ANNÉE (2015)

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

En cotutelle internationale avec
Université de Tunis, Tunisie

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique
Ecole doctorale MATISSE

présentée par

Amira Essaïd

préparée à l'unité de recherche UMR 6074 IRISA
Institut de Recherche en Informatique et Systèmes Aléatoires
ENSSAT

Decision Making for Ontology Matching under the Theory of Belief Functions

Thèse à soutenir à l'université de Rennes 1
le 01/06/2015

devant le jury composé de :

Anne-Laure JOUSSELME
Chercheur au NATO CMRE, Italie / rapporteur

Chantal REYNAUD
Professeur à l'Université Paris Sud, France / rapporteur

Nahla BEN AMOR
Professeur à l'Université de Tunis, Tunisie / examinateur

Arnaud MARTIN
Professeur à l'Université de Rennes1, France / directeur de thèse

Boutheina BEN YAGHLANE
Professeur à l'Université de Carthage, Tunisie / co-directrice de thèse

Grégory SMITS
Maître de conférences, Université de Rennes 1, France / co-directeur de thèse

Abstract

Ontology matching is a solution to mitigate the effect of semantic heterogeneity. Matching techniques, based on similarity measures, are used to find correspondences between ontologies. Using a unique similarity measure does not guarantee a perfect alignment. For that reason, it is necessary to use more than a similarity measure to take advantage of features of each one and then to combine the different outcomes. In this thesis, we propose a credibilistic decision process by using the theory of belief functions. First, we model the alignments, obtained after a matching process, under the theory of belief functions. Then, we combine the different outcomes through using adequate combination rules. Due to our awareness that making decision is a crucial step in any process and that most of the decision rules of the belief function theory are able to give results on a unique element, we propose a decision rule based on a distance measure able to make decision on union of elements (*i.e.* to identify for each source entity its corresponding target entities).

Résumé

L'appariement d'ontologies est une tâche primordiale pour pallier le problème de l'hétérogé néité sémantique et ainsi assurer une interopérabilité entre les applications utilisant différentes ontologies. Il consiste en la mise en correspondance de chaque entité d'une ontologie source à une entité d'une ontologie cible et ceci par application de techniques d'alignement fondées sur des mesures de similarité. Individuellement, aucune mesure de similarité ne permet d'obtenir un alignement parfait. C'est pour cette raison qu'il est intéressant de tenir compte de la complémentarité des mesures afin d'obtenir un meilleur alignement. Dans cette thèse, nous nous sommes intéressés à proposer un processus de décision crédibiliste pour l'appariement d'ontologies. Etant données deux ontologies, on procède à leur appariement et ceci par application de trois techniques. L'ensemble des alignements obtenus sera modélisé dans le cadre de la théorie des fonctions de croyance. Des règles de combinaison seront utilisées pour combiner les résultats d'alignement. Une étape de prise de décision s'avère utile, pour cette raison, nous proposons une règle de décision fondée sur une distance et capable de décider sur une union d'hypothèses. Cette règle sera utilisée dans notre processsus afin d'identifier pour chaque entité source le ou les entités cibles.

Acknowledgements

Besides being an original scientific work, the thesis requires the participation of many people. For that purpose, I would like in these few lines to thank everyone who helped me, who supported me and who believed in me.

I would like to express my deep gratitude to all members of my thesis committee: Pr. Nahla Ben Amor (University of Tunis, Tunisia), Pr. Chantal Reynaud (University Paris Sud, France) and Dr. Anne-Laure Jousselme (NATO CMRE, Italy). Thank you for accepting to assess my dissertation and providing me with valuable suggestions.

I owe deep thanks to Pr. Arnaud Martin (University of Rennes1, France), Pr. Boutheina Ben Yaghlane (University of Carthage, Tunisia) and Dr. Grégory Smits (University of Rennes1, France). I am grateful to your support and continuous encouragement for going through my dissertation. During these years, you provided me a comforting environment of work, you have been very patient and very understanding. Thanking you and expressing my sincere gratitude in only few words are not enough.

I wish to thank all members of my laboratory LARODEC at the Higher Institute of Management of Tunis and all members of my team DRUID at University of Rennes1. I would like also to address my thanks to colleagues in IUT of Lannion.

I dedicate this work to my dear dad and my lovely mum. From the first day I went to school, they have consistently made efforts to push me forward and to help me to be among the first during my studies. If the thesis defense is the final phase of few years work, so I can say it is the fruit of several years of support, monitoring and encouragement. This work is a fulfillment of all your efforts, your sacrifices and your prayers. From the bottom of my heart, I would like to thank you for providing a pleasant environment in which I could make progress. Thank you for believing in me, for being there for me when I encountered a problem and for being patient and comprehensive.

I would like to express my heartfelt thanks to my dear sisters, future doctors, for their love, for being there when I needed them and for their support.

My deep gratitude goes to my husband who never gave up in encouraging me to go through this thesis and supporting me.

I would like to express my love to my son, my sweetheart, whose presence brightened my life. Despite his young age, he made sacrifices and he helped me to go forward in my work.

Prise de décision lors de l'appariement des ontologies dans la théorie des fonctions de croyance

1 Introduction

Le web sémantique, introduit par (Berners-Lee, Hendler, & Lassila, 2001), est une solution pour remédier aux insuffisances du web actuel. En effet, cette nouvelle vision du web permet aux agents logiciels d'accéder au contenu des documents web afin de l'analyser et d'interpréter l'information figurant dans ces documents. Les ontologies ont été reconnues comme un des piliers du web sémantique permettant la représentation des connaissances. (Gruber, 1993) définit une ontologie comme une "*spécification explicite d'une conceptualisation*". La *conceptualisation* n'est autre qu'une vision abstraite d'un domaine de discours où les concepts, les objets et les entités sont identifiés. La *spécification explicite* fait référence au fait que les concepts et les relations entre ces concepts soient définis d'une manière explicite. Formellement, une ontologie est composée de concepts ou classes relatifs à un domaine bien déterminé et qui décrivent une collection d'objets. Ces concepts sont organisés en une hiérarchie taxinomique. En plus de ces concepts, on identifie les relations qui expriment les liens établis entre les instances de classes.

Toutefois, pour chaque contexte applicatif, il n'existe pas d'ontologie de référence partagée entre les membres d'une communauté mais plutôt plusieurs ontologies développées indépendamment les unes des autres et couvrant totalement ou partiellement un domaine de discours. Afin de faire face au problème d'hétérogénéité sémantique, il est nécessaire d'apparier les ontologies (Euzenat & Shvaiko, 2013b). L'appariement consiste en la mise en correspondance de chaque entité d'une ontologie source à une entité d'une ontologie

cible. La recherche des correspondances manuellement est une tâche coûteuse en temps et peut induire des erreurs. Pour cette raison, une panoplie de techniques d'appariement ont été proposées. (Euzenat & Shvaiko, 2013b) présentent un état de l'art exhaustif de ces techniques fondées sur des mesures de similarité.

Utiliser une seule mesure de similarité ne permet pas d'obtenir un alignement parfait vu que chaque mesure a ses propres caractéristiques. Afin d'améliorer le résultat de l'alignement, il est intéressant d'exploiter la complémentarité des différentes mesures. Cependant, l'utilisation de plusieurs mesures fait apparaître un conflit entre les différents résultats produits par chacune de ces mesures. Ce conflit doit être modélisé et résolu.

Dans cette thèse, nous proposons un processus de décision crédibiliste pour l'appariement des ontologies. Ce processus opère principalement en trois étapes. Tout d'abord, on aligne deux ontologies. Pour chaque entité d'une ontologie source, on cherche son correspondant dans une ontologie cible et ceci en utilisant trois différentes techniques. Afin de résoudre le désaccord entre les différents résultats, nous proposons de modéliser les alignements dans le cadre de la théorie des fonctions de croyance et de gérer le conflit par combinaison des différents résultats. Enfin, une étape de prise de décision est effectuée. Afin de choisir pour chaque entité source sa correspondante cible, nous proposons une règle de décision fondée sur une distance. Cette règle est capable d'apparier chaque entité source à plus d'une entité cible. Dans cette thèse, nous testons notre règle de décision sur des bases de données réelles (Bache & Lichman, 2013). Nous montrons que la règle de décision proposée donne de meilleurs résultats que celle proposée par (Appriou, 2005). Afin de tester notre processus de décision, nous utilisons des ontologies relatives à l'organisation des conférences [1].

2 Appariement des ontologies

L'appariement des ontologies est une solution pour pallier le problème d'hétérogénéité sémantique et d'assurer une interopérabilité entre les différentes applications. Selon (Euzenat & Shvaiko, 2013b), l'appariement est défini par une fonction qui tend, à partir de deux ontologies O_1 et O_2, à produire un ensemble de correspondances. Cette fonction peut aussi avoir comme entrée un ensemble d'alignements, un paramètre p et un ensemble de ressources. Apparier deux ontologies revient à chercher pour chaque entité d'une ontologie source son correspondant dans une ontologie cible et ceci par l'utilisation de techniques d'appariement qui sont classées en techniques terminologiques, structurelles, extensionnelles et sémantiques (Euzenat & Shvaiko, 2013b). Dans le cadre de cette thèse, des tech-

[1] http://oaei.ontologymatching.org/2013/

niques terminologiques et structurelles ont été appliquées. Les techniques terminologiques consistent à comparer les chaînes de caractères composant les entités des ontologies. Les techniques structurelles se fondent sur la comparaison des structures des entités. On distingue les techniques de comparaison des structures internes des entités (transitivité, multiplicité, ...) et les techniques de comparaison des structures externes (relations existantes entre les entités d'une même hiérarchie).

3 Théorie des fonctions de croyance

La théorie des fonctions de croyance appelée aussi théorie de Dempster-Shafer est initialement introduite par (Dempster, 1967) et fut reprise par (Shafer, 1976). C'est un outil qui permet de modéliser aussi bien l'incertitude que l'imprécision. Nous présentons dans ce qui suit les concepts de base de cette théorie. Pour un problème donné, la théorie des fonctions de croyance définit un cadre de discernement $\Omega = \{\omega_1, \omega_2, \ldots \omega_n\}$ comme étant l'ensemble des N hypothèses exhaustives et exclusives. A partir du cadre de discernement, on définit $2^\Omega = \{A; A \subseteq \Omega\} = \{\emptyset, \omega_1, \ldots, \omega_n, \omega_1 \cup \omega_2, \ldots, \Omega\}$. 2^Ω est l'ensemble des hypothèses singletons de Ω, toutes les disjonctions possibles de ces hypothèses ainsi que l'ensemble vide. La théorie des fonctions de croyance se fonde sur la manipulation des fonctions de masse. Une fonction de masse est une application des éléments de 2^Ω vers $[0, 1]$ de façon à ce qu'elle assigne une valeur positive entre $[0, 1]$ à une proposition, avec la contrainte:

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{1}$$

Les éléments A tel que $m(A) > 0$ sont appelés les éléments focaux. Il existe plusieurs types de fonctions de masse parmi lesquelles nous citons la fonction de masse catégorique qui admet un élément focal unique tel que $m(A) = 1$. A peut être un singleton ou une disjonction d'hypothèses. Dans le premier cas, la fonction de masses modélise la certitude et la précision. Dans le second, elle modélise plutôt la certitude et l'imprécision. La fonction de croyance (ou de crédibilité) bel mesure à quel point les informations données par une source soutiennent A. Elle est définie pour tout $A \in 2^\Omega$ et pour des valeurs dans $[0, 1]$ par:

$$\text{bel}(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \quad \forall A \subseteq \Omega \tag{2}$$

La fonction de plausibilité pl mesure à quel point les informations données par une source ne se contredisent pas. Elle est définie pour tout $A \in 2^\Omega$ et pour des valeurs dans $[0, 1]$

par:

$$\mathrm{pl}(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{3}$$

En présence d'informations imparfaites, la fusion est une solution pour obtenir une information plus pertinente et plus fiable. La théorie des fonctions de croyance présente l'avantage de combiner, pour un même cadre de discernement, des fonctions de masse élémentaires afin d'en obtenir une et ceci par utilisation d'une règle de combinaison. Pour un état de l'art exhaustif des règles de combinaison, le lecteur peut se référer à (Smets, 2007). Suite à la combinaison, une étape de prise de décision s'avère utile et qui consiste à sélectionner l'hypothèse la plus vraisemblable. Dans le cadre de la théorie des fonctions de croyance, la prise de décision peut se faire sur les hypothèses singletons (Smets, 1989) ou encore sur la disjonction des hypothèses ((Appriou, 2005),(Martin & Quidu, 2008)). Dans cette thèse, nous proposons une règle de décision qui permet de prendre une décision sur une disjonction de singletons.

4 Processus de décision crédibiliste pour l'alignement des ontologies

Le web sémantique permet d'assurer une interopérabilité et un échange de connaissances entre les agents logiciels et les utilisateurs. Récemment, les chercheurs se sont focalisés à tenir compte de l'incertitude dans le web sémantique que ce soit dans la représentation des ontologies ((Ding, 2005), (Costa & Laskey, 2006), (Yang & Calmet, 2005),(Gao & Liu, 2005), (Stoilos, Stamou, Tzouvaras, Pan, & Horrocks, 2005), (Essaid & Ben Yaghlane, 2009)) ou encore dans l'alignement des ontologies ((Ding, 2005), (Mitra, Noy, & Jaiswal, 2005), (Besana, 2006), (Nagy, Vargas-Vera, & Motta, 2007), (Wang, Liu, & Bell, 2007)). Tenir compte de l'incertitude lors de la mise en correspondance permet d'améliorer la détection des correspondances. A cet effet, plusieurs théories mathématiques ont été utilisées parmi lesquels la théorie des fonctions de croyance. Nous présentons dans cette section notre processus de décision crédibiliste qui, à partir de deux ontologies, permet d'une part de gérer le désaccord entre les résultats d'alignement et d'autre part de sélectionner pour chaque entité source une ou plusieurs entités cibles et ceci par application de la règle de décision que nous avons proposée.

4.1 Règle de décision fondée sur une distance

Dans cette thèse, nous proposons une règle permettant de prendre une décision sur un ensemble d'hypothèses. Cette règle, qui a fait l'objet de deux articles ((Essaid, Martin, Smits, & Ben Yaghlane, 2014b),(Essaid, Martin, Smits, & Ben Yaghlane, 2014a)), est inspirée des travaux de (Smarandache, Martin, & Osswald, 2011). Elle est définie par:

$$A = argmin(d(m, m_A)) \tag{4}$$

où A représente la décision à prendre. Cette dernière est obtenue suite au calcul de la distance entre une fonction de masse (m) et une fonction de masse catégorique m_A. Dans le cadre de notre travail, la distance est calculée entre une masse combinée obtenue suite à l'application d'une règle de combinaison et une masse catégorique. Nous optons pour l'utilisation des fonctions de masse catégoriques afin d'ajuster le degré de l'imprécision qui doit être maintenu au moment de la décision. En effet, on peut restreindre à ce que la décision porte sur des éléments focaux de cardinalité 2 ou 3 ou 4,... Une fois, la distance calculée, la décision correspond aux éléments de la fonction de masse catégorique ayant la plus petite distance avec la masse combinée.

La règle proposée opère en trois étapes. Tout d'abord, nous fixons la cardinalité des éléments de 2^Ω pour lesquels nous construisons leur masse catégorique correspondante. Dans cette thèse, nous considérons uniquement les éléments de cardinalité égale à 2. Une fois la fonction de masse catégorique est construite, nous calculons la distance entre la masse combinée et chaque fonction de masse catégorique. La distance de Jousselme est utilisée pour cet effet (Jousselme, Grenier, & Bossé, 2001). L'hypothèse ayant une fonction de masse catégorique très proche à la masse combinée sera considérée comme l'hypothèse la plus vraisemblable.

Etant donné que la classification est un problème de décision, nous utilisons des bases de données réelles de l'U.C.I (Bache & Lichman, 2013) pour évaluer notre règle de décision. Nous comparons les résultats obtenus suite à l'application de notre règle de décision avec ceux obtenus quand la règle d'Appriou est utilisée. Deux types d'expérimentations ont été effectuées. Tout d'abord, nous appliquons l'algorithme k-NN crédibiliste (Denœux, 1995). Ensuite, nous modifions cet algorithme afin qu'il soit capable de combiner les fonctions de masse par la règle de combinaison mixte (Dubois & Prade, 1988b). Une fois la combinaison effectuée, la règle d'Appriou et celle que nous avons proposée seront appliquées pour une prise de décision. Les expérimentations montrent bien que notre règle donne de meilleurs résultats par comparaison à celle d'Appriou et qu'est est capable de décider sur les unions des singletons. Le manuscrit de thèse présente en détail les résultats des expérimentations.

4.2 Description du processus de décision crédibiliste

Ce processus est fondé sur l'utilisation de la théorie des fonctions de croyance comme un outil mathématique pour modéliser l'appariement des ontologies et de résoudre le problème du désaccord entre les mesures de similarité. Le processus considère comme entrée deux ontologies et un ensemble de techniques d'appariement. Comme sortie, le processus délivre un ensemble d'alignements imprécis. Ce processus opère principalement en trois étapes.

4.2.1 Choix des techniques d'appariement

L'appariement des ontologies se fonde sur l'utilisation des techniques. Chaque technique concerne une caractéristique spécifique des entités. Plusieurs études ont été menées pour sélectionner une technique bien déterminée (Euzenat, Ehrig, Jentzsch, Mochol, & Shvaiko, 2006; Huzza, Harzallah, & Trichet, 2006; Mochol, 2009). Dans le cadre de cette thèse, le choix d'une technique se fonde sur la comparaison des résultats des métriques d'évaluation (précision, rappel). Ces métriques, qui ont comme origine le domaine de la recherche d'information, ont été adaptées par (Do, Melnik, & Rahm, 2002) dans le domaine de l'appariement des ontologies. La précision est définie comme étant le rapport du nombre des paires de correspondances pertinentes trouvées par rapport au nombre total des paires obtenues par une technique d'alignement. Le rappel représente le rapport du nombre des paires de correspondances pertinentes trouvées par rapport au nombre total des paires pertinentes. Afin de choisir une méthode terminologique, nous évaluons les méthodes suivantes: *Hamming*, *Jaro*, *Levenshtein*, *Needleman-Wunsch*, *Ngram*, *Monge-Elkan*, *Smith-WaterMan*, *Soundex*. En se fondant sur les résultats d'évaluation, la distance de *Needleman-Wunsch* est sélectionnée comme méthode que nous utilisons dans notre processus. En plus de cette méthode, nous utiliserons *Wu Palmer similarity* et *Gloss Overlap* qui accédent au *WordNet* pour rechercher les similarités entre les concepts de deux ontologies.

4.2.2 Modélisation de l'appariement dans le cadre de la théorie des fonctions de croyance

Etant donné l'ensemble d'alignements obtenus suite à l'application des techniques d'appa riement, on détecte deux types de désaccord. Le premier concerne le fait qu'il n'existe pas un consensus entre les différentes mesures de similarité. En effet, une entité source peut être alignée à différentes entités cibles. Quant au second type de désaccord, il est relatif au fait qu'une entité peut être alignée, par application des techniques d'appariement, à une entité

cible mais avec différentes valeurs de similarité. Nous proposons de gérer ce désaccord en modélisant les résultats d'appariement dans le cadre de la théorie des fonctions de croyance. Pour cela, nous devons définir un cadre de discernement et spécifier comment les fonctions de masses sont construites puis combinées.

1. **Le cadre de discernement** est un ensemble de toutes les hypothèses susceptibles de représenter une solution à un problème donné. Afin de résoudre le désaccord entre les différents résultats d'appariement, nous proposons de définir le cadre de discernement comme étant l'ensemble de toutes les entités cibles identifiées dans les alignements.

2. **La source d'information** : Chaque correspondance établie par une méthode d'apparie ment sera considérée comme une information dont la source est l'application d'une méthode d'appariement sur l'entité de la première ontologie concernée par la correspondance.

3. **Les fonctions de masse** : Une fois que nous obtenons les paires de correspondances, nous ne conservons que celles où l'entité source a un appariement de proposé pour toutes les méthodes d'appariement considérées. Une fois les correspondances retenues, nous construisons pour chaque source sa propre fonction de masse. L'application d'une technique d'appariement permet d'identifier les entités présentant un degré de similarité. Plus les entités sont similaires, plus elles sont proches et par conséquent elles peuvent être appariées. Nous considérons l'hypothèse qu'une entité est proche d'une autre entité si elles sont similaires et donc il y a de forte chance que ces entités soient appariées. Dans le cadre de la théorie des fonctions de croyance, cette distance peut être interprétée comme le degré de croyance d'une mesure de similarité. Afin de construire la fonction de masse et garantir que la somme soit égale à 1, une masse sera allouée à l'ignorance totale.

4. **Combinaison** : Afin de gérer le conflit, nous procédons à la combinaison des fonctions de masse. La combinaison conjonctive, disjonctive et la mixte ont été utilisées.

4.2.3 Prise de décision

Une fois que nous obtenons la masse combinée, il est important de décider pour chaque entité source, le ou les entités cibles à considérer comme correspondantes. Nous avons utilisé notre règle de décision, la règle proposée par Appriou ainsi que la probabilité pignistique. Les différents résultats obtenus sont indiqués dans le manuscrit de thèse.

Afin de valider notre processus de décision crédibiliste pour l'appariement des ontologies, nous avons effectué des expérimentations sur des ontologies relatives à l'organisation des conférences. Le détail des expérimentations est présenté dans le manuscrit de thèse. Des courbes pour la précision et le rappel ont été dressées. Dans ces illustrations, nous comparons les alignements imprécis que nous avons obtenus une fois que notre règle de décision a été appliquée par rapport aux alignements obtenus si l'une des techniques d'appariement est appliquée. Les résultats obtenus sont globalement satisfaisants.

Conclusion

L'objectif de cette thèse est d'apparier les entités de deux ontologies par application d'une règle de décision fondée sur une distance. Pour cette raison, nous avons utilisé la théorie des fonctions de croyance pour modéliser le processus de l'appariement des ontologies. Cette théorie nous a permis de combiner les différents résultats des mesures de similarité utilisées. Au cours de cette thèse, nous avons proposé une règle de décision fondée sur une distance et capable de décider sur une union d'hypothèses. Cette règle a été par la suite utilisée dans notre processus afin de sélectionner pour chaque entité source les entités auxquelles elle pourra être appariée.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

In this chapter, we aim to lay out the main problems that we address and the main contributions. In the following, we give an overview of the context of this thesis, on how ontology matching plays a prominent role in assuring interoperability and how dealing with uncertainty when matching ontologies is crucial. In this chapter, we give a brief description of our approach to ontology matching. The end of this chapter is devoted to the organization of the dissertation document.

1.1 Research context

In his book, *Weaving the Web*, Tim-Berners-Lee considers that *"the web is a more social creation than a technical one"*. In fact, nobody can deny what the web has brought to its users. With just a click, you can talk and share ideas with people who are far from each other. With just a click, you can buy your flight ticket or even you can do your shopping and then be delivered at home. With just a click, you can get access to needed information wherever you are and whenever you want. If one considers the statistics of 2014, then we will understand the importance of the web in our life. Can you imagine that 4 millions of searching on Google are carried out in just a minute?

In the early 90's when it has been created, internet has gained popularity because it is simple to use and it provides many services (email, chat, e-commerce, etc.). But the current web is syntactic. In fact, pages are encoded in a markup language which is rather a data representing language used to better formatting pages content and to establish hyperlinks between different web pages. The main drawback of the current web is its weak ability to interpret the content of the document and to generate new information. Suppose, for example, that you wish to know the date and the place of a conference. Making your request on a search engine gives you a list of web pages that can be related to your request and you have to search in these pages to find the needed information.

Nowadays, establishing only links between web pages and allowing only people to get access to information must be exceeded to give the software agents the possibility to understand the content of documents by analyzing and interpreting the contained information. The current web's limitations prompted to the birth of the *semantic web* (Berners-Lee et al., 2001) which is presented as a large area of resources exchange between humans and software agents allowing the exploitation of large volume of information and releasing users from searching tasks thanks to machines' ability to get access to the documents' content (Laublet, Charlet, & Reynaud, 2007). The semantic web which evolves out of the existing web is represented as a layer cake of technologies where each level takes the advantages of the previous level and presents a basis of the next level. One of the pillars of semantic web are *ontologies* which represent the information into a taxonomy of concepts and relationships between them (Chandrasekaran, Josephson, & Benjamins, 1999). For a domain of discourse, ontologies are described through an ontology representation language providing a vocabulary to define ontologies in formal semantics. The establishment of relations between concepts and their formal representation make the understanding of users' queries easier and provide efficient result.

The open nature of the semantic web tends to encourage the development of heteroge-

neous ontologies which differ from each other either with the representation language used or in the way the domain of discourse is conceptualized. To mitigate the effect of semantic heterogeneity, it is necessary to bring together heterogeneous and distributed ontologies. This is referenced as *ontology matching* which consists in finding correspondences between entities of two ontologies to match (Euzenat & Shvaiko, 2013a). Depending on application needs, these correspondences can be used for various tasks such as merging ontologies, reasoning or data translation. Matching ontologies is carried out through the application of matching techniques. A state of the art of these techniques can be found in (Euzenat & Shvaiko, 2013a). In order to guarantee knowledge sharing and semantic interoperability, two main challenges of the matching process have to be taking into account: robustness and scalability (Su, 2004). The former concerns the fact that minor errors should not have an impact on the matching result whereas the latter quantifies the ability of the matching technique to provide results in a reasonable time even in case of large ontologies.

The semantic web envisions a world where software agents are able to cooperate together and to provide new knowledge based on their interpretation of the documents' content. However, the world is dynamic which makes the web documents stained with imperfect information. According to (Bonissone & Tong, 1985), there are three kinds of imperfection: incompleteness, imprecision and uncertainty. Information is *incomplete* when some data are missing. For example, if we state that as part of his participation in the ESWC 2006, Jérôme Euzenat recorded an interview in which he presented the research area he is working on in his research team [1]. We notice that some information is missing. In fact, we may want to know Euzenat's research area. Information can be *imprecise* when we do not discern the exact value but rather we give several possible choices. Saying that a conference is held early in January, supposes that the date of the conference can correspond to the 1^{st}, 2^{nd}, etc. *Uncertain* information is given by a source expressing its opinion and arises from the lack of information about the real world. For example, a reviewer may hesitate between accepting a paper to be rewritten as a short one or to be presented as a poster and expresses his belief based on this uncertain information. Many mathematical models have been proposed to manage imperfect information. We may cite the fuzzy set theory (Zadeh, 1965), the possibility theory (Dubois & Prade, 1988a) and the theory of belief functions ((Dempster, 1967), (Shafer, 1976)).

Like any other research domain, the semantic web is not deprived of uncertainty mainly with the huge amount of information contained in web documents. Uncertainty in the semantic web became the focus of many works, each of them proposing different approaches. Due to the fact that ontology representation languages are built on crisp logic, some researchers propose to enrich these languages by new constructors able to express the un-

[1] http://videolectures.net/eswc06_euzenat_ije/

certain information and to represent faithfully a domain of discourse (Ding, 2005; Costa & Laskey, 2006; Yang & Calmet, 2005; Gao & Liu, 2005; Stoilos et al., 2005; Essaid & Ben Yaghlane, 2009). Awareness of the importance of uncertainty has not been restricted only to a representational level but it has concerned also the ontology matching area where uncertainty has been considered as one of the main challenges that should be tackled (Shvaiko & Euzenat, 2008). Recently some approaches dealing with uncertainty in ontology matching have emerged (Pan, Ding, Yu, & Peng, 2005; Mitra et al., 2005; Besana, 2006; Nagy et al., 2007; Wang et al., 2007).

1.2 Problem statement and contributions

In the last years, ontology matching has been identified as a crucial step towards semantic interoperability and knowledge exchange between different applications. The process of matching ontologies uses methods from several communities such as knowledge engineering, information retrieval, language process (Euzenat & Shvaiko, 2013a). These methods are based on the use of similarity measures. However:

- Using a similarity measure individually does not give a perfect alignment (an alignment is the output of a matching process) because each measure is related to a particular feature. For instance, the Levenshtein distance is a terminological technique that quantifies the similarity between two entities by comparing their strings. This distance does not take into account if two entities are synonyms or if there is a relationship between their parents or children in their corresponding ontologies. For example, the Levenshtein distance between the two terms *test* and *tent* is 1. Based on this result, *test* and *tent* are considered as similar although they do not belong to the same lexical field. Hence, using only this distance to match ontologies will not give an efficient alignment. For that reason, it is essential to consider the complementarity between the different similarity measures.

- Using several similarity measures and take advantage of each measure's specificity will help to obtain good results. But for a couple of entities, two measures may assign different similarity values. For example, the Jaro measure assigns a value of 0.516 between *ConferenceMember* and *Conference* whereas the Hamming distance assigns a value of 0.625 between these two entities. The difference between the two values shows a disagreement between the two measures. In that case, it is interesting to manage disagreement occurring between similarity measures.

