Data veracity assessment: enhancing Truth Discovery using a priori knowledge

Valentina Beretta

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR D'IMT MINES ALES

En Informatique

École doctorale Information, Structures, Systèmes

Centre de recherche LGI2P de l'IMT Mines Ales

Data veracity assessment: enhancing Truth Discovery using *a priori* knowledge

Présentée par Valentina Beretta
30 Octobre 2018

Sous la direction de Sylvie Ranwez et Isabelle Mougenot

Devant le jury composé de

Catherine FARON ZUCKER, Maître de Conférences (HDR), Université de Nice Sophia Antipolis	Rapporteur
Ollivier HAEMMERLÉ, Professeur, Université de Toulouse	Rapporteur
Laure BERTI-EQUILLE, Directrice de recherche, IRD, UMR 228 Espace Dev	Présidente du jury
Aldo GANGEMI, Professeur, Università di Bologna	Examinateur
Jérôme DAVID, Maître de Conférences, Université Grenoble Alpes	Examinateur
Sylvie RANWEZ, Professeur, IMT Mines Alès	Co-direction de thèse
Isabelle MOUGENOT, Maître de Conférences (HDR), UM, UMR 228 Espace Dev	Co-direction de thèse
Sébastien HARISPE, Maître Assistant, IMT Mines Alès	Encadrant de proximité

IMT Mines Alès
École Mines-Télécom

Abstract

The notion of data veracity is increasingly getting attention due to the problem of misinformation and fake news. With more and more published online information it is becoming essential to develop models that automatically evaluate information veracity. Indeed, the task of evaluating data veracity is very difficult for humans. They are affected by *confirmation bias* that prevents them to objectively evaluate the information reliability. Moreover, the amount of information that is available nowadays makes this task time-consuming. The computational power of computer is required. It is critical to develop methods that are able to automatize this task.

In this thesis we focus on Truth Discovery models. These approaches address the data veracity problem when conflicting values about the same properties of real-world entities are provided by multiple sources. They aim to identify which are the true claims among the set of conflicting ones. More precisely, they are unsupervised models that are based on the rationale stating that true information is provided by reliable sources and reliable sources provide true information. The main contribution of this thesis consists in improving Truth Discovery models considering *a priori* knowledge expressed in ontologies. This knowledge may facilitate the identification of true claims. Two particular aspects of ontologies are considered. First of all, we explore the semantic dependencies that may exist among different values, i.e. the ordering of values through certain conceptual relationships. Indeed, two different values are not necessary conflicting. They may represent the same concept, but with different levels of detail. In order to integrate this kind of knowledge into existing approaches, we use the mathematical models of partial order. Then, we consider recurrent patterns that can be derived from ontologies. This additional information indeed reinforces the confidence in certain values when certain recurrent patterns are observed. In this case, we model recurrent patterns using rules. Experiments that were conducted both on synthetic and real-world datasets show that *a priori* knowledge enhances existing models and paves the way towards a more reliable information world. Source code as well as synthetic and real-world datasets are freely available.

"True genius resides in the capacity for evaluation of uncertain, hazardous, and conflicting information."

— Winston Churchill

Acknowledgements

Through these few lines I would like to express my deepest gratitude to everyone who take part, in one way or another, to these three fantastic years.

First of all, I would like to thank my thesis directors and my supervisor for being always there every time I knocked their doors. Usually, many supervisors mean many problems, instead you have been able to create an excellent working environment. Thanks Sylvie for your supervision, constructive remarks, encouragement and positivity. Thanks to push me outside my comfort zone to present my thesis outside the lab. I really enjoyed it. It has been a pleasure to work with you. Thanks Isabelle for your advice and support (moral and logistic). Especially, thanks for reminding me that real-world examples are often essential when explaining any research topic. I hope that I have finally learned it. Thanks Sébastien for being always cheerful and humorous, but also rigorous, honest and critical about my work. I think that these are important qualities for a researcher. Once again thanks to all of you, without your guidance none of the goals of this thesis would have been reached.

I would like to deeply thank Mme Laure Berti-Equille, Mme Catherine Faron Zucker, M. Jérôme David, M. Aldo Gangemi, and M. Ollivier Haemmerlé to accept to be part of my thesis committee. Special thanks to Mme Catherine Faron Zucker and M. Ollivier Haemmerlé that reviewed this manuscript. Your relevant comments and interesting feedbacks have been precious and stimulating. They have also permitted to highlight new directions that need to be explored.

I also want to thank my PhD colleagues Gildas, Cécile, Behrang, Alexander, Pierre-Antoine, Blazo, Lucie, Frank and all the others, with whom I spent these three important years. Life in the lab would not have been so pleasant without you. I would like also to show my gratitude to all LGI2P members, which share with me scientific discussions that enriched my study, as well as lunches, train and car journeys. Thanks also for showing me the French culture and for being patient when teaching me French. I would like also to mention all the colleagues of the UMR Espace Dev that welcomed me into their group without hesitation during this last year.

A special thanks goes to my family for all the love and incredible support over all these years. Mum and dad, my sister, my grandparents, my uncles and cousins: grazie! You have been excellent examples. You have taught me more than anyone that perseverance and determination are important to reach any goal in life. I known that you will never read this manuscript, but I am sure that you are already looking forward to celebrate this important goal the next time we will be together. Last but not least, a thought for Amos. During these three years of thesis, you have encouraged me, supported me and love me as you always do and will forever do, I hope. Thank you.

Now, no more acknowledgements. It's time to start reading the manuscript... Enjoy!

Contents

Synopsis de thèse

Sommaire

Ce résumé étendu présente le manuscrit écrit en anglais intitulé "Data veracity assessment: enhancing Truth Discovery using *a priori* knowledge". Il reprend les idées principales en détaillant d'abord le travail existant et le positionnement de la thèse par rapport à celui-ci, fixe nos contributions dans le domaine de la recherche de vérité. Il puis introduit l'évaluation de l'approche proposée, qui a été faite sur des jeux de tests synthétiques, puis dans un contexte réel. Enfin, ce synopsis se termine par les conclusions, ainsi que nos perspectives pour l'avenir. Il est à noter que ce synopsis ne se substitue pas au manuscrit mais en présente les lignes principales. Le détail des équations liées aux différentes propositions n'est donc pas donné.

I Contexte général et objectifs de la thèse

Depuis son origine, l'homme laisse des traces de son passage sur Terre. Or, à l'entrée dans le XXIe siècle, nombre de ces traces sont devenues numériques et imprègnent "Internet". Chacun, pour des raisons qui peuvent être sociales, scientifiques, économiques, politiques, militantes ou artistiques, diffuse des informations aussi diversifiées dans leur forme ou dans leur contenu que peuvent l'être nos différentes activités humaines. Après une certaine période d'euphorie engendrée par cet accès massif à différentes informations, l'heure est à la prudence. Les mises en garde sont de plus en plus soutenues auprès des personnes les plus "vulnérables" et en particulier des jeunes générations, afin d'éviter la propagation d'informations fausses (fake news) et l'adhésion à certaines idéologies qui constitueraient une menace pour nos sociétés et les individus qui les composent. Nombre d'évènements de ces derniers mois ont souligné la nécessité d'une telle prudence face à l'information. Des réponses ont été proposées. Ainsi, le site Politifact[1] analyse depuis plusieurs années les discours des responsables politiques américains afin de déterminer leur part de vérité et de mensonge. Dans la même veine, en France, le journal Le Monde propose un outil de vérification de la fiabilité des sources (Décodex[2]). Dans les deux cas, ce sont des acteurs humains (journalistes principalement) qui analysent les contenus et composent des synthèses qui sont restituées au grand public. Mais le volume d'informations est tel que, pour être traité de façon exhaustive, des approches automatisées se révèlent nécessaires. Pour contrer les dangers de la désinformation, un nouveau domaine de recherche a émergé ces dernières années désigné par détection de vérité sur le Web (Truth Discovery). Héritière de la vérification de faits (fact checking) d'une part et des techniques de fusion de données d'autre part, la détection de vérité analyse les assertions émises par plusieurs sources sur un sujet donné, et tente de déterminer parmi toutes ces assertions, celle qui constitue un fait (une vérité objective). Le but est relativement simple : trouver les données qui semblent être probables, et, de façon intimement liée, distinguer les sources d'information les plus fiables. En effet, un des meilleurs indicateurs de la confiance qu'on peut associer à une donnée est sa provenance. Cette étape est particulièrement importante lorsque l'on souhaite

[1] `http://www.politifact.com`
[2] `http://www.lemonde.fr/verification`

enrichir des bases de connaissances à partir de processus d'extraction automatique complexes faisant intervenir plusieurs extracteurs (sources), afin de constituer un support, par exemple, pour l'aide à la décision. C'est ce contexte qui m'a conduite, il y a trois ans, à débuter une thèse sur ce sujet. Les travaux qui sont présentés dans ce manuscrit ont été réalisés au sein du LGI2P (Laboratoire de Génie Informatique et Ingénierie de Production) d'IMT Mines Alès, dans l'équipe de recherche KID (Knowledge and Image Analysis for Decision making). Depuis de nombreuses années, certains chercheurs de cette équipe s'intéressent à l'utilisation des ontologies dans différentes phases (recherche d'information, indexation, analyse, filtrage) d'un processus de prise de décision, avec comme ambition d'assister l'opérateur humain dans ce processus. Récemment, ces recherches se sont centrées sur l'extraction d'information à partir de textes et la constitution de base de connaissances fiables. La détection de vérité s'inscrit pleinement dans cet objectif.

Les techniques actuelles de recherche de vérité se basent principalement sur un postulat : les sources qui ont diffusé majoritairement des assertions vraies sont estimées comme étant fiables et avec une forte propension à dire la vérité. La confiance dans les informations qu'elles diffusent est alors considérée comme d'autant plus élevée (Y. Li et al., 2015). Un processus itératif est utilisé afin de calculer ces degrés de fiabilité et de confiance et ainsi déterminer les assertions qui traduisent des *faits* (*vérités*). Les travaux qui sont présentés dans cette thèse reposent sur une représentation de la connaissance du domaine pour conforter la détection de vérité. Cette connaissance peut avoir été définie et modélisée au préalable dans une ontologie de domaine ou bien transparaître au travers de l'analyse d'une base de connaissance. Dans le premier cas, nous proposons de prendre en compte les liens qui définissent un ordre partiel entre différentes entités de cette ontologie afin d'affiner le calcul de confiance. Dans le second cas, il est possible d'identifier des motifs qui renforcent la confiance accordée à certaines affirmations. Ce sont ces deux approches qui sont présentées dans la suite de cette thèse. Les contributions sont les suivantes : i) proposer une nouvelle formalisation du problème de la détection de vérité qui prenne en compte la connaissance du domaine sous la forme de dépendances entre les valeurs et sous la forme de motifs récurrents ii) décrire les adaptations des modèles

existants nécessaires pour intégrer cette connaissance, iii) proposer une évaluation robuste pour chaque approche avec des jeux de données synthétiques et réels.

La section suivante présente le contexte de notre étude, pose la problématique et revient sur les notations de la littérature mises à contribution.

I.A Découverte de vérité

Par souci de clarté, cette section définit les notations utilisées par la suite. Certaines sont couramment utilisées dans le domaine (Berti-Équille & Borge-Holthoefer, 2015; Y. Li et al., 2015; Yin, Han, & Yu, 2008), alors que les autres sont introduites pour être utilisées ensuite dans la description de notre approche.

Soit e, une entité sujet d'intérêt, par exemple 'Pablo Picasso', appartenant à un ensemble d'entités E; et d, une description[3] de e appartenant à un ensemble de descriptions D, à l'exemple de "Pablo Picasso – bornIn", qui représente une propriété particulière de l'entité "Pablo Picasso". La description d est envisagée comme une propriété particulière de cette entité ou encore un prédicat associé à l'entité sujet. La valeur associée à cette propriété est représentée par le singleton *valeur*, avec $valeur \in V$ où V est l'ensemble de valeurs. Notons que la recherche de vérité envisagée ici ne concerne que des prédicats fonctionnels, c'est-à-dire ceux pour lesquels une seule valeur est admise (e.g. une personne ne peut être née qu'à un seul endroit).

Lors d'un processus d'extraction de connaissances (par exemple à partir d'analyse de textes), plusieurs sources d'information[4] peuvent proposer des valeurs différentes et contradictoires pour une même description d. L'ensemble de ces sources est noté S et on note $V_d \subseteq V$ l'ensemble des valeurs associées par différentes sources à la description d. Pour une description d, chaque proposition d'une valeur $v_d \subseteq V_d$ peut être représentée par un triplet <*entité, prédicat, valeur*>[5], et sera appelée assertion[6] tant qu'elle n'est

[3] Nous employons le terme *description* comme traduction de *data item* couramment utilisé dans la littérature anglaise.

[4] Ici "source d'information" est employé au sens large : il peut d'agir d'un site Internet, d'une base de données, d'une personne (via l'analyse de ses écrits$\cdots$). On simplifiera le propos par la suite en ne parlant que de "source".

[5] *Entité* est utilisée et soit sujet en français, soit subject en anglais.

[6] Une assertion pourra donc être notée indifféremment dans la suite sous la forme d'un

pas validée, c'est-à-dire tant que l'on n'a pas identifié la valeur *vraie* parmi toutes les valeurs associées à la même description. Déterminer cette valeur *vraie* permet de constituer un fait qui pourra être intégré à la base de connaissances. L'ensemble des sources qui font la même assertion est noté $S_{v_d} \subseteq S$ et l'ensemble des assertions proposées pour une source s est noté $V_s \subseteq V$.

Pour résoudre les conflits potentiels entre différentes assertions, il est nécessaire de prendre en compte la fiabilité des sources. On utilise pour ce faire deux fonctions : la *fiabilité* d'une source, que nous noterons t (source trustworthiness), et la *confiance* dans une assertion que nous noterons c (value confidence). Ces fonctions sont définies comme suit :

- $t : S \rightarrow [0,1]$, la *fiabilité* d'une source, représente sa propension à fournir de vraies valeurs (Y. Li et al., 2015). Une source réputée sûre aura un fort degré de fiabilité et sera considérée comme exprimant des valeurs vraies ($t(s) \simeq 1$) alors qu'une source non sûre aura un degré de fiabilité faible ($t(s) \simeq 0$) et sera réputée pour exprimer des valeurs fausses.

