidence level is equal to [1;1], ec-tables and c-tables are equivalent. This interpretation is very important since the whole Dempster-Shafer theory is considered as the generalization of the probability in the discrete case.

Example 63 *We present now a complete example of modeling and querying an evidential database, being first an EDB and transformed then to an ECD. As detailed in Figure 5.4, first of all, we present an example of an evidential database EDB on its compact form. This model is represented on its possible worlds' form. Both forms are queried by applying query Q and the results are verified. Second, the compact EDB is translated into a compact ECD. The latter is modeled into its possible worlds and then both forms are interrogated with the same query Q. Finally, results of EDB and results of ECD are compared.*



ECDB			
ID	Disease	Condition	CL
1	An As	X=1 X≠1	[0,2 ; 0,5] [0,1 ; 0,4]
2	Ca	Y=1	[0,3 ; 0,6]

UW2			
ID	Disease	Condition	CL
1	As	X≠1	[0,1 ; 0,4]
2	Ca	Y=1	[0,3 ; 0,6]

UW1			
ID	Disease	Condition	CL
1	An	X=1	[0,2 ; 0,5]
2	Ca	Y=1	[0,3 ; 0,6]

W1 [0,06 ; 0,3]	
1	An
2	Ca

W2 [0,04 ; 0,28]	
1	As

W3 [0,18 ; 0,54]	
2	Ca

W4 [0,24; 0,63]
ø

Figure 5.3: From the compact ec-table to its possible worlds

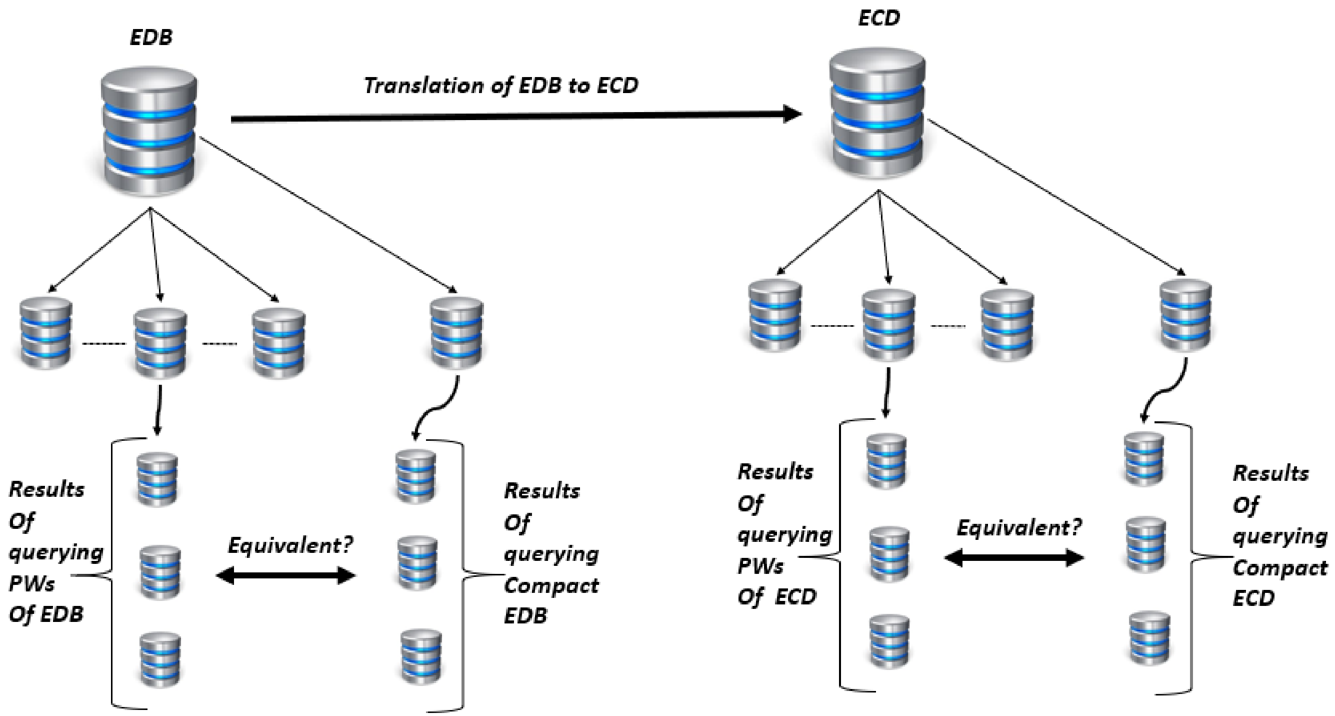

EDB
Translation of EDB to ECD
ECD
Results Of querying PWs Of EDB
Equivalent?
Results Of querying Compact EDB
Results Of querying PWs Of ECD
Equivalent?
Results Of querying Compact ECD

Figure 5.4: The complete process: From EDB to ECD

1. ***Modeling and Querying an EDB:***

   (a) ***Modeling the EDB from its compact form to its possible worlds' form:***

   Table 5.11 is an evidential database EDB that stores medical information about two patients given by a doctor.

ID	Disease	CL
1	{Cancer, Anemia} 0.2 Cancer 0.8	[0.5;0.8]
2	Asthma 0.5 Anemia 0.5	[0.7;1]

Table 5.11: A medical evidential database EDB

This compact EDB (Table 5.11) gives four imperfect possible worlds as shown in Table 5.12.