- Matching ontologies in a certain context supposes that the value given by a simi-

larity measure is just a value obtained after applying an algorithm but what about supposing that this value is none other than a similarity measure's belief? Based on the assumption that if two entities are similar, then they are near to each others, we can admit that there is a chance that these two entities can be aligned. The distance between the two entities can be interpreted as a degree of belief of similarity measure. In fact, we can consider that the value of 0.625 given by Hamming distance between *Conference* and *ConferenceMember* is the belief of the source Hamming distance and thus assigns a value of 0.625. Based on this assumption, it seems to be beneficial to match ontologies under uncertainty.

- Most of the matching approaches that deals with uncertainty in ontology matching, especially those using belief function theory as underlying mathematical model, search for simple matching where each entity in an ontology source has a correspondence an entity in a target ontology. This is because their decision rules identify a unique entity rather than a union of entities. Using another rule able to align each entity to more than one target entity seems to be an interesting idea.

Based on the detailed presentation of our research context and the main problems encountered in the literature, our aim in this thesis is to propose a credibilistic decision process for matching ontologies using the belief function theory. As it has been mentioned earlier, it is expected that using simultaneously many matching techniques will improve the matching results. Each technique gives a set of corresponding entities. For a given entity in an ontology source, we can find either a unique entity in a target ontology or more than a corresponding entity with different similarity values. The different results obtained by the different techniques show a disagreement between them. For that purpose, we suggest to model the obtained alignments under the belief function theory. The modeling is based on a correspondence between matching components and the belief function theory elements. Once, the alignments are represented under uncertainty, we manage the disagreement occurred between the different matching techniques. We suggest to combine these results and to manage the disagreement after the combination. The last and the most important step is to make decision about the corresponding entities for a given entity in an ontology source. With the idea of promoting more than an entity in an ontology target, we propose a decision rule based on a distance measure able to give a result on a union of elements.

1.3 Organization

We outline in the following the organization of our dissertation document.

- Chapter 2 presents a survey on the semantic web field and particularly on the ontology matching field. Ontologies, as key components of the semantic web, are models representing knowledge through the description of concepts related to a domain of discourse as well as the relations between these concepts. For each application context, there is no shared ontology but rather several ontologies developed independently and often partially covering the application context. In order to use these ontologies efficiently, we must alleviate the effect of semantic heterogeneity through matching ontologies. The different steps of the matching process are described in this chapter and a detailed presentation of the main techniques is given.

- Chapter 3 is about the uncertainty in the semantic web. Due to the huge amount of information that the web documents contain and the necessity to faithfully represent a domain of discourse with the different changes that it may knows, it is essential to represent uncertainty in the semantic web. First, we recall the basic concepts of the belief function theory and the main justifications to use this mathematical model. Uncertainty in semantic web concerns ontology representation as well as ontology matching. In this chapter, we present the main approaches that considered that the ontology representation languages are crisp ones and that extending them with adequate constructors helps to take into account the uncertain information. In another section, we present the approaches dealing with uncertainty when matching ontologies. A special focus is devoted to those which used belief function theory as their underlying theory.

- Chapter 4 gives a deep description of our credibilistic decision process as well as the different experimentations we made. Due to our awareness that making decision is a crucial step in any process and that most of the decision rules of the belief function theory are able to give results on a unique element, we propose a decision rule able to make decision on union of elements. We make some experimentation on real data sets and we present results with improved performance compared to the rule proposed by (Appriou, 2005). Then, we give a deep description of our process which is mainly in three steps. First, after selecting the main matching techniques that we will use, we model results of the matching process under the belief function theory. In order to represent all the techniques features and to manage disagreement occurred between techniques' results, we propose to combine all the modeled alignments to get a unique and coherent result. Then, our proposed decision rule is applied. This rule allows to find for each entity in an ontology source more than an entity in a target ontology. At the end of this chapter, the different results of experimentation handled on a set of ontologies are given.

- Chapter 5 concludes the thesis and presents possible future improvements as well as

the main extensions to our credibilistic decision process.

1.4 Publications

The proposed approach has been the subject of four publications. Two are published in international conferences whereas the two others have been presented in national conferences.

[1] Essaid Amira, Ben Yaghlane Boutheina, Martin Arnaud. *Gestion du conflit dans l'appariement des ontologies*. In Atelier Graphes et Appariement d'Objets Complexes, en conjonction avec EGC 2011, Brest, France, January 2011 (p. 50 - 60).

[2] Essaid Amira, Martin Arnaud, Grégory Smits, Ben Yaghlane Boutheina. *Processus de décision crédibiliste pour l'alignement des ontologies*. In Les Rencontres Francophones sur la Logique Floue et ses Applications, Reims, France, October 2013 (p. 59 - 65).

[3] Essaid Amira, Martin Arnaud, Grégory Smits, Ben Yaghlane Boutheina. *Uncertainty in ontology matching: a decision rule-based approach*. In Proceedings of the 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Montpellier, France, July 2014 (Vol. 442, p. 46 - 55).

[4] Essaid Amira, Martin Arnaud, Grégory Smits, Ben Yaghlane Boutheina. *A distance-based decision in the credal level*. In Proceedings of 12th International Conference on Artificial Intelligence and Symbolic Computation, Seville, Spain, December 2014 (Vol. 8884, p. 147 - 156).

Chapter 2

Survey on the semantic web and the ontology matching

Contents

Ontologies are viewed as silver bullet in many fields such as databases, cooperative information systems, electronic commerce, etc.(Fensel, 2004). Using ontologies becomes of a great interest. In fact, they describe a domain of interest with explicit semantics processable by machines. The expansion known by the semantic web has led to the development of disparate ontologies. Heterogeneous ontologies creates a semantic heterogeneity which may be reduced through matching ontologies. The aim of this chapter is to present the OWL

ontology language as a sophisticated language for representing knowledge in ontologies. We give in a second part an overview of the ontology matching field.

2.1 Introduction

The web has expand to a tremendous success. This is due to the huge amount of information available on the web and the increasing number of people using it. In fact, getting access to the web helps users to exploit documents and services. For example, they can communicate with each other, search for information and even buy products or organize a trip. All these activities are only convenient for human users because the web pages' content are in a human readable format. As a consequence, it is difficult for software agents to extract, interpret and process useful information for web users.

In order to give machines the ability to manipulate the information existing on web documents, (Berners-Lee et al., 2001) introduced a new vision of the web, the *semantic web*, which is not an alternative to the existing syntactic web based on HTML documents but rather "*an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation*".

This new generation of the web is underpinned by six basic principles as identified by (Koivunen, 2001), namely:

- identity: Every concept is identified by an Uniform Resource Identifier (URI).
- typing: The current web contains resources and links between these resources. It does not provide any additional information about the resources and the links. This will prevent software agents to be able to guess the content of the document and the nature of the links. For that purpose, the semantic web allows the description and the typing of links existing between the resources.
- partiality: The semantic web allows anyone to say anything about web resources by creating different types of links.
- web of trust: Semantic web does not guarantee an absolute truth of the statements in the web.
- evolution: It is considered as a development principle. In fact, the semantic web has to:
  - allow the combination of the independent work of diverse communities.

  - support the ability to add new information without reconsidering old information modified.
  - be able to resolve ambiguities and clarify consistencies.
  - use conventions that expand as human understanding expands.

- minimalist design: Semantic web enables simple applications based on already existing standards (RSS, Dublin Core) and it will not standardize more than it is necessary.

The main objective of the semantic web is to allow the software agents better understand the web documents' content in order to respond intelligently to users' queries and to carry out sophisticated tasks such as information retrieval, data integration and reasoning. To achieve these ambitious goals, the semantic web relies on ontologies which describe the semantics of data.

During the last decade, research on ontologies has gone through different stages of evolution as described in (Noy & Klein, 2004). At the beginning, the focus was particularly oriented to properly define an ontology and to specify the requirements that it must satisfy. Then, the practitioners were interested in developing expressive and efficient ontology languages for defining and exchanging ontologies. With the appearance of a large number of ontologies, some of them representing similar domains with different terminologies and others describing overlapping domains, the researchers were faced with a new challenge namely *ontology matching* which consists in finding semantic correspondences between entities of different ontologies.

In this chapter, we outline the different phases that ontology research has gone through as well as the changes occurring in this field. In section 2.2, we define the concept of ontology and introduce the OWL as an expressive ontology language for representing a domain of discourse. Then, we present the main scenarios where ontologies can be applied and we expose the most important benefits that an ontology offers. Section 2.3 deals with the ontology matching as a solution to mitigate the effect of semantic heterogeneity. We describe in this section the main steps of the ontology matching process and present the basic techniques to detect correspondences between ontologies.

2.2 Ontologies as knowledge representation models

Ontologies are suggested as a model for representing knowledge. Many languages have been proposed for representing the concepts and relations between them in a standard

way. In this section, we present ontologies and the main advantages of using them and we give a deep description of the OWL language, introducing its predecessors and main components.

2.2.1 The concept of ontology

There is no universal agreement about an adequate definition to the term *ontology* because it comes from the domain of philosophy and it has been adapted across different communities such as psychology, sociology, artificial intelligence and computer science. The term ontology has been taken from a sub-field of philosophy known as metaphysics which studies the nature of being and existence. In fact, an ontology describes the objects that exist in the world and their properties as well as how they are related to each others. The philosophical view of the ontology has been an inspiration for practitioners in artificial intelligence where they develop ontologies to facilitate knowledge sharing and reuse.

The ontology was first defined as "*an explicit specification of a conceptualization*" (Gruber, 1993). Later, a refined definition was provided by (Borst, Akkermans, & Top, 1997) stating that an ontology is a "*formal specification of a shared conceptualization*". These two definitions were built on some relevant concepts that must be recalled in detail:

- Conceptualization refers to an abstract view of a domain in the world we want to represent. Concepts, objects and entities existing in the simplified view are identified.
- Explicit specification means that all the concepts, the constraints on their use as well as the relationships between them must be explicitly defined.
- Formal specification refers to the fact that the description of the domain's concepts must be represented in a formal language understandable by machines.
- Shared conceptualization is related to the fact that an ontology is built upon a consensus between members of a community where they model a specific domain of discourse. An ontology contains knowledge used and reused across different applications.

Based on what is mentioned above, we can summarize that an ontology is a shared model representing a common vocabulary related to a specific domain of discourse. Hence, it is considered as an interesting model for exchanging knowledge and assuring interoperability between automated tools. These tools will be able to give relevant answers to user queries.

2.2.2 OWL - Web Ontology Language

An ontology describes knowledge related to a specific domain of discourse. This knowledge must be structured in a formal language in order to ensure information sharing between different ontologies. For that purpose, a well-defined ontology language is required. On the one hand, this language has to be understandable by human users and on the other hand it should define the relevant concepts related to the domain of discourse and must fit the existing web standards.

There has been a number of languages for representing knowledge in ontologies among them OWL -Web Ontology Language- which is considered as a sophisticated language because it was developed by several communities working on description logics and semantic web technologies (XML, RDF, RDFS).

2.2.2.1 Description Logics (DLs)

The description logics are a standard for the semantic web and are defined as a set of knowledge representation languages able to describe an application domain by representing the knowledge related to this domain in a formal and structured way (Baader, Horrocks, & Sattler, 2005). One of the main advantage of the DLs is their ability to support inference mechanisms and infer implicit knowledge. In DLs, knowledge is represented through a knowledge base that involves two components:

- T-Box: is a terminological box. It refers to the vocabulary related to a domain of discourse. This vocabulary includes the concepts and the roles. The former are used to describe classes of individuals and are organized in a taxonomy of super-concepts and sub-concepts. The latter represents a binary relationship between two concepts. In addition to that, the concepts are described through properties.
- A-Box: is an assertional box. It is a set of assertions on individuals occurring in the domain of discourse.

2.2.2.2 Semantic web technologies

In addition to the DLs, the development of OWL has been influenced by a number of semantic web technologies. During the last decade, the W3C [1] -World Wide Web Consortium-

[1] http://www.w3.org/

has focused on developing a stack of fundamental technologies referred to the semantic web layer cake as illustrated in figure 2.1.

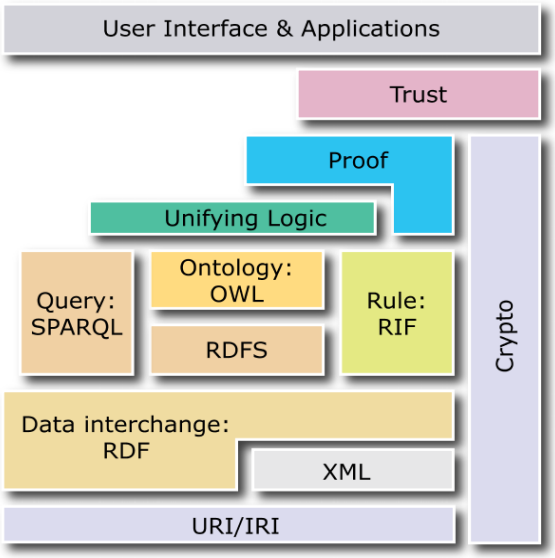

Figure 2.1: Semantic web layer cake

This architecture describes how each layer exploits and uses capabilities of the layer below. The technologies presented in the semantic web stack are organized as follows:

- The bottom layers represent the hypertext web technologies (URI, IRI and XML) which are inherited from the previous web and form the basis for the semantic web.
- The middle layers represent the standardized semantic web technologies (RDF, RDFS, SPARQL, RIF) which contain technologies standardized by the W3C and are able to build the semantic web.
- The top layers contain technologies that are not standardized and it is not clear how these layers will be implemented.

We present in this section the important technologies that have led to the development of OWL namely URI/IRI, XML, RDF and RDFS.

- Uniform/ Internationalized Resource Identifier (URI/IRI)[2]: is the basic of the world wide web because all the hyperlinks on the web are expressed in an URI format. It

[2]http://www.w3.org/Addressing/

is a string of characters which identify a web resource in a unique manner. IRI is a generalization of URI. In fact, IRI takes into account all the alphabets, whatever the language used, for identifying a resource.

- eXtensible Markup Language (XML)[3]: In order to overcome the insufficiency of HTML, XML brings a solution to define the structure of information exchanged on the web. This meta-language is considered as the basic language for the semantic web. It is a tag-based language. It defines its own tags which describe the structure of the web documents in order to facilitate automated processing of the web content.

- Resource Description Framework (RDF)[4]: RDF is a W3C recommendation. It is an XML-based language defined as a data model. It describes the web resources by adding a meta-information. A resource is any object identified by an URI. It can be a simple web page, an image, etc. The resource can be modeled in a RDF statement which is based on the notion of triples. A triple is an association between a subject, a predicate and an object. A subject is a resource described by the RDF statement and identified by an URI. The predicate defines a property of a subject uniquely identified by a URI. Object is a value for the property which can be a resource described with URI or a literal (string or fragment of XML).

- Resource Description Framework Schema (RDFS)[5]: RDFS is dedicated to the representation of ontological knowledge. It extends the RDF vocabulary in order to give a structure to RDF resources. It is based on mechanisms for describing a set of similar resources (classes) and relations between these resources (properties). RDF is able to organize classes and properties in a hierarchy and it defines the subsumption relationships between classes and properties and more concretely through the primitives *"rdfs:subClassOf"* and *"rdfs:subPropertyOf"*. In addition to that, RDFS defines two mechanisms for manipulating properties using for that purpose *"rdfs:domain"* and *"rdfs:range"*. In fact, properties are defined in terms of the classes of resources to which they apply. The subject of a property must belong to the set of instances of the class mentioned by *"rdfs:domain"* whereas the object of a property must belong to the set of instances of the class mentioned by *"rdfs:range"*. The two languages RDF and RDFS, when used together, are referenced by RDF(S).

[3]http://www.w3.org/XML/
[4]http://www.w3.org/RDF/
[5]http://www.w3.org/TR/rdf-schema/

2.2.2.3 DAML+OIL as a predecessor language

DAML+OIL[6] is the fusion of two languages DAML[7] and OIL. This language is an extension of the RDF(S) where it presents additional features and adopts the description logics for handling reasoning mechanisms. Compared to RDF(S), DAML+ OIL presents advantages. In fact, it defines the logical combination between classes and adds specific description to the properties (transitive, symmetric, etc.) in addition to the possibility of adding cardinality restrictions. But the major extension over RDFS is that DAML+OIL is able to provide restrictions on properties through datatypes.

2.2.2.4 Description of OWL

Ontology languages are formal languages used to represent an ontology. There has been a number of these languages but not suitable for the semantic web field. Great efforts are made to propose ontology web languages able to describe an ontology in a formal way and to respond to the semantic web requirements. The OWL [8]-Web Ontology Language- is nowadays the most important language for developing ontologies. It was first recommended in 2004 by the W3C. Then, it was extended as OWL2 [9] and has been a W3C recommendation in 2009. DLs are the basis of OWL which evolved from the DAML+OIL and are developed to fit into the semantic web vision of existing languages namely XML, RDF and RDFS. OWL provides a rich vocabulary for authoring ontologies, facilitating interpretation of documents content as well as inferring additional knowledge. It is an expressive language because it overcomes the lack of expressivity of its predecessors and offers many paradigms for modeling ontologies. In addition to RDFS primitives, OWL is able to express relations between classes through restrictions (*e.g.* disjunction, union, etc.), to specify cardinality and equality. It also offers a way for defining the types of properties as well as their characteristics (symmetry, transitivity, etc.). A detailed description of OWL primitives with examples is given in subsection 2.2.3.

[6] http://www.daml.org/2001/03/daml+oil-index
[7] http://www.daml.org
[8] http://www.w3.org/TR/2004/REC-owl-features-20040210/
[9] http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

2.2.3 OWL ontology

2.2.3.1 OWL Ontology Components

An ontology expressed in OWL describes a domain of discourse through classes, properties, individuals and axioms.

- **Classes**: A class defines a way to put together different resources with similar characteristics. A class can be described through:
  - a *class identifier "rdf:ID"*, for example <owl:Class rdf:ID ="programCommitteeMember"> describes a member of the program committee in a conference.
  - an *exhaustive enumeration* of individuals representing the instances of a class. For example, in a conference it is interesting to list the members of the program committee. OWL helps to represent this information through *owl:OneOf* which lists all the members of a class. It has the following general form:

```
<owl:Class>
<owl:oneOf rdf:parseType="Collection">
  <owl:Thing rdf:about = "member_1" />
  <owl:Thing rdf:about = "member_2" />
  ...
  <owl:Thing rdf:about = "member_N" />
</owl:oneOf>
</owl:Class>
```

  - *Property restriction* describes a class of all individuals that satisfy a restriction. It is introduced in an OWL ontology through *owl:restriction*. A restriction concerns either *value constraints* or *cardinality constraints*. The former puts constraints on the range of the property using the constructors (*owl:AllValuesFrom*, *owl:someValuesFrom*, *owl:hasValue*), whereas the latter puts constraints on the number of values that a property can take via the primitives (*owl:maxCardinality*, *owl:minCardinality*, *owl:cardinality*). In the following, we give two examples of property restriction. The former concerns value constraints. It imposes that a paper has been at least written by a conference_participant. The latter is related to the cardinality constraints. It imposes that a paper has at least 3 reviewers.

```
<owl:Restriction>
 <owl:onProperty>
 <owl:objectProperty rdf:ID = "writtenBy"/>
 </owl:onProperty>
 <owl:someValuesFrom>
 <owl:class rdf:ID = "Conference_Participant"/>
 </owl:someValuesFrom>
</owl:Restriction>
```

```
<owl:Restriction>
 <owl:onProperty>
 <owl:objectProperty rdf:about = "#hasReview"/>
 </owl:onProperty>
 <owl:minCardinality rdf:datatype="http://www.w3.org/
 2001/ XMLSchema#int"> 3 </owl:minCardinality>
</owl:Restriction>
```

– Logical operations defined through the constructors (*owl:intersectionOf*, *owl:unionOf*) are used to describe relations (intersection, union, complement) that may exist between classes.

```
<owl:class>
 <owl:unionOf rdf:parseType = "Collection">
 <owl:class rdf:about ="#Multi-author_volume"/>
 <owl:class rdf:about ="#Programme_Brochure"/>
 <owl:class rdf:about ="#Web_Site"/>
 <owl:class rdf:about ="#Flyer"/>
 </owl:unionOf>
</owl:class>
```

- **Properties** are binary relations. There exist two main categories of properties. Object properties (*owl:objectProperty*) link individuals to individuals and datatype properties (*owl:datatypeProperty*) link individuals to data values. Property axioms can be used to define additional characteristics of properties. In fact, each property has a

domain (*rdf:domain*) and a range (*rdf:range*). The *rdfs:subPropertyOf* is used to arrange properties in a hierarchy. In order to define the relations that may exist between properties, one may use *owl:equivalentProperty* and *owl:InverseOf*. OWL defines property axioms that specify restrictions on property cardinality (*owl:FunctionalProperty*, *owl:InverseFunctionalProperty*) as well as it describes logical features on properties (*owl:TransitiveProperty*, *owl:symmetricProperty*).

```
<owl:objectProperty rdf:ID="read_paper_by">
<rdfs:domain rdf:resource = "#Accepted_paper"/>
<rdfs:range>
 <owl:unionOf rdf:parseType = "Collection">
 <owl:class rdf:about ="#External_Reviewer"/>
 <owl:class rdf:about ="#Secondary_Reviewer"/>
 </owl:unionOf>
</owl:class>
</rdfs:range>
</owl:objectProperty>
```

- **Individual** is defined as a member of a class through *rdf:Type*. To describe relations between individuals, OWL offers constructors such as (*owl:differentFrom*, *owl:AllDifferent*, *owl:sameAs* and *owl:sameIndividualAs*).

- **Axioms**: OWL is a rich language because it is not only limited to describe a class but it offers the possibility to give more information about the characteristics of a class. OWL defines three class axioms that are *rdfs:subClassOf*, *owl:equivalentClass* and *owl:disjointWith*.

```
<owl:class rdf:ID="Accepted_paper">
 <owl:disjointWith">
 <owl:class rdf:ID ="Rejected_Paper"/>
 </owl:disjointWith>
 <rdfs:subClassOf>
 <owl:class rdf:ID ="Evaluated_Paper"/>
 </rdfs:subClassOf>
</owl:class>
```

The example above shows two kinds of relations. *owl:disjointWith* declares a disjoint relation between the two classes Accepted_paper and Rejected_paper where an accepted paper cannot be a rejected one. The class axiom *rdfs:subClassOf* describes a subclass relation between two classes Accepted_paper and Evaluated_paper. In fact, if a paper is accepted then it has been evaluated.

2.2.3.2 OWL sublanguages

OWL presents three sublanguages OWL-Lite, OWL-DL and OWL-Full. Each of these languages is oriented to fulfill some requirements.

- OWL-Lite is a simple language which is easy to implement but it offers a limited expressivity because it describes only a hierarchy of classes with simple constraints.
- OWL-DL is more expressive compared to OWL-Lite. It contains all the OWL constructors but restricts the use of some features. Due to the fact that it is inspired from DLs, the OWL-DL allows for efficient reasoning mechanisms.
- OWL-Full uses all the OWL language primitives and offers the possibility to use these primitives with RDF and RDFS. The main disadvantage of OWL-Full is that its high expressiveness limits its decidability and there is no guarantee of a complete reasoning.

We used in this thesis ontologies expressed in OWL-DL to support our proposed approach.

2.2.3.3 OWL ontology example

Formally, an OWL ontology is defined in (Ehrig & Staab, 2004) as the tuple:

$$O = < C, H_C, R_C, H_R, I, R_I, A>$$

C is a set of concepts or classes (instances of *owl:class*) structured in a subsumption hierarchy H_C (instances of *rdfs:subClassOf*). R_C (instances of *owl:objectProperty*) link concepts. These relations are organized in a subsumption hierarchy H_R (instances of *rdfs:subPropertyOf*). An individual i belonging to the set of individuals I is an instance of a class c such that $c \in C$. Two individuals i and j may be related by a relation R_I. A set of axioms A helps to infer knowledge from existing ones in an ontology. Figure 2.2

is an excerpt of an ontology related to the conference organization[10]. Classes are represented by circles and properties are represented by dotted arrows. For example, *Conference* is a class and $Conference\_volume$ is its subclass. The two classes *Committee* and $Conference\_volume$ are related by an objectProperty $has\_a\_committee$.

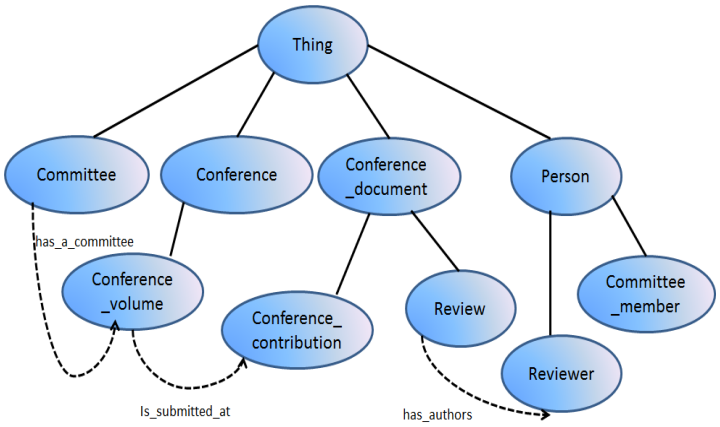

Figure 2.2: Excerpt of an ontology related to conference organization

2.2.4 Benefits of using ontologies

Ontologies are useful in many applications. (Uschold & Grüninger, 2004) identify four main scenarios where using ontologies seems to be the best solution.

- Neutral authoring: An organization's architecture is based on the communication between different target systems, each of them is using different tools. These systems cannot interoperate because the tools used do not have understandable standard format between these systems. For that purpose, an organization may be called to create its own neutral ontology and to develop translators which will play the role of bridges between the created ontology and the terminologies required by the target systems. In order to facilitate the translation, the neutral ontology must include all the common features of the target systems.

- Common access to information: In an application domain, a software system may be faced with the need to operate with another one in order to get access to information. Due to the fact that each system has its own representation, different from each other, it is essential to find a solution to ensure information sharing. Using translators may

[10]http://oaei.ontologymatching.org/2013/conference/index.html

be inconvenient because they can use wrong terms to express shared information. For that purpose, it is crucial to create a common ontology which represents an agreed format able to translate statements of various formats.

- Ontology-based specification: There is a great interest in using ontologies for software engineering. For a domain of discourse, an ontology is created describing the different things that a software must address. The specification and the development of the software are based on the use of this ontology which serves as a partial set of requirements for developing the software.

- Ontology-based search: In addition to its ability to play a role of an agreed standard between systems and a basis of software development, an ontology can be used as a tool for structuring information for a repository. In fact, repositories are organized at a higher level of abstraction and are indexed to an ontology. The use of ontologies as an indexing mechanism facilitates the access to the suitable information.

Based on these main scenarios, we can resume that an ontology as a formal conceptualization of a domain of discourse is used to ensure interoperability, information sharing between systems and to facilitate information retrieval. In addition to that, using ontologies helps to infer implicit knowledge and perform reasoning tasks. We will be restricted in the following to the description of the main benefits of using ontologies such as interoperability, information retrieval and reasoning. Some other benefits can be found in (Stuckenschmidt, 2003).

2.2.4.1 Interoperability

Interoperability is defined as the ability of making systems working together. The interoperation allows information exchange and knowledge sharing. Due to the fact that machines are not able to do reasoning tasks to guess the meaning of terms, a certain degree of formality must be provided in order to facilitate the communication between machines. Ontologies are often used as inter-lingua for providing interoperability, since they use a common vocabulary to describe a domain of discourse in a formal way.

2.2.4.2 Better information access

An ontology can be used as a way to retrieve information efficiently. Information retrieval consists in obtaining an information from a collection of resources. The techniques generally used do not guarantee that the user finds the information he is searching for. In fact, his

query might be inconsistent with the vocabulary used and related to the document. In other words, terms of the query may not match those of the document. An ontology is then proposed as a description of the document vocabulary and it is used as a basis to give the user the desired response.