- $c : V \rightarrow [0,1]$, la confiance dans une assertion, traduit sa propension à être correcte, en fonction de nos connaissances actuelles (contexte). En effet, la vérité absolue n'existe pas et ce que l'on qualifie de vrai, ne l'est souvent qu'à la lumière de nos connaissances du monde (Pasternack & Roth, 2010). Une assertion exacte va avoir un fort degré de confiance ($c(v_d) \simeq 1$) et sera supposée provenir d'une source fiable. Par ailleurs, une assertion inexacte aura un faible degré de confiance ($c(v_d) \simeq 0$) et sera supposée provenir d'une source peu fiable.

On notera dans ces deux définitions, l'étroite relation qui existe entre fiabilité et confiance. À l'aide de ces notations, il est possible de définir la découverte de vérité comme suit – cette définition est une adaptation de celle qui est donnée dans (Y. Li et al., 2015) afin de conserver la cohérence de notation dans la suite de cette thèse.

Definition .1 (Découverte de vérité) *Soit un ensemble de descriptions D, un ensemble de valeurs V, un ensemble de sources S ; l'objectif principal de la découverte de vérité est de trouver pour chaque description $d \in D$, la valeur vraie $v_d^* \subseteq V_d \subseteq V$.*

triplet <entité, prédicat, objet> ou d'une paire (description, valeur) en fonction du contexte.

Ce calcul prend en compte la fiabilité des toutes les sources qui proposent v_d, c'est-à-dire S_{v_d}. Dans le même temps, les méthodes de détection de vérité estiment la fiabilité des sources, $t(s)$ avec $s \in S$, qui pourra influencer la détection de vérité, en tenant compte pour chaque source s de l'ensemble des assertions faites, c'est-à-dire V_s.

Les différentes approches proposées dans la littérature pour l'identification de vérité peuvent être classées en trois catégories que nous désignons par : les approches de référence, basiques et étendues. Nous ne les détaillons pas ici, mais donnons leurs principales caractéristiques. Le lecteur intéressé pourra se reporter à (Berti-Équille & Borge-Holthoefer, 2015) pour un état de l'art plus approfondi.

Les approches de référence utilisent des règles de vote entre les différentes sources (Y. Li et al., 2015). Ces approches font l'hypothèse que chaque source a le même degré de fiabilité. Ainsi, la valeur considérée comme vraie sera celle qui apparaît le plus grand nombre de fois dans les différentes sources. Ce modèle, très simple, possède deux limites majeures : chaque source est considérée de la même façon, y compris celles qui pourraient être qualifiées de non-fiables sur le long terme, et ces approches sont très sensibles à des attaques de type spam.

Les *approches basiques* prennent en compte la fiabilité des sources. Pour cela, elles procèdent suivant le modèle itératif dans lequel les estimations respectives de la confiance des valeurs et de la fiabilité des sources se succèdent jusqu'à la convergence. La confiance dans une assertion est estimée en prenant en compte la fiabilité des sources et pour chaque source, sa fiabilité est mise à jour en fonction de la véracité des assertions qui lui sont associées. Les principales approches de cette catégorie sont : *Sums*, *AverageLog*, *Investment* et *PooledInvestment* décrites dans (Pasternack & Roth, 2010), et *Cosine* et *2-Estimated* décrites dans (Galland, Abiteboul, Marian, & Senellart, 2010). Elles se distinguent par les formulations employées et la procédure itérative utilisée. Certaines approches prennent l'hypothèse d'une totale indépendance entre les assertions (Y. Li et al., 2015), alors que d'autres utilisent des méthodes de vote complémentaires (Galland et al., 2010). A notre connaissance, aucune de ces approches ne considère la connaissance du domaine au cours du processus de détection.

Des *approches étendues* ont donc été proposées. Celles-ci prennent en compte des dépendances possibles entre les assertions exprimées. La plupart de ces approches analysent des dépendances statiques (Blanco, Crescenzi, Merialdo, & Papotti, 2010; X. L. Dong, Berti-Equille, Hu, & Srivastava, 2010; X. L. Dong, Berti-Equille, & Srivastava, 2009a; Pochampally, Das Sarma, Dong, Meliou, & Srivastava, 2014; Qi, Aggarwal, Han, & Huang, 2013; S. Wang et al., 2015) et une approche est proposée pour prendre en compte la dépendance temporelle (X. L. Dong, Berti-Equille, & Srivastava, 2009b). Toutes ces méthodes se basent sur la même intuition que les sources qui partagent les mêmes valeurs fausses sont supposées être interdépendantes. Cette ressemblance entre les sources peut s'observer au niveau des sources elles-mêmes ou d'un groupe de sources. D'autres modèles étendus intègrent une connaissance complémentaire : des similarités entre valeurs, e.g. *TruthFinder* (Yin et al., 2008), des similarités entre descriptions (Meng et al., 2015), une connaissance antérieure (Pasternack & Roth, 2011), ou encore de l'extraction d'information (X. L. Dong et al., 2015).

À notre connaissance, très peu d'approches s'intéressent à des prédicats non-fonctionnels, c'est-à-dire ceux pour lesquels plusieurs valeurs peuvent être possibles simultanément pour une description donnée, par exemple quand plusieurs personnes sont auteur d'un même livre (Pochampally et al., 2014; X. Wang et al., 2016; Zhao, Rubinstein, Gemmell, & Han, 2012). Ces approches considérant de multiples vérités sont évaluées par des mesures de précision et de rappel et partent du postulat qu'une source peut émettre plus d'une assertion pour chaque aspect du monde réel (chaque description).

Les modèles existants ne considèrent pas la connaissance *a priori* que l'on peut avoir sur certaines valeurs. Cette connaissance peut, par exemple, être extraite à partir d'une ontologie, grâce à laquelle il est possible de propager l'information entre ces valeurs. Ainsi il est possible d'utiliser une connaissance de sens commun ou bien des faits déjà reconnus pour s'assurer que les confiances estimées concordent avec la connaissance *a priori* (Pasternack & Roth, 2010). Il est à noter que ces approches sont complètement différentes du contexte d'étude fixé dans la section suivante. En effet, nous considérons dans cette thèse des prédicats fonctionnels, c'est-à-dire pour lesquels il n'y a qu'une seule valeur 'vraie', même si, de par la structuration de la connaissance du domaine, il est possible de définir un ensemble de valeurs 'vraies'

représentant des granularités différentes, des points de vue différents sur cette unique valeur.

L'approche proposée dans la suite se démarque de celles présentées dans l'état de l'art, du fait qu'elle prend en compte la connaissance *a priori* d'un domaine pour calculer la confiance dans une assertion, et de manière induite la fiabilité des sources. Deux formes de connaissance *a priori* sont considérées séparément : l'ordre partiel sur les valeurs et les règles d'association. Pour pouvoir tirer parti des connaissances *a priori*, il faut d'abord introduire des formalismes qui permettent de modéliser ces connaissances.

I.B Modélisation des connaissances

L'obtention de connaissances lisibles et intelligibles par les machines a été largement abordée par les études faites dans le domaine de l'Ingénierie des Connaissances, sous domaine de l'Intelligence Artificielle, dont le but est de modéliser la connaissance sous une forme traitable automatiquement (Guarino, Oberle, & Staab, 2009). De cette manière, des agents logiciels peuvent utiliser cette connaissance pour y appliquer des méthodes de raisonnement automatique. Cela est possible en définissant formellement et rigoureusement des expressions avec de vocabulaires structurés et contrôlés dont la sémantique n'est pas ambiguë. Par conséquent, un modèle de connaissance doit être composé des éléments suivants :

- un *vocabulaire* indiquant les composants de langage ;
- une *syntaxe* définissant quelles configurations des composants du langage sont valides ;
- une *sémantique* spécifiant les faits du monde réel auxquels les *phrases* se réfèrent.

Un exemple de modèle de connaissances répandu est celui des ontologies. Plusieurs définitions formelles ont été proposées. Dans le domaine de l'Ingénierie des Connaissances, la définition la plus connue indique que "*une ontologie est une spécification explicite d'une conceptualisation*" (Gruber, 1993). Cette définition capture plusieurs aspects clés d'une ontologie. Tout d'abord, l'expression *spécification explicite* met en évidence le fait que toutes les connaissances doivent être exprimées, c'est-à-dire spécifiées, dans un format lisible par une machine. Les notions qui ne sont pas clairement énoncées

ne sont pas connues par les machines (ainsi que les notions de *bon sens* que les humains considèrent comme allant de soi). Ensuite, le terme *conceptualisation* indique "*une vue abstraite et simplifiée du monde que l'ontologie veut représenter*" (Gruber, 1993). Une conceptualisation doit être partagée et acceptée par les membres d'une communauté. Une conceptualisation est une entité abstraite qui n'existe que dans leur esprit. Pour être communiquée et partagée cette conceptualisation doit être représentée de façon précise, non ambigüe et synthétique. Un langage partagé doit être spécifié. Dans ce langage, chaque concept est représenté par un symbole qui fait référence à une certaine vision du monde réel. Une ontologie est composée d'un ensemble de concepts, qui chacun représente une classe d'individus partageant certaines propriétés (à l'exemple de *Country*), d'un ensemble d'instances (qui sont des occurrences réelles de concepts, par exemple *France* pour *Country*) et d'un ensemble de relations (qui sont des liens ou des connexions entre instances, ou entre instances et concepts, ou entre concepts, à l'exemple de la relation *isLocatedIn*). Les principaux formalismes utilisés pour représenter les ontologies en Ingénierie des Connaissances sont les graphes conceptuels (Sowa, 1984), les langages de Frames (Minsky, 1974) et les logiques de description (DLs) (Baader, Calvanese, McGuinness, Patel-Schneider, & Nardi, 2003). La définition d'une logique formelle permet une interprétation plus large des connaissances pour déduire automatiquement des faits qui ne sont pas explicitement énoncés. Sur la base de la complexité de la logique définie, plusieurs langages ayant différents niveaux d'expressivité peuvent être obtenus. Évidemment, l'expressivité augmente la complexité algorithmique du raisonnement lié à un langage. Généralement, les ontologies basées sur les DLs sont un bon compromis entre expressivité et efficacité[7]. Dans un contexte d'ontologies s'adossant à la famille des logiques de description, deux composantes sont distinguées : la T-Box (*Terminological* Box) qui intègre la description des concepts et des propriétés liant ces concepts et la A-Box (*Assertion* Box) qui contient les instances des individus qui se conforment aux descriptions de la T-Box. Nous considérons ici les ontologies de domaine construites à l'aide du langage OWL 2 (Group, 2012) qui s'appuient sur des logiques de description spécifiques. Les ontologies basées sur les DLs sont largement utilisés par les communautés scientifiques et industrielles grâce au Web sémantique. En effet, ces logiques sont utilisées

[7] https://www.w3.org/TR/owl-guide/

par le World Wide Web Consortium (W3C) comme fondements formels du langage ontologique du Web (Web Ontology Language). Cela a fortement contribué au développement d'un certain nombre de protocoles standards et de langages basés sur les logiques de description pour promouvoir le Web sémantique et les ontologies.
Dans cette thèse, nous visons à améliorer l'évaluation de la découverte de vérité en utilisant la connaissance *a priori* contenue dans ces types d'ontologies. L'idée est que leur connaissances peut faciliter la compréhension des informations fournies par plusieurs sources de données. Par conséquent, l'évaluation de la véracité peut s'en voir simplifiée. Ceci explique la raison pour laquelle nous avons présenté dans une même section la découverte de vérité et les ontologies. Une vue générale de l'approche que nous proposons est présentée dans la figure 1. Nous y voyons que la connaissance *a priori*

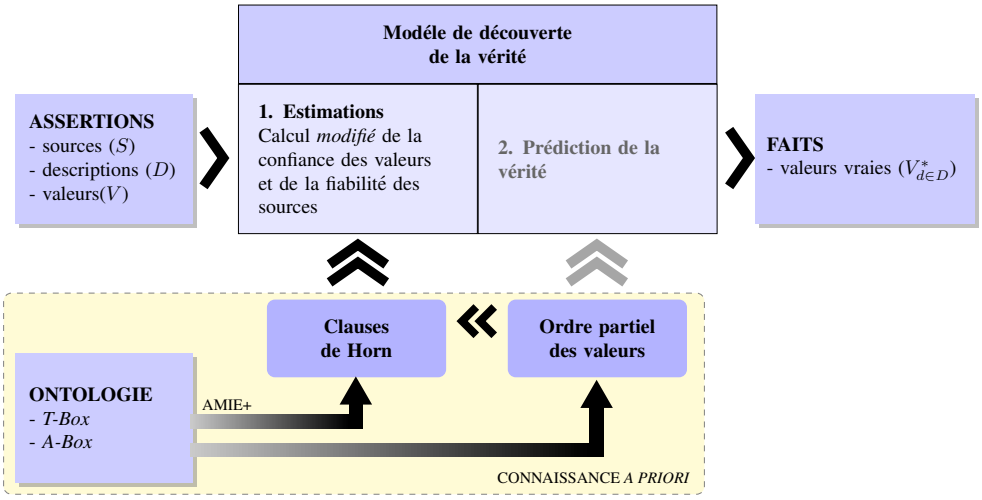


Figure 1: Méthode de découverte de la vérité (TD) intégrant les les relations entre les valeurs et les motifs récurrents au cours de la phase d'estimation.

exprimée dans un'ontologie est utilisée de deux manières différentes : par les relations entre les valeurs (ordre partiel entre ces valeurs, partie droite de la figure) et par la détection de motifs récurrents (partie gauche). Ce sont ces deux aspects qui sont présentés par la suite.

II Contributions de la thèse

Rappelons que nous considérons ici l'analyse d'assertions associées à des prédicats fonctionnels. Afin de sélectionner la valeur vraie associée à une description, tout comme pour estimer la confiance associée à une source, nous considérons que les valeurs proposées par les sources respectent la logique bivalente, et sont donc vraies ou fausses. La notion de vérité peut donc être définie par la fonction binaire $tf : D \times V \rightarrow \{true, false\}$. La formulation du problème telle que nous la proposons vise à représenter de façon plus réaliste les cas réels pour lesquels la dépendance entre plusieurs valeurs est prise en compte. Comme nous allons le voir, cette considération implique des modifications importantes dans la formulation du problème ; cela, aussi bien au niveau des assomptions considérées qu'au niveau des solutions proposées pour résoudre le problème.