IW_1			IW_3		
ID	Disease	CL	ID	Disease	CL
1	{Cancer, Anemia}	[0.5;0.8]	1	Cancer	[0.5;0.8]
2	Asthma	[0.7;1]	2	Asthma	[0.7;1]

IW_2			IW_4		
ID	Disease	CL	ID	Disease	CL
1	{Cancer, Anemia}	[0.5;0.8]	1	Cancer	[0.5;0.8]
2	Anemia	[0.7;1]	2	Anemia	[0.7;1]

Table 5.12: Imperfect worlds of EDB

$m(IW_1) = 0.2 * 0.5 = 0.1$

$m(IW_2) = 0.2 * 0.5 = 0.1$

$m(IW_3) = 0.8 * 0.5 = 0.4$

$m(IW_4) = 0.8 * 0.5 = 0.4$

$m(IW_1) = m(\{UW_1, UW_2\}) = 0.1$

$m(IW_2) = m(\{UW_3, UW_4\}) = 0.1$

$m(IW_3) = m(\{UW_1\}) = 0.4$

$m(IW_4) = m(\{UW_3\}) = 0.4$

Imperfect worlds themselves give four uncertain possible worlds as shown in Table 5.13

$m(IW_1) = \{UW_1, UW_2\}$ *= 0.1*

UW_1			UW_2		
ID	Disease	CL	ID	Disease	CL
1	Cancer	[0.5;0.8]	1	Anemia	[0.5;0.8]
2	Asthma	[0.7;1]	2	Asthma	[0.7;1]

UW_3			UW_4		
ID	Disease	CL	ID	Disease	CL
1	Cancer	[0.5;0.8]	1	Anemia	[0.5;0.8]
2	Anemia	[0.7;1]	2	Anemia	[0.7;1]

Table 5.13: Uncertain worlds of EDB

$m(IW_2) = \{UW_3, UW_4\}$*= 0.1*
$m(IW_3) = \{UW_1\}$*= 0.4*
$m(IW_4) = \{UW_3\}$*= 0.4*

We compute now confidence levels of uncertain worlds by taking into consideration their inherited mass functions.

$UW_1 \in \{IW_1, IW_3\}$ *then* $UW1$ *[0.4;0.5]*
$UW_2 \in \{IW_1\}$ *then* $UW2$ *[0;0.1]*
$UW_3 \in \{IW_2, IW_4\}$ *then* $UW3$ *[0.4;0.5]*
$UW_4 \in \{IW_2\}$ *then* $UW4$ *[0;0.1]*

UW_1			UW_2		
ID	Disease	CL	ID	Disease	CL
1	Cancer	[0.2;0.4]	1	Anemia	[0;0.1]
2	Asthma	[0.28;0.5]	2	Asthma	[0;0.1]

UW_3			UW_4		
ID	Disease	CL	ID	Disease	CL
1	Cancer	[0.2;0.4]	1	Anemia	[0;0.1]
2	Anemia	[0.28;0.5]	2	Anemia	[0;0.1]

Table 5.14: Uncertain worlds of EDB

Each uncertain world gives a set of possible worlds as shown in Table 5.15.

(b) ***Querying the compact form of EDB:***

Let us process query Q on the compact EDB (Table 5.11).

```
Q: SELECT * FROM EDB WHERE (Disease=‘Cancer’);
```

UW_1	CL
W_{11}={(1,Cancer);(2,Asthma)}	[0.056;0.2]
W_{12}={(1,Cancer)}	[0.1;0.288]
W_{13}={(2,Asthma)}	[0.168;0.4]
W_{14}={∅}	[0.3;0.576]

UW_2	CL
W_{21}={(1,Anemia);(2,Asthma)}	[0;0.01]
W_{22}={(1,Anemia)}	[0;0.1]
W_{23}={(2,Asthma)}	[0;0.1]
W_{24}={∅}	[0.81;1]

UW_3	CL
W_{31}={(1,Cancer);(2,Anemia)}	[0.056;0.2]
W_{32}= {(1,Cancer)}	[0.1;0.288]
W_{33}={(2,Anemia)}	[0.168;0.4]
W_{34}={∅}	[0.3;0.576]

UW_4	CL
W_{41}={(1,Anemia);(2,Anemia)}	[0;0.01]
W_{42}={(1,Anemia)}	[0;0.1]
W_{43}={(2,Anemia)}	[0;0.1]
W_{44}= {∅}	[0.81;1]

Table 5.15: Possible worlds of EDB

Querying the compact EDB gives a compact result $Q(EDB)$ as shown in Table 5.16. This table generates possible answers R'.

ID	Disease	CL
1	{Cancer,Anemia} 0.2 Cancer 0.8	[1;1]

Table 5.16: Table result Q(EDB)

R'_1= *{1,Anemia} [0;0.2]*

R'_2=*{1,Cancer} [0.8;1]*

(c) ***Querying the possible worlds' from of EDB:***

Let's now process the same query Q over the non compact form.