2.2.4.3 Reasoning

An ontology is able to describe new concepts and relationships in a given domain as well as the instances of these concepts and relations. The need to perform reasoning tasks and to infer implicit information from what is described in an ontology is important. Among these tasks, we can mention:

- Ontology consistency: In an ontology research field, one cannot talk about a static ontology because it is a representation of knowledge for a given domain of discourse and users may have new requirements to represent or even some modifications may be done. These changes can have an effect on the ontology's consistency. (Haase & Stojanovic, 2005) distinguish three forms of ontology consistency.
  - Structural consistency: This notion of consistency ensures that the ontology is conform to constraints imposed by this language. Structural consistency can be enforced by verifying a set of structural conditions related to the concerned ontology language. As an example of structural conditions we can state "The complement of a class must be a class".
  - Logical consistency: An ontology is logically consistent if it does not contain contradicting information, it conforms to the underlying formal semantics of the ontology language.
  - User-defined consistency: A lack of definitions of consistency may not be captured by the ontology language which leads to additional conditions defined explicitly by some applications or users to ensure the ontology consistency. As an example users could require that classes can only be defined as a subclass of at most one of the other classes.
- Concept satisfiability: It verifies whether a concept does not necessarily represent an empty concept. In fact, it verifies whether this concept admits at least one individual.
- Concept subsumption: It checks whether a concept (the subsumer) is considered more general than another one (the subsumee).

- Concept equivalence: A concept A is equivalent to concept B if A and B subsume each other.
- Concept disjointness: Two concepts are disjoint if they do not share any common instance.

2.3 Ontology-based semantic integration

Ontologies have gained popularity as a mean for providing an explicit formal vocabulary that describes a domain of discourse. The open nature of the semantic web and the need to ensure information sharing, make the adoption of common ontology unrealistic solution for three main reasons (Hameed, Preece, & Sleeman, 2004). First, the construction of a shared ontology leads to a competing choice of what the ontology must represent. Second, an ontology is a conceptualization of a particular view of the world so its design is based on subjective features. Finally, due to the fact that knowledge domains are not static, a common ontology needs to evolve over time in order to fit with the world's dynamicity. These reasons make it impossible to adopt a single ontology but rather disparate ontologies. These ontologies may be developed independently from each others and can cover totally or partially the same scope. Their entities can be defined according to different levels of granularity or even they can be described with different representation languages. In order to enable interoperability and to ensure information sharing, it is mandatory to mitigate the effect of *semantic heterogeneity* through a *semantic integration*. We present in this section *ontology matching* as a key application area where discovering correspondences between ontologies is a solution to the problem of semantic heterogeneity.

2.3.1 Motivating example

In order to highlight the importance of matching ontologies, we present in this section a detailed example related to conference organization. Suppose that the organizers of two different conferences decide to cooperate together. They consider that bringing together two different and complementary research areas can have benefits. This cooperation is performed technically by the integration of two ontologies O_1 and O_2 such that O_1 stores information related to the first conference and O_2 stores those related to the second conference. The integration goes first by identifying the candidates to be aligned. This is the objective of an ontology matching process which consists in searching for an entity of O_1 its corresponding entity in O_2. Figure 2.3 illustrates the two ontologies matching O_1 and O_2. For example, the entity *Conference* in O_1 has *Conference* and $Conference\_volume$ as

corresponding entities in O_2 whereas the entity *document* in O_1 has to be aligned to the entity $conference\_document$ in O_2.

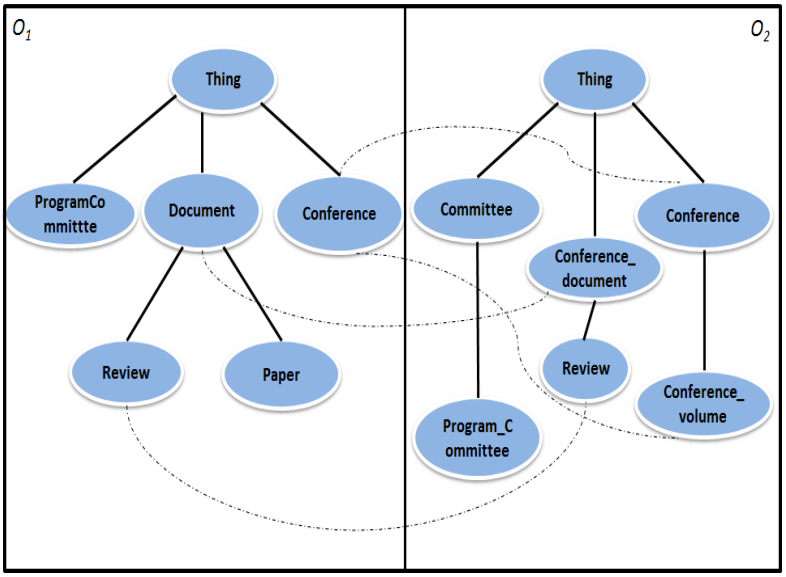

Figure 2.3: Excerpt of matching ontologies O_1 and O_2

2.3.2 Classification of ontology mismatches

The objective of matching ontologies is to dwindle heterogeneity between them. (Predoiu et al., 2004) define ontology mismatch as "a difference between two ontologies that contradicts the semantic correspondence between the ontology entities at hand". It is very important to detect mismatches between individual ontologies, analyze them and try to resolve them. Many researches and practitioners have focused on analyzing the origin of ontology mismatches on the semantic web like in (Smart & Engelbrecht, 2008). Others were rather interested in classifying the different types of ontology mismatches encountered during the semantic integration process as in (Klein, 2001).

According to (Klein, 2001), mismatches may occur in two main levels namely the *language level* and the *ontology level*. The former concerns the features and the constructors used to describe the terms of an ontology. The latter focuses on the differences occurring with overlapping ontologies in the sense that ontologies do not describe exactly a same domain but rather overlapping domains where some features existing in one ontology may not exist in the second one. Below, we give in detail the different types of mismatches that can occur at each of these two levels.

- **Language level mismatches** occur when two ontologies are described with different ontology languages. In this level, four types of mismatches are identified.

  – *Syntax*: Generally, different ontology languages use different syntaxes. For example, the class of conferences is defined in RDFS as *<rdfs:Class ID = "Conference">* while in LOOM, the same class is expressed through *(defconcept Conference)*.

  – *Logical representation*: A logical notion can be represented in different ways. For example, if we want to represent the acceptance of a paper or its rejection, we may use the disjoint classes accepted_paper and rejected_paper. To represent this disjointness, in one language it is possible to define it as (disjoint accepted_paper rejected_paper) while in another language we have to use negation in subclass statements (accepted_paper subclass-of (NOT rejected_paper), rejected_paper subclass-of (NOT accepted_paper)).

  – *Semantics of primitives*: Although the same name is used for a primitive construct in two languages or even the same syntax is used, the semantics may differ. For example, RDFS interprets multiple statements <rdfs:range ...> as the union of ranges while DAML+OIL uses intersection semantics.

  – *Language expressivity*: Some languages are able to express notions that other languages can not. For example, in a conference an author can submit up to two articles. This precision of qualifying a cardinality restriction is not possible with the RDFS but it can be expressed through maxCardinality with the OWL language.

- **Ontology level mismatches**: Even if two ontologies are represented with a same ontology language, ontology mismatches can occur. In this level, a distinction is made between *conceptualization mismatches* and *explication mismatches*. The former are differences in the way a domain is conceptualized, on how we identify ontological concepts and the different relations between them. The latter concerns the differences in the specification, on how concepts and constraints as well as relationships between them are defined.

  Conceptualization mismatches are divided into scope and model coverage.

  – *Scope*: Two classes seem to represent the same concept but do not have exactly the same instances, although they intersect.

  – *Model coverage and granularity*: It concerns the part of the domain covered or the level of detail used to model the domain. For example, for the organization of conferences, we may find an ontology representing the contributions of authors

as long paper or abstract. Another ontology may specify in which session this contribution has been presented (industrial session, demo session, poster session, etc.).

Explication mismatches are divided into terminological, modeling style and encoding.

- *Terminological mismatches*
  - Synonym terms: Different names are used for a same concept. For example, the term "paper" is used in one ontology while the term "article" is used in another one.
  - Homonym terms: The same term can have different meanings depending on the context used. For example, the term "article" has a different meaning in a conference organization than it has in grammar. In the former it identifies the paper submitted in a conference while in the latter it means a grammatical element used to indicate definiteness or indefiniteness.
- *Modeling style*
  - Paradigm: Different paradigms can be used to represent concepts such as time, temperature, plans, etc.
  - Concept description: Modeling concepts in an ontology can differ depending on how the domain described through an ontology is modeled. For example, in an ontology O_1 the concept *paper* is represented as a subclass of the concept *Document* ($paper < document$) whereas in O_2 it is represented as the subclass of $regular\_contribution$ and thus through a subclassOf hierarchy described by ($paper < Regular\_contribution < Written\_contribution < Conference\_contribution < ConferenceDocument$)
- *Encoding*: Values in ontologies can be encoded in different formats. A conference's date for example may be encoded as "dd/mm/yyyy" or as "mm-dd-yy".

2.3.3 Ontology matching process

As already mentioned, developing different ontologies independent from each other creates a semantic heterogeneity. Ontology matching is a solution to handle heterogeneity and to ensure an efficient interoperability.

Definition 2.1. *Ontology matching is a function f which from a pair of ontologies to match O_1 and O_2, an input alignment A, a set of parameters p and a set of parameters and resources r, returns an alignment A' between these ontologies (Euzenat & Shvaiko, 2013a):*

$$A' = f(O_1, O_2, A, p, r)$$

Parameters and resources refer to thresholds and external resources respectively. The output of an ontology matching is an alignment which is a set of correspondences between entities belonging to matched ontologies. Alignments can be simple or complex. The former concerns the alignments with cardinality 1:1 (one-to-one) where only one entity of O_1 is matched with only one entity of O_2. The latter concerns the cardinalities 1:m (one-to-many) where one entity of O_1 is matched with multiple correspondences of entities in O_2 or n:1 (many-to-one) where multiple entities of O_1 were matched with only one entity of O_2 or even the cardinality can be of the form n:m (many-to-many) where multiple entities in O_1 can be matched with multiple entities of O_2. We will be restricted in this thesis to one-to-one and one-to-many correspondences.

Definition 2.2. *(Euzenat & Shvaiko, 2013a) define a correspondence as a 5-tuple <id, e_1, e_2, n, R> where:*

- *id is a unique identifier of a correspondence.*
- *e_1 and e_2 are entities belonging respectively to a source ontology O_1 and a target ontology O_2. These entities can be concepts, properties or instances.*
- *n is a confidence measure for a correspondence. It is the result of the application of a matching technique, $n \in [0, 1]$.*
- *R is the relation between two entities. R can be equivalence, subsumption, disjointness.*

Based on the figure 2.3, *Conference* has as a correspondence $Conference\_volume$. This is represented as follows:

```
<map>
 <Cell cid='10'>
 <entity1 rdf:resource="http://cmt#Conference"/>
 <entity2 rdf:resource="http://conference#Conference_Volume"/>
 <measure rdf:datatype="xsd:float">1.0</measure>
 <relation>=</relation>
 </Cell>
 </map>
```

The example of the correspondence presented above respects the alignment format proposed in (Euzenat, 2004). The example shows that the equivalence relation *(=)* holds between entities $Conference$ and $Conference\_Volume$ with a confidence measure equal to *1.0*.

(Ehrig, 2007) presents a general ontology matching process as illustrated in figure 2.4. This process consists of six different steps that can be found in the majority of ontology matching approaches excepting some cases where some steps are merged or a change is made in the order of these steps.

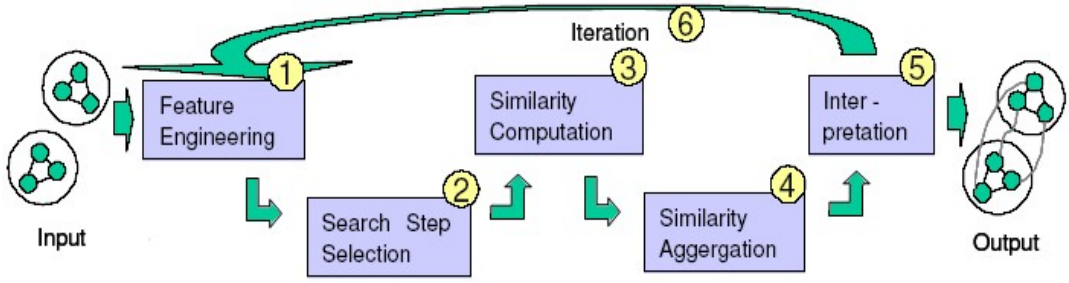

Figure 2.4: Ontology matching process

(Ehrig, 2007)

The input of this process is two or more ontologies to be aligned. In addition to these ontologies, the input can include initial alignments.

1. *Feature engineering*: Comparing two entities from two given ontologies is the basis of the matching process. Each entity is described through its features which have to be taken into account for the comparison because ontologies are not only viewed as a graph of concepts and the different relations between them but also they hide semantics of each individual feature that have to be exploited. Possible characteristics to be considered for a comparison are identifiers, RDFS primitives, OWL primitives, etc. as identified in (Ehrig & Sure, 2004). For example, for an alignment, we may consider the OWL primitives related to the taxonomy or the label of an entity.

2. *Search step selection*: It consists in selecting the pairs of entities to be compared during the matching process. It is possible to select a subset of pairs and to ignore others. The selection of the candidate pairs to be compared can be handled following two different strategies. First, we can choose to compare all the entities of the first ontology O_1 with all the entities of the second ontology O_2. Second, we can choose to compare only the entities of the same type. For instance, we compare all concepts

of O_1 with all concepts of O_2 or all properties of O_1 with all properties of O_2 or all instances of O_1 with all instances of O_2.

3. *Similarity computation*: The entities of each pair selected in the previous step are compared based on the feature chosen for this comparison. The comparison focuses on computing similarities between entities through the application of similarity measures. This measure returns a degree of similarity between the entities.

4. *Similarity aggregation*: At the previous step, we calculated for each pair of entities several similarity values where each individual value is related to a specific feature. At this step the different similarity values for a candidate pair must be aggregated in order to get a unique and possibly more relevant similarity value.

5. *Interpretation*: Based on the individual or aggregated similarity values previously calculated, the aim of this step is to assign an alignment for each entity. An entity can have either a corresponding entity or multiple corresponding entities. The assignment is based on a threshold. In fact if the calculated similarity value is above the cut-off then an alignment is retained.

6. *Iteration*: The similarity of a pair of entities has an influence on the similarity of the pairs of entities neighbor to it. For that purpose, the computation of similarity values is performed through iterations where the similarity value is recomputed in each iteration based on the similarity values of the neighboring pairs of entities.

Several matching approaches have adopted the described process. They use for identifying the alignments a number of matching techniques. These techniques will be our focus in the next subsection where we give a detailed description.

2.3.4 Basic techniques for ontology matching

The ontology matching tends to discover relations between entities of two different ontologies O_1 and O_2. We are interested in this thesis in matching ontologies based on equivalence relations which focus on finding for each entity in O_1 its similar entity in O_2. There is a plethora of basic techniques able to detect similar entities (Euzenat & Shvaiko, 2013a). Each of these techniques focus on a particular feature of entities and are used as the basis of most of the ontology matching methods. In the following, we are restricted to the presentation of name-based techniques and structure-based techniques as the main techniques used in our thesis.

2.3.4.1 Name-based techniques

In order to find, for a given entity in O_1 its corresponding entity in O_2, the name-based techniques, called also *terminological techniques*, compare the names, the labels and the comments used to describe entities. We distinguish two main categories of name-based techniques: string-based techniques and language-based techniques.

- *String-based techniques* are based on the comparison of the structure of strings. Often, these techniques quantify their similarity by calculating a distance between two strings. There are several techniques for comparing strings depending on the way the string is viewed (a set of letters, a set of words, etc.). Among them, we may cite:

  – Hamming distance calculates the number of positions in which two strings differ. It is a dissimilarity measure $\delta_{hamming}$: $S \times S \rightarrow [0, 1]$ such that:

  $$\delta_{hamming}(s,t) = \frac{\left(\sum_{i=1}^{\min(|s|,|t|)} s\,[i] \neq t\,[i] \right) + ||s| - |t||}{\max(|s|, |t|)} \tag{2.1}$$

  where S is a set of strings, s and $t \in S$ and $|s|$ is the length of the string. Suppose that we have two strings s = "paper" and t = "abstract" then $\delta_{hamming}(paper, abstract) = \delta_{hamming}(s,t) = 0.875$.

  – Edit distance is the minimal cost of edition operations to be applied to one string in order to obtain the other string. These operations consist on insertion, deletion and substitution (replacement of a character by another).

  The Levenshtein distance is an edit-distance with all costs equal to 1. It is the minimum number of edition operations to transform a string into another.

  For example, the Levenshtein distance between *paperAbstract* and *Abstract* is equal to 0.615.

  – Jaro measure is based on the number and order of the common characters between two strings. It is defined as σ_{jaro} $S \times S \rightarrow [0, 1]$ such that:

  $$\sigma_{jaro}(s,t) = \frac{1}{3}\left(\frac{|com(s,t)|}{|s|} + \frac{|com(t,s)|}{|t|} + \frac{|com(s,t)| - |transp(s,t)|}{|com(s,t)|} \right) \tag{2.2}$$

  where $s\,[i] \in com(s,t)$ iff $\exists j \in [i - \min(|s|,|t|)/2i + (\min(|s|,|t|)/2]$ and *transp(s,t)* are the elements of *com(s,t)* which occur in a different order in s and t.

For instance, if we suppose that s = "paper" and t = "abstract" then the *common*(s,t) is equal to 2 representing the number of common letters and transp(s,t) is equal to 1 representing the number of transposed common letters then
$\sigma_{jaro}(s,t) = \sigma_{jaro}(paper, abstract) = 0.38$.

- *Language-based techniques* use natural language processing techniques (NLP) to determine the similarity between two terms. String-based methods, previously described, consider strings as a sequence of characters and determine similarity between strings unlike the language-based techniques where the comparison is held between terms used for labeling concepts in ontologies. For example, $conference\_fees$ is a term. These techniques can be intrinsic or extrinsic.

  - Intrinsic techniques are based on algorithms and consist on handling a linguistic normalization to transform a given entity into a standardized form. We may cite:

    * Tokenization consists in segmenting a term into a set of tokens where punctuation, blank characters are omitted. For example, the term *program_committee_chair* becomes <program, committee, chair>.
    * Lemmatisation: Tokens are morphologically analyzed in order to transform them into normalized forms. For example *abstract_of_paper* is a variant of *paperAbstract*.
    * Term extraction: Similar terms are identified based on repetition of morphologically similar terms in a text and the use of patterns ($noun^1 noun^2 \rightarrow noun^2 on noun^1$). Based on this pattern for example, the term *web conference* becomes *conference on web*.
    * Stopword elimination discards meaningless tokens such as conjunctions, articles, prepositions, etc. For example, *reviewer of a paper* becomes *reviewer paper*.

  - Extrinsic techniques use external linguistic resources (dictionaries, thesauri, etc.) to determine the similarity between terms. The similarity is identified through the semantic links (synonyms, hyponyms, etc.) existing on the resources. Among these resources, we may cite WordNet [11] which is a lexical database for English developed by Princeton University. Nouns, verbs, adjectives and adverbs are organized in synsets (sets of synonyms) and the synsets are organized into senses (different meanings of the same concepts). The

[11] http://wordnet.princeton.edu/

synsets are related to other synsets via semantic relationships such as hypernym/hyponym denoting the relation superConcept/subConcept and the relation meronym/holonym denoting part of relations.

2.3.4.2 Structure-based techniques

Terminological techniques previously described, compare entities based on their names and their identifiers. Exploiting the structure of entities can be useful to detect similarities between concepts of two ontologies. We distinguish between the comparison of the internal structure of an entity and the comparison of relational structure. The former focuses on the entity itself without reference to other entities whereas the latter is based on the comparison between an entity with other entities to which it is related to.

- Internal structure: The comparison between two concepts goes through the comparison of the information contained on their internal structure. This information includes the properties of entities, their range, their cardinality as well as their characteristics (transitivity, symmetry, etc.). Based on the internal structure, we may find several entities sharing similar properties. For that reason, these methods are generally combined with name-based techniques in order to reduce the number of candidate correspondences. This is handled through the creation of correspondence clusters rather than discovering similar concepts.

- Relational structure: This kind of technique considers an ontology as a graph where the relation names label edges. Finding correspondences between entities of two ontologies goes through comparing entities they are related to. The more two entities are similar, the more their related entities are similar too. The taxonomic structure (*i.e.* graph made with the subClassOf relations) is intensively used for comparing classes. Wu-Palmer similarity has been proposed to calculate the distance between two classes. This distance considers that two classes can be semantically different even if they are near to the root of the hierarchy in terms of edges. But they can also be closer semantically although they are separated with a large number of edges.

  The Wu-Palmer similarity σ is defined for $o \times o \rightarrow R$ as:

$$\sigma(c_1, c_2) = \frac{2 \times \delta(c_1 \wedge c_2, \rho)}{\delta(c_1, c_1 \wedge c_2) + \delta(c_2, c_1 \wedge c_2) + 2 \times \delta(c_1 \wedge c_2, \rho)} \tag{2.3}$$

  where ρ is the root of the structure, $\delta(c_1, c_2)$ is the number of intermediate edges between two classes c_1 and c_2 and $c_1 \wedge c_2 = \{c_3 \in o; c_1 \leq c_2 \wedge c_1 \leq c_2\}$.

2.4 Conclusion

In this state of the art chapter, we emphasized the prominence of ontologies as the backbone of the semantic web through listing the main benefits of using them. In order to describe a domain of interest, ontologies are specified in a formal representation language such as the OWL deemed as the most expressive language. Hence, a detailed description with examples was given in this chapter. The evolution that the semantic web knows has led to the existence of different ontologies for a given domain. To reduce heterogeneity occurring between disparate ontologies, a matching process is performed. The main milestones of this process as well as the basic techniques used for detecting alignments are given in detail in this chapter. But, what about uncertainty in the matching? Can we consider that there is a degree of uncertainty on the resulting alignments especially that sometimes we get results far from each other? Can we consider that the similarity value calculated by a matching technique is none other than its degree of confidence to match an entity e_1 to an entity e_2? Not long ago, practitioners have considered that dealing with uncertainty in ontology matching is one of the challenges to be addressed (Shvaiko & Euzenat, 2008). This research area is not mature enough. For that purpose, few approaches have been proposed as a solution to the uncertainty in matching. A state-of-the art of the main approaches is presented in the following chapter.

Chapter 3

Uncertainty in the Semantic Web Community

Contents

The dynamicity of semantic web and the huge amount of shared information which, sometimes, is imprecise or vague, make dealing with uncertainty in semantic web one of the challenge to be tackled (Euzenat & Shvaiko, 2013a). Due to its importance, we devote a chapter where we give an overview of different approaches that managed uncertainty whether in representing ontologies or in matching them.

3.1 Introduction

Imperfection is a general term involving various concepts. Among them, we may cite imprecision and uncertainty.

- Imprecision is related to the content of information. It *"covers cases where the value of a variable is given but not with the precision required"* (Smets, 1991).
- Uncertainty is *"partial knowledge of the true value of the data. It results in ignorance (etymologically not knowing). It is essentially, if not always, an epistemic property induced by a lack of information. A major cause of uncertainty is imprecision in the data"* (Smets, 1996).

In order to represent accurately a real world, imperfection with its various aspects must be taken into account through an appropriate mathematical model. This model has to be chosen carefully because every aspect of imperfection has its own and appropriate model. The probability theory has gained popularity in representing uncertainty but failed in modeling other aspects of imperfection. For that purpose, many numerical models have been developed in the last years to cope with imperfection and to reason with uncertainty such as the fuzzy sets theory (Zadeh, 1965), the possibility theory (Dubois & Prade, 1988a; Zadeh, 1999), the imprecise probabilities (Walley, 1991) and the theory of belief functions (Dempster, 1967; Shafer, 1976).

The semantic web allows interactions between humans and computers, knowledge sharing, interoperability and re-usability among different sources of information. Dealing with uncertainty in the semantic web has been recently of great interest. To emphasize the importance of uncertainty in the semantic web, the World Wide Web Consortium (W3C) has created the World Wide Web Incubator Group[1] (URW3-XG) in 2006 which explored and defined the challenges of reasoning and representing uncertain information in the context of the World Wide Web.

As mentioned in URW3-XG's reports, managing uncertainty in the semantic web was motivated by a set of possible use cases where uncertainty representation and reasoning are mandatory. We may cite information fusion where the contained information in the semantic web can be incorrect or partially correct or even contradictory since it is provided by different sources of information. This situation introduces a problem of trust and credibility. In order to manage the conflict occurring, it is possible to assign a value to every source of information describing its degree of reliability. The aggregation of

[1] http://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/

information from multiple sources may reveal some uncertainty aspects and helps to infer new knowledge.

Ontologies, as the core of the semantic web, are not primarily designed to deal with uncertainty because their specification of a domain is strict and has to be well-defined. But, due to the fact that an ontology is an instantiation of a system, we may be faced with different types of uncertainty (lack of knowledge about the domain to represent, unreliability of the sources of information, etc.) (Hois, 2009). This has incited researchers to specify syntax and semantics for modeling uncertainty in an ontology by extending ontology languages with constructors able to give expressive power for representing uncertain knowledge and imprecise information. Modeling uncertainty in ontologies has been the focus of many works. Some use the probability theory to extend the OWL with additional constructors and use Bayesian networks as a graphical model to do reasoning tasks (Ding, 2005; Costa & Laskey, 2006; Yang & Calmet, 2005). In (Gao & Liu, 2005; Stoilos et al., 2005), the authors adopt the fuzzy set theory as an uncertainty approach for ontology representation. The belief function theory has been used in (Essaid & Ben Yaghlane, 2009) as a framework for enriching an OWL ontology with uncertain statements and to translate the obtained ontology into an evidential network.

Ontology matching is another area in which uncertainty must be taken into account. For example, an entity of a source ontology can find partial matches to one or more entities in a target ontology. Recently, matching ontologies has been viewed as a process not only limited to find correspondences between ontologies but also a process to cope with imperfect information (Euzenat & Shvaiko, 2013a). Modeling uncertainty during the matching process helps to improve the detection of correspondences and better characterize the obtained results. For that purpose, a number of studies used mathematical models to handle uncertainty in the ontology matching process: the probability theory(Pan et al., 2005; Mitra et al., 2005) and the belief function theory (Besana, 2006; Nagy et al., 2007; Wang et al., 2007).

The remainder of this chapter is organized as follows. In section 3.2, we present the belief function theory as the mathematical model we used in our study. This presentation is given based on the transferable belief model (Smets & Kennes, 1994). We describe the belief functions used for representing knowledge, the main rules of combination to get a new piece of evidence and finally we detail the main strategies for making decisions under the belief function theory. A special concern will be devoted in section 3.3 to the main studies that extend the OWL ontology to make it able to express uncertain knowledge. A deep description of works integrating the Dempster-Shafer theory when mapping ontologies will be the focus of section 3.4.

3.2 Belief function theory

The belief function theory, also known as *the theory of evidence* or *Dempster-Shafer theory* was originally introduced by (Dempster, 1967) and then further developed by (Shafer, 1976). It is a general mathematical framework for representing belief and reasoning under uncertainty. The work on this theory was first inspired from upper and lower probabilities (Dempster, 1967; Shafer, 1976). Then, a subjective interpretation has been given to this theory through a non-probabilistic model, namely *the Transferable Belief Model (TBM)*, proposed by Smets and Kennes (Smets, 1990; Smets & Kennes, 1994). It represents the quantified beliefs held by a source of information without using probabilistic measures. Accordingly, the TBM includes two levels:

- credal level ("*credo*" means I believe) is composed mainly of two steps. The first corresponds to the static part where the beliefs are quantified by belief functions and the second step represents the dynamic part of the model where beliefs are combined in order to obtain a new piece of evidence.
- pignistic level ("*pignus*" means a bet) in which decisions are made.

Compared to the probability theory, the Dempster-Shafer theory presents many benefits. Unlike the probability theory that is based on the use of singletons as possible solutions for a given problem, the Dempster-Shafer theory allocates beliefs to elementary hypotheses as well as to composite ones allowing then a better knowledge modeling and complex problem solving. The probabilistic approach is additive which means that an event exists or not. Since the sum of the probabilities must be equal to one, then the probability of an event determines the probability of its negation. The additivity constraint does not allow ignorance representation which is well modeled within the belief function theory. In this latter, the belief assigned to an event does not determine the beliefs of other events. As opposed to the probability theory, the theory of evidence can model the degree of ignorance making it possible to distinguish between uncertainty and ignorance. Finally, one of the strongest point of the evidence theory is its ability to combine evidences from different sources of information in order to get a global and new piece of evidence. The obtained evidence helps to make decision which can be handled through different strategies depending on application's needs.

In the following, we present the belief function theory's concepts based on the TBM interpretation.

3.2.1 Representation of uncertainty by belief functions

3.2.1.1 Frame of Discernment

Let Ω be a finite non empty set of n elementary hypotheses representing possible atomic solutions for a given problem. The set Ω, called the *frame of discernment*, is exhaustive and the hypotheses are mutually exclusive.