II.A Utilisation de l'ordre partiel pour la détection de vérité

Face à des prédicats fonctionnels, la plupart des modèles existants partent du postulat qu'une seule valeur peut être vraie parmi celles proposées par différentes sources. Pourtant, généralement, les valeurs proposées ne sont pas indépendantes. Un ordre partiel sur ces valeurs peut exister. Par exemple parmi les propositions suivantes deux valeurs seulement entrent en conflit :

- <Pablo Picasso, bornIn, *Spain*>
- <Pablo Picasso, bornIn, *Málaga*>
- <Pablo Picasso, bornIn, *Europe*>
- <Pablo Picasso, bornIn, *Granada*>

En effet, *Granada* et *Málaga* étant deux villes distinctes, elles ne peuvent être considérées toutes les deux comme étant vraies. Or avec une connaissance ontologique du domaine, et en particulier certaines de ses relations, il est possible de déterminer que *Málaga* et *Granada* sont toutes les deux des villes d'*Espagne* et donc d'*Europe*. La connaissance exprimée par ce type de relation est particulièrement pertinente dans notre problématique et profitable pour l'identification des valeurs vraies. Ainsi la formulation du problème que

nous proposons vise à représenter de façon plus réaliste les cas réels pour lesquels la dépendance entre plusieurs valeurs est prise en compte.

La première approche que nous proposons exploite une portion réduite de l'ontologie, surtout constituée des définitions des classes[8] contenues dans la T-Box. Plus précisément, nous modélisons la dépendance entre les différentes valeurs, qui s'exprime *a priori* dans une ontologie, sous la forme d'un ordre partiel $O = (V, \preceq)$ défini par certaines relations transitives. Cet ordre partiel O précise les relations de l'ontologie qui sont prises en compte entre les valeurs, c'est-à-dire les relations qui précisent les valeurs qui subsument d'autres valeurs. Ainsi, pour les valeurs $x, y \in V^2$, écrire $x \preceq y$ signifie que x implique y. Par exemple $Espagne \preceq Europe$ signifie que dire que quelqu'un est né en *Espagne* implique de dire que cette personne est née en *Europe*. Nous nous focalisons sur les ordres partiels des ressources formés par la structuration des classes (e.g., rdfs:subClassOf), le typage des ressources (e.g., rdf:type) et d'éventuels liens entre les ressources exprimés par des prédicats transitifs supplémentaires (e.g., dbo:isPartOf). A noter que pour chaque prédicat différentes relations peuvent être considérées pour composer l'ordre. L'important est de préserver la transitivité de l'ordre. Dans tous les cas, cet ordre partiel pourra être intégré à l'analyse des assertions exprimées par les sources étudiées, comme connaissances supplémentaires sur les valeurs considérées. En effet, si une source exprime une valeur, elle supporte aussi de façon implicite l'ensemble des valeurs qui la subsume. L'ontologie de domaine contient également d'autres types d'information qui peuvent être considérés tels que le contenu informationnel (IC pour Information Content) qui est rattaché à chaque classe (Seco et al. 2004). Cet indicateur permet d'estimer la spécificité d'une classe et donc représente son degré d'abstraction/concrétude par rapport à la connaissance d'un domaine. Une propriété particulièrement intéressante de l'IC est que sa valeur croit de façon monotone de la racine jusqu'aux feuilles de la hiérarchie de classes. Ainsi, si $x \preceq y$, alors $IC(x) \geq IC(y)$, ($IC(root) = 0$). Ainsi la spécificité d'une valeur est un bon indicateur de son caractère informatif. Plus une valeur est abstraite, moins elle est informative, du fait que l'ensemble des valeurs qu'elle subsume est grand. Par exemple, *Málaga* s'avère plus informative (car plus précise, plus spécifique) que *Europe*. Ainsi, prendre

[8] Le terme classe est utilisé ici plutôt que concept, en conformité avec le vocabulaire défini dans RDFS/OWL, langages standards dans le domaine du Web Sémantique.

en compte l'IC, c'est-à-dire le degré de spécificité, permettra de contraindre l'ensemble des valeurs vraies potentielles. Cet indicateur sera utilisé par la suite pour sélectionner la valeur vraie.

II.A.a Exploitation de l'ordre partiel des valeurs dans le processus de découverte de vérité

Plus formellement, une source exprimant une assertion v_d avec $d \in D$ supporte aussi l'ensemble des assertions v'_d associées à la description d qui correspondent à des valeurs plus générales que v_d, c'est-à-dire $\forall v'_d \in V_d$ si $v_d \preceq v'_d$, alors $v_d \rightarrow v'_d$ (v_d implique v'_d). En effet quand d est connu, un ordre partiel sur les assertions peut être considéré à partir de l'ordre partiel défini sur les valeurs. Dans la suite, par abus de langage, nous utiliserons indifféremment assertion ou valeur quand la description d est connue et fixe. Si l'on se place dans ce contexte, la valeur de vérité ne peut être réduite à une valeur unique mais se compose plutôt d'un ensemble de valeurs. Par exemple, les deux assertions <Pablo Picasso, bornIn, *Granada*> et <Pablo Picasso, bornIn, *Málaga*> supportent les deux assertions <Pablo Picasso, bornIn, *Spain*> et <Pablo Picasso, bornIn, *Europe*>. En d'autres termes, les assertions plus génériques qu'une assertion considérée comme vraie seront nécessairement, elles aussi, toujours vraies ; donc plusieurs valeurs peuvent être considérées comme vraies pour une description d particulière. Cela signifie que si une source exprime un fait, la source exprime également de façon implicite l'ensemble des faits plus généraux que le fait exprimé.
La modélisation de la solution proposée repose sur les fonctions de croyance introduites dans (Shafer et al., 1976). Ces fonctions permettent de représenter l'ignorance et l'incertitude contenues dans des informations contradictoires. Pour faciliter la lecture, nous présentons notre approche en nous appuyant sur une adaptation des notations habituelles en théorie des croyances. L'unité atomique manipulée par ces fonctions est la fonction de masse qui représente la portion de preuve allouée à une valeur particulière. Elle peut être utilisée pour définir la croyance (belief en anglais) qui peut être associée à une valeur donnée. Cette théorie mathématique permet de sommer l'information apportée par l'observation d'une valeur. Dans notre cas, la fonction de croyance propage l'information véhiculée par une assertion aux assertions qui lui sont plus générales en considérant l'ordre partiel défini

par l'ontologie. À titre illustratif, nous proposons d'adapter le modèle de découverte de vérité *Sums*, défini dans (Pasternack & Roth, 2010), en y intégrant notre reformulation du problème et la prise en compte du modèle de propagation présenté. La méthode *Sums* adopte une procédure itérative dans laquelle le calcul de la fiabilité associée à une source et le calcul de la confiance associée à une assertion sont alternés jusqu'à atteindre une convergence. La fiabilité associée à une source est ensuite évaluée en sommant les confiances sur les assertions qui lui sont associées. De façon similaire, la confiance associée à une assertion est évaluée en sommant les fiabilités des sources qui expriment cette assertion. L'approche *Sums* peut être adaptée à notre problématique en modifiant le calcul de la confiance d'une assertion. Au lieu de ne considérer que l'ensemble des sources qui expriment une assertion, nous tenons compte de la transitivité de l'ordre partiel et modifions l'ensemble des sources considérées. Il est composé des sources qui proclament une assertion donnée et des sources qui proclament des assertions plus spécifiques, et donc implicitement l'assertion considérée.

Comme prévu, une conséquence importante de cette modification concerne le nombre de valeurs vraies. Ainsi l'adaptation de la méthode *Sums*, ou de toute autre méthode, nécessite la définition d'une stratégie permettant de distinguer l'ensemble des valeurs vraies après convergence.

Normalement les approches existantes pour la détection de vérité identifient pour chaque description d'une entité donnée, la valeur qui a la plus grande confiance et qui est donc considérée comme étant vraie. Cette stratégie ne peut s'appliquer à notre contexte où l'on considère un ordre partiel sur les valeurs à partir d'un modèle de connaissance du domaine (e.g. relations de subsomption d'une ontologie). En effet, dans ce cas, les valeurs les plus génériques vont être associées à un fort degré de confiance. Les sources qui proposent une valeur supportent également de façon implicite toutes ses généralisations. Ne seraient donc considérées comme vraies (c'est-à-dire ayant le plus fort degré de confiance), que des valeurs hautement génériques (voire même la racine de l'ontologie). Sur notre exemple, une source proposant l'assertion <Pablo Picasso, bornIn, *Málaga*> soutient de façon implicite les assertions plus génériques telles que <Pablo Picasso, bornIn, *Spain*>, <Pablo Picasso, bornIn, *Europe*>, etc. La valeur qui aurait donc la confiance maximum, c'est-à-dire <Pablo Picasso, bornIn, *Location*> ne serait pas forcé-

ment d'un grand intérêt.

II.A.b Sélection des valeurs vraies

Nous avons mis en place une stratégie de sélection des valeurs vraies qui prend en compte la définition d'un ordre partiel entre les valeurs et, pas à pas, raffine la granularité de la valeur vraie associée à chaque description. À partir de la valeur la plus générique, implicitement cautionnée par toutes les valeurs candidates, le processus de sélection a pour objectif de détecter la ou les valeurs les plus spécifiques susceptibles d'être vraies. Ce processus, en partant de la racine, parcourt le graphe composé des valeurs candidates reliées par les relations existantes dans l'ordre partiel considéré. À chaque étape, il sélectionne les meilleures alternatives parmi les valeurs descendantes directes d'une valeur considérée, jusqu'à atteindre la valeur vraie. L'hypothèse que nous considérons est que les valeurs qui ont la plus haute confiance parmi les valeurs proches considérées ont le plus de chances d'être vraies. Le choix du nœud qui doit être considéré à l'étape suivante est donc fait en fonction de la comparaison des scores de confiance des fils du nœud considéré. La sémantique de chaque nœud sélectionné prend en compte le fait que ce nœud subsume la valeur vraie (c'est-à-dire la valeur attendue). Le dernier nœud considéré doit correspondre à la valeur la plus spécifique et avec un fort degré de confiance parmi celles proposées. Deux situations particulières peuvent se présenter au cours du processus : i) devoir choisir une valeur alors que son degré de confiance est trop faible et donc sa pertinence discutable, et ii) devoir choisir entre deux alternatives qui ne diffèrent que faiblement au niveau de leur degré de confiance. C'est pour répondre à ces difficultés que deux seuils ont été introduits : γ et δ.
Le paramètre γ permet de spécifier un seuil de confiance minimal en deçà duquel la valeur ne sera pas considérée comme candidate possible à la valeur vraie.
Le paramètre δ représente la différence minimale exigée entre les scores de confiance de deux nœuds. En particulier, si cette différence est inférieure ou égale à δ, alors, le choix entre les deux alternatives est difficile, car peu significatif. Cette comparaison concerne les valeurs qui descendent d'une même valeur.
Une fois les valeurs et leurs ancêtres sélectionnés il y a la phase d'ordonnan-

cement afin d'identifier la valeur vraie attendue pour chaque description. Nous avons procédé à plusieurs expérimentations. Le premier choix s'est porté sur une sélection basée sur l'IC des différentes valeurs candidates. Un autre mode de sélection consiste à ordonner les valeurs en utilisant la moyenne des fiabilités de leurs sources. De plus, lors de la phase de filtrage suivante, nous pouvons filtrer toutes les valeurs retournées, qu'elles aient ou non été ordonnées. Des expériences sur des ensembles de données synthétiques ont été réalisées. Plus précisément, nous avons généré 60 jeux de données pour cinq prédicats différents. Pour évaluer le modèle en fonction de l'ontologie utilisée, nous avons utilisé deux ontologies différentes (DBpedia (Auer et al., 2007) et Gene Ontology (Ashburner et al., 2000)) pour obtenir les prédicats et les informations relatives aux valeurs possibles assumées par ces prédicats. Ces expériences montrent l'efficacité de notre première proposition.

II.B Utilisation de règles pour la détection de vérité

Le deuxième volet de notre approche concerne l'exploitation d'une autre forme de connaissance *a priori* d'un domaine et intègre une analyse plus large de la A-Box. L'idée consiste à prendre en compte tous les types de relation et les faits qui la composent. En effet, en étudiant les cooccurrences entre ces faits, il est possible d'identifier des motifs qui peuvent être ensuite utilisés pour conforter notre jugement *a priori* sur certaines assertions. Prenons l'exemple très simplifié représenté dans la Figure 2. Une analyse de la base permet de déduire que la majorité des personnes qui parlent espagnol sont nées en *Espagne*. Cette observation peut être prise en compte dans un processus de recherche de vérité concernant le lieu de naissance de Pablo Picasso, par exemple. Si on observe que Pablo Picasso parle couramment espagnol, la confiance attribuée à l'assertion <Picasso, bornIn, *Spain*> doit être renforcée, ainsi que les assertions qui contiennent des valeurs plus génériques. Si à notre connaissance cela n'a jamais été appliqué dans le contexte de la détection de vérité, il serait pertinent d'exploiter les cooccurrences de faits par l'identification de règles d'association. Comme mentionné dans la synthèse sur les règles d'association présentée dans (Maimon & Rokach, 2005), il est difficile d'avoir une vue exhaustive des travaux dans ce domaine. Pour des applications en lien avec le Web Sémantique, on peut toutefois se

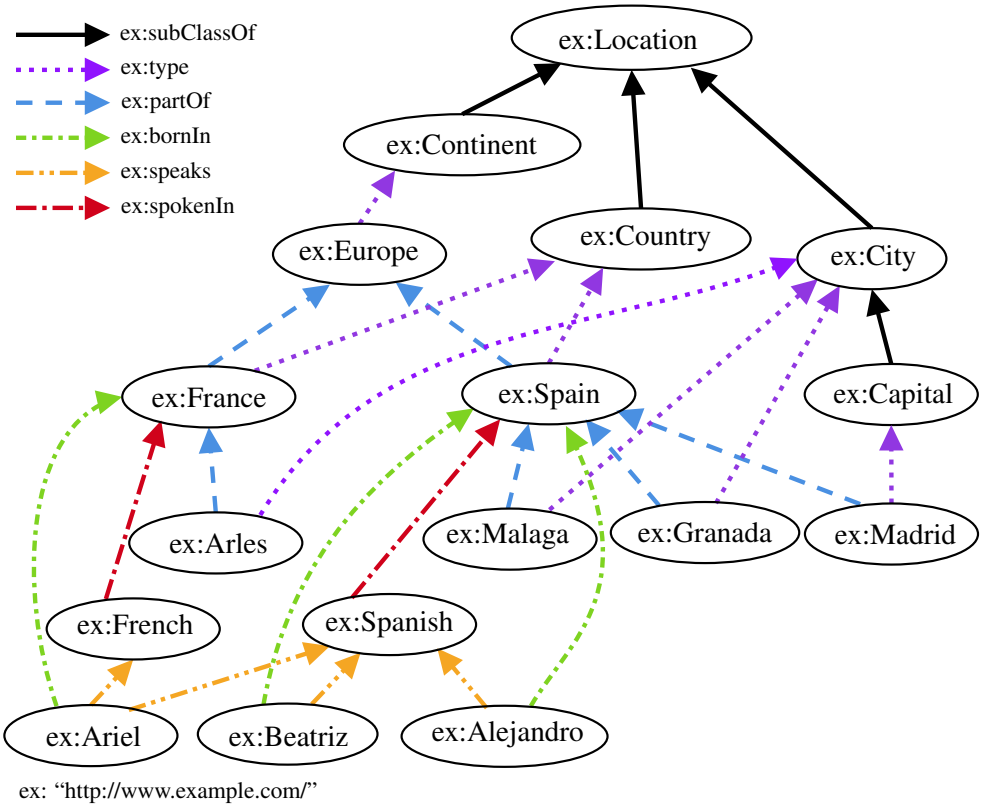


Figure 2: Extrait d'une base de connaissance intégrant certains éléments de l'ontologie (concepts, relations), mais également des instances et leurs relations avec les concepts de cette ontologie.

référer à (L. Galárraga, Teflioudi, Hose, & Suchanek, 2015; Z. Wang & Li, 2015). Certains problèmes sont particuliers à ce contexte : la quantité de données, l'assomption du monde ouvert et les données manquantes (Quboa & Saraee, 2013).