$Q(W_{11})$= *{1,Cancer} with[0.056 ; 0.2]* = R_1

$Q(W_{12})$= *{1,Cancer} with[0.1 ; 0.288]* = R_1

$Q(W_{13})$= {∅} *with [0.168 ; 0.4]* = R_2

$Q(W_{14})$= {∅} *with[0.3 ; 0.576]* = R_2

$Q(W_{21})$= $\{\emptyset\}$with [0 ; 0.01] = R_2
$Q(W_{22})$= $\{\emptyset\}$ with [0 ; 0.1] = R_2
$Q(W_{23})$= $\{\emptyset\}$ with [0 ; 0.1] = R_2
$Q(W_{24})$= $\{\emptyset\}$ with [0.8 ; 1] = R_2
$Q(W_{31})$= $\{$1,Cancer$\}$ with [0.056 ; 0.2] = R_1
$Q(W_{32})$= $\{$1,Cancer$\}$ with [0.1 ; 0.288] = R_1
$Q(W_{33})$= $\{\emptyset\}$ with [0.168 ; 0.4] = R_2
$Q(W_{34})$= $\{\emptyset\}$ with [0.3 ; 0.576] = R_2
$Q(W_{41})$= $\{\emptyset\}$ with [0 ; 0.01] = R_2
$Q(W_{42})$= $\{\emptyset\}$ with [0 ; 0.1] = R_2
$Q(W_{43})$= $\{\emptyset\}$ with [0 ; 0.1] = R_2
$Q(W_{44})$= $\{\emptyset\}$ with [0.81 ; 1] = R_2

Applying query Q over the possible worlds gives the set of possible answers $\{R_1, R_2\}$. These answers are redundant. In fact, answer R_1=$\{$(1,Cancer)$\}$ has different CLs coming from W_{11}, W_{12}, W_{31} and W_{32}. Answer R_2=$\{\emptyset\}$ has also different CLs coming from worlds W_{13}, W_{14}, W_{21}, W_{22}, W_{23}, W_{24}, W_{33}, W_{34}, W_{41}, W_{42}, W_{43} and W_{44}. The redundancy of answers is treated using Definition 30 such that:
R_1 = $\{$(1,Cancer)$\}$ = [1-(1-0.056)(1-0.1)*(1-0.056)*(1-0.1);1-(1-0.2)*(1-0.288)*(1-0.2)*(1-0.288)] = [0.27; 0.67]*

R_2 = $\{\emptyset\}$ = [1-(1-0.168)(1-0.3)*(1-0)*(1-0)*(1-)*(1-0.8)*(1-0.168)*(1-0.3)*(1-0)*(1-0)*(1-0)*(1-0.81); 1- (1-0.4)*(1-0.576)*(1-0.01)*(1-0.1)*(1-1)*(1-0.2)*(1-0.288)*(1-0.4)*(1-0.576)*(1-0.01)*(1-0.1)*(1-1)] = [0.98 ; 1]*
Thus, the final answer when applying the query Q over the possible worlds of EDB:
R_1= $\{$(1,Cancer)$\}$ [0.27;0.67]
R_2= $\{\emptyset\}$ = [0.98 ; 1]

(d) ***Checking the equivalence:*** *It is obvious that both results; ie. the one coming from querying the compact form and the one coming from the possible worlds' form, are not equivalent.*
$\Rightarrow \{R'_1, R'_2\} \neq \{R_1, R_2\}$.

2. ***Translating an EDB to an ECD:*** *The compact evidential database EDB is translated into the compact evidential database ECD of Table 5.17. Confidence levels are attributed to each hypothesis by combining its CL in attribute level and its CL in tuple level.*

ID	Disease	CL	Condition
1	Cancer	[0.4;0.8]	$x = 1$
1	Anemia	[0;0.16]	$x \neq 1$
2	Asthma	[0.35;0.5]	$y = 1$
2	Anemia	[0.35;0.5]	$y \neq 1$

Table 5.17: The Obtained evidential database ECD after translation

3. ***Modeling and Querying an ECD:***

(a) ***Modeling the ECD from its compact form to its possible worlds' form:***

The compact ECD generates four uncertain worlds as shown in Table 5.18.

UW_1				UW_2			
ID	Disease	CL	Condition	ID	Disease	CL	Condition
1	Cancer	[0.4;0.8]	$x = 1$	1	Cancer	[0.4;0.8]	$x = 1$
2	Asthma	[0.35;0.5]	$y = 1$	2	Anemia	[0.35;0.5]	$y \neq 1$
UW_3				UW_4			
ID	Disease	CL	Condition	ID	Disease	CL	Condition
1	Anemia	[0;0.16]	$x \neq 1$	1	Anemia	[0;0.16]	$x \neq 1$
2	Asthma	[0.35;0.5]	$y = 1$	2	Anemia	[0.35;0.5]	$y \neq 1$

Table 5.18: Uncertain worlds of ECD

The uncertain worlds generate the set of possible worlds as shown in Table 5.19.

(b) ***Querying the compact from of ECD:***

Let's apply the same query Q on the compact ECD. The result is shown in Table 5.20.

```
Q: SELECT * FROM ECD WHERE (Disease='Cancer');
```

Two possible answer are derived from the compact ECD:

R'_1=*{(1,Cancer)}*

R'_2={∅}

- *The first answer* R'_1=*{(1,Cancer)} is coming from different cases:*
  - *When the first tuple with condition (x=1) and the second tuple with the condition (y=1) exist together, but the latter does not respond to the query Q. Its CL is [0.14 ; 0.4].*

UW_1	CL
W_{11}={(1,Cancer);(2,Asthma)}	[0.14;0.4]
W_{12}={(1,Cancer)}	[0.2;0.52]
W_{13}={(2,Asthma)}	[0.07;0.3]
W_{14}={$\emptyset$}	[0.1;0.39]

UW_2	CL
W_{21}={(1,Cancer);(2,Asthma)}	[0.14;0.4]
W_{22}={(1,Cancer)}	[0.2;0.52]
W_{23}={(2,Asthma)}	[0.07;0.3]
W_{24}={$\emptyset$}	[0.1;0.39]