The set Ω is defined as:

$$\Omega = \{\omega_1, \omega_2, \dots \omega_n\} \tag{3.1}$$

From the frame of discernment, we define the power set denoted by 2^Ω as the set containing singleton hypotheses of Ω, all the disjunctions of these hypotheses as well as the empty set.

$$2^\Omega = \{A; A \subseteq \Omega\} = \{\emptyset, \omega_1, \dots, \omega_n, \omega_1 \cup \omega_2, \dots, \Omega\} \tag{3.2}$$

We will use the notation ω_i for representing a singleton hypothesis and A for designating any subset of Ω.

Example 3.1. *In order to participate in a conference, authors submit their papers which will be reviewed by a committee. There are three types of contributions: long paper (LP), short paper (ShP) and poster (PS). The frame of discernment can be represented as:*

$$\Omega = \{LP, ShP, PS\} \tag{3.3}$$

The corresponding power-set 2^Ω is defined as:

$$\Omega = \{\emptyset, LP, ShP, PS, LP \cup ShP, LP \cup PS, ShP \cup PS, LP \cup ShP \cup PS\} \tag{3.4}$$

3.2.1.2 Basic Belief Assignment

The *basic belief assignment (bba)*, denoted by m, is a mass function able to represent imperfect knowledge. It is a mapping from elements of 2^Ω to $[0, 1]$ such that it assigns a positive value belonging to $[0, 1]$ to any proposition. A *bba* satisfies the constraint:

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{3.5}$$

The value $m(A)$ is the basic belief mass *(bbm)* given to A. It represents the part of belief exactly committed to the proposition A and to none other subset of A (Smets & Kennes, 1994).

The *focal elements* are the subsets A of 2^Ω such that $m(A)$ is not null. The union of focal elements forms the *core*.

To work under the *closed-world assumption*, (Shafer, 1976) added a constraint ($m(\emptyset) = 0$) to make the frame of discernment exhaustive where the possible decisions to be taken are within the frame. With the introduction of the TBM, (Smets, 1990) advocates the *open-world assumption* where he supposes that the frame can be incomplete and that a decision can be outside the frame through $m(\emptyset) \geq 0$.

Example 3.2. *(Continued) Let us consider* $\Omega = \{LP, ShP, PS\}$. *Reviewers can express their beliefs, e.g. when a paper is reviewed. The reviewer may suggest that a paper has to be rewritten as a short paper. He expressed his beliefs through the following mass distribution:* $m(LP) = 0.3, m(LP \cup ShP) = 0.5, m(LP \cup ShP \cup PS) = 0.2$.

Special bbas are defined. We may cite:

- normal bba: a bba is normalized when $m(\emptyset) = 0$. To get a normalized bba from an unnormalized bba, a normalization process must be performed which is defined as:

$$m(A) = \begin{cases} \frac{m(A)}{1 - m(\emptyset)} & \forall A \subseteq \Omega,\ A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \tag{3.6}$$

- categorical bba: a bba is categorical when it has a unique focal element A such that $m(A) = 1$. If A is a singleton, then the bba models certainty and precision and if A is a disjunction of hypotheses then the bba models certainty and imprecision.
- vacuous bba: It is a categorical bba with Ω as a unique focal element ($m(\Omega) = 1$). It models the total ignorance.
- dogmatic bba: It is the case when Ω is not a focal element such that $m(\Omega) = 0$.
- consonant bba: A bba is consonant when all the focal elements are nested.
- Bayesian bba: A bba is Bayesian when all the focal elements are singletons. In that case, the bba represents a probability distribution such that $\sum_{\omega \in \Omega} m(\omega) = 1$.

- simple support function bba: a bba is of simple support if it has two focal elements where one of them is Ω. This bba is defined for $\alpha \in [0,1]$ as:

$$\begin{cases} m(A) = 1 - \alpha & A \subset \Omega \\ m(\Omega) = \alpha \end{cases} \tag{3.7}$$

3.2.1.3 Transformations of belief functions

Based on the aforementioned basic belief assignment, other functions (credibility function, plausibility function and commonality function) can be deduced where they represent with different semantics the same information and are used especially to make easier their computation.

3.2.1.3.1 *Belief Function*

Unlike the *basic belief mass* that quantifies the part of belief exactly committed to a proposition A, the belief function, noted as bel, takes into account all the belief allocated to A by summing all the masses of subsets of A. The belief function is a mapping from elements of 2^Ω to $[0,1]$ such that:

$$\text{bel}(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \quad \forall A \subseteq \Omega \tag{3.8}$$

The belief function verifies the following properties:

- $\text{bel}(\emptyset) = 0$ and $\text{bel}(\Omega) = 1$ under the closed world assumption.
- $\text{bel}(A_1 \cup \ldots \cup A_n) \geq \sum_{\emptyset \neq I \subset \{1,\ldots,n\}} (-1)^{|I|+1} \text{bel}(\cap_{i \in I} A_i), \forall n > 0, \forall A_1 \cup \ldots \cup A_n \subseteq \Omega$

As mentioned previously, the belief function represents the same information as the mass distribution m but differently. This is noticeable through the Möbius transformation that allows to get the mass distribution from the belief one using the following equation:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} \text{bel}(B) \tag{3.9}$$

where $|A \backslash B|$ represents the cardinality of the set of elements of A which do not belong to B.

3.2.1.3.2 *Plausibility Function*

The plausibility function, noted pl, is defined as:

$$\begin{cases} \text{pl} : 2^{\Omega} \rightarrow [0,1] \\ \text{pl}(A) = \sum\limits_{A \cap B \neq \emptyset} m(B) \end{cases} \tag{3.10}$$

pl is a dual measure of bel and it can be written as:

$$\text{pl}(A) = 1 - \text{bel}(\overline{A}) \tag{3.11}$$

The plausibility function measures the maximum amount of belief that supports the proposition A by taking into account all the elements that do not contradict A. The plausibility function should verify the following properties:

- $\text{pl}(\emptyset) = 0$ and $\text{pl}(\Omega) = 1$ under the closed world assumption.
- $\text{pl}(A_1 \cap \ldots \cap A_n) \leq \sum\limits_{I \subset \{1,\ldots,n\}} (-1)^{|I|+1} \text{pl}(\cup_{i \in I} A_i), \forall n > 0, \forall A_1 \cup \ldots \cup A_n \subseteq \Omega$

Like the belief function, the basic belief assignment can be obtained from the plausibility function through the following equation:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A-B+1|} \text{pl}(\overline{B}) \tag{3.12}$$

3.2.1.3.3 *Commonality Function*

It is especially used for facilitating the computation and demonstration of some theorems. It is computed as the sum of the masses allocated to the super sets of A.

$$\text{q}(A) = \sum_{B \supseteq A} m(B) \tag{3.13}$$

The basic belief assignment can be recovered from the commonality function through the following equation:

$$m(A) = \sum_{A \subseteq B} (-1)^{|B-A|} \text{q}(B) \tag{3.14}$$

Example 3.3. *(Continued) Suppose that a reviewer gives his mass distribution. In table 3.1, we present this bba as well as its corresponding belief, plausibility and commonality functions.*

Table 3.1: Belief, plausibility and commonality functions

	m	bel	pl	q
$\emptyset$	0	0	0	1
LP	0.03	0.03	0.47	0.47
ShP	0.3	0.3	0.9	0.9
$LP \cup ShP$	0.2	0.53	0.97	0.4
PS	0.03	0.0.3	0.47	0.47
$PS \cup LP$	0.04	0.1	0.7	0.24
$PS \cup ShP$	0.2	0.53	0.97	0.4
$PS \cup ShP \cup LP$	0.2	1	1	0.2

3.2.2 Combination of belief functions

The combination of imperfect data (uncertain, imprecise and inconsistent) presents a solution to obtain aggregated information. The theory of belief function is a useful tool for data fusion. In fact, for a given problem and for the same frame of discernment, it is possible to get a mass function synthesizing knowledge from separate and independent sources of information using a combination rule. Mainly, there exists three modes of combination: conjunctive combination, disjunctive combination and mixed combination.

3.2.2.1 Conjunctive Combination

This mode of combination is used when the two sources of information to combine are distinct and independent. The normalized conjunctive rule of combination was initially introduced by (Dempster, 1967) and then used by (Shafer, 1976). It combines mass functions by taking into account the intersection of the elements of 2^{Ω}. This rule, noted as $\oplus$, allows to combine two distinct mass functions m_1 and m_2 as follows:

$$m_{1\oplus 2}(A) = \begin{cases} \dfrac{\sum\limits_{B\cap C=A} m_1(B) \times m_2(C)}{1- \sum\limits_{B\cap C=\emptyset} m_1(B) \times m_2(C)} & \forall A \subseteq \Omega,\ A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \tag{3.15}$$

where $\sum\limits_{B\cap C=\emptyset} m_1(B) \times m_2(C)$ represents the global conflict and the rule is normalized *via*

$1 - \sum_{B \cap C = \emptyset} m_1(B) \times m_2(C)$. This normalization is used to hide the conflict by reallocating the mass of the conflict onto all the focal elements. In addition to that, this rule is interesting under the closed-world assumption.

In order to solve the problem enlightened by (Zadeh, 1965) where he demonstrated that the normalization step leads to unsatisfactory results, many fusion operators have been proposed (Yager, 1987; Smets, 1990; Dubois & Prade, 1988b). Under a conjunctive combination, (Smets, 1990) considered that the conflict is rather due to the fact that the frame of discernment is non exhaustive and proposed to work under the open-world assumption where a non null mass can be allocated to the empty set. For that purpose, he suggested a non-normalized conjunctive combination rule noted as $\textcircled{\cap}$ and defined through:

$$m_{1 \textcircled{\cap} 2}(A) = \sum_{B \cap C = A} m_1(B) \times m_2(C) \tag{3.16}$$

(Yager, 1987) proposed to manage the conflict under the closed-world assumption and affected the conflict measure to the frame of discernment. The conflict $m(\emptyset)$ is then interpreted as the total ignorance. The combination rule proposed by (Yager, 1987) is defined for two bbas m_1 and m_2 as:

$$\begin{cases} m_Y(A) = m_{1 \textcircled{\cap} 2}(A) & \forall A \in 2^\Theta,\ A \neq \Theta \text{ and } A \neq \emptyset \\ m_Y(\Theta) = m_{1 \textcircled{\cap} 2}(\Theta) + m_{1 \textcircled{\cap} 2}(\emptyset) & \\ m_Y(\emptyset) = 0 & \end{cases} \tag{3.17}$$

3.2.2.2 Disjunctive Combination

The conjunctive combination is generally used when the two sources of information are reliable. The disjunctive rule of combination (DRC), noted as $\textcircled{\cup}$, has been proposed by (Smets, 1990) and is used when we doubt of the reliability of at least one of the sources. This rule takes into account the unions between focal elements and is defined for two bbas m_1 and m_2 as:

$$m_{1 \textcircled{\cup} 2}(A) = \sum_{B \cup C = A} m_1(B) \times m_2(C) \tag{3.18}$$

3.2.2.3 Mixed Combination

(Dubois & Prade, 1988b) proposed a mixed combination that takes advantage of the conjunctive combination and the disjunctive as well. This rule is expressed as follows:

$$m_{DP}(A) = m_{1\bigcirc\!\!\!\!\cap 2}(A) + \sum_{B\cap C=\emptyset, B\cup C=A} m_1(B)m_2(C) \quad \forall A \in 2^\Theta \tag{3.19}$$

For an exhaustive state of the art of combination rules in the theory of belief functions, the reader may refer to (Smets, 2007).

Example 3.4. *(Continued) Suppose now that a paper submitted to a conference is evaluated by two reviewers. Each of them expresses his degree of belief via a bba. The combination of bbas is illustrated in table 3.2. This combination is performed with different combination rules.*

Table 3.2: Combination of two bbas through different combination rules.

	m_1	m_2	$m_{1\oplus 2}$	$m_{1\bigcirc\!\!\!\!\cap 2}$	m_Y	$m_{1\bigcirc\!\!\!\!\cup 2}$	m_{DP}
$\emptyset$	0	0	0	0.2751	0	0	0
LP	0.03	0.4	0.2846	0.2063	0.2063	0.0120	0.2063
ShP	0.3	0.02	0.2662	0.1930	0.1930	0.0060	0.1930
$LP \cup ShP$	0.2	0.1	0.1380	0.1000	0.1000	0.2576	0.2206
PS	0.03	0.1	0.0872	0.0632	0.0632	0.0030	0.0632
$PS \cup LP$	0.04	0.01	0.0199	0.0144	0.0144	0.0360	0.0294
$PS \cup ShP$	0.2	0.07	0.1214	0.0880	0.0880	0.0917	0.1186
$PS \cup ShP \cup LP$	0.2	0.3	0.0828	0.0600	0.3351	0.5937	0.1689

3.2.3 Decision making

Belief combination helps to make decision which consists in selecting, for a given problem, the most suitable action to handle. Under the belief function theory, there exist two main types of decision processes: decision on singleton hypotheses and decision on composite ones. The focus of this section is to describe in detail these two decision processes.

3.2.3.1 Decision process on singleton hypotheses

As described previously, in the credal level, information is modeled as belief functions which can be synthesized into a coherent one taking into account all the available information. Based on the obtained piece of evidence, decision is made on the pignistic level to select the best hypothesis. In this level, beliefs are transformed into a probability function, named *pignistic probability*. This probability quantification is based on the *"Insufficient Reason Principle"* which supposes, for a lack of information, an equi-probability between hypotheses instead of privileging a specific hypothesis (Smets, 1989). The pignistic probability, noted BetP is defined for $\omega \in \Omega$ by:

$$\mathrm{BetP}(\omega) = \frac{1}{1 - m(\emptyset)} \sum_{\omega \in A} \frac{m(A)}{|A|} \tag{3.20}$$

where $|A|$ is the cardinality of $A \subseteq \Omega$. The obtained solution equally distributes the mass $m(A)$ among the elements of A.

Example 3.5. *(Continued) Let us consider the results of combination obtained once the Dempster's rule of combination is used. To decide about how the paper should be submitted (long or short or poster), the pignistic probability can be used. The obtained results are:* $\mathrm{BetP}(LP) = 0.5252$, $\mathrm{BetP}(PS) = 0.3112$, $\mathrm{BetP}(ShP) = 0.6084$.
According to the obtained probabilities, the author should rather submit his paper as a short one.

Once we obtain the probability distribution, we select the most suitable hypothesis which is the one with a maximum BetP, applying decision theory. Suppose A is a finite set of possible actions $A = \{a_1, a_2, \ldots, a_n\}$ and Ω a finite set of hypotheses, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. An action a_j corresponds to the action of choosing the hypothesis ω_j. But, if we select a_i as an action whereas the hypothesis to be considered is rather ω_j then the loss is $\lambda(a_i|w_j)$. The expected loss associated with the choice of the action a_i is defined as:

$$R_{\mathrm{betP}}(a_i) = \sum_{w_j \in \Omega} \lambda(a_i|w_j)\mathrm{BetP}(w_j) \tag{3.21}$$

Based on a probabilistic reasoning, the decision consists in selecting the action which minimizes the expected loss. In the case of $\{0, 1\}$, we have to choose the hypothesis with a maximum BetP. In addition to minimizing pignistic risk, lower (equation 3.22) and upper expected loss (equation 3.23) can be determined (Denœux, 1997):

$$R_*(a_i) = \sum_{A \in \Omega} m(A) \min_{w_j \in A} \lambda(a_i|w_j) \tag{3.22}$$

$$R^*(a_i) = \sum_{A \in \Omega} m(A) \max_{w_j \in A} \lambda(a_i|w_j) \tag{3.23}$$

If we consider the three equations (3.21, 3.22, 3.23), we note that they are linked as follows:

$$R_*(a_i) \leq R_{\text{betP}}(a_i) \leq R^*(a_i) \tag{3.24}$$

(Denœux, 1997) considers that, based on the expected loss minimization, three distinct decision strategies may be defined:

- D_* is used for minimizing the lower expected loss and is defined as:

$$D_*(\omega) = a_* \text{ knowing that } R_*(a_*) = \min_{a \in A} R_*(a|\omega) \tag{3.25}$$

- D^* is used for minimizing the upper expected loss and is defined as:

$$D^*(\omega) = a^* \text{ knowing that } R^*(a^*) = \min_{a \in A} R^*(a|\omega) \tag{3.26}$$

- D_{betP} is used for minimizing the expected loss relative to BetP and is defined as:

$$D_{\text{betP}}(\omega) = a_{\text{betP}} \text{ knowing that } R_{\text{betP}}(a_{\text{betP}}) = \min_{a \in A} R_{\text{betP}}(a|\omega) \tag{3.27}$$

3.2.3.2 Decision process on composite hypotheses

Depending on application needs, we may be called to choose a solution for a given problem by considering all the elements in 2^Ω rather than considering only the elements of Ω. (Appriou, 2005) proposed a rule where he weights the plausibility function by a utility function depending on the cardinality of the set. The rule is defined for each $A \in 2^\Omega$:

$$A^* = \arg\left(\max_X(m_d(X)\text{pl}(X))\right) \tag{3.28}$$

where m_d is a mass defined by:

$$m_d(X) = K_d \lambda_X \left(\frac{1}{|X|^r}\right) \tag{3.29}$$

The value r is a parameter in $[0, 1]$ helping to choose a decision which varies from a total indecision when r is equal to 0 and a decision based on a singleton when r is equal

1. λ_X helps to integrate the lack of knowledge about one of the elements of 2^Ω. K_d is a normalization factor ad $\text{pl}(X)$ is a plausibility function. We note that instead of using plausibility function one may use credibility function or pignistic probability.

(Martin & Quidu, 2008) were interested in deciding on a union of hypotheses when it is impossible to decide between two classes and also to take a decision when the belief on a singleton is too weak. For that purpose, they proposed a decision rule operating in two steps:

- The decision rule of the maximum of credibility with reject proposed in (Le Hégarat-Mascle, Bloch, & Vidal-Madjar, 1997) makes decision on singletons and is applied to determine the elements that do not belong to the learning classes. This pessimistic decision rule consists in attributing the class ω_k for a new pattern x if:

$$\begin{cases} bel(\omega_k)(x) = \max\limits_{1\leq i\leq n} bel(\omega_i)(x), \\ bel(\omega_k)(x) \geq bel(\omega_k^c)(x) \end{cases} \tag{3.30}$$

- The decision rule presented in equation (3.28) is then applied on the non-rejected elements.

3.3 Approaches supporting imperfection in ontology representation

Ontologies have proved to be a powerful tool to capture knowledge about concepts and their relations thanks to the OWL which is a sophisticated language able to describe faithfully a domain of discourse. Yet, OWL is built on crisp logic where all the encoded information is supposed to be true. The dynamic that the semantic web is knowing and the huge amount of shared information between applications require more expressivity in the sense that it should be possible to deal with uncertainty. Making the OWL able to represent different aspects of imprecision is one of the research area that has interested many practitioners who combine mathematical theories and OWL to create a language able to represent uncertain information. Probability theory is the most common one to represent uncertainty and more specifically the Bayesian network which is powerful for holding probability reasoning tasks (Ding, 2005; Costa & Laskey, 2006; Yang & Calmet, 2005). We notice also that other approaches make use of the fuzzy sets theory (Gao & Liu, 2005; Stoilos et al., 2005) and the Dempster-Shafer theory (Essaid & Ben Yaghlane, 2009). This section goes briefly on the most promising approaches supporting uncertainty in ontology representation.

3.3.1 Ontology representation under the probability theory

3.3.1.1 BayesOWL

(Ding, 2005) proposed an approach to annotate the OWL ontology with additional probabilistic markups. Then, based on a set of structural rules, the obtained ontology is translated into a Bayesian network. To represent probabilistic information, (Ding, 2005) considers classes in an ontology as random binary variables (true and false) and treats the probability as a resource. For this purpose, the author defines two OWL classes *"PriorProb"* and *"CondProb"* to represent prior probability and conditional probability respectively. For example $P(A = a)$ is interpreted as the prior probability that an arbitrary individual belongs to class A.

- *"PriorProb"* has two properties "hasVariable" and "hasProbValue".
- *"CondProb"* has three properties "hasCondition", "hasVariable" and "hasProbValue".

Once the OWL ontology is enriched with probabilistic information, it is converted into a Bayesian network according to a set of structural translation rules. The quantitative aspect of the Bayesian network consists in assigning conditional probability tables (CPT) to each node of the Bayesian network. The obtained network, which preserves the semantics of the original ontology and which is consistent with all the given probability constraints, can support ontology reasoning, both within and across ontology as Bayesian inferences.

3.3.1.2 OntoBayes

OntoBayes is an ontology-driven uncertainty model able to represent uncertain knowledge (Yang & Calmet, 2005). It operates mainly in three steps. First, it annotates OWL ontology with Bayesian probabilities then it specifies dependency relationships to finally construct the model.

- Annotating OWL with probabilities: In order to make OWL ontologies able to represent uncertain information, the authors propose to annotate OWL ontologies with probabilities using three OWL classes: *"PriorProb"*, *"CondProb"* and *"FullProbDist"*. The first two classes are defined to identify the prior probability and conditional probability respectively. The probabilistic value is defined through a datatype property *"ProbValue"*. The class *"FullProbDist"* defines the full disjoint probability distribution. It has two disjoint object properties: *"hasPrior"* and *"hasCond"*

which establish the relation between *"FullProbDist"* and *"PriorProb"* and between *"FullProbDist"* and *"CondProb"* respectively.

- Annotating OWL with dependency relations: In order to facilitate the construction of a Bayesian network, the authors propose a property element *<rdfs:dependsOn>* to markup dependency properties which can be datatype properties or object properties.

- Graphical representation of OntoBayes goes through the construction of two graphs: The OWL graph and the Bayesian graph. The former is a directed graph built on the graph data model of RDF where nodes consist of classes and datatypes and the second is extracted from the OWL graph in order to show dependency relations. Properties represent nodes of this graph.

  The construction of the Bayesian network relies on two steps. First the dependency triples are identified and extracted from the OntoBayes ontology. A dependency triple consists of a subject, a predicate and an object where the predicate is constantly the primitive *<rdfs:dependsOn>*. The subject and object are properties. Then, all triples are merged where all nodes with a same identifier are composed.

3.3.1.3 PR-OWL

To face the inability of semantic web technologies to represent and reason under uncertainty, (Costa & Laskey, 2006) introduced PR-OWL as a probabilistic extension of OWL that provides a framework for creating probabilistic ontologies. The probabilistic semantics of PR-OWL are based on Multi-Entity Bayesian Networks (MEBN) (Costa & Laskey, 2005). MEBN is an extension of Bayesian networks that brings together classical first-order logic and Bayesian networks. MEBN represents the world as a collection of entities related to each others and described through their attributes. Knowledge about attributes of entities and their relationships is represented as a collection of MEBN fragments (MFrags) which describe probabilistic knowledge as a conditional probability distribution. A MEBN theory (MTheory) is a set of MFrags that collectively satisfies first-order logical constraints ensuring a unique joint probability distribution.

Probabilistic OWL (PR-OWL) is an OWL upper ontology for probabilistic ontologies. It is a set of classes, subclasses and properties. It extends OWL by adding new definitions and presenting its formal semantics based on MEBN. To create a probabilistic ontology, one has to import PR-OWL definitions (classes, subclasses, properties) into an OWL editor (*e.g.* Protégé). Then, one has to construct domain-specific concepts by enriching the ontology with uncertain aspects based on the MEBN model.

The initial version of PR-OWL fails in ensuring compatibility with OWL. For this

reason, PR-OWL2 has been proposed as an improvement for bridging the gap to OWL semantics (Carvalho, 2011). In fact, PR-OWL2 focuses on formalizing the relationship between OWL properties and PR-OWL random variables. In other words, given a concept in OWL, uncertainty definition goes through adding PR-OWL constructors so that OWL semantics are maintained and vice-versa to represent a random variable, already defined in PR-OWL, uncertainty should be maintained when it is represented in OWL.

3.3.2 Ontology representation under the Dempster-Shafer theory

Probability theory cannot deal with all the facets of uncertainty. In addition to that, the development of Bayesian networks as graphical models to represent uncertainty cannot handle situations where the representation of ignorance is crucial. For that purpose, (Essaid & Ben Yaghlane, 2009) proposed BeliefOWL as a model for representing uncertainty based on Dempster-Shafer theory. It takes as input an OWL ontology and produces an evidential network as output. The approach is handled mainly in four steps. First, OWL ontology is extended with belief constructors in order to make the ontology able to represent evidential information. For that purpose *"Prior evidence"* and *"Conditional evidence"* are added to represent prior belief masses and conditional belief masses respectively. Second, the evidential network is constructed where the qualitative level concerns the creation of nodes and relations between them based on a set of rules. These rules are applied to translate evidential information in an ontology into corresponding nodes and edges. Third, once the directed acyclic graph of the evidential network is constructed, masses are assigned to each node depending on the kind of node. If it represents an OWL class, then prior evidence and conditional evidence are attributed. If a node results from a relation existing between classes (union, intersection), then an adequate combination rule is applied. Finally, once the evidential network is constructed and masses assigned to each node, an inference process can be performed.

3.3.3 Ontology representation under the fuzzy sets theory

(Stoilos et al., 2005) proposed fuzzy OWL (f-OWL) as a method for extending OWL with fuzzy sets theory in order to represent and reason with imprecise information in the semantic web. The fuzzy extension of OWL DL focuses on OWL facts by adding degrees. For that purpose the f-OWL introduces an additional element <*owlx:degree*> to express the degree of fuzziness added to the facts.

Once the uncertain information is represented, a reasoning task must be provided for f-OWL which will be realized through the combination of syntactical extensions with f-SHOIN. F-SHOIN extends SHOIN to the fuzzy case by letting concepts and roles denote fuzzy sets of individuals and relations among them respectively. In f-SHOIN, the fuzzy knowledge base contains:
- Fuzzy TBox: a finite set of fuzzy concept axioms,
- Fuzzy RBox: a finite set of fuzzy role axioms,
- Fuzzy ABox: a finite set of fuzzy assertions.

The work presented in (Gao & Liu, 2005) extends the OWL language by encoding fuzzy constructors, axioms and constraints in order to map them to fuzzy DL. The extended OWL can represent fuzzy ontology as well as resolving fuzzy inference questions by constraint propagation calculus. In addition to the vocabularies, the authors present some rules to translate OWL to FOWL, as from the viewpoint of fuzzy set, some common OWL concepts are also special fuzzy concepts.

3.4 Approaches supporting uncertainty in ontology matching

Uncertainty becomes more crucial when matching ontologies. It is often the case that an entity defined in one ontology can only find partial matches to one or more entities in another ontology (Ding, 2005). Handling the uncertainty aspect began to emerge in a number of works in the last years. Many researchers focused on clarifying the main reasons leading to uncertainty. (Madhavan, Bernstein, Domingos, & Halevy, 2002) argue for the need to incorporate inaccurate correspondences and to handle uncertainty about them because in most of the cases there is no precise mapping. According to them, inaccuracy in mappings may come from the mapping language itself (*e.g.* relational data, XML, RDF, DAML+OIL), that is generally too limited to express more precise mappings, or from the concepts that do not match up precisely in the two ontologies. (Cross, 2003) underlined the fact that matching ontologies induces a degree of uncertainty. According to her, the use of syntactic or element-level matching to discover correspondences between names without the use of a thesaurus for checking synonyms and homonyms may lead to inaccuracies.

Most of the time and during the matching process, a combination of different matchers is required in order to discover correctly the semantic correspondences between entities (Ngo, Bellahsene, & Todorov, 2013). (Besana, 2006) thinks that the matchers to be combined have a partial view of the relations between entities or even may miss important

information. In addition to that, he believes that uncertainty can be due to the incompleteness of a thesaurus (it may not contain a term used in an ontology).

Considerable efforts have been devoted to matching ontologies under uncertainty. Mainly two mathematical models have been used: the probability theory ((Ding, 2005), (Mitra et al., 2005)) and the Dempster-Shafer theory ((Besana, 2006), (Nagy et al., 2007), (Wang et al., 2007)). These approaches will be described in the following.