De façon formelle, appelons KB une base de connaissances supposée n'être composée que de faits (vérités objectives) représentés par un ensemble de triplets RDF de la forme <sujet, prédicat, objet>. Plusieurs types de prédicats sont autorisés, ainsi que différents types pour les entités sujets et les objets. Dans la suite, la notation utilisée pour les règles sera celle de Datalog (Boley, 2000; Nenov et al., 2015). Une règle est une implication d'un ensemble d'atomes reliés par un opérateur de conjonction, appelé corps (aussi appelé antécédent ou prémisse), vers un autre ensemble appelé tête (conséquence). Formellement, la règle r pourra s'écrire : $r : B_1 \wedge B_2 \wedge \cdots \wedge B_n \implies H$ qui

est équivalent à $\widehat{B} \rightarrow \widehat{H}$.

Dans notre approche, nous considérons uniquement des clauses de Horn, c'est-à-dire qui n'ont qu'un singleton dans la tête. Ici, un atome est assimilé à une assertion constituée d'un prédicat défini et d'entités sujet et objet qui peuvent être variables. L'identification des règles par l'analyse de la base de connaissances est réalisée avec AMIE+ (L. Galárraga et al., 2015).

Plusieurs métriques ont été proposées pour évaluer la qualité d'une règle, dont les plus répandues sont le support et la confiance. Le support indique la proportion d'entités vérifiant à la fois le corps et la tête de la règle. La confiance, quant à elle, indique la proportion d'entités vérifiant la tête, parmi celles qui vérifient le corps. Cette valeur peut être vue comme une estimation de la probabilité de la tête de la règle si $\widehat{B}$. Cette mesure de confiance a été définie dans un contexte de raisonnement en monde fermé, où l'on considère comme fausses les assertions qui ne sont pas exprimées dans la base. Or dans le contexte du Web sémantique, celui qui nous concerne dans cette étude, c'est l'hypothèse d'un monde ouvert qui est envisagée selon les principes qui ont cours dans les logiques de description. À cet effet, les auteurs de (L. Galárraga et al., 2015) ont introduit la mesure de PCA confidence qui repose sur l'hypothèse de complétude partielle (PCA pour Partial Completeness Assumption) qui considère que si la base de connaissances contient au moins une assertion qui concerne une description d=(sujet, prédicat), alors toutes les valeurs possibles pour cette description sont connues. Autrement dit, si une description n'apparait jamais dans la base, elle n'est considérée ni comme étant vraie, ni comme étant fausse.

Dans notre proposition, nous avons défini un coefficient propulseur (booster), calculé à partir de l'identification de règles, qui représente les cooccurrences récurrentes entre différents faits, et leur mesure de qualité afin de renforcer la confiance dans certaines valeurs pendant le processus de détection de vérité. Le coefficient propulseur représente le degré de soutien (ou caution) apporté pour cette assertion par les informations contenues dans KB. Ainsi le postulat de base du processus de recherche de vérité présenté en introduction en sera modifié et nous considérerons désormais que les *faits* (*vérités*) sont des assertions proposées par des sources fiables et/ou qui sont renforcées par un coefficient *booster* élevé, en considération de règles d'association extraites de KB. Comme dans les approches traditionnelles, la

fiabilité d'une source dépendra, quant à elle, du nombre de vérités qu'elle a proposées. Il est à noter que les motifs récurrents n'ont pas tous le même degré d'expressivité et ne doivent donc pas avoir le même impact sur le processus de détection de vérité. L'influence du coefficient *booster* sur le calcul de confiance dans une assertion sera donc paramétrable afin d'accorder plus d'importance à la fiabilité des sources ou au contraire à l'information contenue dans KB en fonction du contexte et/ou de la qualité de la base.
Nous avons également considéré l'ordre partiel afin de propager l'information donnée par le coefficient propulseur aux valeurs plus générales.

Aussi dans ce cas, pour évaluer le potentiel de l'approche proposé, des expériences ont été réalisées sur les mêmes jeux de données synthétiques générés précédemment. Après avoir obtenu des résultats satisfaisants, nous avons décidé d'évaluer également l'approche avec des jeux de données réels.

II.C Application des méthodes proposées sur des jeux des données réels

Afin d'évaluer les approches proposées dans un scénario réel, nous avons décidé de les exploiter dans un processus de Slot-filling à partir du données Web. Il s'agit d'une sous-tâche d'un processus plus général de peuplement de bases de connaissances (KBP) ayant pour objectif d'identifier les vraies valeurs manquantes dans des bases de connaissances existantes, et ce pour chaque paire (*sujet*, *propriété*) à partir d'une collection de textes. Chaque paire correspond à un *data item* dans le contexte de la recherche de vérité et représente un aspect (c'est-à-dire une propriété) d'une entité du monde réel (c'est-à-dire le sujet), dont la valeur est manquante dans la base de connaissances considérée. Dans ce contexte du Web, les noms de domaine sont considérés comme des sources d'information. Alors que différentes sources peuvent fournir des valeurs contradictoires pour un même data item, les méthodes de détection de vérité peuvent être utilisées pour distinguer les vraies valeurs des fausses. Nous pouvons identifier deux avantages principaux pour l'utilisation de la détection de vérité dans ce contexte. Premièrement, l'approche proposée est basée sur des données Web. Ainsi, une collection de textes n'est plus nécessaire pour le remplissage des slots. Deuxièmement, aucune phase d'entraînement n'est nécessaire pour ces méthodes. En effet, les modèles de recherche de vérité sont des techniques non super-

visées.

Puisque ces modèles nécessitent de disposer d'assertions structurées en entrée et qu'il est difficile d'avoir ce contenu directement à partir du Web, il est nécessaire de définir une phase de pré-traitement pour extraire ces assertions à partir du texte brut trouvé sur le Web. Pour cela nous avons proposé l'utilisation d'un processus d'extraction naïf. Les assertions structurées et leurs sources respectives sont fournies en entrée d'une procédure de recherche de vérité pour compléter la tâche de slot-filling et identifier les valeurs vraies manquantes.

Nous avons comparé les différentes performances obtenues, sur les jeux de données réels, par les modèles de détection de vérité proposés et les méthodes de détection de vérité existants. Une évaluation étendue avec les systèmes de slot-filling traditionnels dépasse le cadre de notre étude. Ici, nous nous concentrons sur la comparaison entre les différents modèles de détection de vérité lorsqu'ils sont appliqués à la réalisation de cette tâche. En effet, l'objectif principal est d'évaluer l'impact de la prise en compte des connaissances *a priori* lors de l'utilisation des modèles de recherche de vérité dans un scénario réel, qui a ses propres caractéristiques. L'approche qui a obtenu le meilleur résultat est celle qui considère les deux types de connaissances *a priori* (ordre sur les valeurs et motifs récurrents). Nous montrons notamment que les modifications proposées du modèle *Sums* permettent d'obtenir des gains de performance de l'ordre de 16% par rapport au modèle *Sums* traditionnel. Cette augmentation des performances permet à ce modèle 'simple' de l'état de l'art d'obtenir des performances comparables à celles obtenues par les modèles les plus raffinés du domaine du recherche de vérité. Nous faisons ainsi la démonstration de la pertinence d'intégrer la prise en compte de connaissances *a priori* pour la définition et l'amélioration de modèles de recherche de vérité.

III Synthèse et élargissement

À l'heure où la détection de vérité devient de plus en plus cruciale pour nombre d'applications, il nous semble indispensable de développer des approches de recherche de vérité qui tiennent compte d'une modélisation de connaissance sous forme d'ontologies.

Cette thèse propose différentes approches permettant la détection de vérité dans une base d'assertions, en tenant compte de la modélisation de la connaissance d'un domaine (ontologie). Notons que nous restons dans les travaux proposés dans le cas de prédicats fonctionnels, c'est-à-dire pour lesquels il n'existe dans l'absolu qu'une seule valeur vraie, mais où cette valeur peut être considérée à différents degrés de précision. En effet, afin de mieux répondre à des problématiques du monde réel, il est nécessaire de considérer que différentes valeurs associées à des descriptions de certaines entités, ne sont pas nécessairement concurrentes, mais s'expliquent plutôt dans certains cas par des variabilités en terme de précision de réponse. Ainsi pour une entité donnée et une description qui y est rattachée, nous proposons d'étendre le cadre classiquement considéré par les approches de détection de vérité étudiées en considérant non plus une valeur vraie unique mais plutôt un ensemble de valeurs vraies (valeurs non conflictuelles). Cet ensemble est construit en utilisant la propagation de *confiance*, inspirée par les approches de la théorie des croyances, appliquée à des méthodes traditionnelles (*Sums* dans ce manuscrit). Dans ce cas, nous avons exploité principalement la T-Box et la propriété algébrique de transitivité qui s'applique à l'ordre partiel qui est sous-tendu par les relations entre les valeurs. Ce modèle exprime donc une approche déductive grâce aux propriétés mathématiques de l'ordre partiel. Nous avons également proposé une approche inductive, et donc complémentaire, qui généralise la connaissance des cas individuels. Il est basé sur l'intégration de règles d'association collectées après l'analyse de la A-Box. Une évaluation au travers des jeux de données a été menée sur les modèles proposés. Les résultats montrent que, considérer une connaissance *a priori* pendant l'estimation des confiances de valeur, apporte une vraie plus-value. Par ailleurs, la recherche rapportée dans ce manuscrit montre que les approches proposées se révèlent efficaces, et amènent notamment une amelioration des performances des modèles adaptés. Nous avons validé nos travaux sur la base de jeux de données et de développements spécifiques. Les différentes méthodes implémentées en Python et autres données utilisées dans le cadre de la thèse sont partagées librement sur Internet.

Les perspectives envisageables pour étendre nos travaux sont nombreuses. Cette étude ouvre aussi la place à de nouvelles pistes qui pourront être explorées à court terme. Par exemple, dans le monde réel, beaucoup de pro-

priétés d'entités sont non fonctionnelles, c'est-à-dire que plusieurs valeurs vraies attendues existent. Donc, il serait bien de modifier les modèles proposés pour faire face à ce genre de situations. Aussi, plusieurs propriétés de valeur vraie sont dynamiques, c'est-à-dire qu'elles changent au cours du temps. Par exemple, l'affirmation «Le président des États-Unis est Donald Trump» est actuellement vraie, mais à un certain moment dans le futur, elle sera fausse. Une autre considération importante est que le Web sémantique est basé sur l'assomption du monde ouvert, donc il est important de prendre en compte cette hypothèse lors de l'extraction de la connaissance *a priori*. À l'avenir, nous prévoyons de modifier la génération de l'ordre partiel des valeurs en faisant la distinction entre une relation qui n'existe pas et une relation inconnue. L'idée est d'utiliser l'information de disjonction entre les concepts pour détecter quand une relation n'existe pas de manière sûre. Dans tous les autres cas, une relation pourrait simplement être inconnue. Lorsque c'est le cas, un support faible peut être propagé entre les valeurs qui partagent cette relation. Dans les expérimentations menées, nous avons toujours considéré les connaissances *a priori* exprimées au sein d'une seule ontologie. Considérer plusieurs ontologies pour extraire des connaissances *a priori* pourrait augmenter la probabilité d'identifier des dépendances entre les valeurs et les autres éléments de données.

En tenant compte de la modélisation de la connaissance, cette thèse contribue à l'évolution du domaine de la recherche de vérité et plus généralement de la véracité des données. Nous espérons que ce travail et les perspectives qui en résultent, inspireront d'autres travaux et seront à l'origine de nouvelles idées. Au cours de ces années de recherche, il nous a semblé essentiel de sensibiliser à cette thématique. Cette prise de conscience face aux flots d'information qui sont produits, et publiés en particulier sur le Web est très importante. Chacun doit être vigilant, tant que des solutions efficaces ne sont pas mises en place pour garantir de la qualité du savoir collectif. Mais gageons que ces dernières soient bientôt proposées et appliquées à grande échelle.

Chapter 1

Introduction

Contents

1.1 General context

The digital revolution has highly impacted today's society through the development of microprocessors, computers, internet, smart-phones and network technologies (Perez, 2010). These advances enable to easily generate, process and disseminate any kind of information. In recent years, the pervasiveness of these technologies makes even possible to create an unprecedented amount of information. As a result, nowadays people can potentially exploit all of it.

At the beginning, the Web was a medium dedicated to share documents where users were mainly allowed to search and read resources. At that time, the information was limited and users could check information veracity by themselves. With the evolution of the Web, available information has increased. New technologies and tools have been designed to support collaboration among users and gather collective intelligence[1]. The main advantage of these technologies and tools is that they enabled creation and diffusion

[1]Wisdom that emerges from a group. It is a kind of knowledge that does not exist on the individual level.

of content without necessity of any external control. In this context, social networks, blogs and wikis had their growth. As a consequence, the amount of information on the Web has exploded. Statistics show that currently Facebook users create 3.3 millions of posts each 60 seconds[2]. Moreover, the same statistics indicate that there are 3.8 millions of Google searches each minute. This means that users both create and access online content compulsively. Whilst users can access and use this great amount of information to potentially satisfy their needs, they are not able to easily benefit from it. Indeed, users have difficulty to criticize this information.