UW_3	CL
W_{31}={(1,Anemia);(2,Asthma)}	[0;0.08]
W_{32}= {(1,Anemia)}	[0;0.11]
W_{33}={(2,Asthma)}	[0.3;0.5]
W_{34}={$\emptyset$}	[0.42;0.65]

UW_4	CL
W_{41}={(1,Anemia);(2,Anemia)}	[0;0.08]
W_{42}= {(1,Anemia)}	[0;0.11]
W_{43}= {(2,Anemia)}	[0.3;0.5]
W_{44}={$\emptyset$}	[0.42;0.65]

Table 5.19: Possible worlds of EDB

1	Cancer	[0.4;0.8]	$X = 1$

Table 5.20: The compact result Q(ECD)

  - *When the first tuple with (x=1) exists and the second tuple with (y=1) does not. Its CL is [0.2 ; 0.52].*
  - *When both tuples with condition (x=1) and ($y \neq 1$) exist but the latter does not respond to Q. Its CL is [0.14 ; 0.4].*
  - *When the first tuple with (x=1) exists and the second tuple with ($y \neq 1$) does not exist. Its CL is [0.2 ; 0.52]*

- *The second answer R_2'={$\emptyset$} is also coming from different cases, computed the same way; i.e, considering all the cases where R_2' can appear based on the existence of tuples in the compact database of Table 5.17. Computed CLs are:*
  - *R_2'=[0.07 ; 0.3]*
  - *R_2'=[0.1 ; 0.39]*
  - *R_2'=[0.07 ; 0.8]*

- R_2'=*[0.1 ; 0.39]*
- R_2'=*[0 ; 0.08]*
- R_2'=*[0 ; 0.4]*
- R_2'=*[0.3 ; 0.5]*
- R_2'=*[0.42 ; 0.65]*
- R_2'=*[0 ; 0.08]*
- R_2'=*[0 ; 0.11]*
- R_2'=*[0.3 ; 0.5]*
- R_2'=*[0.42 ; 0.65]*

Redundancy of answers R_1' *and* R_2' *are treated such that:*

- R_1' = {*(1,Cancer)*} = *[1-(1-0.14)*(1-0.2)*(1-0.14)*(1-0.2);1-(1-0.4)*(1-0.52)*(1-0.4)*(1-0.52)] = [0.52; 0.91]*

- $R_2' = \{\emptyset\}$ = *[1-(1-0.07)*(1-0.1)*(1-0.07)*(1-0.1)*(1-0)*(1-0)*(1-0.3)*(1-0.42)*(1-0)*(1-0)*(1-0.3)*(1-0.42); 1- (1-0.3)*(1-0.39)*(1-0.8)*(1-0.39)*(1-0.08)*(1-0.4)*(1-0.5)*(1-0.65)*(1-0.08)*(1-0.11)*(1-0.5)*(1-65)] = [0.88 ; 0.999]*

Thus, answers derived from applying query Q over the compact ECD are:

R_1' = {*(1,Cancer)*} =*[0.52; 0.91]*

R_2' = $\{\emptyset\}$ = *[0.88 ; 0.999]*

(c) ***Querying the possible worlds' from of ECD:***

Let's now apply the same query Q over the possible worlds' form:

$Q(W_{11}) = \{1, Cancer\}$ *with [0.14 ; 0.4]* $= R_1$

$Q(W_{12}) = \{1, Cancer\}$ *with [0.2 ; 0.52]* $= R_1$

$Q(W_{13}) = \{\emptyset\}$ *with [0.07 ; 0.3]* $= R_2$

$Q(W_{14}) = \{\emptyset\}$ *with [0.1 ; 0.39]* $= R_2$

$Q(W_{21}) = \{1, Cancer\}$ *with [0.14 ; 0.4]* $= R_1$

$Q(W_{22}) = \{1, Cancer\}$ *with [0.2 ; 0.52]* $= R_1$

$Q(W_{23}) = \{\emptyset\}$ *with [0.07 ; 0.3]* $= R_2$

$Q(W_{24}) = \{\emptyset\}$ *with [0.1 ; 0.39]* $= R_2$

$Q(W_{31}) = \{\emptyset\}$ *with [0 ; 0.08]* $= R_2$

$Q(W_{32}) = \{\emptyset\}$ *with [0 ; 0.11]* $= R_2$

$Q(W_{33}) = \{\emptyset\}$ *with [0.3 ; 0.5]* $= R_2$

$Q(W_{34}) = \{\emptyset\}$ *with[0.42 ; 0.65]* $= R_2$

$Q(W_{41}) = \{\emptyset\}$ *with [0 ; 0.08]* $= R_2$

$Q(W_{42}) = \{\emptyset\}$ *with [0 ; 0.11]* $= R_2$

$Q(W_{43}) = \{\emptyset\}$ *with [0.3 ; 0.5]* $= R_2$

$Q(W_{44}) = \{\emptyset\}$ *with [0.42 ; 0.65]* $= R_2$

Applying query Q over the possible worlds of ECD gives a set of possible answers. Due to redundancy, multiple confidence levels are given. To combine the selected tuples, we use Definitions 29 and 30.