3.4.1 Ontology matching through the probability theory

3.4.1.1 BayesOWL

BayesOWL, as presented in the previous section, is a probabilistic framework developed to model uncertainty in semantic web. Based on a set of rules, it translates an annotated OWL with probabilistic constructors into a Bayesian network. In (Pan et al., 2005), an ongoing research on matching ontologies is presented. The proposed methodology, based on BayesOWL, operates in four steps. First, probabilistic information (prior probability about concepts, conditional distribution for relation between concepts in the same ontology and joint probability distribution for semantic similarity between concepts in two ontologies O_1 and O_2) is learned using a naive Bayes text classification technique where each concept is represented by a set of sample documents retrieved automatically from the WWW. Second, the learned probabilistic information related to concepts and relations is represented as probabilistic constraints on their corresponding ontologies. Third, BayesOWL is used to translate O_1 and O_2 into Bayesian networks BN_1 and BN_2 respectively and conditional probability tables are constructed based on the learned probabilities. Finally, mapping ontologies relies on the computation of semantic similarity between two concepts C_1 in O_1 and C_2 in O_2 which is obtained using the joint probability distribution P(C_1, C_2). To determine this distribution, the authors propose to build for C_1 a classifier based on the statistical information in the exemplars into the model of O_1. Then, C_2 is classified with respect to C_1 by feeding its exemplars into the model of O_1. Similarity between C_1 and C_2 is quantified by a Jaccard coefficient computed from the joint probability distribution. Concept mapping is processed as some form of probabilistic evidential reasoning between BN_1 and BN_2. For this reason, three probability spaces are defined: S_{C_1} and S_{C_2} for BN_1 and BN_2 respectively and $S_{C_1C_2}$ for P(C_1, C_2). Mapping C_1 to C_2 amounts to determine the distribution of C_2 in S_{C_2}, given the distribution P(C_1) in S_{C_1} under the constraint P(C_1, C_2) in $S_{C_1C_2}$. To propagate probabilistic influences across spaces S_{C_1}, S_{C_2} and $S_{C_1C_2}$, Jeffrey's rule is used (Pearl, 1990).

3.4.1.2 OMEN

OMEN is a semi-automatic ontology matching tool based on a Bayesian network for enhancing existing ontology mappings (Mitra et al., 2005). The enhancement is performed through creating missed mappings and discarding existing false mappings. OMEN takes as input two ontologies O_1 and O_2 and initial probability distributions on the root nodes of the Bayesian network graph. The probability distribution is determined by the use of element-level techniques. Probability matching process occurs mainly in three steps. First, the construction of the Bayesian network graph goes through the creation of nodes. When the initial probability of a matching is above a given threshold, then a root node representing the match is created. Other nodes are created such that each node represents a mapping between pairs of classes and properties of the source ontologies. Only the nodes with a distance k of a root node are kept. Edges between nodes are created. They represent influences between the nodes in the Bayesian network graph. The quantitative construction of the Bayesian network consists in generating the conditional probability tables based on a set of meta-rules. These rules capture the influence of the structure of the input ontologies and semantics of ontology relations and match nodes that are neighbors of already matched nodes in the input ontologies. One of the most frequently rule used when creating mappings is that if two concepts C_1 from O_1 and C_2 from O_2 match and there is a matching relationship between r and r' such that r relates C_1 and C_2 and r' relates C'_1 and C'_2 then the probability to match C_2 and C'_2 increases. The authors use other kinds of rules that rely more heavily on the semantics of the ontology language. Finally, probabilistic inferences are made in order to generate a *posteriori* probabilities for each node. Only probabilities higher than a threshold are chosen to create the alignments

3.4.2 Ontology matching through the Dempster-Shafer theory

3.4.2.1 Paolo Besana's ontology matching framework

(Besana, 2006) proposed a framework based on Dempster-Shafer theory for matching ontologies. For that purpose, he suggested to combine the outcomes of different matchers in order to get better results. Besana depicted four issues to be considered when matching ontologies. Some of the issues justify the importance of dealing with uncertainty.

- Combining matchers: It is mandatory to **combine the outcomes of different matchers** because each one analyzes only some aspects of the relation that may exist between entities. For example, if the matching is based on comparing entities as strings then this comparison fails to consider the meaning of each term.

- Interpreting matchers' results: The results returned by matchers are of different types. In order to perform the combination, a uniform interpretation of different results must be handled. Besana proposed that **each matcher's result is interpreted as a measure** that denotes the plausibility of the relation between entities.

- Indistinguishable results: When using a matcher, **an entity may be matched to different entities**. These matchings may lead to the same numerical values. Based on the previous issue, these results will be interpreted as the pairs of entities that must have the same plausibility. When the obtained results are very close, then the same plausibility can be given. For that purpose, the author proposed to define intervals whose values correspond to the same plausibility.

- Ignorance and reliability: Besana accentuates the importance to **express ignorance and reliability of matchers**. Ignorance occurs when a matcher has no sufficient information to evaluate the degree of similarity between two entities. For example, when a matcher uses a thesaurus to search for similarity between two words and it happens that one of these words is not found in the thesaurus. This lack of information must be represented. In addition to ignorance, representing the different degrees of reliability of matchers is another important issue that must be addressed when modeling matching processes under uncertainty.

Based on the issues presented above, the author proposed a mathematical framework for handling uncertainty in ontology matching which consists in comparing each entity of the source ontology with all the entities of the target ontology. In order to match ontologies, the author used name-based techniques and structure-based techniques.

In order to model matching process under uncertainty, Besana rejects the closed-world assumption because he thinks that it is possible that an entity in a source ontology can have no corresponding entity in a target ontology and thus proposes to work under the open-world assumption. The different elements of his modeling are:

Belief Functions Representation

- The frame of discernment represents the Cartesian product $e \times O_{target}$ where e is an entity of the source ontology O_{source} and O_{target} represents all the entities of the target ontology . Each hypothesis of the frame is the couple $< e, e_i >$ such that e is an entity of the source ontology and e_i is an entity of the target ontology.

- An information represents each correspondence established by a matcher (*i.e.* the matching method used to detect the similarity between entities). A source of infor-

mation is the application of a matcher on an entity belonging to O_{source} and concerned by the correspondence.

- The mass function is deduced from the similarity measure obtained when applying a matching method. Depending on the matcher, the obtained result may be different from a classical value between $[0, 1]$. In that case, it must be converted into a bba as it is seen in the previous issues.
- Ignorance is represented through allocating a mass to the frame of discernment. Ignorance is due to an inability of a matcher to associate correctly a pair of entities.
- Reliability is represented through discounting the mass distributed by a matcher by a reliability factor. The discounted mass is allocated to the frame of discernment.

Belief Functions Combination

The combination of the mass distributions generated by the matchers is performed through the application of the Dempster's rule of combination where the open-world assumption holds.

Decision Making

It consists in choosing for each entity in a source ontology the most similar entity in the target ontology based on the combined results. For that purpose, the plausibility for each entity is calculated. The pairs of entities are ordered by plausibility and pairs with plausibility and belief below a given threshold are discarded.

3.4.2.2 DSSim: multi-agent approach for uncertain matching

DSSim is an agent-based ontology matching framework. It takes the Dempster-Shafer theory as its basis for matching large scale OWL ontologies. It is designed to be used in different domains such as question answering. (Nagy & Vargas-Vera, 2010) proposed to integrate it with the AQUA Question Answering System (Vargas-Vera, Motta, & Domingue, 2003) which answers user queries over heterogeneous data sources described by their own ontologies. The proposed system envisions to achieve *"machine intelligence"* on the semantic web through considering collective intelligence produced by combining agents' beliefs in order to match ontologies.

To match ontologies, *DSSim* operates as follows: Initially, ontologies are partitioned into fragments. Each concept or property taken from a first ontology O_1 is viewed as a query fragment that would normally be asked by a user in the AQUA system. Then, *WordNet*

is used in order to retrieve different hypernyms related to the concept or property. These hypernyms represent the possible query concept that can appear in the second ontology O_2. The matching consists in searching for correspondences between the query fragment and the ontology fragments of the second ontology based on the retrieved hypernyms. Syntactic similarity and semantic similarity are used to establish these correspondences.

We mentioned earlier that *DSSim* is an approach based on a multi-agent framework. There are different agents. Some manage users' queries and decompose them in fragments in order to send these fragments to mapping agents which are responsible of the matching process itself.

DSSim considers the uncertain aspect for matching ontologies because "each *agent* carries only partial knowledge of the domain and can observe it from its own perspective where available prior knowledge is generally uncertain". For matching ontologies, authors use syntactic-based techniques and semantic-based techniques. Each of these techniques evaluates the similarity between concepts and properties of two ontologies and draws up a similarity matrix. In the context of Dempster-Shafer theory, the authors consider each of the similarity measure used as an "expert" who gives his subjective evaluation on the matching through a similarity matrix. They define the frame of discernment as a set of all possible correspondences that have been detected by a particular expert. For a given expert, its similarity values represented by a matrix are converted into belief mass functions. We have to note here that the authors have not specified how the belief mass functions have been really constructed. They did not mention how the sum of 1 is obtained. They did not specify if there is a normalization or not. The converted similarity values are combined into a single belief function in order to create a mapping. The best mapping for a given concept is selected based on the highest belief.

3.4.2.3 Wang et al.'s approach

In (Wang et al., 2007; Wang, Liu, & Bell, 2009), the authors integrated uncertainty when matching ontologies using two different methods. The first one searches on simple correspondences (one-to-one) and the second one focuses rather on complex correspondences of the form (m:1 or 1:m or m:n).

3.4.2.3.1 Uncertain Simple Matching

(Wang et al., 2007) proposed to improve matching results by combining the outputs of three different matchers (two name-based matchers and one structure-based matcher). The aim of this method is to detect simple correspondences where each entity of the first ontology

O_1 is aligned to one entity of the second ontology O_2. In addition to that, the authors opted for dealing with uncertainty in matching ontologies the Dempster-Shafer theory and the possibility theory. They think that considering uncertainty is an important issue to be addressed because *"automatic ontology matching tools use heuristics or machine learning techniques which are imprecise by their very nature"*. We detail in this section how the Dempster-Shafer theory is used in detecting simple matching.

The authors used an edit-distance-based technique and a linguistic-based technique which calculate similarity between pairs of words. Due to the fact that names of ontology entities can be composed of several words, the authors suggested to preprocess these names with words splitting. The obtained words are then put into sets. Similarity computation is then performed as follows:

- For every word in a set, the similarity between this word and each word in the other set is calculated. The largest similarity value is retained to be attached to the word. This calculation is repeated until all the words have their own attached values.
- The sum of similarity values of all words in both sets is divided by the total number of all words. The obtained value reflects the final degree of similarity of names.

Suppose that we have to calculate the similarity between two entities *ConferenceMember* and *CommitteeMember*. Then, we have to create two sets: $set_1 = \{Conference, Member\}$ and $set_2 = \{Committee, Member\}$. Having these two sets, we calculate the similarity value between *Conference* in set_1 and *Committee* in set_2 and then between *Conference* in set_1 and *Member* in set_2. Once, we get these two values, we choose the largest one to be attached to the word *Conference* of set_1. The calculation is repeated until all the words have their own attached value. To get the degree of similarity between the entities *ConferenceMember* and *CommitteeMember*, the calculated similarity values obtained previously are summed and divided by four (cardinality of words in the two sets). This preprocessing step is used to apply the named-based techniques (edit-distance-based matcher and linguistic-based matcher). In addition to these techniques, the authors used the structure-based techniques. Based on the obtained results, the proposed approach focuses on combining the different results in order to improve the overall matching. This combination relies on the Dempster-Shafer theory. For this purpose, they specify the frame of discernment as a set of pairs of entities. For each entity e from the first ontology O_1, its mappings with all the entities in the second ontology O_2 are formed such that each pair is formed from an entity: $\Theta = e \times O_2$. For each pair of entities, we will have three normalized similarity values considered as mass functions. In order to get a unified mapping result for a pair of entities, the mass functions are combined using the Dempster's combination rule.

3.4.2.3.2 Uncertain Complex Matching

First, (Wang et al., 2009) propose a set-inclusion based approach for dealing with complex matching. Due to the fact that concepts are structured in a hierarchy, a concept is represented as a set containing the concept itself and all the concepts along the path between this concept and the root node. As a result, each entity is represented by a set of words. Similarity computation between entities consists in computing similarity between the sets of words, each one representing an entity. The computation is handled as described previously in the simple matching subsection. Then, a set S_1 is obtained. It contains all the mapping candidates pairs where each pair involves an entity from O_1 and an entity from O_2. After that, for each entity in O_1, the best mapping entity in O_2 is selected and the best mapping pair is added to a set S_2. The latter may contain multiple source entities mapped to the same target entity. To make a decision on the number of source entities from O_1 that should be matched to the same entity from O_2, an algorithm based on the Apriori is applied.

To deal with uncertainty in complex matching, (Wang et al., 2009) propose a clustering-based approach. First, they applied the average-linkage clustering algorithm to partition entities of O_1 into clusters and used for that purpose Lin's matcher (Lin, 1998) and a structure-based matcher. Having the clusters, the main objective is to choose the most appropriate one for each entity in O_2. The cluster with the largest similarity value is chosen. To calculate the similarity between an entity from O_2 and a cluster, four different matchers are used. The outputs of these matchers are combined using the Dempster-Shafer theory.

3.4.2.4 Comparison between the three approaches

The three systems described previously have some points in common. In fact they define mapping between OWL ontologies taking into account the uncertainty aspect which is modeled by the Dempster-Shafer theory. In addition to that they focus on mapping concepts and properties of two ontologies by using different kinds of ontology matching techniques.

Table 3.3 summarizes the major differences between the three systems based on the following criteria (*context, mapping, uncertainty theory, handling uncertainty, conflict management*).

Table 3.3: Major differences between the three systems

Criterion	**Besana**	**Nagy and al.**	**Wang and al.**
Context	general domain	particular domain	general domain
Mapping	simple matching	simple matching	simple, complex matching
Uncertainty theory	evidential theory	evidential theory	evidential theory, possibility theory
Relations	equivalence	equivalence, subsumption	equivalence
Handling uncertainty	uncertain mapping, combining matchers outputs	combining matchers outputs	combining matchers outputs
Conflict management	no	yes	no

3.5 Conclusion

Practically, it is impossible to guarantee the consistency of ontologies because semantic web is knowing a spectacular evolution for ensuring knowledge sharing and interoperability between applications. As the amount of shared information grows, the need to deal with uncertainty in semantic web becomes mandatory. In this chapter, we outlined the importance to deal with uncertainty in different ontology research tasks (ontology representation, ontology matching and ontology reasoning). After presenting in detail the mathematical formalism of Dempster-Shafer theory, we gave a survey on the different approaches which have been interested in resolving uncertainty problem applying for that purpose different mathematical models. We argue that considering uncertainty in ontology matching is one of the most important research area to tackle especially if one considers the fact that each matching technique focuses on a particular aspect of an entity. To the best of our knowledge, rare are the works that focus on uncertainty in ontology matching as it has been presented in subsection 3.4. For that purpose, we propose a new approach that differs from existing ones first on how the results of matching techniques are modeled under uncertainty and also on the way the best correspondences for a given entity are selected. A deep description of our proposal will be the focus of the following chapter.

Chapter 4

Credibilistic Decision Process for Ontology Matching

Contents

As it has been presented in chapter 3, dealing with uncertainty in ontology matching is an interesting research area to tackle. In this chapter, we describe our credibilistic decision process which focuses mainly in managing disagreement between similarity measures as well as making imprecise decision (a source entity may be matched to more than a target entity). We propose a decision rule based on a distance measure to make imprecise decision.

4.1 Introduction

In chapter 2, we introduced ontology matching as a solution to lessen the effect of semantic heterogeneity. Recently, dealing with uncertainty in ontology matching has been considered as an issue to be addressed especially that discovering alignments can be fed by a degree of uncertainty. Chapter 3 has been devoted to a deep presentation of this challenge by giving a survey of different matching approaches dealing with uncertainty.

Convinced that managing uncertainty in a matching process is an important task, we propose a credibilistic decision process for modeling the matching under the belief function theory on one hand and on the other hand for making decision on the alignments to be kept. Finding mappings whether simple or complex ones is dealt through the application of matching techniques which are mainly based on the use of similarity measures. Since no similarity measure applied individually is able to give a perfect alignment, the exploitation of the complementarity of different similarity measures may yield to a better alignment. Combining these similarity measures may also raise disagreement between the different results which should be modeled and resolved.

The approach that we suggest is based on three main steps:

- First, ontologies are matched by using three main techniques (a string-based matcher, a linguistic based matcher and a structure-based matcher).
- Second, the matching is modeled under the theory of belief functions and the different results of alignments are combined in order to manage the disagreement between the similarity measures.
- Finally, once we get the alignment (*i.e.* a set of correspondences), some ultimate questions can be asked: what are the correspondences that will be kept? which target entity to choose if a source entity has more than one corresponding entity? and in case a decision has to be made on a set of entities, how it is performed? To respond to these questions, we propose a decision rule based on a distance measure. The particularity of this rule is its ability to make a decision on a set of hypotheses. In our case, using this rule makes it possible to have more than a target entity for a source one.

In the sequel, section 4.2 presents our decision rule based on a distance measure. This rule is able to decide on a set of hypotheses rather than on a singleton. In this section, we demonstrate that our proposed rule is a particular case of that rule defined by equation 3.21. Experiments made on real databases are given in this section in order to evaluate

our proposed rule and to compare it with Appriou's rule (section 3.2.3). Section 4.3 is devoted to a deep description of our credibilistic decision process for matching ontologies under uncertainty. Finally, in section 4.4, we present results of experiments made on a set of ontologies.

4.2 Decision rule based on a distance measure

As mentioned in the subsection 3.2.3, depending on application needs, decision can be made either on a singleton or on a union of singletons. In this section, we present our rule which is able to make decision on a set of singletons.

4.2.1 Decision rule based on a distance principle

We propose a rule based on a distance measure. This rule has been the subject of two papers in (Essaid et al., 2014b) and in (Essaid et al., 2014a) where in the latter we gave experiments performed on a set of mass functions generated randomly as well as on real databases. This rule, inspired from the contradiction measure given in (Smarandache et al., 2011), is defined as:

$$A = \arg\min(d(m, m_A)) \tag{4.1}$$

A is the decision to take according to the information available. This decision is obtained through calculating the distance between a bba m and a categorical bba m_A. In this thesis, this distance is calculated between a combined bba m_{Comb} and a categorical one. m_A is used to adjust the degree of imprecision that has to be kept when deciding: by the use of categorical bba, we can specify the cardinality of focal elements to be considered. These elements can be a singleton or a union of two elements or three, etc. The minimum distance between m_{Comb} and m_A is kept and the decision corresponds to the categorical bba's element having the lowest distance with the combined bba.

Example 4.1. *Let us consider the frame of discernment that we worked with throughout this dissertation* $\Omega = \{LP, ShP, PS\}$. *Table 4.1 gives for each element of* 2^Ω *its corresponding categorical bba.*

Suppose now that we have two bbas:
bba$_1$*:* $m_1(PL) = 0.2$, $m_1(ShP) = 0.5$ *and* $m_1(PS) = 0.3$.
bba$_2$*:* $m_2(PS \cup ShP) = 0.1$, $m_2(PL \cup ShP) = 0.4$, $m_2(PL \cup PS) = 0.3$ *and* $m_2(PL \cup ShP \cup PS) = 0.2$.

Table 4.1: Categorical bbas construction.

Elements of 2^Ω	Corresponding categorical bba
LP	$m(\text{LP}) = 1$
ShP	$m(\text{ShP}) = 1$
PS	$m(\text{PS}) = 1$
LP $\cup$ ShP	$m(\text{LP} \cup \text{ShP}) = 1$
LP $\cup$ PS	$m(\text{LP} \cup \text{PS}) = 1$
PS $\cup$ ShP	$m(\text{PS} \cup \text{ShP}) = 1$
LP $\cup$ ShP $\cup$ PS	$m(\text{LP} \cup \text{ShP} \cup \text{PS}) = 1$

Due to imprecise context, Ω will not be considered. In other words, all the elements of 2^Ω can be selected except Ω.

Table 4.2 presents the resulting distances d_1 and d_2 where d_1 is the distance between bba_1 and a categorical bba and d_2 is the distance between bba_2 and a categorical bba. From this table, we remark that when a bba is constructed on singletons like bba_1 then the decision is made on a singleton (ShP) and when it is constructed on a union of singletons then an imprecise decision is rather kept ($LP \cup ShP$). In order to ensure an imprecise decision and to guarantee that this imprecision is always obtained, we fix the cardinality of elements of 2^Ω for which we construct their corresponding categorical bba. Depending on the cardinality of 2^Ω, we may choose to work on only some elements of 2^Ω. This filtering helps to limit the number of elements to be considered. In this thesis, we choose to work with categorical bbas whose cardinality is equal or below to 2.

Table 4.2: Distances between a combined bba and categorical bbas.

Elements of 2^Ω	Corresponding categorical bba	d_1	d_2
LP	$m(\text{LP}) = 1$	0.7	0.635
ShP	$m(\text{ShP}) = 1$	**0.436**	0.709
PS	$m(\text{PS}) = 1$	0.624	0.744
LP $\cup$ ShP	$m(\text{LP} \cup \text{ShP}) = 1$	0.583	**0.392**
LP $\cup$ PS	$m(\text{LP} \cup \text{PS}) = 1$	0.663	0.469
PS $\cup$ ShP	$m(\text{PS} \cup \text{ShP}) = 1$	0.539	0.594
LP $\cup$ ShP $\cup$ PS	$m(\text{LP} \cup \text{ShP} \cup \text{PS}) = 1$	0.597	0.294

The proposed rule, as detailed in algorithm 1, is in three steps. First, we fix the cardi-

nality of elements of 2^Ω for which we want to construct their corresponding categorical bba. Second, for each selected element, we construct its corresponding categorical bba. Finally we calculate the distance between the combined bba and each categorical bba. Jousselme distance (Jousselme et al., 2001) is used for that purpose. The most likely hypothesis to maintain is the hypothesis whose categorical bba is the nearest to the combined bba.

Jousselme distance is specific to the theory of belief functions because of the matrix D defined on 2^Ω. This distance has the advantage of taking into account the cardinality of the focal elements. This distance is defined for two bbas m_1 and m_2 as follows:

$$d(m_1, m_2) = \sqrt{\frac{1}{2}(m_1 - m_2)^t \underline{\underline{D}}(m_1 - m_2)} \tag{4.2}$$

where $\underline{\underline{D}}$ is a matrix based on Jaccard distance as a similarity measure between focal elements. This matrix is defined as:

$$D(A, B) = \begin{cases} 1 & \textit{if } A{=}B{=}\emptyset \\ \frac{|A \cap B|}{|A \cup B|} & \forall A, B \in 2^\Omega \end{cases} \tag{4.3}$$

Algorithm 1 Decision Rule

Require: n: cardinality of an element of 2^Ω, N: cardinality of 2^Ω, m_{Comb}: combined bba obtained after combining several matchers.
  i = 1
  while i $\leq N$ **do**
    if $|A| \leq n$ **then**
      Construct categorical bba m_A of *element*
      Compute distance between m_{Comb} and m_A
    end if
    i = i + 1
  end while
  return the element corresponding to the lowest distance.

Our rule is easy to use compared to Appriou's rule. To make decision, we have only to fix the maximum cardinality of the elements considered. If the decision must be made on a singleton, then we have to consider only elements with a cardinality equal to 1 and if the decision is based on union of elements, then the cardinality must be upper to 1. The use of Appriou's rule is complex and difficult because it depends mainly on three parameters: λ, the cardinality of elements of 2^Ω and the parameter r. This latter must be calibrated in order to decide on singletons or on unions.

Let us consider again the results of table 4.2. We recall that d_1 is the distance between a bayesian bba and a categorical one. We already mentioned that the decision to be taken

is the element with the minimum distance and it corresponds to *ShP*. If we want that the decision is rather a union of elements then we can force, for example, the cardinality of elements to 2 and in that case the decision corresponds to $PS \cup ShP$. In table 4.3, we show the results of the comparison between our proposed rule and Appriou's rule.

Table 4.3: Comparison between our proposed rule and Appriou's rule

Our decision rule	
Elements of 2^Ω	d_1
LP	0.7
ShP	0.436
PS	0.624
LP ∪ ShP	0.583
LP ∪ PS	0.663
PS ∪ ShP	0.539
LP ∪ ShP ∪ PS	0.597

Appriou's rule	
r	decision
$[0, 0.550]$	LP ∪ ShP ∪ PS
$[0.551, 0.678]$	PS ∪ ShP
$[0.679, 1]$	ShP

Considering these two tables, we remark that we obtain the same results as with the Appriou's rule whether when the decision is on singletons or on unions of elements. In fact, $PS \cup ShP$ is the solution when we decide on union of elements. This result is obtained when we fix the cardinality of elements to be considered to 2 whereas with Appriou's rule, we were bringing to vary the parameter r each time. In this example, for r $\in [0.551, 0.678]$ we obtain the same solution as given by our rule.

We recall that the distance d_2 is calculated between a bba based on union of elements and a categorical bba. The results of comparison between the two rules when we consider the distance d_2 is illustrated in table 4.4. Through this table, we note that we are able to give an uncertain result ($LP \cup ShP \cup PS$) which is not obtained when Appriou's rule is applied. Like the previous comparison tables, we obtain the same results as Appriou's rule. For example, if decision is based on union of elements, we get $LP \cup ShP$ as a decision which is obtained when $r \in [0, 0.15]$.

Based on these different comparisons, the application of our rule is easier than Appriou's rule because we have to only fix from the beginning the cardinality of elements to be considered for decision.

In the following, we give an example using two distinct decision rules based on different combination rules.

Table 4.4: Comparison between our proposed rule and Appriou's rule

Our decision rule	
Elements of 2^Ω	d_2
LP	0.635
ShP	0.709
PS	0.744
LP ∪ ShP	0.392
LP ∪ PS	0.469
PS ∪ ShP	0.594
LP ∪ ShP ∪ PS	0.294

Appriou's rule	
r	decision
$[0, 0.15]$	LP ∪ ShP
$[0.16, 1]$	LP

Example 4.2. *Let us continue with the example 3.2. We recall in table 4.5 the results of combining two bbas m_1 and m_2 by using the Dempster rule of combination $\oplus$, the disjunctive rule $\textcircled{\cup}$ and the mixed rule (DP).*

Table 4.5: Combination of two bbas through combination rules (excerpt of table 3.2).

	m_1	m_2	$m_{1\oplus 2}$	$m_{1\textcircled{\cup}2}$	m_{DP}
$\emptyset$	0	0	0	0	0
LP	0.03	0.4	0.2846	0.0120	0.2063
ShP	0.3	0.02	0.2662	0.0060	0.1930
$LP \cup ShP$	0.2	0.1	0.1380	0.2576	0.2206
PS	0.03	0.1	0.0872	0.0030	0.0632
$PS \cup LP$	0.04	0.01	0.0199	0.0360	0.0294
$PS \cup ShP$	0.2	0.07	0.1214	0.0917	0.1186
$PS \cup ShP \cup LP$	0.2	0.3	0.0828	0.5937	0.1689

Based on the results of table 4.5, it remains to apply our proposed rule to make decision. This is performed through constructing for each element of 2^Ω, except Ω its corresponding categorical bba and to compute:

- $Distance_{DS}$ is the distance between a categorical bba and a combined bba obtained after the application of Dempster's rule of combination.
- $Distance_{Disj}$ is the distance between a categorical bba and a combined bba obtained after the application of the disjunctive rule of combination.

- $Distance_{DP}$ is the distance between a categorical bba and a combined bba obtained after applying the mixed rule.

Table 4.6 shows the different results. The decision consists in selecting the element whose categorical bba has the minimum distance with the combined bba. Whatever the combination rule used, the decision corresponds, in our case, to the element $LP \cup ShP$.

Table 4.6: Results of our proposed decision rule.

Element	$Distance_{DS}$	$Distance_{Disj}$	$Distance_{DP}$
LP	0.5591	0.7276	0.5858
ShP	0.5293	0.7124	0.5584
$LP \cup ShP$	**0.4336**	**0.4256**	**0.3854**
PS	0.7199	0.7882	0.7330
$PS \cup LP$	0.5969	0.5748	0.5839
$PS \cup ShP$	0.5457	0.5442	0.5368

The comparison between our proposed rule with that defined by Appriou in equation (3.28) is presented in table 4.7. These results are obtained whe, the parameter r is equal to 0.5.

From this table, we notice that Appriou's rule gives a decision on a union of singletons when Dempster's rule of combination is used and a decision on a singleton when the disjunctive rule or the mixed one is used. These results are different from what we obtain when our proposed decision rule is applied. In fact, it promotes a decision union of singletons and thus whatever the combination rule used. The obtained results seems to be convenient especially that the disjunctive and the mixed rules help to get results on unions of singletons.

Table 4.7: Decision results comparison

	Appriou rule	Rule based on distance measure
Dempster rule	$LP \cup ShP$	$LP \cup ShP$
Disjunctive rule	ShP	$LP \cup ShP$
Mixed rule	ShP	$LP \cup ShP$

4.2.2 Decision rule based on distance analysis

At this stage, we presented our decision rule based on a distance measure and gave an example to illustrate its ability to give a decision on a set of hypotheses. In this subsection, we demonstrate that our proposed rule can be seen as a particular case of that proposed in equation (3.21).