The problem of information overload and information veracity

When assessing online information, two major problems arise for users. First of all, users have to discern relevant information according to their needs, i.e. *information overload problem*. Search engines support users trying to solve this issue. Second, once relevant information is identified, users need to *evaluate its veracity*. The lack of control over what gets published online can often lead to dissemination of unreliable information. Therefore, a best practice is to check the reliability of the collected information.

Humans and confirmation bias

Evaluating the reliability of information is a critical task even for humans. Psychological studies show that human judgement is often biased (Plous, 1993). When looking for information, humans tend to find, read and accept information that is in accordance with their viewpoints and beliefs, and to reject information that is not. For instance, imagine that a person holds a belief that introverts are more creative than extroverts. Whether this person meets an individual that is both introverted and creative, this person will give a high importance to this evidence that supports his/her beliefs. On the contrary, this person will highly discount evidence against his/her beliefs. This phenomenon is called *confirmation bias* and plays a pivotal role in decision-making processes (Nickerson, 1998). It may influence the decision people make, leading to poor or faulty choices. Indeed, it may prevent people from evaluating objectively the reliability of information. Nowadays, confirmation bias is empathised on the Web by the filter bubble phenomenon (Pariser, 2011). An increasing number of search engines perform personalized search based on the user preferences, user click behaviour, user browsing history and so on. The returned results foster confirmation bias since

[2]Statistics published in February 2017 and available at `https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/`

these results are usually in accordance with user beliefs. Indeed, these beliefs are captured by the user profile that is used by the personalized Web search. In addition, to complicate information evaluation, *confirmation bias* is also the basis for spreading fake news (Lazer et al., 2018). This kind of news are usually created to manipulate public opinion, political motivation, incite mass protest and so on. Fake news usually embrace crowd will in order to be credible. Confirming people beliefs, fake news have a high probability to be considered true.

Humans and time-consuming tasks

Evaluating information reliability is even more critical considering the amount of information available today. It is time-consuming if not impossible for a human to read through this large amount of information and identify reliable data. Even reputed editors and journalists face significant challenges in pursuing the good practice of moderating and verifying news before publishing them. With more and more news content created online, this editorial oversight is often absent. As a consequence, influential newspapers have started warning readers about the possibility of finding false news on the Web, and provide them with tools to help them assessing and questioning the veracity of news. As an example of an active fact-checking project, the French daily "Le Monde" launched *Les décodeurs*[3] in conjunction with the beginning of a French political campaign in 2017. *Les décodeurs* is a fact-checking service that aims to highlight false information and provide, through a specific tool, i.e. Décodex, public information about website reliability to help people obtaining trustworthy information. This service is maintained by a team of journalists; the computational power of computers is clearly necessary to apply such a service at Web scale. Other popular projects in this domain are mentioned in the Duke University Reporters' Lab database[4]. Researchers of this laboratory also report that a growing number of initiatives have promoted fact checking[5] in recent years (64 projects in 2015, 114 in 2017). The most famous actions are *PolitiFact*, *PundiFact*, *Snopes* and *FullFact*. Moreover, big companies such as Facebook and Google have started to take advantage of these services by generating fake alerts on news

[3] `www.lemonde.fr/les-decodeurs` [Accessed:2017-06].

[4] `www.reporterslab.org/fact-checking` [Accessed:2017-06].

[5] `www.reporterslab.org/international-fact-checking-gains-ground` [Accessed:2017-06].

that are disputed by fact-checkers[6,7].

The growing interest in Data Veracity is related to the possibility of exploiting its results in several tasks

The creation of an increasing number of applications based on Web data has made the problem of obtaining reliable information even more remarkable. For instance, Information Retrieval (IR) systems need to rank information based on their degree of reliability. The higher its reliability, the higher its position in the rank. Other applications are also affected by the reliability of information they use. For example, business intelligence applications, as well as all domain-specific applications that are specifically built to address a particular range of problems within a specific domain (Eiermann et al., 2010; Rocha, Zucker, & Giboin, 2018). In these cases, using unreliable information will lead to erroneous decisions that will negatively impact the application performances. Due to the wide range of applications that can benefit from reliable information, a lot of research is still necessary to automatize information processing aimed at evaluating information reliability.

1.2 Thesis contributions

Among the different types of information whose reliability can be criticized, this thesis addresses the problem of data veracity of factual claims. A factual claim, related to an entity property, is a statement whose veracity is unknown and that can be verified in an unambiguous way. For instance, a statement indicating the height of a mountain is a factual claim. It can be confirmed or not measuring it with a dedicated instrument. Moreover, in this thesis, we focus on the analysis of functional properties for which a single true value exists. When the factual claim is true, it is called fact[8].

Prior knowledge provides useful resources that can facilitate data veracity

To tackle the problem of factual claim veracity, we propose to use *a priori* knowledge contained in ontology to strengthen Truth Discovery models. Truth Discovery approaches evaluate the veracity of factual claims comparing information provided by different sources, eventually resolving conflicts that may arise. Their idea is to benefit from the abundance of available information. The problem is that, when aggregating information from multiple

[6] `www.independent.co.uk/voices/facebook-fake-news-fact-check-google-ad-save-journalism-a7645706.html` [Accessed:2017-06].

[7] `www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world` [Accessed:2017-06].

[8] Some previous studies have called a generic *claim* with the term *fact*, but we prefer to use an unambiguous terminology in this manuscript.

sources, conflicts may occur. Indeed, different values may be provided for the same property of a real-world entity. Truth discovery models are able to figure out which are the true claims among the conflicting ones using the following assumption. True information is provided by reliable sources and reliable sources provide true information. Therefore, both the reliability of claims and reliability of sources are very important aspects for these models.

More precisely, we study ontology-based approaches to enhance Truth Discovery performance. The general assumption is that injecting additional knowledge into Truth Discovery models could make them more effective. We show how semantics expressed by reliable Linked Data can be exploited to identify true information and reliable source of information. Linked Data is a result of the knowledge reuse principle that has been introduced by Semantic Web. Linked Data consists of a set of datasets and vocabulary, including ontologies, that are interconnected.

Three main contributions can be distinguished:

First contribution. As additional knowledge, we decide to consider semantic dependencies that may exist between provided values to improve Truth Discovery models. Two values that completely differ in their syntax may not differ in their semantics. For instance, several properties can be described with different level of granularities. In such a case, a true value can be also represented by its generalizations. This means that, also in case of functional predicate, considering true a value does not imply that the others are systematically false, i.e. multiple true values may exist. This is an important consideration that makes not trivial incorporating this knowledge into existing Truth Discovery models. The semantic dependencies between values are modelled using value partial order that can be extracted from an existing reliable ontology. This relation indicates when a value is subsumed by another one. This enables to propagate the reliability associated with a claim to its generalizations. Moreover, these semantic dependencies are also adopted to identify the most true specification among the set of true claims that are identified by the proposed approach for a certain functional property. The first contribution of this thesis is a method that exploits prior knowledge in the form of partial order among values within Truth Discovery process. To evaluate the proposed approaches several synthetic datasets have been generated. They contain conflicting claims and partial order relationships that

exist among values.

Second contribution. We also propose to exploit another type of knowledge that can be extracted from a reliable ontology. This time, it concerns the recurrent patterns that can be observed from available facts. The idea is that when real-world entities are similar they should have similar values for their properties. The approach we designed models recurrent patterns as rules. Rules enable us to easily establish when entities are similar (they share a subset of properties and the corresponding values). Then, rules also suggest a set of values that should be associated with a considered property. Indeed, inspired by Leibniz Law (Feldman, 1970), two entities are identical if and only if they share all and only the same properties. Finally, we also propose to combine rules with the partial order of values to further improve Truth Discovery performance. Evaluation of this second contribution has been done using the same synthetic datasets.

Third contribution. In order to evaluate the proposed approaches in a real-world scenario, we collected a real-world dataset. The analysis of this dataset coupled with the experimental results highlights the main limitations of the proposed approaches when they are applied in a real-world setting. Several improvements have been proposed to overcome the drawbacks we have identified.

Synthetic datasets, real-world datasets and source codes implementing the proposed approaches are open source, documented and freely available online.

1.3 Thesis outline

The remainder of this thesis is organised as follows.

Chapter 2 introduces the concepts of data veracity and knowledge representation. It first explores the state-of-the-art approaches of Truth Discovery highlighting that different types of *a priori* knowledge may be used to facilitate the task. Then, it moves to describe ontologies, one of the most effective examples of knowledge representation. This chapter ends with the proposition of using *a priori* knowledge represented by ontologies to enhance Truth Discovery performances.

Chapter 3 illustrates how *a priori* knowledge in the form of partial order of values can be integrated into existing Truth Discovery models. It formally introduces partial order. Then, it specifies several considerations that this integration implies making the proposed model not straightforward. This chapter also shows that the proposed model results to be effective on synthetic datasets.

Chapter 4 describes how *a priori* knowledge in the form of rules can also be useful for Truth Discovery. This chapter firstly introduces rules from a formal point of view. Then, it describes the method we proposed. Studies on synthetic datasets show that also this kind of *a priori* knowledge expressed into ontologies results to be useful.

Chapter 5 describes how real-world datasets have been collected. Then, it presents how the models proposed in the previous chapters behave on these real-world datasets.

Finally, chapter 6 summarises the main findings, identifies advantages and disadvantages of the proposed method and suggests future directions for further developments.

Chapter 2

State of the art: data veracity and knowledge modelling

Contents

This chapter begins with a general introduction to the problem of data veracity and puts a special emphasis on the evaluation of truth values of factual

claims. Among the different strategies that can be used to address this problem, we focus on Truth Discovery (TD) models. Presenting an overview of the approaches that have been proposed in this domain, we highlight that, to the best of our knowledge, existing models do not use prior knowledge expressed in ontologies to enhance TD performance. In this perspective, we introduce ontologies that are a formal paradigm that is used to model *a priori* knowledge.

2.1 Data veracity

Veracity is typically perceived as the truthfulness of any obtained information[1] to be consistent with reality. The Merriam-Webster dictionary[2] defines the term "veracity" as follows:

- conformity with truth or fact: accuracy
- devotion to the truth: truthfulness
- power of conveying or perceiving truth
- something true, e.g. makes lies sound like veracities

As confirmed by these definitions, veracity is a characteristic that is related to the truth. The term *truth* is usually too loaded with philosophical meaning (questions such as "Does an absolute truth exists?" often arise). Since any philosophical debate is out of the scope of this study, in the rest of the manuscript the term *truth* will be interpreted as the correspondence between an information and the reality. In order to avoid further ambiguities, it is important to highlight that the term *reality* refers to the actual state of facts given our knowledge of the world.

2.1.1 Data veracity in Computer Science

When dealing with information stored and processed on technological devices, the notion of data veracity is a relevant aspect. A lot of studies have been conducted at this regard in the database community. In recent years, data veracity is gaining importance also in other research communities such

[1] Data and information will be used interchangeably in this manuscript.
[2] https://www.merriam-webster.com/dictionary/veracity

Table 2.1: Examples of important aspects related to data veracity.

Dimension	Definition
Accuracy	percentage of data matching the knowledge we have of the real-world
Consistency	percentage of semantic rules violated
Freshness	time elapsed since data was created
Accessibility	degree to which data can be accessed
Minimality	percentage of data that does not contain redundancies
Completeness	percentage of real-world objects modelled in data
Provenance	origin of data
Volatility	frequency of change of data

as artificial intelligence and complex systems. Indeed, it is becoming a primary concern to ensure the effectiveness of numerous applications and services based on Web data. For their success, they require the consumption of data whose veracity has been verified.

Data veracity is a multidimensional problem

Data veracity is strictly related to inconsistency and data quality problems (Berti-Équille & Borge-Holthoefer, 2015). It is a multidimensional problem that depends on several aspects that can be associated with data. Indeed, high quality data corresponds to data that are consistent with reality and also easily accessible (accessibility), up-to-date (freshness), not repetitive, as much as possible exhaustive (completeness) and so on. These and other interesting characteristics that can be considered to evaluate data veracity and their definitions are reported in Table 2.1. The complexity of data veracity problem is mainly due to the fact that it is very difficult to consider all the different dimensions at the same time. The majority of the studies facing this problem only consider a limited subset of these attributes. The creation of this subset depends on the application scenario where data is used. For instance, data volatility is a primary concern during a surgery, but not during an annual check up. In the first case, patient parameters such as blood pressure may vary suddenly. Therefore they must be monitored frequently to provide doctors with high quality information. On the contrary, in the second case, it is highly improbable that bloody pressure changes frequently. Therefore, it is sufficient to monitor the bloody pressure only once to provide doctors with high quality information.

2.1.2 Data veracity as conformity to reality

In this thesis, we consider the aspect of veracity related to data accuracy, i.e. whether data conforms to real-world. Indeed, the final aim is to obtain a set of facts that may be used to populate a knowledge base. More formally, in this study, veracity is the *property that an assertion truthfully reflects the aspect it makes a statement about* (Krotofil, Larsen, & Gollmann, 2015).

Data veracity of factual claims

More precisely, this study focuses on the veracity assessment of factual claims. A factual claim is an assertion on a certain entity stating the value of a property whose true value can be verified with respect to reality in an unambiguous and objective way. For instance, the claim "Usain Bolt is 190 cm" is a factual claim since the actual height of an adult can be easily measured and verified using a measuring tape. Assessing the reliability of opinions, beliefs and impressions is out of the scope of this manuscript. If the reader is concerned about discovering opinion veracity, please refer to the study made by Samadi et al. (Samadi, Talukdar, Veloso, & Blum, 2016) where they explicitly intend to analyse non-factual claims, and the study of Wan et al. in which for the first time the expression "trustworthy opinion discovery" has been mentioned (Wan et al., 2016).

The truth value associated with a claim is either true or false according to bi-valence principle

Evaluating the veracity of factual claims consists of establishing their truth value with respect to reality. According to the principle of bi-valence (basic assumption of classical logic) introduced by Chrisippus, the truth value of a factual claim is either true or false when the claim, respectively, corresponds or not to reality[3]. Considering this basic assumption of classical logic, nothing may be 60% true and 40% false. Often, people make this kind of statements improperly. Indeed, their actual intention is to express the level of confidence they have on a truth value associated with a claim. This means that a claim is true/false and a confidence score can be associated with the evaluation of the truth value that is considered, i.e. the confidence that a claim is true is equal to 60%. Note that other types of logics, such as fuzzy logic, admit to associate degrees of being true and false with a statement in order to deal with the vagueness of natural language. However, in this study, we consider classical logic and its bi-valence principle.