- $R_1 = \{$*(1,Cancer)*$\}$ = *[1-(1-0.14)*(1-0.2)*(1-0.14)*(1-0.2);1-(1-0.4)*(1-0.52)*(1-0.4)*(1-0.52)] = [0.52 ; 0.91]*

- $R_2 = \{\emptyset\}$ = *[1-(1-0.07)*(1-0.1)*(1-0.07)*(1-0.1)*(1-0)*(1-0)*(1-0.3)*(1-0.42)*(1-0)*(1-0)*(1-0.3)*(1-0.42); 1- (1-0.3)*(1-0.39)*(1-0.8)*(1-0.39)*(1-0.08)*(1-0.4)*(1-0.5)*(1-0.65)*(1-0.08)*(1-0.11)*(1-0.5)*(1-65)] = [0.88 ; 0.999]*

Thus, the answers derived from applying query Q over the possible worlds' form are:

$R_1 = \{$*(1,Cancer)*$\}$ = *[0.52 ; 0.91]*

$R_2 = \{\emptyset\}$ =*[0.88 ; 0.999]*

(d) ***Checking the equivalence:***

As shown both results, coming from the compact ECD and from the non compact ECD, are equivalent, namely:

$\{R'_1, R'_2\} = \{R_1, R_2\}$

4. ***Discussing results:***

Even though the classical evidential model, denoted EDB can be more informative than the ECD model, it does not provide a strong representation system under relational operators. In the other side, we showed that the ECD model is an advanced step towards a strong representation system. Translating the EDB model to the ECD model is then very beneficial. Thus, the EDB can explicitly illustrate the imperfect information within the belief functions' theory tools and the ECD manages very well the querying methods under relational operators. In fact, we showed throughout Example 62 that querying the EDB model into its both forms provides non equivalent results, but transforming this model into an ECD and querying it into its both forms provide equivalent ones.

5.5 Conclusion

A representation system may be queried over its compact or its possible worlds forms. While the possible worlds form is not an implementable model, proving that a representation system is strong may reduce computational complexity by querying directly the equivalent compact form.

Even though, evidential databases as introduced in (Bell et al., 1996; Lee, 1992a; Lee, 1992b) are more informative than ec-tables. They do not represent a strong representation system. Thus, we proved with a counterexample that this model do not provide the same results when querying its compact form and when querying its possible worlds' form. To solve this problem, we introduced the ec-tables that are based on singleton hypothesis and conditions. Querying the compact and the non compact forms of ec-tables generates equivalent results. Added to that, we showed how to transform any classical evidential database to ec-tables, in order to benefit from its querying methods' efficiency.

Implementing the proposed evidential conditional model and evaluating other types of queries like skyline (Bousnina et al., 2017b; Elmi et al., 2017; Yong et al., 2014) and top-k over ec-tables, constitutes a very interesting and promising future work.

Conclusion

Generally, modeling imperfect information can be a very challenging process. Thus, several theories were introduced to handle imperfect information depending on contexts, domains and the nature of information. Theory of belief functions (or the evidence theory) is one of these theories that offered efficient tools to model, manage and combine imperfect data. In fact, it provides a good framework to represent uncertainty, imprecision and ignorance. Evidential databases store information modeled via the belief functions theory. It has two equivalent representations: the compact and the possible worlds. Results of querying both forms should be equivalent to consider the database model as a strong representation system.

Our aim in this thesis is to investigate all techniques and methods of modeling and querying evidential databases in order to set a strong representation system. Indeed, we started by developing the possible worlds' representation for evidential databases EDB. While generating the non compact form, we treated the tuple level uncertainty and the attribute level uncertainty. This step is essential to validate querying methods over the compact form. Then, we developed an object-oriented implementation for evidential databases EDB in order to prepare the querying step. Thanks to the object-oriented design, we applied several types of querying over the compact form: evidential relational queries, evidential top-k queries and evidential skyline queries. Results were satisfying in terms of execution time for all applied queries. Added to that, we used a real data that we extracted from the Tripadvisor platform in order to construct the evidential database. To construct this database, we used the mathematical tools provided by the belief functions theory. We also proposed a formalism to rank evidential results and select the best ones. Studying in depth the evidential database EDB guided us to conclude that this representation system is not strong. Thus, we prove with a counter example that querying the compact EDB and querying the non compact EDB does not provide the same results, neither the same confidence levels. Therefore, we introduced a new evidential database system: the ec-tables ECD. This database model combines the strengths of conditional tables and the specificities of the evidence theory. After modeling the evidential ECD from its compact form to its possible worlds' form, we proved that this system offers a strong representation system under select-project operators.

Our major contributions are summarized as follows:

(1) Modeling and querying the compact evidential database (EDB):

- Object-Relational Evidential Implementation (Bousnina et al., 2016): We presented an object relational model for the evidential database on its compact form (Lee, 1992b; Lee, 1992a; Bell et al., 1996). Then we presented the implementation of this object relational model using SQL3 and Java. Finally, we evaluated the select and the project operators under the belief functions framework.

- Evidential Top-K query (Bousnina et al., 2017a; Bousnina et al., 2018b): We introduced a new ranking querying formalism for the evidential data. Its aim is to select the best k responses when querying an evidential database on its compact form. We made use of the object-relational evidential implementation to implement and evaluate the evidential Top-k query. Added to that, we introduced new semantics for this kind of queries.

- Evidential skyline query (Bousnina et al., 2017b): We treated the aggregation of information coming from different sources to construct an evidential database where information are coming from a real platform (*the TripAdvisor platform*). Then, we applied the evidential skyline query as introduced in (Elmi et al., 2014).

(2) Modeling and querying the possible worlds of the evidential database (EDB):

Modeling Evidential databases as Possible worlds' (Bousnina et al., 2018a): We modeled the evidential database (Bell et al., 1996) on its possible worlds' form. We used a previous contribution as a basis (Bousnina et al., 2015) for this work. Thus, we modeled the non compact form of the evidential database by treating the two levels of uncertainties (the tuple level uncertainty and the attribute level uncertainty). This contribution was a very important step towards the evaluation and the validation of querying methods of an EDB (Lee, 1992b; Lee, 1992a; Bell et al., 1996).