We recall that our proposed rule is defined as:

$$A = \arg\min(d(m_{Comb}, m_A)) \tag{4.4}$$

This rule uses the Jousselme distance, defined in equation (4.2) and which can be rewritten as:

$$d(m_{Comb}, m_A) = \frac{1}{2}\sum_{Y\subseteq\Omega}\sum_{X\subseteq\Omega}\frac{|X\cap Y|}{|X\cup Y|}m_{comb}(X)m_A(Y) \tag{4.5}$$

Consequently, our proposed rule can be reformulated as:

$$A = \arg\min(\frac{1}{2}\sum_{Y\subseteq\Omega}\sum_{X\subseteq\Omega}\frac{|X\cap Y|}{|X\cup Y|}m_{comb}(X)m_A(Y)) \tag{4.6}$$

By analogy to what was presented by Denoeux in section 3.2.3, we propose to rewrite the decision to be taken as

$$A = \arg\min_{A\subseteq\Omega} R_d(A) \tag{4.7}$$

where

$$R_d(A) = \frac{1}{2}\sum_{X\subseteq\Omega}\frac{|X\cap Y|}{|X\cup Y|}m_{comb}(X)\text{when } m_A(Y) = 1 \tag{4.8}$$

Throughout this new reformulation, we remark that $\frac{|X\cap Y|}{|X\cup Y|}$ is no one either than the jaccard coefficient and it can be interpreted as an expected loss.

Our objective is to demonstrate that the equation (4.4) is equal to that defined in equation (3.21) for a value of λ.

The equation (3.21) is defined as:

$$R_{\text{betP}}(a_i) = \sum_{Y\in\Omega}\lambda(a_i|Y)\text{BetP}(Y) \tag{4.9}$$

The pignistic probability can be written as:

$$\text{BetP}(Y) = \sum_{X \in \Omega} \frac{|X \cap Y|}{|X|} \frac{m(X)}{1 - m(\emptyset)} \tag{4.10}$$

If we consider again the equation (4.9), then it can be formulated as:

$$\frac{1}{2} \sum_{Y \subseteq \Omega} \frac{|X \cap Y|}{|X \cup Y|} m_{comb}(X) = \sum_{Y \in \Omega} \sum_{X \in \Omega} \lambda(a_i|X) \frac{|X \cap Y|}{|X|} \frac{m_{comb}(X)}{1 - m_{comb}(\emptyset)} \tag{4.11}$$

Consequently, our rule is a particular case of that defined by Denoeux when:

$$\lambda(a_i|X) = \frac{1}{2} \sum_{Y \subseteq \Omega} \frac{|X|}{|X \cup Y|} (1 - m_{comb}(\emptyset)) \tag{4.12}$$

4.2.3 Experiments

We did some experiments on data sets in order to evaluate our proposed decision rule and to compare its results with those given by Appriou's rule presented in equation (3.28). These experiments are based on data sets selected from the U.C.I. machine learning repository (Bache & Lichman, 2013). Since the classification is a decision problem, we think it is possible to use the U.C.I data sets for testing our proposed decision rule.

Table 4.8 presents the data sets chosen for evaluation. For each data set, we give its number of instances, number of attributes as well as the number of classes.

Table 4.8: Description of data sets

Data set	#instances	#attributes	#classes
Iris	150	4	3
Seeds	210	7	3
Statlog (vehicule silhouettes)	946	18	4

For tests, two kinds of experiments are handled:

- First, we applied the belief k-NN algorithm (Denœux, 1995).

- Second, we modified the belief k-NN algorithm. We used the mixed rule for combination. Once the combined bba is obtained, decision can be made either by Appriou's rule or by our proposed decision rule. A comparison between results given by these two rules is made.

For evaluation, we construct for each data set a confusion matrix (Provost & Kohavi, 1998) which contains information about actual classes and predicted ones. For simplification, we note the classes by ω_i. For example, the classes of the data set Iris are noted as $(\omega_1, \omega_2, \omega_3)$.

Table 4.9: Confusion matrices for Iris

k-NN classifier			
	ω_1	ω_2	ω_3
ω_1	14	0	0
ω_2	0	10	0
ω_3	0	4	12

Appriou's rule based on modified k-NN						
	ω_1	ω_2	$\omega_1 \cup \omega_2$	ω_3	$\omega_1 \cup \omega_3$	$\omega_2 \cup \omega_3$
ω_1	14	0	0	0	0	0
ω_2	0	10	0	0	0	0
ω_3	0	4	0	12	0	0

Our decision rule based on modified k-NN						
	ω_1	ω_2	$\omega_1 \cup \omega_2$	ω_3	$\omega_1 \cup \omega_3$	$\omega_2 \cup \omega_3$
ω_1	14	0	0	0	0	0
ω_2	0	9	0	0	0	1
ω_3	0	0	0	8	0	8

Table 4.9 presents the results of classification for the data set Iris. For tests, we choose randomly 40 instances. The use of k-NN algorithm and the modified k-NN algorithm with Appriou's rule, give 90% as a rate of good classification. In fact, all instances originally belonging to either class ω_1 or ω_2 are well classified and among the 16 sets having ω_3 as their corresponding class, only 4 are not well classified. Although, Appriou's rule provides a good classification, it does not give a result on a set of singletons which is different from the results that we obtain with the modified k-NN algorithm where our proposed rule is applied. A first look at the results makes us think that the classification is not good comparing it to the two previous classifiers. In fact, among the 16 sets originally belonging to ω_3 there are 8 sets that have $\omega_2 \cup \omega_3$ as a corresponding class. We have to recall that we aim to make decision on a union of singletons. For that purpose, we consider that obtaining 8 sets with $\omega_2 \cup \omega_3$ as a corresponding class is a good result because the set of singletons $\omega_2 \cup \omega_3$ contains ω_3 which is the original belonging class. This interpretation makes the rate of good classification equal to 100%.

Table 4.10: Confusion matrices for Seeds

k-NN classifier			
	ω_1	ω_2	ω_3
ω_1	16	0	2
ω_2	2	14	0
ω_3	1	0	15

Appriou's rule with r $\in [0.0248, 1[$						
	ω_1	ω_2	$\omega_1 \cup \omega_2$	ω_3	$\omega_1 \cup \omega_3$	$\omega_2 \cup \omega_3$
ω_1	16	1	0	2	0	0
ω_2	1	15	0	0	0	0
ω_3	3	0	0	12	0	0

Our decision rule						
	ω_1	ω_2	$\omega_1 \cup \omega_2$	ω_3	$\omega_1 \cup \omega_3$	$\omega_2 \cup \omega_3$
ω_1	14	0	0	1	3	0
ω_2	0	12	4	0	0	0
ω_3	0	0	0	12	4	0

In table 4.10, we present the results of classification of the data set Seeds. 50 sets are chosen for tests. Both the belief k-NN algorithm and the modified one based on the use of Appriou's rule give the same results where only two sets are not well classified as ω_1 and among the sets originally belonging to ω_2, only 2 are misclassified and finally among the 16 sets having ω_3 as their actual class only one set has ω_1 as a predicted class. All these results give a rate of good classification equal to 90%. As in the confusion matrix for Iris, the rule proposed by Appriou does not make decision on a set of singletons as it is supposed to do. If we consider now the obtained results where our proposed rule is used, then we notice that on one hand we have results on union of classes and on the other hand, we improved the rate of good classification which becomes equal to 98% where only one set originally belonging to class ω_1 is misclassified as ω_3.

Table 4.11 illustrates the obtained confusion matrices for the data set Statlog. Due to the fact that Appriou's rule and our proposed decision rule give results on 2^Ω, we will be limited in presenting the results of only predicted classes where we have a value different to 0. 146 sets are chosen for tests. Both the belief k-NN algorithm and the modified one based on Appriou's rule give the same results and a rate of a good classification equal to 67,12%. As with the previous two data sets, the rule proposed by Appriou did not give a classification on a union of singletons contrary to what we obtained with our proposed decision rule where we can have a union of predicted classes containing the actual ones. By using our rule, the classification is ameliorated to 93.15%.

Based on the different experiments that we handled, our proposed decision rule is able to make decision on a union of singletons. This rule will be used in our credibilistic decision process for matching ontologies as a way to designate for each source entity its

Table 4.11: Confusion matrices for Statlog

k-NN classifier				
	ω_1	ω_2	ω_3	ω_4
ω_1	36	1	2	0
ω_2	2	17	3	12
ω_3	1	4	29	2
ω_4	3	14	4	16

Appriou's rule based on modified k-NN							
	ω_1	ω_2	$\omega_1 \cup \omega_2$	ω_3	...	ω_4	...
ω_1	36	1	0	2	0	0	0
ω_2	2	17	0	3	0	12	0
ω_3	1	4	0	29	0	2	0
ω_4	3	14	0	4	0	16	0

Our decision rule based on modified k-NN				
	ω_1	ω_2	ω_3	ω_4
ω_1	20	0	0	1
ω_2	0	1	0	0
$\omega_1 \cup \omega_2$	5	1	0	3
ω_3	0	0	16	0
$\omega_1 \cup \omega_3$	5	0	2	0
$\omega_2 \cup \omega_3$	0	1	8	0
$\omega_1 \cup \omega_2 \cup \omega_3$	1	1	0	2
ω_4	0	0	0	1
$\omega_1 \cup \omega_4$	5	1	0	0
$\omega_2 \cup \omega_4$	0	20	1	18
$\omega_1 \cup \omega_2 \cup \omega_4$	2	3	0	5
$\omega_3 \cup \omega_4$	0	1	2	0
$\omega_1 \cup \omega_3 \cup \omega_4$	0	0	2	1
$\omega_2 \cup \omega_3 \cup \omega_4$	1	5	5	6

target entities. The next section is devoted to a deep description of this process through detailing its different steps.

4.3 Credibilistic decision process

We present in this section our credibilistic decision process for matching ontologies. This process has been the subject of two papers (Essaid, Ben Yaghlane, & Martin, 2011; Essaid, Martin, Smits, & Ben Yaghlane, 2013). This process is based on the use of the theory of belief functions as a tool to model the matching under uncertainty and to make decision on which target entities to match with a source entity.

4.3.1 Process description

The process, as illustrated in figure 4.1, requires as an input two ontologies to match as well as a list of matchers. We choose to use only three matchers each belonging to a specific type of matching techniques. The process produces as an output a set of imprecise correspondences. Depending on the selected matchers, some matching techniques require external resources (*e.g.* dictionaries, thesauries, etc.) in order to detect correspondences.

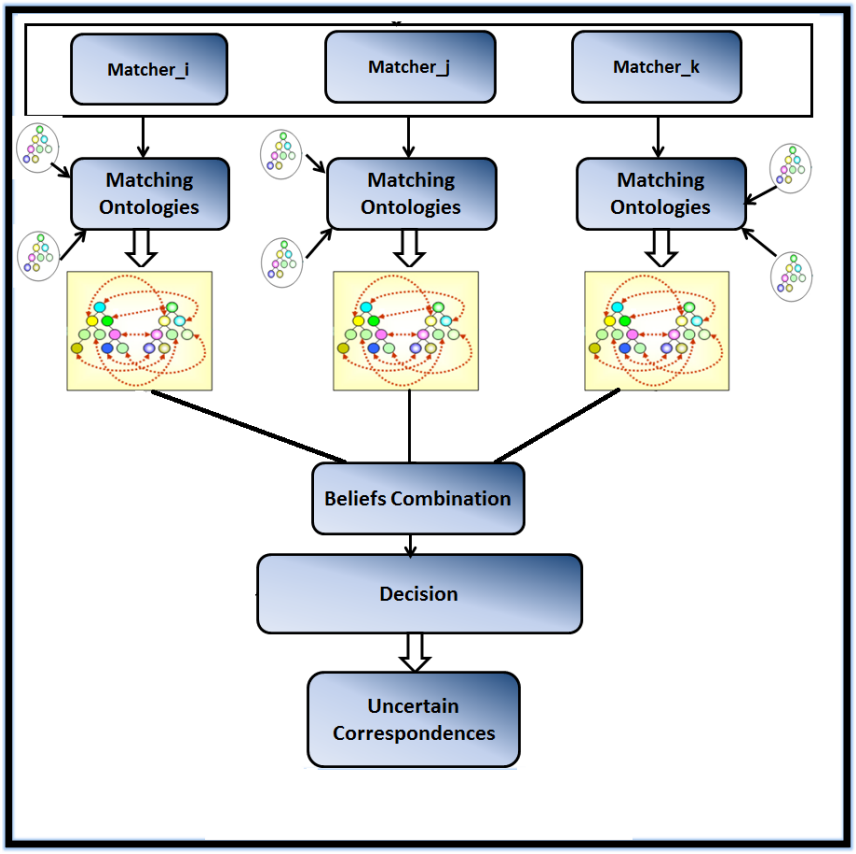

Figure 4.1: Credibilistic decision process

Our credibilistic decision process involves mainly in three steps:

- *Selecting matchers* consists in selecting a matcher among the existing ones. In section 2.3.4, we mentioned that there are a great number of matchers. The selection of a specific matcher must be held carefully where it is recommended to use a method allowing to decide which matcher to use.

- *Beliefs Combination* is performed once we obtain for each matcher its corresponding alignments. We propose to model the matching under the theory of belief functions and to combine beliefs of the different matchers.
- *Making decision* is based on the use of our proposed decision rule (section 4.2). In fact, once we obtain for each entity source in a source ontology the possible corresponding entities in a target ontology, we are called to apply our rule in order to designate for each source entity its corresponding target entities. The result of this step is a set of belief alignments.

4.3.2 Matcher selection

To match ontologies, one has to find a matcher among a panoply of matching techniques (Euzenat & Shvaiko, 2013a). Each of these matchers concerns a specific feature of entities. Depending on the characteristics of ontologies and the application requirements, many studies have been proposed to guide a developer on selecting a suitable matcher to use (Euzenat et al., 2006; Huzza et al., 2006; Mochol, 2009). In this thesis, we select a matcher based on its quality evaluation results. For evaluation, three metrics can be calculated: Precision *(Prec)*, Recall *(Rec)* and F-measure *(Fm)*. These metrics originate from information retrieval domain and consist in comparing the expected results with the obtained ones. (Do et al., 2002) proposed to adapt these metrics in ontology matching field through comparing alignments determined by a system to evaluate and a reference alignment. Let us denote the alignment returned by a matcher as A and the reference alignment by R. (Do et al., 2002) define these metrics as follows:

- *Precision* is the proportion of correctly shared correspondences over the total number of found correspondences.

$$Prec = \frac{|A \cap R|}{|A|} \tag{4.13}$$

- *Recall* represents the proportion of correctly shared correspondences over the total number of referenced correspondences.

$$Rec = \frac{|A \cap R|}{|R|} \tag{4.14}$$

- *F-measure* represents the harmonic mean of precision and recall and is determined as:

$$Fm = 2 * \frac{(Prec * Rec)}{(Prec + Rec)} \tag{4.15}$$

To evaluate the matchers' performance, we did experiments on ontologies provided by the Ontology Alignment Evaluation Initiative campaign (*OAEI*)[1] (Euzenat, Meilicke, Shvaiko, Stuckenschmidt, & Dos Santos, 2011). *OAEI* is a coordinated international initiative which aims to evaluate ontology matching systems. Its goal is to give researchers the possibility to compare their own matching algorithms with other ones and to select the best matching strategies. This campaign is handled every year and provides benchmarks and data sets for evaluating matching algorithms. In this thesis, we use the *Conference track* [2] which contains 16 ontologies related to conference organization. In this track, there exists only 21 reference alignments corresponding to the complete alignment space between 7 ontologies (cmt, $Conference$, $ConfOf$, $Edas$, $Ekaw$, $Iasted$, $SigKdd$). Table 4.12 gives some characteristics of these ontologies used for evaluation.

Table 4.12: Conference Track

Name	*Number of classes*	*Number of Datatype Properties*	*Number of Object Properties*
cmt	36	10	49
$Conference$	60	18	46
$ConfOf$	38	23	13
$Edas$	104	20	30
$Ekaw$	74	0	33
$Iasted$	140	3	38
$SigKdd$	49	11	17

The main string-based matchers used for evaluation are: *Hamming*, *Jaro*, *Levenshtein*, *Needleman-Wunsch*, *Ngram*, *Monge-Elkan*, *Smith-WaterMan*, *Soundex*. The evaluation is handled as follow: For each couple of ontologies, we apply a matching technique (among those of the same category) providing then a similarity value for a pair of entities. A comparison between the obtained alignment and a reference alignment is done through calculating evaluation metrics. In our case the harmonic mean of the precision values are computed over all the ontologies of the Conference track. The technique with the best evaluation results is kept as our matcher for our credibilistic process. To improve the matching results and to keep only the significant correspondences, a filtering may be used. Different evaluation results are obtained depending on the threshold used.

[1] http://oaei.ontologymatching.org/2013/

[2] http://oaei.ontologymatching.org/2013/conference/index.html

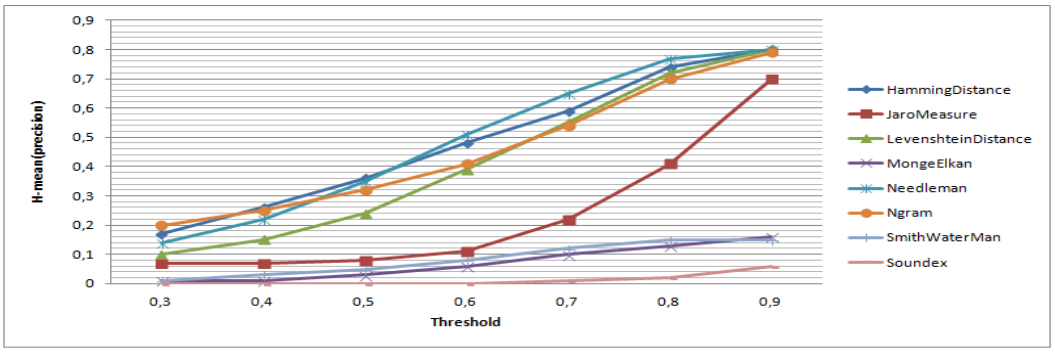

Figure 4.2: Evaluation of some string-based matchers

Figure 4.2 represents the comparison between some string-based matchers. Given the high number of existing methods, we chose to compare the precision values of 8 methods each focusing on a particular feature of a string as described in chapter 2. Figure 4.2 shows that all the string-based methods improve the precision when the threshold value increases but not with the same degree. For example, the *soundex* gives results equal to 0 when the threshold value is between 0.3 and 0.6 and even when it increases, the precision is less than 0.1. This is explained by the fact that the considered ontologies are heterogeneous where a same concept can be labeled differently. The edit distance methods (*Needleman-Wunsch distance*, *Hamming distance* and *Levenshtein distance*) improve the results of precision when the threshold value increases because these methods are based on the number of edit operations to get a string from a first one where a minimum operations are handled to transform a string into another one. Based on the results given in figure 4.2, the *Needleman-Wunsch distance* is chosen as our string-based method for our decision process.

We will be limited in this thesis in evaluating only some string-based methods. Convinced that using only these methods is not sufficient to get good results when matching ontologies, we suggest to use the *Wu-Palmer similarity* and the *GlossOverlap* which use *WordNet*[3] for searching similarities between concepts of two ontologies.

4.3.3 Modeling matching under the belief function theory

To highlight the modeling matching, we will use in this section an excerpt of two ontologies *cmt* and *Conference* as shown in figure 4.3. Throughout this section, we will note these two ontologies as O_1 and O_2 respectively.

[3]http://wordnet.princeton.edu/

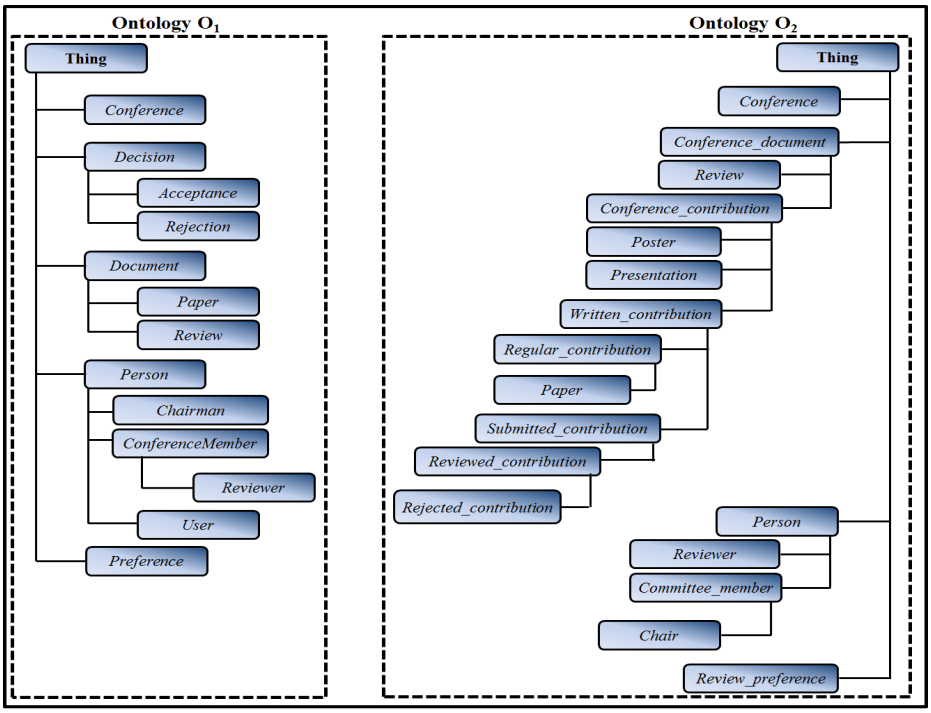

Figure 4.3: Excerpt of two ontologies *Cmt* and *Conference*

We select the ontology O_1 as a reference ontology. For each entity in the reference ontology, we search its corresponding entity(ies) in the target ontology O_2. The results of applying our selected matching methods (*Needleman-Wunsch*, *Wu-Palmer similarity* and *Gloss Overlap*) are presented in table 4.13. For example, the entity *Chairman* of the ontology O_1 is aligned to *Chair* when one of the matching methods is applied.

The results presented in table 4.13 are obtained with a filter threshold equal to 0.3. In other words, we keep only he alignments with a similarity measure's value equal or up to 0.3. This threshold will allow that an entity can have different correspondences. In fact, considering a threshold (> 0.8) will, on one hand, keep only the identical correspondences whatever the matching technique applied. In other words, for each source entity we will have always the same target entity. On the other hand, considering this threshold will allow us to show the interest of our proposed decision rule (in getting imprecise correspondences).

Based on the results obtained in table 4.13, we depict two kinds of disagreement:

- The first one concerns the obtained target entities. For example, if we consider the entity *Rejection* which, depending on the similarity measure used, is aligned to *Organization*, *Presentation* and *Rejected_Contribution*. There is no consensus on

a particular entity to be matched to the entity *Rejection*.

- The second one is noticeable in the obtained similarity values. For example, for the *Chairman* entity, the similarity measures agree in aligning *Chairman* to the entity *Chair* but the application of the technique *Gloss Overlap* assigns a value of *0.45* which is less than the value assigned by the two others similarity measures.

Given the disagreement in the obtained alignments, we propose to manage it through modeling the matching results under the theory of belief functions. For that purpose, we have to define our frame of discernment, how the bbas are constructed and how the combination is handled.

1. **The frame of discernment** is a set of all possible hypotheses susceptible to represent a solution for a given problem. Then the frame contains all the target entities identified in the alignments. In our example, the frame of discernment is:

   $\Omega = \{\, Organization, Person, Chair, Presentation, Rejected\_Contribution,$
   $Reviewer, Conference, Poster, Conference\_fees, Program\_Committee, \dots \}$

2. **Source of information**: In order to construct the bbas, one has to identify the source of information. We define an information given by a source, every correspondence established by one of the matching techniques. For example, matching the two entities *Decision* and *Organization* with a degree of *0.667* by the *Wu-Palmer similarity* is an information.

3. **Basic belief assignments (bbas)**: Once we obtain all the correspondences, we keep only those where an entity source $e_1 \in O_1$ has a correspondence when applying the three techniques. For each selected correspondence, we construct its corresponding mass function. Entities are matched when they present a degree of similarity according to a matching technique. The more they are similar, the more the distance between them is small and thus they can be matched. We start from the assumption that an entity e_1 is near to an entity e_2 if they are similar and thus there is a chance that they can be matched. Under the probability theory, this distance can be interpreted as true if the two entities are matched to each other and false otherwise. Under the theory of belief functions, this distance can be interpreted as a degree of belief of a similarity measure. This degree of belief is related to the fact of matching e_1 to e_2 and reflects if the two entities are far from each other or near to each other. Based on this assumption, we consider, for example, the value *0.667* is none other than the degree of belief of *Wu-Palmer similarity* in considering *Decision* and *Organization* as an alignment. In addition to that, the obtained similarity values are in $[0, 1]$, hence we do not have to convert these values but rather interpret them as

masses and include them in the construction of bbas. To construct a bba, the sum of mass functions must be equal to 1. For that purpose, a mass is allocated to the total ignorance.

Let us consider the entity *Decision* for which we want to construct its mass function. If we note that $S^{e_1}_{wupalm}$ as the source of information based on the use of *Wu-Palmer similarity*, then its related mass function is:

$m^{e_1}_{S_{wupalm}}$ (Organization) = 0.667 and $m^{e_1}_{S_{wupalm}}(\Omega)$ = 1 - 0.667 = 0.333.

Tables 4.14 and 4.15 present for each source entity its corresponding mass functions.

Table 4.13: Results of matching O_1 and O_2.

$e_1 \in O_1$	*Matching techniques*	$e_2 \in O_2$	*similarity value*
Decision	*Wu-Palmer similarity*	Organization	0.667
	Gloss Overlap	Person	0.4
	Needleman-Wunsch	Organization	0.4
Chairman	*Wu-Palmer similarity*	Chair	1
	Gloss Overlap	Chair	0.45
	Needleman-Wunsch	Chair	1
Rejection	*Wu-Palmer similarity*	Organization	0.7962
	Gloss Overlap	Presentation	0.4667
	Needleman-Wunsch	Rejected_Contribution	0.4
Acceptance	*Wu-Palmer similarity*	Organization	0.7619
	Gloss Overlap	Conference	0.3
	Needleman-Wunsch	Presentation	0.5217
Person	*Wu-Palmer similarity*	Person	1
	Gloss Overlap	Person	1
	Needleman-Wunsch	Person	1
User	*Wu-Palmer similarity*	Person	0.8571
	Gloss Overlap	Poster	0.3750
	Needleman-Wunsch	Paper	0.444
Reviewer	*Wu-Palmer similarity*	Reviewer	1
	Gloss Overlap	Reviewer	1
	Needleman-Wunsch	Reviewer	1
Conference	*Wu-Palmer similarity*	Conference	1
	Gloss Overlap	Conference	1
	Needleman-Wunsch	Conference	1
Paper	*Wu-Palmer similarity*	Paper	1
	Gloss Overlap	Paper	1
	Needleman-Wunsch	Paper	1
Review	*Wu-Palmer similarity*	Review	1
	Gloss Overlap	Review	1
	Needleman-Wunsch	Review	1
Preference	*Wu-Palmer similarity*	Organization	0.7
	Gloss Overlap	Conference	0.7
	Needleman-Wunsch	Review_Preference	0.7407

Table 4.14: Construction of mass functions.

$e_1 \in O_1$	*Matching Techniques*	*masses functions*
Decision	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$(Organization) = 0.667, $m^{e_1}_{S_{wupalm}}(\Omega)$ =0.333
	Gloss Overlap	$m^{e_1}_{S_{glover}}$(Person) = 0.4, $m^{e_1}_{S_{glover}}(\Omega)$ =0.6
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Organization) = 0.4, $m^{e_1}_{S_{nwunsch}}(\Omega)$ =0.6
Chairman	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$ (Chair) =1
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Chair) = 0.45, $m^{e_1}_{S_{glover}}(\Omega)$ =0.55
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Chair) = 1
Rejection	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$(Organization) = 0.796,$m^{e_1}_{S_{wupalm}}(\Omega)$ =0.204
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Presentation) = 0.4667, $m^{e_1}_{S_{glover}}(\Omega)$ =0.533
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$(*Rejected_Contribution*) = 0.4, $m^{e_1}_{S_{nwunsch}}(\Omega)$ =0.6
Acceptance	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$(Organization) = 0.7619,$m^{e_1}_{S_{wupalm}}(\Omega)$ =0.2381
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Conference) = 0.3, $m^{e_1}_{S_{glover}}(\Omega)$ =0.7
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Presentation) = 0.5217,$m^{e_1}_{S_{nwunsch}}(\Omega)$ =0.4783
Person	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$ (Person) = 1
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Person) = 1
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Person) = 1

Table 4.15: Construction of mass functions (cont'd).