[3]Be careful to distinguish truth value from true value. The truth value of a claim is true when the claim contains the correct (true) value for the considered property. It is false when the claim contains the erroneous (false) value for the considered property.

Moreover, we deal with functional properties whose values do not change over the time[4]. A property $p(x, y)$ is functional if for any x there is a unique y for which $p(x, y)$ is true. For instance the birth location of a person is a functional property, i.e. a person can be born only in a single location.

Functional predicate

2.1.3 Assessing data veracity

Several strategies can be applied to establish veracity of factual claims. For instance, content-based models use linguistic features that can be derived analysing the text where the claim appears, as well as its context (Nakashole & Mitchell, 2014). The use of an objective language should indicate that the claim is not a speculation or an opinion. Therefore, the probability of obtaining a true claim will be higher in the case that the text does not contain subjective and sentiment words. However, sometimes it is not possible to obtain the text from which a claim has been extracted. Alternatively, source-based models take advantage of source meta-information. For instance, they analyse source freshness, i.e. update frequency and last update time of a source. The higher the update frequency of a source is, the higher its reliability should be. Also, source graphical interface is a meta-information that may suggest the reliability of a source. The presence of a lot of advertisement usually is an indicator of unreliable source. Moreover, the clearance level needed to access the source is another aspect that can be considered. Source that required an authentication to be consulted should be more reliable. However, it is difficult to obtain source meta-information. Approaches that do not require context and source meta-information have been proposed. These alternative methods compare claims provided by multiple sources on the same subject to decide which claims are true and which ones are false (Bleiholder & Naumann, 2009; Y. Li et al., 2015; C. Li, Sheng, Jiang, & Li, 2016). Indeed, due to the amount of data available nowadays, it is likely that a topic is discussed by more than one source. Numerous scientific communities contribute to studying this problem, most notably data integration in information systems and databases (Berti-Équille & Borge-Holthoefer, 2015).

Text-based approaches

Source meta-information-based models

Comparing information provided by multiple sources

Data integration is a process that aims to fuse an information provided

[4]We therefore consider synchronic setting. It means that properties and corresponding values exist at one point in time without reference to the history of the considered world. This is the contrary of diachronic setting where changes between successive points in time are considered.

Data integration

by several sources in order to obtain a more complete and concise representation of available information (Berti-Équille & Borge-Holthoefer, 2015; Guzman-Arenas, Cuevas, & Jimenez, 2011; Knap, Michelfeit, & Necaskỳ, 2012; D. Wang, Abdelzaher, & Kaplan, 2015). It consists of three steps. First of all, schema mapping is performed. It identifies the correspondence among properties expressed by different sources in order to obtain a common representation of them. For instance, the *gender* of a person may be indicated also as *sex*. Then, *duplicate detection* is done. It determines which representations refer to the same real-world entity. Approaches that deal with both of these problems have been proposed in the fields of ontology alignment and instance matching (David, 2007; Euzenat & Shvaiko, 2013). Once representations have been aligned, sources may not agree on the predicate value of the same real-world entity (Anokhin & Motro, 2001). Resolving this inconsistencies is the goal of *data fusion*, i.e. the third step of data integration process that aims to fuse the different values into a single one. When inconsistencies occur at value level they are called value conflicts. They can be handled using several strategies. They can be ignored, avoided or resolved (X. L. Dong & Naumann, 2009):

- Conflict ignoring, this strategy does not include any action. Therefore, conflicts remain in the data. All decisions are left to end-users. This solution is not acceptable in the majority of application contexts.
- Conflict avoiding, this strategy consists of making decision on which predicate value to keep without any reasoning on the different values. In this case, pre-defined rules are specified. A straightforward solution is to remove all predicates having conflicting values. Alternatively, only values provided by certain sources can be considered as true. In this case, users have to express their preference in advance.
- Conflict resolving, this strategy consists of selecting the correct values among the provided alternatives, e.g. the most frequent ones, or proposing intermediate correct values resulting from the provided ones, e.g. average (Knap et al., 2012).

Conflict resolving is the most difficult alternative, although the most likely to be applied in real-world scenario where the data entry process can at most only partially be controlled. Considering this kind of strategy, the

most straightforward approach that can be used is voting. All sources are considered equally reliable (Guzman-Arenas et al., 2011; Y. Li et al., 2015) and, for each property, the value provided with the highest frequency is considered as the true one. This method is based on the "wisdom of the crowds" concept. Groups of people can be equally smarter than few experts (Surowiecki, 2004). This concept has proved to be powerful in several scenarios such as crowdsourcing (Smyth, Fayyad, Burl, Perona, & Baldi, 1995; Whitehill, Wu, Bergsma, Movellan, & Ruvolo, 2009). The problem of this baseline approach is that it is unable to deal with spam-based attacks, or duplicated errors that are common on the Web. In these cases, indeed, true information can be provided by few but reliable sources. This principle is called "wisdom of minority" (H. Li, Zhao, & Fuxman, 2014; Y. Li et al., 2015). Therefore, it is better to weigh sources based on their reliability. Indeed, reliable information should be provided by reliable sources (Berti-Équille & Borge-Holthoefer, 2015; D. Wang et al., 2015). When reliability of sources is not known *a priori*, it can be evaluated based on source reputation, e.g. inferred from network structure, or source content. Intuitively, well-reputed sources (hubs) should be reliable. In the same manner, if source content is true, then the corresponding source should be reliable. While source reputation information may not be available, source content should always be. Indeed, the aim of data veracity is to evaluate the veracity of claim content.
The idea behind using source content to evaluate source reliability is that reliable sources should provide reliable information with higher probability than unreliable one. Intuitively, the more reliable a source is, the more reliable the information it provides will be.

Solving conflicts with voting

Solving conflicts weighing each source differently

2.2 Truth Discovery

Summarising the considerations made in the previous section, veracity of claims can be evaluated based on source reliability and, in turn, source reliability can be evaluated on veracity of its claims.
This rationale is used by Truth Discovery (TD) to solve conflicts that may occur when multiple sources provide different values for a given property of a real-world entity (D. Wang et al., 2015), e.g. the place of birth of a person. Other research areas address the same problem using different names, e.g. truth-finding, information trustworthiness, information credibil-

ity, information corroboration, data fusion or fact-checking (Berti-Équille & Borge-Holthoefer, 2015). In this section, several application domains are described. Then, problem settings are formalized introducing all notations. An overview of the state-of-the-art approaches is also presented.

2.2.1 Application domains

During the last 10 years there has been an increasing interest for TD. Because of this, the effectiveness of TD has been tested in various fields of humans activity (or domains).

Social sensing. This domain aims at collecting huge amount of data from a large group of individuals. Indeed, individuals have become sensors since they started to use wearable devices and mobile phones. However, sensors that collect this data are error-prone. Thus the veracity of collected data has to be assessed. TD is used to this aim (Su et al., 2014; D. Wang, Kaplan, Le, & Abdelzaher, 2012).

Crowdsourcing It is the process of completing some tasks (such as answering a set of questions) by requiring participation of a large group of people, i.e. workers. Since workers are usually non-experts, errors may occur. TD models help to identify which are the most reliable workers and which are the correct information that they provide (Gao, Li, Zhao, Fan, & Han, 2015; Ouyang, Srivastava, Toniolo, & Norman, 2016).

Online health communities. The idea, here, is to distinguish reliable and unreliable information provided by users in health forums. The distinction is based on the use of TD models to estimate trustworthiness of users (Mukherjee, Weikum, & Danescu-Niculescu-Mizil, 2014).

Knowledge base population. Truth discovery models result to be useful to enhance the completeness of existing knowledge bases such as Google Knowledge Graph[5] and YAGO[6] (X. Dong et al., 2014; X. L. Dong et al., 2014, 2015). In this case, the information that can be added to existing knowledge

[5]http://www.google.com/insidesearch/features/search/knowledge.html.

[6]http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

bases is provided by multiple websites. TD models are applied to solve conflicts when they occur.

Slot filling validation. For each attribute, TD models can be used to solve conflicts that may occur among the outputs returned by different information extractions techniques (Yu et al., 2014).

In the rest of the manuscript, we will often consider Web-based scenarios as a key example. However, it is important to highlight that sources are not limited to websites, as shown by the different examples of application domains listed above. They can be persons, experts, trained systems and so on. The important point is that they provide information on the same topics. In this way, information can be compared in order to discriminate reliable and unreliable sources, as well as true and false information.

2.2.2 Basic elements of Truth Discovery

Given as input a set of conflicting claims, the aim of TD is to solve these conflicts returning as output a set of facts, see Figure 2.1. This result is obtained identifying the truth value of each claim. Note that, for TD, the Unique Name Assumption holds. This means that each real-world entity in claims is always identified with a unique name by all sources. Moreover, all values and properties are assumed to be disambiguated.
Formally, let D be a set of data items where each $d \in D$ refers to a property *predicate* of an entity *subject*. Let V be the set of values that can be assigned to these data items and S be a set of sources. Each source $s \in S$ can state a value $v \in V$ for a data item $d \in D$, hence providing a claim v_d where

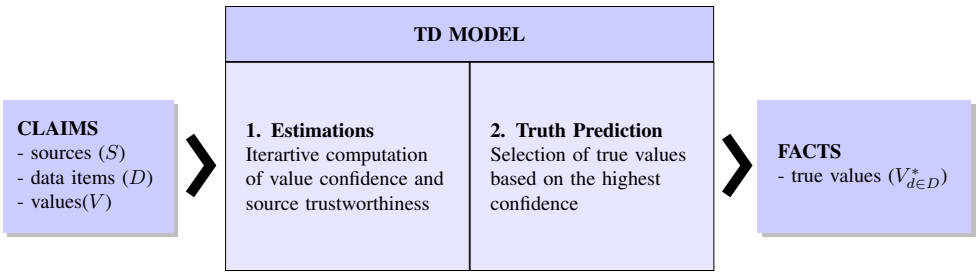


Figure 2.1: Overall Truth Discovery (TD) procedure.

V_d is the set of values stated for data item $d \in D$.[7] The truth value of a factual claim v_d can be defined, in accordance with the bivalent logic, as the following binary truth function tf:

$$tf : D \times V \rightarrow \{true, false\} \tag{2.1}$$

To establish the truth value of claims, TD methods use the assumption stated previously. To model this rationale they introduce the concepts of *source trustworthiness* and *value confidence*[8]. These concepts are commonly manipulated through the following functions:

- $t : S \rightarrow [0,1]$ the trustworthiness of a source: this value indicates, for each source, its propensity to provide facts (true claims). In the literature this notion is also sometimes called source weight (Y. Li et al., 2015) or source reliability. Intuitively it is assumed that a reliable source has a high trustworthiness and it is likely to provide accurate claims ($t(s) \simeq 1$). On the contrary, an unreliable source has a low trustworthiness ($t(s) \simeq 0$) and, consequently, it is likely to provide inaccurate claims.

- $c : V \rightarrow [0,1]$ the confidence of a value: denotes the propensity of a value contained in the claim of being correct. It is assumed that an accurate claim has a high confidence and it is likely to be provided by trustworthy sources ($c(v_d) \simeq 1$). On the other hand, an inaccurate claim has a low confidence and, consequently, it is likely to be provided by untrustworthy sources ($c(v_d) \simeq 0$).

Hence, *per* definition, source trustworthiness and value confidence are two functions highly correlated. Indeed, the confidence of a value is estimated based on the trustworthiness of the sources claiming it. This set of sources is represented by S_{v_d}. In the same way, the trustworthiness of a source is evaluated based on the confidence of the values this source claims. This set of claims is represented by V_s.

[7] A claim v_d can be seen as an RDF triple $\langle subject, predicate, object \rangle$ where $d = (subject, predicate)$ and $v = object$.

[8] Value and claim are used interchangeably. Indeed, when a value is evaluated, it is always assigned to a data item.

2.2.3 Problem setting

TD models are considered as unsupervised approaches. No labelled data on source trustworthiness and value confidences is needed. The only information they require is the correspondence between sources and claims (i.e. which are the claims provided by each source). Otherwise, the iterative reasoning cannot be applied. Moreover, they assume that the claims given as input are structured. Based on these preliminary considerations and on the study proposed in (Y. Li et al., 2015), we consider the following formal definition of the TD task:

Definition 2.1 (Truth Discovery) *Considering a set of data items D, a set of values V, a set of sources S, the first goal of Truth Discovery methods is to find for all $d \in D$, $V_d^* \subseteq V_d \subseteq V$, the true value set for d among the set of values that are claimed for this specific data item. Meanwhile, Truth Discovery methods estimate source trustworthiness $t(s)$ for each source $s \in S$.*

The notations used in the previous definition as well as in the rest of this manuscript are summarized in Table 2.2. Most of them are well admitted in the literature (Berti-Équille & Borge-Holthoefer, 2015; Y. Li et al., 2015; Waguih & Berti-Equille, 2014; Yin et al., 2008). New ones are introduced to present the proposed models.

Given this set of basic elements, TD models adopt an iterative technique to propagate source trustworthiness and value confidence. Many methods have been proposed since the seminal work of Yin et al. in 2008 (Berti-Équille

Table 2.2: Notations.

Notation	Definition
$d \in D$	the data item d composed of a pair $(object, predicate)$
$v \in V$	the value v
$V_d \subseteq V$	set of values provided for data item d
$v_d \in V_d$	the claim assigning the value v to the data item d
$V_d^* \subseteq V$	set of true values associated to data item d
$s \in S$	the source s
$V_s \subseteq V$	set of values provided by source s
$D_s \subseteq D$	set of data item for which source s provides a value
$S_{v_d} \subseteq S$	set of sources providing a specific claim (d, v)

& Borge-Holthoefer, 2015; Y. Li et al., 2015; Yin et al., 2008). They mainly differ in the way they compute confidence of claims and trustworthiness of sources. Some of them do not use any additional information, while others attempt to improve TD performance using external support such as extractor information (i.e. confidence associated with extracted triples), temporal dimension, hardness of claims, common sense reasoning or dependencies. Models that take relationships into account can be divided based on the kinds of dependencies they consider: source dependencies, value dependencies and data item dependencies.