(3) Modeling and querying the evidential conditional database (ECD):

Evidential Conditional Tables: We proved that the most used evidential database EDB (Lee, 1992b; Lee, 1992a; Bell et al., 1996) is not a strong representation system. Then, we introduced a new evidential database model, named ec-tables (ECD). We discuss how the ec-tables present a vigorous basis towards a strong representation system under relational operators. We used several detailed examples to support the presented formalism.

As future works, we intend to improve and extend our researches in other directions:

- In the last part of this thesis, we developed the ec-tables model under select-project. We intend to validate the rest of querying method (Cartesian product, join and union) for evidential conditional databases. In fact, checking if the ec-tables are strong representation system is a very complex mathematical process and needs to evaluate all relational operators.
- We started to collect real data in the medical domain where doctors give their diagnoses and their beliefs about patients' diseases. To evaluate the queries and their results, we started to implement the evidential conditional database model. Conditions are essential in such application because some hypotheses (like diseases, blood types, pharmaceuticals, etc,.) can not coexist at the same time.
- Using the evidential conditional databases in aeronautics can be a very promising future work. Thus, in air crashes for example, hypotheses relative to reasons of the crash can not coexist simultaneously. Added to that, hypotheses are given by experts and can associate degrees of truthfulness about each hypothesis. Therefore, we intend to collect real data and evaluate the relational queries that we intend to apply over this data.
- We presented the evidential top-k queries (Bousnina et al., 2017a) and the evidential skyline queries (Elmi et al., 2014; Bousnina et al., 2017b; Abidi et al., 2018; Elmi et al., 2016) for the EDB model. It will be interesting to extend these queries using the ec-tables model.
- Data-mining methods applied earlier over evidential databases (Bach Tobji and Ben Yaghlane, 2011; Bach Tobji et al., 2008; Samet and Dao, 2015; Samet et al., 2014; Samet et al., 2016) can be extended to be applied over the ECD model.
- It would be also interesting to extend the evidential models to the case of Nosql datasets framework.

Bibliography

Abidi, A., Elmi, S., Tobji, M. A. B., Hadjali, A., and Yaghlane, B. B. (2018). Skyline queries over possibilistic RDF data. *International Journal of Approximate Reasoning*, 93:277–289.

Abiteboul, S., Hull, R., and Vianu, V. (1995a). *Foundations of Databases.*

Abiteboul, S., Hull, R., and Vianu, V. (1995b). *Foundations of databases: the logical level.* Addison-Wesley Longman Publishing Company Inc.

Agrawal, P., Benjelloun, O., Sarma, A. D., Hayworth, C., Nabar, S., Sugihara, T., and Widom, J. (2006). Trio: A system for data, uncertainty, and lineage. In *32nd International Conference on Very Large Data Bases*, pages 1151–1154, Seoul, Korea.

Agrawal, P. and Widom, J. (2010). Generalized uncertain databases: first steps. Technical report.

Amato, G., Rabitti, F., Savino, P., and Zezula, P. Region proximity in metric spaces and its use for approximate similarity search. *Transactions on Information Systems.*

Ayoun, A. and Smets, P. (2001). Data association in multi-target detection using the transferable belief model. *International Journal of Intelligent Systems*, 16(10):1167–1182.

Bach Tobji, M. A. and Ben Yaghlane, B. (2011). Extraction des itemsets fréquents à partir de données imparfaites : application à une base de données éducationnelles. *Revue des Nouvelles Technologies de l'Information, Fouille de données complexes-Complexité liée aux données multiples*, pages 211–231.

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

Barbara, D., Garcia-Molina, H., and Porter, D. (1992). The management of probabilistic data. *IEEE Transactions on knowledge and data engineering*, 4(5):487–502.

Basir, O. and Yuan, X. (2007). Engine fault diagnosis based on multi-sensor information fusion using dempster shafer evidence theory. *Information Fusion*, 8(4):379–386.

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

Ben Yaghlane, A., Denœux, T., and Mellouli, K. (2008). Elicitation of expert opinions for constructing belief functions. *Uncertainty and Intelligent Information Systems*, pages 75–88.

Börzsönyi, S., Kossmann, D., and Stocker, K. (2001). The skyline operator. In *the 17th International Conference on Data Engineering*, pages 421–430, Heidelberg, Germany.

Bosc, P., Duval, L., and Pivert, O. (2003). An initial approach to the evaluation of possibilistic queries addressed to possibilistic databases. *Fuzzy Sets and systems*, 140(1):151–166.

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

Bosc, P. and Pivert, O. (2010). Modeling and querying uncertain relational databases: A survey of approaches based on the possible worlds semantics. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 18(5):565–603.

Bousnina, F. E., Bach Tobji, M. A., Chebbah, M., and Ben Yaghlane, B. (2018a). Modeling evidential databases as possible worlds. *International Journal of Intelligent Systems*, 33(6):1146–1164.

Bousnina, F. E., Bach Tobji, M. A., Chebbah, M., Liétard, L., and Ben Yaghlane, B. (2015). A new formalism for evidential databases. In *the 22nd International Symposium on Methodologies for Intelligent Systems*, pages 31–40, Lyon, France.