$e_1 \in O_1$	*Matching Techniques*	*masses functions*
User	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$(Person) = 0.8571,$m^{e_1}_{S_{wupalm}}(\Omega)$ =0.1429
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Poster) = 0.3750,$m^{e_1}_{S_{glover}}(\Omega)$ =0.625
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Paper) = 0.444,$m^{e_1}_{S_{nwunsch}}(\Omega)$ =0.556
Reviewer	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$ (Reviewer) = 1
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Reviewer) = 1
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Reviewer) = 1
Conference	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$(Conference) = 1
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Conference) = 1
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Conference) = 1
Paper	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$ (Paper) = 1
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Paper) = 1
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Paper) = 1
Review	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$(Review) = 1
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Review) = 1
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Review) = 1
Preference	*Wu-Palmer similarity*	$m^{e_1}_{S_{wupalm}}$(Organization) = 0.7 ,$m^{e_1}_{S_{wupalm}}(\Omega)$ =0.3
	Gloss Overlap	$m^{e_1}_{S_{glover}}$ (Conference) = 0.7,$m^{e_1}_{S_{glover}}(\Omega)$ =0.3
	Needleman-Wunsch	$m^{e_1}_{S_{nwunsch}}$ (Review_Preference) = 0.7407,$m^{e_1}_{S_{nwunsch}}(\Omega)$ =0.2593

4. **Combination**: Through this modeling, we aim to manage conflict occurring between similarity measures. For that reason, it is essential to combine the different bbas. We give the results of using Dempster's rule of combination, the conjunctive rule, the disjunctive rule and the mixed rule in tables (4.16, 4.17, 4.18) respectively.

Table 4.16: Managing conflict with Dempster's rule of combination.

$e_1 \in O_1$	Combined bba	
Decision	$m(Organization)$	0.7059
	$m(Person)$	0.1176
	$m(\Omega)$	0.1765
Chairman	$m(Chair)$	1
Rejection	$m(Rejected\_Contribution)$	0.1135
	$m(Presentation)$	0.1489
	$m(Organization)$	0.5674
	$m(\Omega)$	0.1702
Acceptance	$m(Organization)$	0.5595
	$m(Conference)$	0.0749
	$m(Presentation)$	0.1907
	$m(\Omega)$	0.1748
Person	$m(Person)$	1
User	$m(Person)$	0.7143
	$m(Paper)$	0.0952
	$m(Poster)$	0.0714
	$m(\Omega)$	0.1190
Reviewer	$m(Reviewer)$	1
Conference	$m(Conference)$	1
Paper	$m(Paper)$	1
Review	$m(Review)$	1
Person	$m(Person)$	1
Preference	$m(Review\_Preference)$	0.3352
	$m(Conference)$	0.2737
	$m(Organization)$	0.2737
	$m(\Omega)$	0.1173

Table 4.17: Managing conflict with conjunctive rule of combination.

$e_1 \in O_1$	Combined bba	
Decision	$m(\emptyset)$	0.32
	$m(Organization)$	0.48
	$m(Person)$	0.08
	$m(\Omega)$	0.12
Chairman	$m(Chair)$	1
Rejection	$m(\emptyset)$	0.5662
	$m(Rejected\_Contribution)$	0.0492
	$m(Presentation)$	0.0646
	$m(Organization)$	0.2462
	$m(\Omega)$	0.0738
Acceptance	$m(\emptyset)$	0.5441
	$m(Organization)$	0.2551
	$m(Conference)$	0.0342
	$m(Presentation)$	0.087
	$m(\Omega)$	0.0797
Person	$m(Person)$	1
User	$m(\emptyset)$	0.5833
	$m(Person)$	0.2976
	$m(Paper)$	0.0397
	$m(Poster)$	0.0298
	$m(\Omega)$	0.0496
Reviewer	$m(Reviewer)$	1
Conference	$m(Conference)$	1
Paper	$m(Paper)$	1
Review	$m(Review)$	1
Person	$m(Person)$	1
Preference	$m(\emptyset)$	0.8011
	$m(Review\_Preference)$	0.0667
	$m(Conference)$	0.0544
	$m(Organization)$	0.0544
	$m(\Omega)$	0.0233

Table 4.18: Managing conflict with disjunctive rule and mixed rule.

$e_1 \in O_1$	Combined bba	
Decision	$m(Organization \cup Person)$	0.1067
	$m(\Omega)$	0.8933
Chairman	$m(Chair)$	0.4545
	$m(\Omega)$	0.5455
Rejection	$m(Organization \cup Presentation \cup Rejected\_Contribution)$	0.1436
	$m(\Omega)$	0.8564
Acceptance	$m(Organization \cup Conference \cup Presentation)$	0.1193
	$m(\Omega)$	0.8807
Person	m(Person)	1
User	$m(Person \cup Poster \cup Paper)$	0.1429
	$m(\Omega)$	0.8571
Reviewer	$m(Reviewer)$	1
Conference	$m(Conference)$	1
Paper	$m(Paper)$	1
Review	$m(Review)$	1
Preference	$m(Organization \cup Conference \cup Review\_Preference)$	0.3630
	$m(\Omega)$	0.6370

4.3.4 Making decision

Obtaining for each entity source its corresponding combined mass function is an input for a decision making process. Deciding which target entity(es) to match with the source entity is an important step in ontology matching process. In this section, we will give the results of making decision when one of the decision rules is applied namely the pignistic probability, Appriou's rule and our proposed rule.

Table 4.19 describes the results of decision when the pignistic probability is used. The obtained results are the same whatever the applied combination rule. We notice that the pignistic probability promotes simple matching where each source entity is aligned to a unique target entity. For example, this rule considers that the entity *Rejection* should be aligned to the entity *Organization* rather than matching it with *Presentation* or *Rejected_Contribution*.

Table 4.19: Making decision with pignistic probability

$e_1 \in O_1$	$e_2 \in O_2$
Decision	Organization
Chairman	Chair
Rejection	Organization
Acceptance	Organization
Person	Person
User	Person
Reviewer	Reviewer
Conference	Conference
Paper	Paper
Decision	Organization
Preference	Review_Preference

The application of our proposed rule gives the results as presented in table 4.20. The results are obtained when the mixed rule is used. We choose to work with elements of 2^{Ω} such that their cardinality is equal to 2.

Table 4.20: Making decision with our proposed rule

$e_1 \in O_1$	$e_2 \in O_2$
Decision	Organization∪ Person
Chairman	Chair
Rejection	Organization∪Rejected_Contribution
Acceptance	Organization∪Presentation
Person	Person
User	Person∪Paper
Reviewer	Reviewer
Conference	Conference
Paper	Paper
Decision	Organization
Preference	Review_Preference∪Organization

4.4 Results

In the previous section, we presented the different steps of our credibilistic decision process and we gave a detailed example of managing disagreement between two ontologies *cmt* and *Conference*. In this section, we present results of experiments handled on the *Conference track* (see section 4.3.2). This set of experiments concerns a comparison between the precision and recall of two types of alignments: belief alignment and certain alignment. We call a belief alignment, an imprecise alignment obtained once the constructed bbas are combined and decision is made based on our proposed decision rule. A certain alignment is an alignment obtained when a matching technique is applied without taking into account the disagreement. The obtained imprecise results, as it has been shown in table 4.20, will be rendered as alignment. To express the fact that the source entity *Preference* has as target entities *Review_Preference* or *Organization*, we propose the following:

```
<Alignment>
 <map>
 <Cell cid='1'>
 <entity1 rdf:resource='http://cmt#Preference'/>
 <entity2 rdf:resource='http://conference#Organization'/>
 <measure rdf:datatype='xsd:float'>0.0</measure>
 <relation>=</relation>
 </Cell>
 </map>
 <map>
 <Cell cid='2'>
 <entity1 rdf:resource='http://cmt#Preference'/>
 <entity2 rdf:resource='http://conference#Review_Preference'/>
 <measure rdf:datatype='xsd:float'>0.0</measure>
 <relation>=</relation>
 </Cell>
 </map>
 </Alignment>
```

Although this example shows two possible correspondences for the entity *Preference*, the number of found correspondences will be recorded as one correspondence including the two found correspondences. An adequate format is proposed in Chapter 6. At this stage, we will simply represent alignments as in the example. The number of shared correspondences is needed to calculate the *recall* measure. When we find that at least, there is one shared correspondence, in that case this correspondence (although it contains two correspondences) is seen as correct. In the following, we adopt these notations for designating ontologies (1: Conference, 2: ConfOf, 3: Ekaw, 4: Edas, 5: Iasted, 6: Sigkdd). In all sets of comparisons, we consider a source ontology O_1 and X as a target ontology where X can be (1: Conference, 2: ConfOf, 3: Ekaw, 4: Edas, 5: Iasted, 6: Sigkdd). In the different sets, we show results of comparing between our alignment based on a decision rule and those obtained when one of these matchers (*Wu-Palmer similarity*, *Gloss OverLap* and *NeedlemanWunsch*) is used.

In these different tests, we fix a threshold of 0.3. Abscissa axis represents the number of the target ontology and the ordinate axis represents the precision and recall values. In the different sets where the precision measure is used for comparison, we remark that the obtained results when our rule is used are good especially in the set Conference - X. In

this latter, we obtained the best results in comparison with other methods.

- cmt - X: In this set of comparison, we consider the *cmt* as the source ontology and X as a target ontology. The comparison between the different alignments in term of precision is illustrated in figure 4.4. Figure 4.5 gives the results of comparing the recall of the different methods. We notice that applying *Wu-Palmer similarity* and *NeedlemanWunsch* for aligning *cmt* and *Conference* provides a precision equal to 0 whereas we obtain a precision equal to 0.36 when our method is used. The best obtained result of our method in term of precision is when we align *cmt* to *Ekaw* although the precision value is lower than that obtained when *Gloss Overlap* is applied. If we consider the recall results. Most of the time *NeedlemanWunsch* gives the best result except when we align *cmt* with *Conference*. We remark that in three cases our method has the same result as the *Gloss Overlap* and *NeedlemanWunsch*.
- Conference - X: In this set of comparison, we consider the *Conference* as the source ontology and X as a target ontology. Figure 4.6 illustrates the obtained precision results whereas figure 4.7 gives the recall results. In term of precision, we notice that our method gives the best result when *Conference* is aligned to the ontologies *ConfOf*, *Edas* and *Iasted*. The recall results are not good enough in addition to that most of the time our method gives the same recall value to that obtained when *Wu-Palmer similarity* is applied except when *Conference* and *Sigkdd* are aligned.

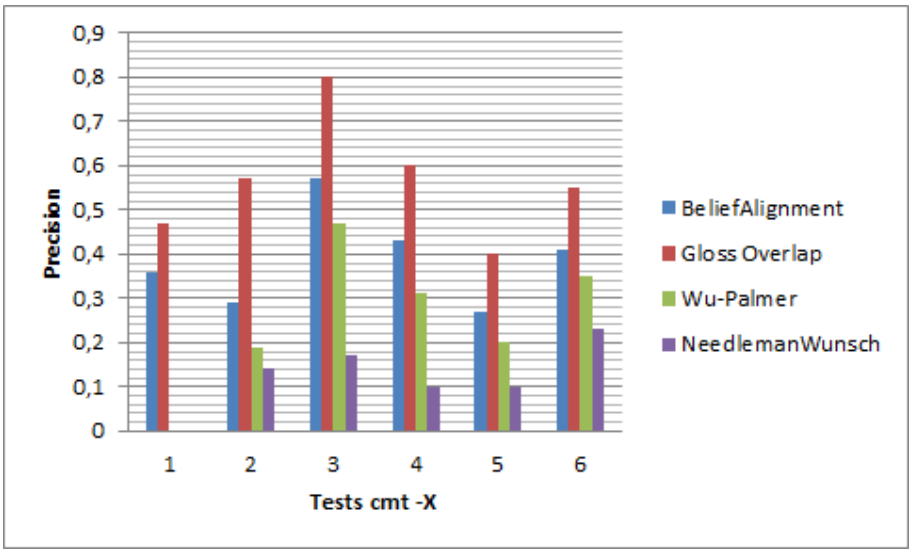

Figure 4.4: Precision results between *cmt* and *X*

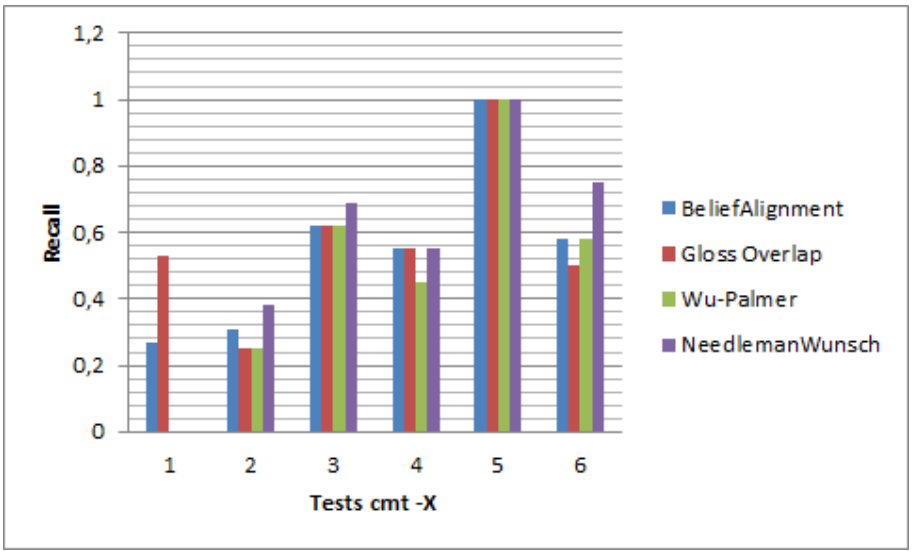

Figure 4.5: Recall results between *cmt* and *X*

- ConfOf - X: In this set of comparison, we consider the *ConfOf* as the source ontology and *X* as a target ontology. Figure 4.8 and figure 4.9 illustrate the precision and

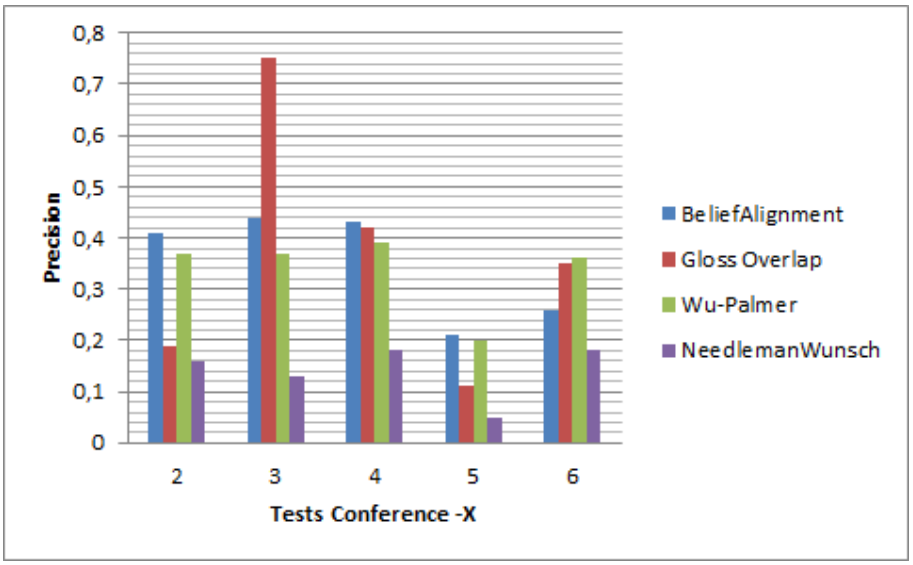

Figure 4.6: Precision results between *Conference* and *X*

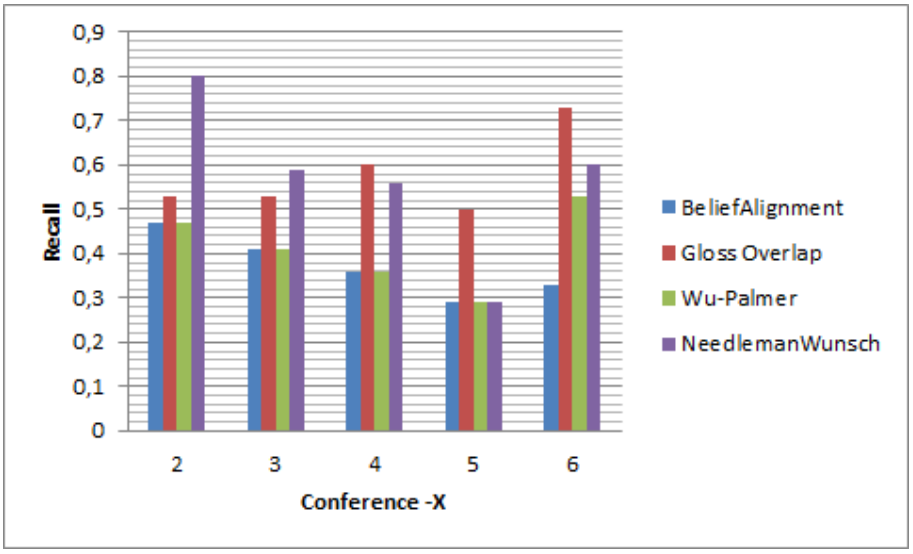

Figure 4.7: Recall results between *Conference* and *X*

recall results respectively. When the comparison is based on precision results, *Basic Gloss Overlap* gives the best results. In terms of recall, our method gives good results especially when *ConfOf* is aligned to *Ekaw* and *Iasted*. Aligning *ConfOf* to *Sigkdd* gives the same recall value of 0.57 when one of this method is applied (our method, *Gloss Overlap* and *Wu Palmer similarity*).

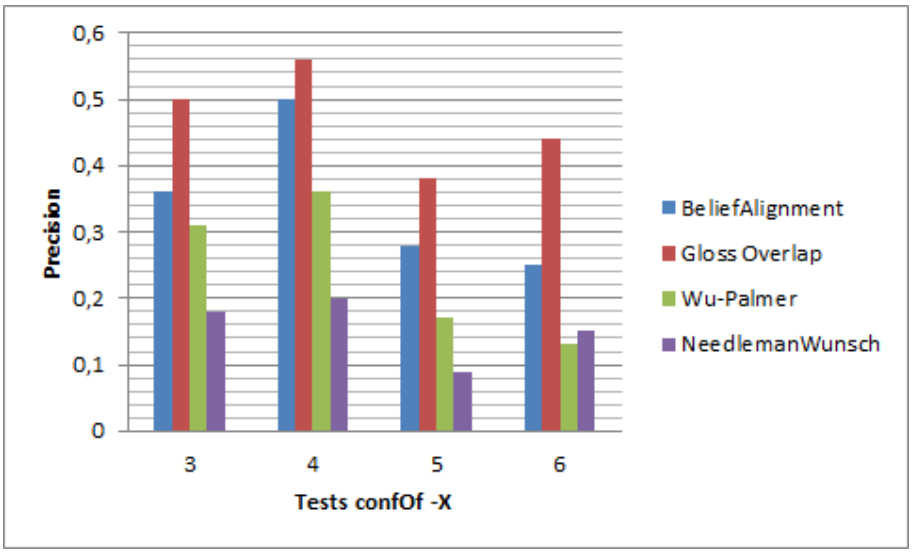

Figure 4.8: Precision results between *ConfOf* and *X*

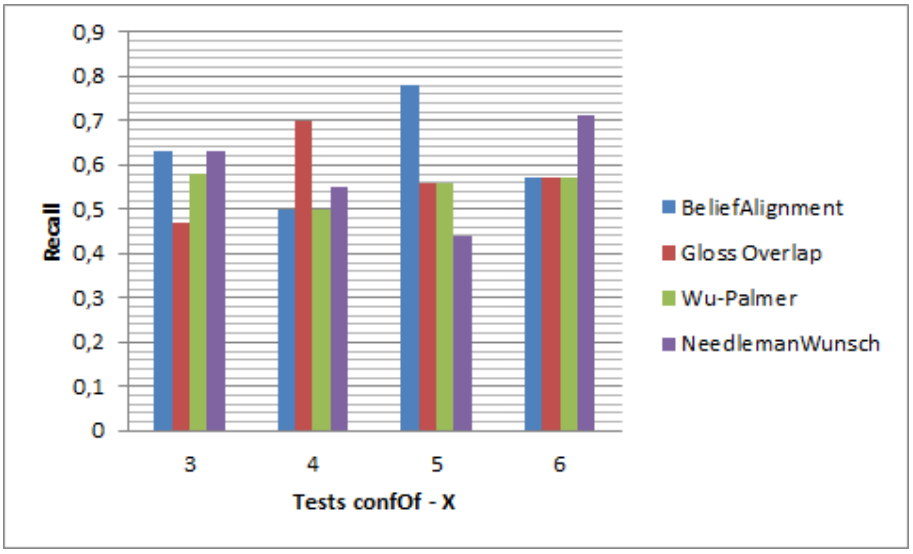

Figure 4.9: Recall results between *ConfOf* and *X*

- Edas - X: In this set of comparison, we consider the *Edas* as the source ontology and *X* as a target ontology. The comparison between the different alignments in term of precision is illustrated in figure 4.10. Figure 4.11 gives the results of comparing the recall of the different methods. We notice that *Basic Gloss Overlap* has the best precision whereas *NeedleMan Wunsch* gives the best recall by comparison to other methods.

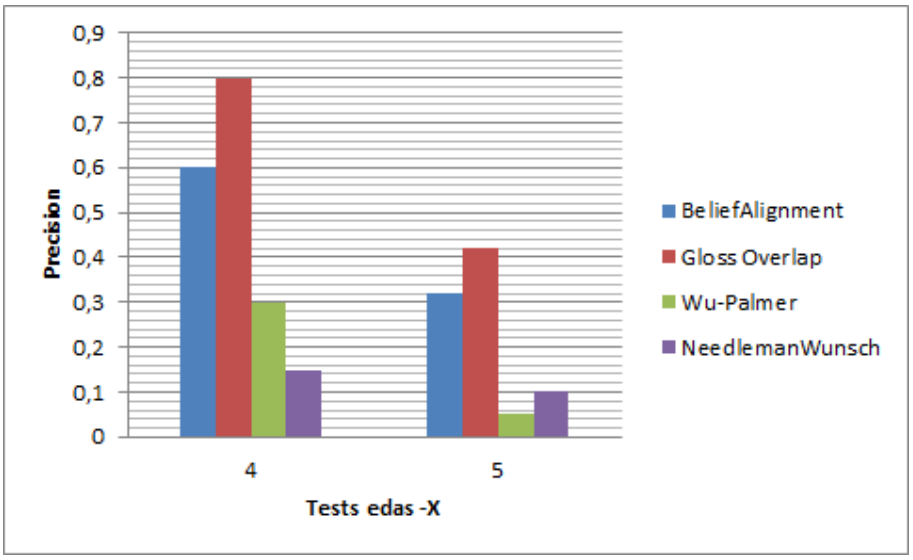

Figure 4.10: Precision results between *Edas* and *X*

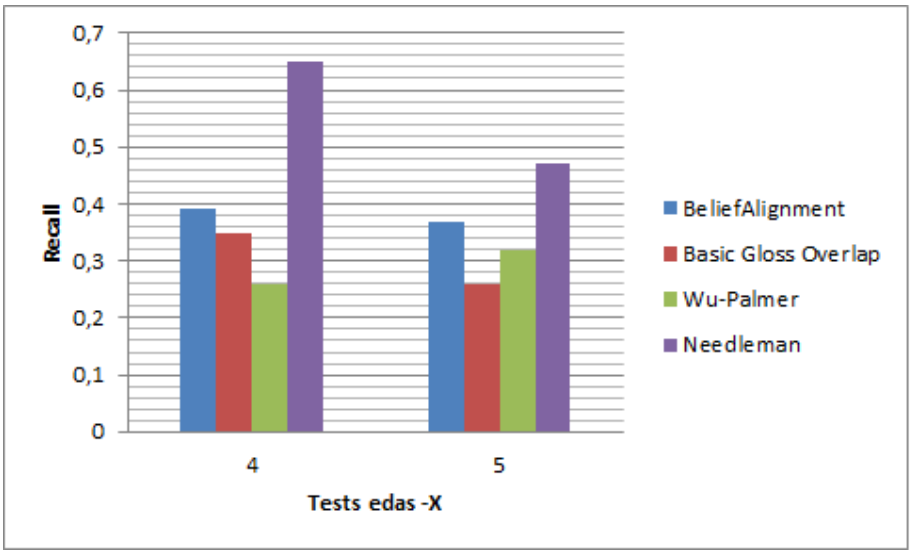

Figure 4.11: Recall results between *Edas* and *X*

- Iasted - X: In this set of comparison, we consider the *Iasted* as the source ontology and *X* as a target ontology which corresponds to the ontology Sigkdd. Figure 4.12 illustrates the obtained precision results whereas figure 4.13 gives the recall results. In term of precision, we notice that our method gives the best result. The recall results are not bad compared to other methods.

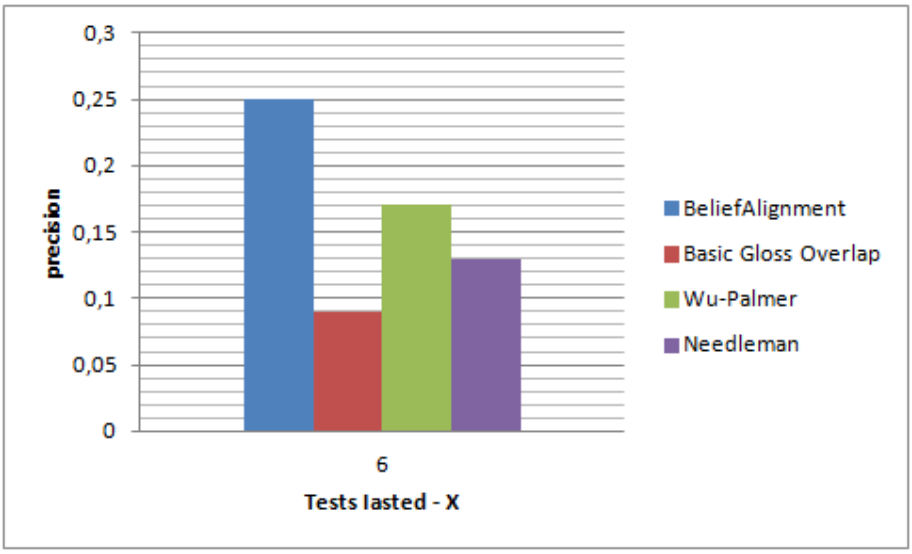

Figure 4.12: Precision results between *Iasted* and X

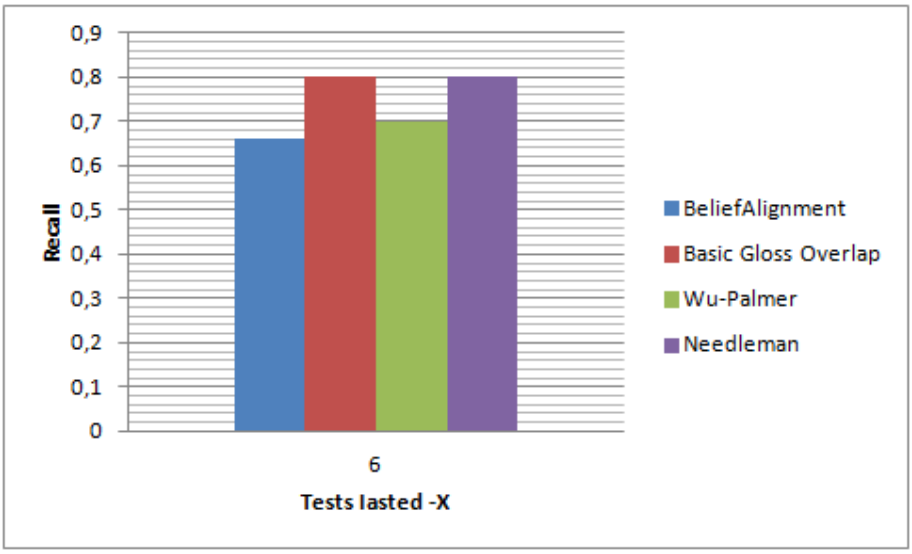

Figure 4.13: Recall results between *Iasted* and X

We remark that in some cases, our method gives the best results in terms of precision and recall. First, we have to notice that our method is used to manage the uncertainty aspect and is obtained by combining the outcomes of three different matchers (Needleman-Wunsch, Gloss Overlap and Wu Palmer similarity). Based on the results illustrated in the different figures, one can consider that an uncertain result is correct if the reference alignment belongs to the suggested ones, then our method improves the precision and recall. As

conference ontologies present concepts related to the domain of conference organization, then these ontologies present synonyms which can be easily detected by the Gloss Overlap matcher. This explained why this matcher gives the best results. We have to precise also that for calculating the recall, we compare our belief alignment with the reference one. Or this latter is a certain alignment and the results obtained when our method is used are of the form of union of two concepts which will affect the recall results. We notice also that in figure 4.9, we obtained good results because the obtained ones are rather certain. In fact most of the source entities are aligned to a unique target entity.