2.2.4 Existing models

This section describes the state-of-the-art TD approaches. We distinguish existing models based on the type of claims they deal with, the number of true values that are expected for each data item, and the relationships that are considered to enhance the performance. Table 2.3 reports the state-of-the-art TD models indicating for each approach its characteristics. In this table, models are also ordered by publication year in order to observe the evolution of research focus over the time. It is clear that, currently, researchers are more interested by studying ad-hoc models to deal with non-functional predicates and exploiting dependency that may exist among data items. All the approaches are detailed based on the characterizing aspects in the rest of this section.

2.2.4.1 Input data

First of all, different types of values can be associated with predicates: continuous, categorical, or both of them. Continuous values are contained in numerical claims such as the ones on the population size of a city. Categorical data represents one or more text strings from a finite set of choices.

Categorical Data

Example of categorical data are as the occupation of a person or the authors of a book (Y. Li et al., 2016). The most of existing methods take categorical data as input (Galland et al., 2010; Ma et al., 2015; Ouyang, Srivastava, et al., 2016; Pasternack & Roth, 2010; D. Wang et al., 2012; Zhao et al., 2012).

Continuous Data

Approaches designed specifically for continuous data have also been proposed (Meng et al., 2015; Ouyang, Kaplan, et al., 2016; Zhao & Han, 2012). There are also models able to deal with both types. In this case, models use differ-

Table 2.3: Existing Truth Discovery models ordered by publication year. They are characterized based on type of claims they deal with (input data), truth cardinality for each data item and types of dependencies that are considered.

Model	Input Data		Output Cardinality		Dependencies		
	Continuous	Categorical	Single	Multi	Sources	Values	Data items
TruthFinder (Yin et al., 2008)	x	x	x			x	
Accu (X. L. Dong et al., 2009a)	x	x	x		x		
AccuSim (X. L. Dong et al., 2009a)	x	x	x		x	x	
Sums, Investment, PooledInvestment (Pasternack & Roth, 2010)		x	x				
2-Estimates, 3-Estimates (Galland et al., 2010)		x	x				
Semi-Supervised Truth Finding[a] (SSTF) (Yin & Tan, 2011)	x	x	x			x	
Latent Truth Model (LTM) (Zhao et al., 2012)		x		x			
Gaussian Truth Model (GTM) (Zhao & Han, 2012)	x		x				
Maximum Likelihood Estimation (MLE) (D. Wang et al., 2012)		x		x			
Latent Credibility Analysis (LCA) (Pasternack & Roth, 2013)	x	x	x				
Multi-Source Sensing (MSS) (Qi et al., 2013)		x	x		x		
Confidence-Aware Truth Discovery (CATD) (Q. Li, Li, Gao, Su, et al., 2014)	x	x	x				
Conflict Resolution on Heterogeneous data[b] (CRH) (Q. Li, Li, Gao, Zhao, et al., 2014)	x	x	x				
PrecRecCorr (Pochampally et al., 2014)		x	x		x		
TD_corr (Meng et al., 2015)	x		x				x
EM_Cat (S. Wang et al., 2015)		x		x			x
Multi-truth Bayesian Model (MBM) (X. Wang et al., 2015)		x		x	x		
FaitCrowd (Ma et al., 2015)		x	x				x
Quantitative Truth Finder (QTF) (Ouyang, Kaplan, Toniolo, Srivastava, & Norman, 2016)	x		x				x
Empowering Truth Discovery (X. Wang et al., 2016)		x		x		x	

[a] It requires labelled data

[b] It handle hetereougeous data with jointly estimations

ent types of distance function to capture the features of different data types (Pasternack & Roth, 2013; Pochampally et al., 2014; Yin et al., 2008; Yin & Tan, 2011). Note that no TD approach requires labelled data. The only exception is the Semi-Supervised Truth Finding (SSTF) model (Yin & Tan, 2011). This approach tries to use labelled data to improve value confidence and source trustworthiness estimations.

2.2.4.2 True value cardinality

The general output of any TD method is commonly the truth label and confidence score of each valued claim, and the trustworthiness score of each source. Existing methods can differ in the cardinality of the true values that are considered for each data item. The majority of existing approaches consider a single truth for each data item (case of functional predicate). In this case, for each data item, the value having the highest confidence is selected as truth. Only a small subset of extended approaches deals with non-functional predicates for which multi-truth values could exist for each claim (Pochampally et al., 2014; X. Wang et al., 2015; Zhao et al., 2012) – i.e. values like *Málaga* and *Granada* are true values for the following data item $(Spain, hasCity)$. Different strategies have been developed to deal with non-functional predicate. Some of these approaches identify the set of true values by computing precision and recall assuming that a source can provide more than one claim for each predicate. In these models, no prior knowledge related to the expected true values is generally employed. Moreover, each claim is transformed into a binary claim. Other approaches instead try to establish the number of expected true values during the estimation phase (X. Wang et al., 2016). When the convergence is reached, the approach selects the first top-k values having the highest confidence as solutions. All proposed models assume there is at least one source providing the true value. Only one approach assumes the non existence of the true value. An extra value "unknown" is added for each data item. The idea is that when the relations among values are represented as constraints, these constraints may contradict each other. Thus when no solution (truths for all the objects) can satisfy all the constraints, "unknown" is returned.

Single-truth

Multi-truth

2.2.5 Relationship-based approaches

Among existing models, there are approaches that do not contain any additional information. Example of these approaches are *Sums*, Investement and PooledInvestement (Pasternack & Roth, 2010), and 2-Estimates and 3-Estimates (Galland et al., 2010). They iteratively update the estimated source reliability and value confidence. Then, they aggregate the results. They are differentiated by the various formulas they use. For example, *Sums* (Pasternack & Roth, 2010), inspired by Hubs and Authorities algorithm (Kleinberg, 1999), computes source trustworthiness and value confidence using the following formulas at each iteration i:

$$t^i(s) = \frac{\sum_{v_d \in V_s} c^{i-1}(v_d)}{max_{s' \in S} \left(\sum_{v_d \in V_{s'}} c^{i-1}(v_d) \right)} \tag{2.2}$$

$$c^i(v_d) = \frac{\sum_{s \in S_{v_d}} t^i(s)}{max_{v'_d \in V} \left(\sum_{v'_d \in S_{v'_d}} t^i(s) \right)} \tag{2.3}$$

where i represents the iteration number (not a power factor). The trustworthiness of a source $s \in S$ is evaluated summing up all confidences of claims that are provided by s. Similarly, the confidence of a value $v_d \in V_d$ for a given data item $d \in D$ is computed summing up all trustworthiness of sources that claim v_d. Both denominators represent normalization factors ensuring that trustworthiness and confidence scores range between 0 and 1. A uniform distribution was chosen as prior to initialize $c(v_d)$ with $v_d \in V$. More precisely, value confidences were set to 0.5. Moreover, the maximum number of iterations was fixed at 20 because empirical experiments showed that it was enough.

As shown by the listed formulas, no prior knowledge is used. *Sums* and the other models of this kind make several assumptions that enable them not to consider any additional information:

- context from which claims are extracted is completely ignored;
- existence of an accurate information extraction system to extract structured claims from the sources;

- sources are independent;
- values are independent;
- data items are independent;
- unreliable sources do not make the same errors, leading to different false claims;
- optimistic scenario: there are more reliable sources than unreliable ones;
- a single true value exists for each claim;
- truth is static.

Several TD ameliorations have been proposed to enhance the final performance relaxing different assumptions. They incorporate additional aspects beyond source trustworthiness and claim confidence. For instance, they may estimate source reputation via trust assessment to better initialize source trustworthiness. The rationale is that the higher the reputation of a source, the higher the prior on its trustworthiness. Other models may consider *common sense knowledge* to better compute the estimations. For instance, it is commonly recognized that the population size of a city ranges between 1500 and 50000 inhabitants. Thus, when evaluating claims on city population size, claims stating values that are lower than 1500 should be penalized. Alternative approaches may model the concept of evolving truth considering the temporal dimension associated with claims. For instance, the true value associated with claims on the American President changes over the time. At the time of writing, Donald Trump is the American President. This claim will be false in 20 years. Other ones consider the difficulty of knowing the true value of certain claims. Indeed the veracity of some claims is sometimes more difficult to estimate. Knowing the population size of New York is easier than knowing the population size of Lambertville, i.e. a small village in New Jersey. Sources stating the true value for difficult claims should be rewarded.

Among all of these enhanced methods, we focus our attention on those that use, as additional knowledge, dependencies that may exist among sources (Pochampally et al., 2014; Qi et al., 2013; X. Wang et al., 2015), among data items (Meng et al., 2015; D. Wang et al., 2015; S. Wang et al., 2015) or among

values (X. L. Dong et al., 2009a; Yin et al., 2008). A dependence is a relationship that exists between two entities, whether they are sources, values or data items, when an entity influences the other. The models that take dependencies into account are called relationship-based approaches. We can distinguish three main classes based on the type of entities that are considered. In each class, the nature of dependence varies based on the aspects that is analysed to identify dependencies. For instance, when considering sources, the number of overlapping claims is often considered to identify their dependence. Differently, when considering data items, spatial relations may be considered. In the rest of this section, we describe some of the aspects that can be considered to identify dependencies among sources, data items and values.

2.2.5.1 Source relationships

The definition of false claims is not restricted to the ones that are false by accident but includes also claims that are false on purpose. Therefore, models that take this distinction into account have been proposed. In real-world, sources may be not independent of each other . Indeed, malicious sources may copy from other sources in order to consciously spread false values. For instance, during the American political campaign in 2010, Donald Trump said that Obama was born in Kenya (Allcott & Gentzkow, 2017). This fake news were spread by a lot of Republicans in order to discourage Obama supporter. Taking source dependencies into account can reduce the risk of overestimating claims made by dependent sources, especially when these claims are false. As first attempt to fix this problem, a dampening factor have been modelled in TruthFinder (Yin et al., 2008), see Section 2.2.5.3. It aims at compensating the high confidence assigned to claims that have been copied by a source from another. The main limitation of this factor is that it does not vary based on sources providing a claim under examination. It is fixed *a priori* and it is the same for all sources. A more refined study proposes to identify source dependence characterizing each source by its precision (probability that a source provides a true value) and recall (probability of true values to be provided by a source) (Pochampally et al., 2014). Other studies propose an alternative method (X. L. Dong et al., 2009a). They usually assume that sources sharing common false values are more likely to

Copying sources exist in real-world setting

be dependent than sources sharing common true values. Indeed, it is difficult to identify a dependence between sources stating different false values (Berti-Equille et al., 2009). Another signal indicating a copying relation is when a source has a trustworthiness evaluated on a subset of claims it share with another source that is significantly different from its trustworthiness evaluated on the rest of the claims. Several methods such as DEPEN and its extensions take advantage of source dependence (X. L. Dong et al., 2009a; Pochampally et al., 2014; Qi et al., 2013; X. Wang et al., 2015). DEPEN is a voting method that penalizes the weight of a source if the source results to be a copier of another one. It compute the confidence of a value as:

DEPEN model

$$c^i(v_d) = \sum_{s \in S_{v_d}} IND^i(s) \tag{2.4}$$

where i the iteration number and $IND^i(s)$ the probability that source s provides v_d independently from any other source that has already been considered during the computation. The set of sources that have already been considered, before s, is denoted $pre(s)$. Note that, when relaxing the assumption for which the sources have the same trustworthiness (ACCU model), the source trustworthiness is computed as $t^i(s) = \sum_{v_d \in V_s} c^i(v_d)/|V_s|$, the value confidence formula becomes:

ACCU model

$$c^i(v_d) = \sum_{s \in S_{v_d}} \left(\ln \frac{n t^{i-1}(s)}{1 - t^{i-1}(s)} \right) IND^i(s) \tag{2.5}$$

with $n = |V_d| - 1$ the number of false values provided for data item d when considering functional predicates. This formula weights each source based on its trustworthiness level (initialized to 0.8), i.e. first factor of the equation, penalizing sources that copy from others, i.e. second factor. $IND^i(s)$ is computed using the following formula:

$$IND^i(s) = \prod_{s' \in pre(s)} (1 - uP^i(s \sim s')) \tag{2.6}$$

with u the probability that a value provided by a copier is copied and $P^i(s \sim s') = P^i(s \to s') + P^i(s' \to s)$ the probability of the two sources of being dependent. To ease the reading, in the next formulas, we omit the superscript specifying the iteration number. The probability that s copies from s' is:

$$P(s \to s') = \frac{\alpha Dep(s, s')}{\alpha Dep(s, s') + \alpha Dep(s', s) + (1 - 2\alpha) Indep(s, s')} \tag{2.7}$$

with α prior probability of s and s' to be dependant. Two sources are independent when:

- they both provide a true value independently, i.e. $p_t = t(s)t(s')$;
- they both provide a false value independently, i.e. $p_f = \frac{(1-t(s))(1-t(s'))}{n}$ with n the number of false values;
- they provide different values independently, i.e. $p_{diff} = 1 - p_t - p_f$.

Therefore the probability of two sources to be independent is

$$indep(s, s') = p_t^{|V_d^{true}|} \; p_f^{|V_d^{false}|} \; p_{diff}^{|V_d^{diff}|} \tag{2.8}$$

where, considering the current iteration, V_d^{true} is the set of common true values provided by sources, V_d^{false} is the set of common false value and V_d^{diff} is the set of different values.
In addition, two sources are dependent when:

- s provides a true value copying from s' or s provides a true value independently, i.e. $dep\_p_t = ut(s) + (1-u)p_t$;
- s provides a false value copying from s' or s provides a false value independently, i.e. $dep\_p_f = u(1-t(s)) + (1-u)p_f$;
- s provides a different value independently, i.e. $dep\_p_{diff} = (1-u)p_{diff}$

Therefore the probability of two sources to be independent is

$$dep(s, s') = dep\_p_t^{|V_d^{true}|} \; dep\_p_f^{|V_d^{false}|} \; dep\_p_{diff}^{|V_d^{diff}|} \tag{2.9}$$

Source dependence in dynamic setting

As the model just reported, most of the studies focusing on source dependencies only analyse static correlations. To the best of our knowledge, time-course dependency relationship patterns has been only considered in (X. L. Dong et al., 2009b). In this case, dependence among sources is captured by studying the similarity between patterns of updates associated with sources (Berti-Equille et al., 2009).

Source dependence at group level

Moreover, the models just presented consider dependencies between pairs of sources. This implies that if many dependent sources repeat the same false claims, estimations may be overestimated. For this reason, MSS (Multi-Source Sensing) (Qi et al., 2013) identifies groups of dependent sources

through a graphical model. Since groups are not equally reliable, it is important to estimate the trustworthiness of each group, and minimize the negative impact of unreliable sources. Each group is characterised by a general trustworthiness and data-item specific trustworthiness. General trustworthiness measures the overall performance of a group by aggregating each specific trustworthiness over the entire set of data items. The data item-specific trustworthiness is estimated considering that a generally reliable group is likely to be reliable on an specific data item and *vice versa*.