Bousnina, F. E., Chebbah, M., Bach Tobji, M. A., Hadjali, A., and Ben Yaghlane, B. (2017a). On top-k queries over evidential data. In *the 19th International Conference on Enterprise Information Systems*, volume 1, pages 106–113.

Bousnina, F. E., Chebbah, M., Bach Tobji, M. A., Hadjali, A., and Ben Yaghlane, B. (2018b). Evidential top-k queries evaluation: Algorithms and experiments. In *the17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 407–417, Cádiz, Spain.

Bousnina, F. E., Elmi, S., Bach Tobji, M. A., Chebbah, M., Hadjali, A., and Ben Yaghlane, B. (2016). Object-relational implementation of evidential databases. In *the 1st International Conference on Digital Economy*, pages 80–87, Carthage, Tunisia.

Bousnina, F. E., Elmi, S., Chebbah, M., Bach Tobji, M. A., Hadjali, A., and Ben Yaghlane, B. (2017b). Skyline operator over tripadvisor reviews within the belief functions framework. In *the 2nd International Conference on Digital Economy*, pages 186–197, Sidi Bou Said, Tunisia.

Cavallo, R. and Pittarelli, M. (1987). The theory of probabilistic databases. In *the 13th Very Large Data Bases Conference, Brighton*, pages 71–81.

Chebbah, M., Martin, A., and Ben Yaghlane, B. (2015). Combining partially independent belief functions. *Decision Support Systems*, 73:37–46.

Choenni, S., Blok, H. E., and Leertouwer, E. (2006). Handling uncertainty and ignorance in databases: A rule to combine dependent data. In *the 11th International Conference on Database Systems for Advanced Applications*, pages 310–324.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387.

Codd, E. F. (1972). *Relational completeness of data base sublanguages.* IBM Corporation.

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

Dubois, D., Foulloy, L., Mauris, G., and Prade, H. (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable computing*, 10(4):273–297.

Dubois, D. and Prade, H. (1987). *Théorie des possibilités.* Masson, Paris.

Dubois, D. and Prade, H. (1988). Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4:244–264.

Elmi, S., Bach Tobji, M. A., Hadjali, A., and Ben Yaghlane, B. (2016). Efficient skyline maintenance over frequently updated evidential databases. In *the 16th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 199–210. Springer.

Elmi, S., Bach Tobji, M. A., Hadjali, A., and Ben Yaghlane, B. (2017). Selecting skyline stars over uncertain databases: Semantics and refining methods in the evidence theory setting. *Applied Soft Computing*, 57:88–101.

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

Elmi, S., Benouaret, K., Hadjali, A., Bach Tobji, M. A., and Ben Yaghlane, B. (2015). Requêtes skyline en présence des données évidentielles. In *Extraction et Gestion des Connaissances*, pages 215–220.

Ennaceur, A., Elouedi, Z., and Lefevre, E. (2014). Multi-criteria decision making method with belief preference relations. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 22(04):573–590.

Essaid, A., Martin, A., Smits, G., and Ben Yaghlane, B. (2014). A distance-based decision in the credal level. In *12th International Conference on Artificial Intelligence and Symbolic Computation*, pages 147–156. Springer.

Fagin, R. (1996). Combining fuzzy information from multiple systems. In *the 15th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 216–226.

Fagin, R. (1998). Fuzzy queries in multimedia database systems. In *17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 1–10.

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

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

Fuhr, N. (1992b). Probabilistic models in information retrieval. *The Computer Journal*, 35:243–255.

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

Hau, H.-Y. and Kashyap, R. L. (1990). Belief combination and propagation in a lattice-structured interference network. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(1):45–57.

Hong, S. J. and May, G. S. (2004). Neural network-based real-time malfunction diagnosis of reactive ion etching using in situ metrology data. *IEEE transactions on semiconductor manufacturing*, 17(3):408–421.

Ilyas, I. F., Beskales, G., and Soliman, M. A. (2008). A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys*, 40(4):11.

Imielinski, T. and Lipski, W. (1984). Incomplete information in relational databases. *Journal of the ACM*, 31(4):761–791.

Janez, F. (1997). *Rappels sur la théorie de l'évidence.* Office national d'études et de recherches aérospatiales.

Kamel, H. (1954). Relational algebra and uniform spaces. *Journal of the London Mathematical Society*, 1(3):342–344.

Laplace, P. S. (1812). Théorie analytique des probabilités. *Courcier Corporation.*

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

Lee, S. K. (1992b). Imprecise and uncertain information in databases : an evidential approach. In *the 8th International Conference on Data Engineering*, pages 614–621, Washington, USA.

Martin, A. (2019). *Conflict management in information fusion with belief functions.* Information Quality in Information Fusion and Decision Making.

Moore, R. L. (1932). *Foundations of point set theory*, volume 13. American Mathematical Society.

Pawlak, Z. (1982). Rough sets. *International journal of computer & information sciences*, 11(5):341–356.

Pfeiffer, P. E. (2013). *Concepts of probability theory.* Courier Corporation.

Prade, H. and Testemale, C. (1984). Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries. *Information sciences*, 34(2):115–143.

Re, C., Dalvi, N., and Suciu, D. (2007). Efficient top-k query evaluation on probabilistic data. In *the 23rd International Conference on Data Engineering*, pages 886–895. IEEE.

Samet, A. and Dao, T. (2015). Mining over a reliable evidential database: Application on amphiphilic chemical database. In *14th International Conference on Machine Learning and Applications*, pages 1257–1262, Miami, USA. IEEE.

Samet, A., Lefèvre, E., and Yahia, S. B. (2014). Classification with evidential associative rules. In *the 15th International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 25–35, Montpellier, France.