4.5 Conclusion

Dealing with uncertainty in ontology matching is an important task. For that purpose, we propose a credibilistic decision process for managing disagreement in ontology matching. This process operates in three steps. First, matchers are selected. Then, ontology matching is performed. Based on the obtained alignments, we detect conflict that we manage by using the theory of belief functions. Finally, the most important step consists in finding for each source entity its corresponding target entities. In order to obtain an imprecise results, we proposed a decision rule based on a distance measure. Through experiments that we handled, we notice that applying this rule leads to better results by comparison to that rule proposed by Appriou. In the next chapter, we conclude this dissertation through listing the main improvements that can be made as well as the main future work on how extending the proposed process.

Chapter 5

Conclusion and perspectives

Contents

The objective of this chapter is to summarize our contributions and to give the main future works that can be handled. We recall the different steps of our credibilistic decision process for matching ontologies. Deciding on which target entity to align with a source entity is handled by a decision rule that we proposed. This rule is based on a distance measure and is able to give results on unions of entities. Nevertheless, the proposed approach is still subject of improvements and extensions. In this chapter, we give the main improvements that we can make. In addition to that, we present how the obtained alignments can be extended. First, we propose an alignment format able to represent imprecise alignments. Second, we propose ontology merging as a solution for using the obtained alignments.

5.1 Synthesis

The semantic web, as an open and dynamic system, is based on the use of heterogeneous ontologies. This heterogeneity may be due to a difference in conceptualizing a domain of interest or in the use of different representation languages for modeling knowledge in ontologies. To assure an interaction between applications using different ontologies, the effect of heterogeneity must be minimized through matching ontologies. If we suppose that an entity may have more than a target entity, then we can consider that this illustrates an imprecision state where an entity has no precise correspondences. For example, if we consider the entity $ConferenceMember$ from a source ontology where it can be matched to $Conference$ or $Conference\_fees$ from a target ontology then we can suppose that this situation describes an imprecision where it is interesting to model it. In addition to that, matching ontologies is based on the use of similarity measures where a disagreement may occur between these measures.

This dissertation focuses on managing disagreement between similarity measures and choosing for each source entity its target entity based on the imprecise results through using a decision rule based on a distance measure. For that purpose, we used the theory of belief functions for modeling the matching process under uncertainty and for combining the outcomes of different similarity measures. Combining information leads to a conflict which can be managed under the theory of belief functions. Aware that using a unique similarity measure does not help to obtain a result able to take into account all the features of the entities of the two ontologies, we suggest to use three different matching techniques. Giving the results of alignments, a disagreement is detected either in the obtained similarity values assigned to a given couple of entities or in aligning an entity to different target entities. To model the conflict under the theory of belief functions, we suggest a credibilistic decision process which is based on a correspondence between matching components and the theory of belief functions elements. In fact, the frame of discernment is represented as the set of all target entities identified in the alignments. Every correspondence established by a similarity measure is defined as an information given by a source. Under the theory of belief functions, each source gives its mass function. Our mass functions are constructed on singletons. To guarantee a sum of mass functions equal to 1, a mass is allocated to the total ignorance. Based on the different constructed masses functions, decision is made in order to select for each source entity its corresponding target entities.

Due to the importance of making decision in any process, we proposed a decision rule based on a distance measure which is able to decide on union of elements. This rule calculates the distance between a combined mass function and a categorical one. The choice to work with categorical mass functions allows to adjust the degree of imprecision

that has to be kept when making decision. In this thesis, we proposed to decide on union of two entities. Once the distance is calculated, the minimum one is kept and the decision corresponds to the categorical mass function's element that has the lowest distance with the combined mass function. Then, we were interested in demonstrating that our rule can be seen as a particular case of that rule proposed in (Denœux, 1997). The proposed rule is used with different combination rules and is tested with datasets. The obtained results are satisfactory compared to those obtained with the rule proposed by (Appriou, 2005).

5.2 Perspectives

Like any work, our proposed approach needs some improvements and can be subject of extensions especially that the ontology matching area is a rich field and a dynamic one. In this section, we present the improvements that we can make as well as in proposing two main future works that can extend our credibilistic decision process.

5.2.1 Credibilistic decision process improvements

In this thesis, we presented our credibilistic decision process as an approach for managing disagreement in alignments. To make decision, we proposed a rule based on a distance measure able to align a source entity with a union of target entities. We think that there exist some improvements that can be done. In the following, we list some of them.

- In this thesis, we use only three matching techniques (a terminological matcher, a linguistic-based matcher and a structure-based matcher). We think that it seems to be interesting to test the matching process with more than three matchers either belonging to a same category of matchers or to different ones.
- In this work, we used the *Conference track* to evaluate our approach. The ontologies of this track are simple and not with huge number of entities. It will be interesting to test the proposed approach with other ontologies where the number of entities can reach 1000. This test will allow us to measure the performance of our algorithm and then to improve it if it is necessary.
- Our proposed rule calculates the distance between a combined bba and a categorical one. What about using another kind of bbas? For example we can allocate a mass value of λ to a given mass function. If we consider the frame of discernment related to the two ontologies *cmt* and *Conference* in section 4.3.3, then we can construct our

mass functions as $m(Organization \cup Person) = \lambda$ and $m(\Omega) = 1 - \lambda$. Then, we can test the approach with different values of λ and select λ for which we can obtain a decision on union of entities.

- Matching ontologies consists in finding for each entity its target entities. In the evaluation that we made, we obtained only a unique target entity for a given source entity. But, we can be faced with a particular case where a similarity measure aligns a source entity to more than a target entity. This case should be taken into account especially in modeling the matching under the theory of belief functions.

- In this dissertation, we focused on managing disagreement once we got the alignments. In other words, we managed the disagreement after matching ontologies. Another way to manage the conflict consists in dealing with it during or before the matching process itself. In that case, the frame of discernment is the set of all entities of the target ontologies.

5.2.2 Credibilistic decision process extensions

In the sequel, we sketch some possible ways on how the obtained results by our credibilistic decision process can be used and extended for future works. In fact, the obtained imprecise results can be represented as belief alignments or even can be used for constructing an uncertain ontology. A deep description of these two extensions is given in the following.

5.2.2.1 Alignments representation

Alignments have their own life cycle (Euzenat, Mocan, & Scharffe, 2008) as illustrated in figure 5.1. First, they are *created* by a matching process. Then, they roll by an iterative phase of *evaluation* and *enhancement* where modifications can be made on the resulting alignments through discarding, for example, correspondences with a calculated similarity value above a threshold. This iterative phase continues to occur until we get the desired alignments. In that case, they are stored and *communicated* to other parties interested in such an alignment. At last, these alignments can be *exploited* by applications for other purposes such as ontology merging, data translation, etc.

In order to allow a syntactic expression of these alignments as well as an efficient manipulation over applications, a set of representation formats have been suggested (Maedche, Motik, Silva, & Volz, 2002; Bouquet, Giunchiglia, van Harmelen, Serafini, & Stuckenschmidt, 2004; Horrocks et al., 2004; Euzenat, 2004). The format presented in (Euzenat, 2004) expresses alignments through metadata. We may cite:

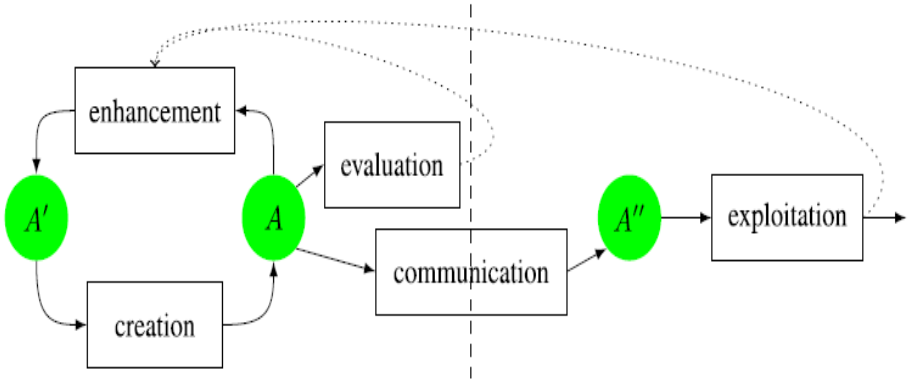

Figure 5.1: Alignment life cycle.

- *references* correspond to URIs of the two ontologies to match.
- *set of correspondences* describes the relation holding between entities of the source ontology and entity of the target ontology.
- *level* corresponds to the level of alignment. It can take the values 0, 1 and 2. For example the level 0 is used for characterizing simple correspondences between named entities while level 2 is used for more complex relations the kind of correspondence.
- *arity* denotes the type of correspondence.
- *entity1* corresponds to the first matched entity.
- *entity2* corresponds to the second matched entity.
- *relation* expresses the relation holding between entities (equivalence, subsumption, etc.).
- *strength* denotes the confidence measure provided by a matching technique.
- *id* is the identifier of a correspondence.

The alignment format offers several alignment levels which correspond to different possibilities for expressing entities.

- *Level0* does not depend on a specific ontology language. In this level, aligned entities are identified by URIs and can be classes, properties or individuals. In the following, we provide an excerpt of an alignment between two ontologies.

- *Level1* is independent of an ontology language and the correspondence concerns pairs of sets of entities and not pairs of entities like in level0.
- *Level2* depends on the language used to express entities and correspondences are described in a more complex way (formulas, queries ...)

We give in the following an excerpt of using the *Wu-Palmer similarity* to match the two ontologies *cmt* and *confOf* referenced respectively as $http://oaei.ontologymatching.org/2013/conference/data/cmt.owl$ and $http://oaei.ontologymatching.org/2013/conference/data/confOf.owl$. One of the resulting alignment of an equivalence relation holding between the two entities *email* and *location* has a strength equal to 0.625 and is represented in the given example.

```
<Alignment>
 <xml>yes</xml>
 <level>0</level>
 <type>??</type>
 <onto1>
 <Ontology rdf:about="http://cmt">
 <location>http://nb.vse.cz/ svabo/oaei2010/cmt.owl</location>
 </Ontology>
 </onto1>
 <onto2>
 <Ontology rdf:about="http://conference">
 <location>http://nb.vse.cz/ svabo/oaei2010/confOf.owl</location>
 </Ontology>
 </onto2>
 <map>
 <Cell cid='1'>
 <entity1 rdf:resource='http://cmt#email'/>
 <entity2 rdf:resource='http://confOf#location'/>
 <measure rdf:datatype='xsd:float'>0.625</measure>
 <relation>=</relation>
 </Cell>
 </map>
 </Alignment>
```

This alignment representation format is adequate to represent certain alignments. As it

has been mentioned in chapter 3, few are the works that dealt with uncertainty in ontology matching. These works are based on the use of the theory of belief functions to combine the outcomes of different similarity measures. In these works, the pignistic probability is used as a decision rule to specify for each source entity its corresponding target entity. Hence, the obtained results can be rendered by respecting the format given in the previous example. The use of our decision rule based on a distance measure does not allow us to render our results in the format previously described because for each source entity we obtain an imprecise correspondence (*i.e.* a union of target entities). This situation must be rendered in an adequate format able to express this imprecision. For that purpose, we suggest as an extension a format able to represent belief alignments.

Let us consider again the example of matching the entity *email*. Once the different bbas are combined, we apply our decision rule which proposes to align *email* with ($location \cup hasEmail$). The obtained distance between these entities is 0.8316. In this thesis, we represented this alignment as follow:

```
<Alignment>
 <map>
 <Cell cid='1'>
 <entity1 rdf:resource='http://cmt#email'/>
 <entity2 rdf:resource='http://confOf#location'/>
 <measure rdf:datatype='xsd:float'>0.0</measure>
 <relation>=</relation>
 </Cell>
 </map>
 <map>
 <Cell cid='2'>
 <entity1 rdf:resource='http://cmt#email'/>
 <entity2 rdf:resource='http://confOf#hasEmail'/>
 <measure rdf:datatype='xsd:float'>0.0</measure>
 <relation>=</relation>
 </Cell>
 </map>
 </Alignment>
```

At this stage, we consider that representing target entities in two separate cells will reflect the union notion but it would be better that this union will be represented in an

adequate format. We suggest to add metadata for expressing union of entities. The measure value corresponds to the obtained distance. For example, the alignment representation related to the entity *email* can be similar to the following:

```
<Alignment>
 <map>
 <Cell cid='1'>
 <entity1 rdf:resource='http://cmt#email'/>
 <unionEntity1 rdf:resource='http://confOf#location'/>
 <unionEntity2 rdf:resource='http://confOf#hasEmail'/>
 <measure rdf:datatype='xsd:float'>0.8316</measure>
 <relation>=</relation>
 </Cell>
 </map>
 </Alignment>
```

We remark through this excerpt that *unionEntity1* represents the first target entity and the *unionEntity2* represents the second target entity that can be aligned to the entity *email*.

To evaluate a matching algorithm, it is important to compare the obtained alignment with a reference one. Considering an alignment represented as we suggested will not allow us to make the evaluation. For that purpose, the evaluation algorithm must be ameliorated in order to make the comparison between a belief alignment and a reference one possible. Since the evaluation is based on the calculation of metrics such as *precision* and *recall*, then an obtained correspondence is considered as correct if it contains at least one of the correspondence given in the reference alignment. For example, if we consider the reference alignment between the two ontologies *cmt* and *confOf*, we find that *email* should be aligned to *hasEmail*. If we consider the represented alignment in the previous excerpt then *hasEmail* is one of the union entity related to *email*. In that case, we consider this correspondence as a correct one even if *email* as another union entity (*i.e. location*) and the couple (email, location) is not a correct correspondence in the reference alignment.

5.2.2.2 Ontology merging as a use of alignments

Finding correspondences between two ontologies can be seen as an input for other processes such as ontology merging which is a first natural use of ontology matching (Euzenat &

Shvaiko, 2013a). Based on the two matched ontologies O_1 and O_2 as well as on the resulting alignments, axioms can be generated helping then in constructing a single coherent ontology O_3. The process of merging ontologies is illustrated in figure 5.2. The construction of the

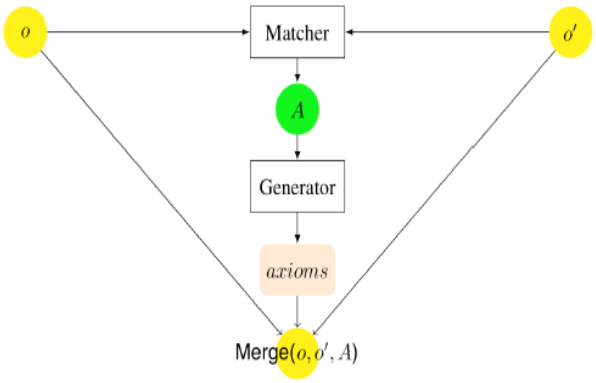

Figure 5.2: Ontology merging process.
(Euzenat & Shvaiko, 2013a)

merged ontology is based on the use of axioms (generated from alignments). We have to note that entities for which we have no correspondences will be included in the merged ontology.

Let us take the same example of the previous subsection where *email* can be aligned to *hasEmail* or *location*. Suppose that we are working on a certain context where *email* is matched to *hasEmail*. Then a possible generated axiom helping for constructing a merged ontology is:

```
<owl:Class rdf:about="http://cmt#email">
<owl:equivalentClass rdf:resource="http://confOf#hasEmail"/>
</owl:Class>
```

Merging ontologies under uncertainty will allow to model that *email* can be matched to either *hasEmail* or *location*. For example, we can propose to add a constructor union

```
<owl:Class rdf:about="http://cmt#email">
<owl:equivalentClass>
<owl:unionOf>  <owl:class rdf:resource="http://confOf#hasEmail"/>
<owl:class rdf:resource="http://confOf#location"/>
< owl:unionOf"/>
</owl:Class>
```

Many approaches have been suggested to match ontologies but few are those that dealt with uncertainty under the theory of belief functions. In this thesis, we proposed a credibilistic decision process able to match ontologies under uncertainty. To identify correspondences, we proposed a decision rule based on a distance measure. This rule is able to give an imprecise result. In this concluding chapter, we presented the main improvements that can be made and we described some possible extensions on how to use the obtained alignments.

References

Appriou, A. (2005). Approche générique de la gestion de l'incertain dans les processus de fusion multisenseur. *Traitement du signal*, *22*, 307 - 319.

Baader, F., Horrocks, I., & Sattler, U. (2005). Description logics as ontology languages for the semantic web. In D. Hutter & W. Stephan (Eds.), *Mechanizing mathematical reasoning* (Vol. 2605, p. 228-248).

Bache, K., & Lichman, M. (2013). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284(5)*, 34-43.

Besana, P. (2006). A framework for combining ontology and schema matchers with dempster-shafer. In *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006) collocated with the 5th International Semantic web Conference (ISWC-2006), Athens, Georgia, USA* (Vol. 225).

Bonissone, P., & Tong, R. (1985). Editorial: reasoning with uncertainty in expert systems. *International Journal of Man Machine Studies*, *22*, 241 - 250.

Borst, P., Akkermans, J. M., & Top, J. L. (1997). Engineering ontologies. *International Journal on Human Computer Studies*, *46(2/3)*, 365 - 406.

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., & Stuckenschmidt, H. (2004). Contextualizing ontologies. *Journal of Web Semantics*, *1(1)*, 325 - 334.

Carvalho, R. (2011). *Probabilistic ontology: Representation and modeling methodology* (Unpublished doctoral dissertation). Fairfax, VA, George Mason University.

Chandrasekaran, B., Josephson, J., & Benjamins, V. (1999). What are ontologies and why do we need them? *IEEE Intelligent Systems*, *14*, 20 - 26.

Costa, P., & Laskey, K. (2005). *Multi-entity bayesian networks without multi-tears.* Retrieved from `http://hdl.handle.net/1920/456` (Draft,, Department of systems engineering and operations research, George Mason University: Fairfax, VA, USA)

Costa, P., & Laskey, K. (2006). Pr-OWL: A framework for probabilistic ontologies. In *Proceeding of the Fourth International Conference on Formal Ontology in Information Systems (FOIS), Baltimore, Maryland, USA* (Vol. 150, p. 237 - 249).

Cross, V. (2003). Uncertainty in the automation of ontology matching. In *4th international symposium on uncertainty modeling and analysis (isuma' 03)* (p. 135 - 140).

Dempster, A. (1967). Upper and Lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, *38*, 325 - 339.

Denœux, T. (1995). A K-nearest neighbor classification rule based on Dempster-Shafer Theory. *IEEE Transactions on Systems, Man and Cybernetics*, *25(5)*, 804 - 813.

Denœux, T. (1997). Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, *30(7)*, 1095 - 1107.

Ding, Z. (2005). *BayesOWL: a probabilistic framework for semantic web* (Unpublished doctoral dissertation). University of Maryland, Baltimore Country.

Do, H., Melnik, S., & Rahm, E. (2002). Comparison of schema matching evaluations. In *Web, Web-Services, and Database Systems, NODe 2002 Web and Database-Related Workshops, Erfurt, Germany* (Vol. 2593, p. 221 - 237).

Dubois, D., & Prade, H. (1988a). *Possibility theory.* Plenum Press - New York.

Dubois, D., & Prade, H. (1988b). Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, *4*, 244 -264.

Ehrig, M. (2007). *Ontology alignment: bridging the semantic gap* (Vol. 4). Springer.

Ehrig, M., & Staab, S. (2004). QOM - quick ontology mapping. In *International semantic web conference* (Vol. 3298, p. 289 - 303). Springer.

Ehrig, M., & Sure, Y. (2004). Ontology mapping - an integrated approach. In *the first european semantic web symposium, esws 2004* (p. 76 - 91).

Essaid, A., & Ben Yaghlane, B. (2009). BeliefOWL: An evidential representation in OWL ontology. In *Proceedings of the Fifth International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2009), collocated with the 8th International Semantic Web Conference, Washington DC, USA* (Vol. 527, p. 77 - 80).

Essaid, A., Ben Yaghlane, B., & Martin, A. (2011). Gestion du conflit dans l'appariement des ontologies. In *Atelier Graphes et Appariement d'Objets Complexes, en conjonction avec EGC 2011, Brest, France* (p. 50 - 60).

Essaid, A., Martin, A., Smits, G., & Ben Yaghlane, B. (2013). Processus de décision crédibiliste pour l'alignement des ontologies. In *Les Rencontres Francophones sur la Logique Floue et ses Applications, Reims, France* (p. 59 - 65).

Essaid, A., Martin, A., Smits, G., & Ben Yaghlane, B. (2014a). A distance-based decision in the credal level. In *Proceedings of 12th International Conference on Artificial Intelligence and Symbolic Computation, Seville, Spain* (Vol. 8884, p. 147 - 156).

Essaid, A., Martin, A., Smits, G., & Ben Yaghlane, B. (2014b). Uncertainty in ontology matching: a decision rule-based approach. In *Proceedings of the 15th Inter-*

national Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Montpellier, France (Vol. 442, p. 46 - 55).

Euzenat, J. (2004). An api for ontology alignment. In *Proceedings of 3rd International Semantic Web Conference, Hiroshima, Japan* (Vol. 3298, p. 698 - 712).

Euzenat, J., Ehrig, M., Jentzsch, A., Mochol, M., & Shvaiko, P. (2006). *Case-based recommendation of matching tools and techniques* (Tech. Rep.). Retrieved from `ftp://ftp.inrialpes.fr/pub/exmo/reports/kweb-126.pdf"`

Euzenat, J., Meilicke, C., Shvaiko, P., Stuckenschmidt, H., & Dos Santos, C. (2011). Ontology alignment evaluation initiative: six years of experience. *Journal on Data Semantics*, *15*, 158 -192.

Euzenat, J., Mocan, A., & Scharffe, F. (2008). Ontology alignments: an ontology management perspective. In M. Hepp, P. De Leenheer, A. De Moor, & Y. Sure (Eds.), *Ontology management: semantic web, semantic web services, and business applications* (p. 177 - 206).

Euzenat, J., & Shvaiko, P. (2013a). *Ontology matching* (2nd ed.). Springer.

Euzenat, J., & Shvaiko, P. (2013b). Ontology matching: State of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, *25*(1), 158 - 176.

Fensel, D. (2004). *Ontologies: a silver bullet for knowledge management and electronic commerce* (2nd ed.). Springer.

Gao, M., & Liu, C. (2005). Extending OWL by fuzzy description logic. In *Proceedings of 17th IEEE International Conference on Tools with Artificial Intelligence, Hong-Kong, China* (p. 562 - 567).

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199 - 220.

Haase, P., & Stojanovic, L. (2005). Consistent evolution of OWL ontologies. In *Proceedings of the Second European Semantic Web Conference, Heraklion, Crete, Greece* (Vol. 3532, p. 182 -197).

Hameed, A., Preece, A., & Sleeman, D. (2004). *Handbook on ontologies* (S. Steefen & R. Studer, Eds.). Springer.

Hois, J. (2009). A semantic framework for uncertainties in ontologies. In *Twenty-second international flairs conference.*

Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: a semantic web rule language combining OWL and RuleML.* (http://www.w3.org/Submission/SWRL/)

Huzza, M., Harzallah, M., & Trichet, F. (2006). OntoMas: a tutoring system dedicated to ontology matching. In *Proceedings of the 1st International Workshop on Ontology Matching Collocated with the 5th International Semantic Web Conference, Athens, Georgia, USA* (Vol. 225, p. 228 - 323).

Jousselme, A., Grenier, D., & Bossé, E. (2001). A new distance between two bodies of evidence. *Information Fusion*, *2*(2), 91 - 101.

Klein, M. (2001). Combining and relating ontologies: an analysis of problems and solutions. In *Proceedings of the International Workshop on Ontologies and Information Sharing at the 17th International Joint Conference on Artificial Intelligence, Seattle, WA, USA* (p. 53 - 62).

Koivunen, M. (2001). *W3C Semantic Web Activity.* (www.w3.org/Talks/2001/1102-semweb-fin)

Laublet, P., Charlet, J., & Reynaud, C. (2007). Sur des aspects primordiaux du web sémantique. In *La redocumentarisation du monde* (chap. 6).

Le Hégarat-Mascle, S., Bloch, I., & Vidal-Madjar, D. (1997). Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE T. Geoscience and Remote Sensing*, *35*(4), 1018 - 1031.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference of Machine Learning, Madison, WI, USA* (p. 296 - 304).

Madhavan, J., Bernstein, P. A., Domingos, P., & Halevy, A. Y. (2002). Representing and reasoning about mappings between domain models. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, Edmonton, Alberta, Canada* (p. 80 - 86).

Maedche, A., Motik, B., Silva, N., & Volz, R. (2002). MAFRA - a mapping framework for distributed ontologies. In *Proceedings of the 13th International Conference on knowledge engineering and knowledge management (ekaw)* (Vol. 2473, p. 235 - 250).

Martin, A., & Quidu, I. (2008). Decision support with belief functions theory for seabed characterization. In *11th International Conference on Information Fusion, Cologne, Germany* (p. 1 - 8).

Mitra, P., Noy, N., & Jaiswal, A. (2005). OMEN: A probabilistic ontology mapping tool. In *Proceedings of the 4th International Semantic Web Conference, Galway, Ireland* (Vol. 3729, p. 537 - 547). Springer.

Mochol, M. (2009). *The methodology for finding suitable ontology matching approaches* (Unpublished doctoral dissertation). Freie Universität Berlin.

Nagy, M., & Vargas-Vera, M. (2010). Towards an automatic semantic data integration: multi-agent framework approach. In G. Wu (Ed.), *Semantic web* (p. 107 - 134).

Nagy, M., Vargas-Vera, M., & Motta, E. (2007). DSSIM - managing uncertainty on the semantic web. In *Proceedings of the 2nd International Workshop on Ontology Matching collocated with the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, busan, korea* (Vol. 304, p. 160 - 169).

Ngo, D., Bellahsene, Z., & Todorov, K. (2013). Opening the black box of ontology matching. In *Proceedings of The Semantic Web: Semantics and Big Data, 10th Interna-*

tional Conference, Montpellier, France (Vol. 7882, p. 16 - 30).

Noy, N., & Klein, M. (2004). Ontology evolution: Not the same as schema evolution. *Knowledge and Information Systems*, *6*(4), 428 - 440.

Pan, R., Ding, Z., Yu, Y., & Peng, Y. (2005). A bayesian network approach to ontology mapping. In *Proceedings of the Fourth International Semantic Web Conference, Galway, Ireland* (Vol. 3729, p. 563 - 577).

Pearl, J. (1990). Jeffrey's rule, passage of experience, and neo-bayesianism. In H. Kyburg, R. Loui, & G. Carlson (Eds.), *Knowledge representation and defeasible reasoning* (p. 245 - 265).

Predoiu, L., Martin-Recuerda, F., Polleres, A., Feier, C., Mocan, A., Bruijn, J., ... Zimmermann, K. (2004). *Framework for representing ontology networks with mappings that deal with conflicting and complementary concept definitions* (Tech. Rep.).

Provost, F., & Kohavi, R. (1998). On applied research in machine learning. In *Editorial for the special issue on applications of machine learning and the knowledge discovery process* (Vol. 30).

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton University Press.

Shvaiko, P., & Euzenat, J. (2008). Ten challenges for ontology matching. In *On the Move to Meaningful Internet Systems: Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE, Monterrey, Mexico* (Vol. 5332, p. 1164 - 1182).

Smarandache, F., Martin, A., & Osswald, C. (2011). Contradiction measures and specificity degrees of basic belief assignments. In *Proceedings of the 14th International Conference on Information Fusion, Chicago, Illinois, USA* (p. 1 - 8).

Smart, P., & Engelbrecht, P. (2008). An analysis of the origin of ontology mismatches on the semantic web. In *Proceedings of Knowledge Engineering: Practice and Patterns, 16th International Conference, Acitrezza, Italy* (Vol. 5268, p. 120 - 135).

Smets, P. (1989). Constructing the pignistic probability function in a context of uncertainty. In *Proceedings of the Fifth annual conference on uncertainty in artificial intelligence, windsor, ontario, canada* (p. 29 - 39). Elsevier Science.

Smets, P. (1990). The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12(5)*, 447 - 458.

Smets, P. (1991). Varieties of ignorance and the need for well-founded theories. *Information Sciences*, *57*, 135 -144.

Smets, P. (1996). Imperfect information: Imprecision and uncertainty. In A. Motro & P. Smets (Eds.), *Uncertainty Management in Information Systems* (p. 225 -254).

Smets, P. (2007). Analyzing the combination of conflicting belief functions. *Information Fusion*, *8(4)*, 387 - 412.

Smets, P., & Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, *66(2)*, 191 -234.

Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J., & Horrocks, I. (2005). Fuzzy OWL: uncer-