Samet, A., Lefèvre, E., and Yahia, S. B. (2016). Evidential data mining: precise support and confidence. *Journal of Intelligent Information Systems*, 47(1):135–163.

Sarma, A. D., Benjelloun, O., Halevy, A., and Widom, J. (2006). Working models for uncertain data. In *the 22nd International Conference on Data Engineering*, pages 7–7. IEEE.

Shafer, G. (1976). A mathematical theory of evidence. *Princeton University Press.*

Smets, P. (1988). Belief functions. In *Non-standard logics for automated reasoning*, pages 253–286. Academic Press, London.

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

Smets, P. (1995). The canonical decomposition of a weighted belief. In *the 14th International Joint Conference on Artificial Intelligence*, volume 95, pages 1896–1901, Montréal, Canada.

Smets, P. (1996). Imperfect information: imprecision and uncertainty. In *Uncertainty Management in Information Systems: From Needs to Solution*, pages 225–254. Kluwer Academic Publishers.

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

Smets, P. (1998b). The transferable belief model for quantified belief representation. In *Quantified Representation of Uncertainty and Imprecision*, pages 267–301. Springer.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66(2):191–234.

Soliman, M. A., Ilyas, I. F., and Chang, K. C.-C. (2007). Top-k query processing in uncertain databases. In *the 23rd International Conference on Data Engineering*, pages 896–905, Istanbul, Turkey. IEEE.

Suciu, D., Olteanu, D., Ré, C., and Koch, C. (2011). *Probabilistic databases*, volume 3. Morgan & Claypool Publishers.

Theobald, M., Schenkel, R., and Weikum, G. (2005). An efficient and versatile query engine for topx search. In *the 31st International Conference on Very large Data Bases*, pages 625–636, Trondheim, Norway. VLDB Endowment.

Wang, Y.-M., Yang, J.-B., and Xu, D.-L. (2005). A preference aggregation method through the estimation of utility intervals. *Computers & Operations Research*, 32(8):2027–2049.

Willink, R. (2006). Principles of probability and statistics for metrology. *Metrologia*, 43(4).

Yager, R. R. (1987). On the dempster-shafer framework and new combination rules. *Information Sciences*, 41(2):93–137.

Yang, B.-S. and Kim, K. J. (2006). Application of dempster–shafer theory in fault diagnosis of induction motors using vibration and current signals. *Mechanical Systems and Signal Processing*, 20(2):403–420.

Yang, J., Huang, H.-Z., He, L.-P., Zhu, S.-P., and Wen, D. (2011). Risk evaluation in failure mode and effects analysis of aircraft turbine rotor blades using dempster–shafer evidence theory under uncertainty. *Engineering Failure Analysis*, 18(8):2084–2092.

Yong, H., Lee, J., Kim, J., and Hwang, S.-w. (2014). Skyline ranking for uncertain databases. *Information Sciences*, 273:247–262.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

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

Zhang, S.-y., Pan, Q., and Zhang, H.-c. (2001). Conflict problem of dempster-shafer evidence theory (n). *Acta Aeronautica Et Astronautica Sinica*, 22(4):369–369.

Modeling and Querying Evidential Databases

An evidential database has two equivalent representations: (1) the compact representation represented as a set of attributes and a set of tuples; (2) the possible worlds' representation modeled as a distribution of candidate databases. Querying the possible worlds' form is a fundamental step in order to check the querying methods over the compact one. In fact, a model is said to be a strong representation system when results of querying its compact form are equivalent to results of querying its non compact form. Throughout this thesis, we study the foundations of evidential databases in both modeling and querying via three major parts: (i) first by modeling and querying the compact form of the evidential database (EDB); (ii) second by modeling the possible worlds' form of the evidential database (EDB) through treating the tuple level uncertainty and the attribute level uncertainty; (iii) finally by modeling and querying the evidential conditional database (ECD) in its both forms (the compact and the non compact).

Keywords: Databases Management; Imperfect databases; Dempster-Shafer theory; Evidential Databases; Possible worlds; Evidential Conditional databases; Representation system; Querying; Evidential Top-k.

Modélisation et Exploitation des Bases de Données Crédibilistes

Une base de donnée crédibiliste a deux représentations équivalentes: (1) la représentation compacte caractérisée par un ensemble d'attributs et un ensemble de tuples; (2) la représentation des mondes possibles représentée par une distribution de base de données candidates. Interroger la représentation des mondes possibles est une étape fondamentale pour valider les méthodes d'interrogation sur la base compacte crédibiliste. En effet, un modèle de base de donnée est dit système fort si le résultat de l'interrogation de sa représentation compacte est équivalent au résultat de l'interrogation de sa représentation des mondes possibles. Tout au long de cette thèse, nous étudions les fondements des bases de données crédibilistes à travers trois parties majeures: (i) premièrement par la modélisation et l'interrogation de la base de donnée crédibiliste (EDB) sous sa forme compacte; (ii) deuxièmement par la modélisation de la base de données crédibiliste (EDB) sous sa forme des mondes possibles en traitant les niveaux incertitudes par attributs et par tuples; (iii) finalement par la modélisation et l'interrogation de la base de données crédibiliste (ECD) sous ces formes compacte et non compacte.

Mots clés: Bases de données – Gestion; Bases de données imparfaites; Théorie de Dempster-Shafer; Bases de données crédibilistes; Bases de données crédibilistes conditionnelles; Mondes possibles; Système de représentation; Bases de données – Interrogation; Top-k évidentiel.