2.2.5.2 Data item dependencies

The second relationship class is related to data items. The first body of works in this context proposed to deal with the social sensing problem. In crowd sensing, humans coupled with their smartphones become sensors that explicitly or implicitly provide observations about their physical environment. Then it becomes necessary to understand the validity of data sent by sensors. TD models applied to this domain take advantage of both physical (D. Wang et al., 2013) and temporal (S. Wang, Wang, Su, Kaplan, & Abdelzaher, 2014) correlations as well as causal relationships (S. Wang et al., 2015). For physical correlations, they assume that co-located data items should have similar values. For instance, gas stations located in the same area should have similar gas prices. For temporal correlations, the assumption is that two temporally close observations cannot have very different values. This kind of correlation is especially useful when analysed data has a long-tail characteristic, i.e. many data items observed by few sources and few data items observed by many sources. Indeed, in this case the estimations can easily deteriorate if the few sources that provide claims for a data item are also unreliable. Using correlations, information associated with data items having a high number of observations provided by reliable sources can be propagated to data items having only a few claims associated with them. The findings of the two studies (D. Wang et al., 2013) and (S. Wang et al., 2014) permit to partition data items into small groups without considering any dependence among groups, but the complexity of their solutions is exponential w.r.t. (with respect to) the maximum group size. Alternative models have been proposed to overcome this limitation to be able to deal with a large number of dependencies, e.g. (Meng et al., 2015; S. Wang et al., 2015). The former approach, called TD_corr,

Physical and temporal dependence

regards the problem as an optimization problem. Considering continuous claims, it aims to minimize the difference between truths[9] and claims provided by reliable sources knowing that dependent data items should have similar true values. In order to exploit data item dependencies, this model identifies independent data items. It splits the entire data item set D into a disjoint sets $\{\mathcal{D}', \cdots, \mathcal{D}^{n-th}\}$, with n the number of disjoint sets, such as $D = \cup_{D' \subset D} \mathcal{D}'$. Each disjoint set contains data items that are independent with each other. The idea is that each claim is influenced by all dependent data items belonging to the other disjoint set. First of all, the model computes source trustworthiness using a formula that rewards source providing values similar to the truth. The lower the distance between provided values and truths, the higher the trustworthiness of a source is:

TD_corr model

$$t^i(s) = -\log\left(\frac{\sum_{\mathcal{D}' \subset D} \sum_{v_d \in D' \cap D_s} ||v_d^{*i-1} - v_d||^2}{\sum_{s' \in S} \sum_{D' \subset D} \sum_{v_d \in D' \cap D_{s'}} ||v_d^{*i-1} - v_d||^2}\right) \tag{2.10}$$

with $\mathcal{D}'$ a set of independent data items. Then, the model computes the true values for each data item by averaging over the claimed values weighted by trustworthiness of sources providing them, and the true values of dependent data items weighted by the similarity between them and the considered data item.

$$v_d^{*i} = \frac{\sum_{v_d \in V_d} \sum_{s \in S_{v_d}} t^i(s) v_d + \alpha \sum_{d' \in corr(d)} sim(d, d') v_{d'}^{*i}}{\sum_{s \in S} t(s) + \alpha \sum_{d' \in corr(d)} sim_d(d, d')} \tag{2.11}$$

with $sim_d(d, d')$ a function that returns the similarity degree between data items and $corr(d)$ a function that return the set of data items that are dependant with d. Note that $\forall d \in \mathcal{D}' \subset D, \forall d' \in corr(d), d' \notin \mathcal{D}'$. Moreover, the dependencies among data items are defined *a priori*. Considering the previous example, the city where a gas station is located can be used to detect dependencies among different gas stations.

Another approach that consider data item dependence is EM_cat (S. Wang et al., 2015). It models the problem as a Bayesian network exploiting potential conditional independence among data items. The main limitation of this approach is that the Bayesian network has to be known or empirically learned

Causal dependence

[9] In this case, the true value is directly estimated. Its confidence is implicitly considered weighting a value for the trustworthiness of sources providing it.

from historical data by specific algorithms.

Category-based dependence

Data item correlations can be also expressed by categories to which a data item belongs. Using this kind of data item dependencies, the assumption stating that source reliability is consistent across different data items can be relaxed. Source reliability may varies based on data item categories. For instance, a doctor that is not interested by car engines should be more reliable when providing information on drugs than when providing claims on cars. FaitCrowd using this rationale accurately estimates source reliability for each topic when expertise differs w.r.t. the topics (Ma et al., 2015).

2.2.5.3 Value dependencies

There are approaches that consider values independent. For these models, no relation exists between values. Given this consideration the complementary vote may be applied. If a source provides a value, all the other values are considered false. Other models relax the value independence assumption. In this case, two dependant values should support each other. If one of them is considered true, then the other one has a high probability to be true as well. Previous studies use similarity among values to identify value dependencies. Examples of these approaches are TruthFinder (Yin et al., 2008), AccuSim (X. L. Dong et al., 2009a), and SSTF (Yin & Tan, 2011). They compute value dependence based on the edit distance of strings, similarity among sets, or difference among numerical values. For instance, TruthFinder

TruthFinder model

(Yin et al., 2008) takes value dependencies into account defining $\sigma^{*^i}(v_d)$ as a function that aggregates the basic confidence $\sigma^i(v_d)$ associated with v_d and the support provided by other values to v_d. More precisely, $\sigma^i(v_d)$ and $\sigma^{*^i}(v_d)$ are computed as

$$\sigma^i(v_d) = -\sum_{s \in S_{v_d}} ln(1 - t^{i-1}(s)) \tag{2.12}$$

$$\sigma^{*^i}(v_d) = \sigma^i(v_d) + \rho \sum_{v'_d \in V_d} \sigma^i(v'_d)\ sim(v'_d, v_d) \tag{2.13}$$

with $1 - t^{i-1}(s)$ the probability that a value provided by s is false (trustworthiness was initialized to 0.8 for each source), and $\rho \in [0, 1]$ a parameter that controls the influence of related values when computing the confidence score. This influence is evaluated by $sim(v'_d, v_d)$ that returns a score included

in $[-1, 1]$. Positive (negative) scores indicate that v'_d is supporting v_d (conflicting with v_d). The authors of TruthFinder specify that this function is domain dependant. They propose an example of this function that indicates the support given by v'_d to v_d considering claims on authors of books. Given v_d with x authors, v'_d with y authors and z the number of shared ones, $sim(v'_d, v_d) = z/x - base\_sim$ with $base\_sim = 0.5$ a threshold for positive implication between values. Alternatively, they suggest to use edit distance among strings. In any case, negative scores are admitted by $sim(v'_d, v_d)$. A logistic function is thus required to ensure positive confidences. The final formulas to evaluate value confidence and source trustworthiness are:

$$c^i(v_d) = \frac{1}{1 + e^{-\gamma \sigma^{*i}(v_d)}} \tag{2.14}$$

$$t^i(s) = \frac{\sum_{v_d \in V_s} c^i(v_d)}{|V_s|} \tag{2.15}$$

with $\gamma \in (0, 1)$ a dampening factor that try to compensate error propagation of dependant sources. As stopping criteria, TruthFinder checks cosine similarity of source trustworthiness between two successive iterations to be less than or equal to a given threshold. For each data item, the value having the highest confidence is then selected as true value.

Using ontologies to enhance TD models

To the best of our knowledge, no existing work takes advantage of *a priori* knowledge expressed in ontologies to detect relationships between entities and use them to enhance TD models. Using ontologies, TD has the advantage of bridging semantic gaps between provided data. Ontologies are able to encode formal semantics in a well structured fashion which is easy for the machine to read and process. In the next section, we describe the main elements of ontologies to better understand how they can model *a priori* knowledge that can be exploited by several tasks.

2.3 Knowledge modelling

Knowledge Representation

Obtaining machine-readable and machine-understandable knowledge has been extensively addressed by previous studies in Knowledge Representation (KR). KR is a sub-field of Artificial Intelligence (AI) whose aim is to model knowledge in a computer tractable form (Guarino et al., 2009). In this way, AI agents can automatically use this knowledge for reasoning purposes.

The idea is to formally and rigorously define expressions whose semantics is not ambiguous of structured and controlled vocabularies. Therefore, a KR is characterized by the following aspects:

- a *vocabulary* indicating the components of this language;
- a *syntax* defining which configurations of the components of the language are valid;
- a *semantics* specifying which facts in the world the sentences refer to.

For instance, assuming the arithmetic as KR language, we can characterized it with these aspects. Components of this language are $x, y, \leq$. The arithmetic syntax specifies that a valid statement is $x \leq y$, but not $x\ y \leq\leq$. Then, the arithmetic semantics consider that $x \leq y$ is true if and only if x is lower than or equal to y. An example of well-defined KR is given by ontologies that are introduced hereinafter.

2.3.1 Ontologies

Several formal definitions have been proposed for the term ontology. In AI, the most popular one states that "*an ontology is an explicit specification of a conceptualization*" (Gruber, 1993). This definition captures several key aspects of an ontology. First of all, the expression *explicit specification* highlights the fact that all the knowledge has to be expressed, i.e. specified, in a machine-readable format. Notions that are not clearly stated are not known by machines (also common sense notions that are taken for granted by humans). Then, the term *conceptualization* indicates "*an abstract, simplified view of the world that the ontology wants to represent for some purpose*" (Gruber, 1993). A conceptualization has to be shared and agreed by people of a community. A conceptualization is an abstract entity that only exists in their mind. To communicate and share it, it must be captured by some concrete artefact. A shared language is thus necessary to represent it in a concise, complete and unambiguous way. As a result, each concept is represented by a symbol that refers to a certain real-world view. The relations among conceptualization, reality and language are well-explained by the semiotic triangle of Richard and Ogden (Richards & Ogden, 1923). As reported in Figure 2.2, the word "house" refers to specific occurrences of houses in the real-world. When you

Ontology definition

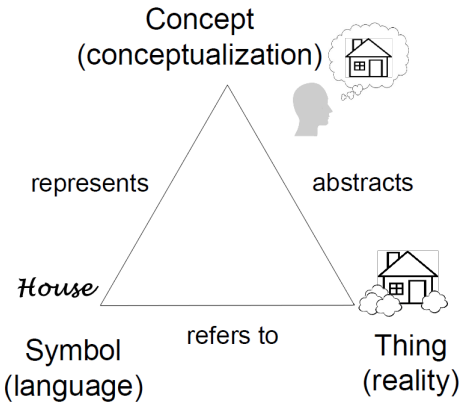


Figure 2.2: The semiotic triangle of Richard and Ogden depicts the relations between a thing in reality, its conceptualization and a symbolic representation of this conceptualization (Richards & Ogden, 1923).

say to someone the word house, this term invokes, in his mind, the idea of house that, in turn, identify a certain real-world entity.

An ontology consists of the following basic elements:

- Concepts, they represent class of individuals sharing some properties, i.e. *Country*;
- Instances, they are actual occurrences of concepts, i.e. *France*;
- Relations, they are links or connections between instances, and between instances and concepts, i.e. *isLocatedIn*.

Based on these elements, an ontology can be formally defined as a pair (S, A), where S is the signature (vocabulary) of the ontology and A is the set of ontological axioms, which specify the interpretation of the signature (Kalfoglou & Schorlemmer, 2003). An axiom may specify a concept in terms of others previously defined concepts or it may define inclusion relation between concepts. A more refined definition divides S into three sets, the set of concepts C, the set of relations R and the set of instances I. It defines an ontology as a 4-tuple (C, R, I, A) (De Bruijn, Martin-Recuerda, Manov, & Ehrig, 2004).

Ontologies intend to represent knowledge in the most formal and reusable way possible. Figure 2.3 shows an overview of the various formalisms that can be adopted to model and manipulate knowledge based on their expres-

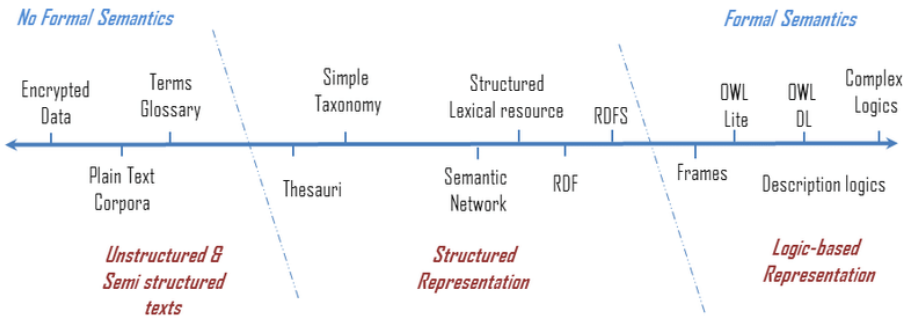


Figure 2.3: Overview of the different knowledge representations that can be used (Harispe, 2014; Staab & Studer, 2009).

siveness: from weak semantics describing only terms and linguistic relationships, to strong semantics of more refined and complex conceptualization. The main formalisms used in KR are conceptual graphs (Sowa, 1984), frame languages (Minsky, 1974) and description logics (DLs) (Baader et al., 2003). The definition of a formal logic enables a wider interpretation of knowledge to automatically infer facts which are not explicitly stated. Based on the complexity of the defined logic, several languages having different levels of expressiveness can be obtained. Obviously, expressiveness comes at a price. It increases the computational complexity of reasoning related to a language. Usually ontologies based on DL are a good trade-off between expressiveness and efficiency[10]. In this study, we are mainly interested in DLs as the formal foundations of the representation of knowledge. DLs are widely used by both scientific and industrial communities. Indeed, DLs have been used by the World Wide Web Consortium (W3C) as formal foundations of the ontological language of the Web (Web Ontology Language), in the context of Semantic Web. This has strongly contributed to the developments of DLs.

Formalisms to model knowledge

Considering ontologies based on DL, we can distinguish two types of knowledge representation:

- *terminological knowledge* or *T-Box*, knowledge that includes assertions about concepts and relations;
- *assertional knowledge* or *A-Box*, knowledge that includes assertions re-

[10]https://www.w3.org/TR/owl-guide/

lated to instances of the concepts and relations; the assertions should be in accordance with the *T-Box*.

Ontologies versus Knowledge Bases

This distinction makes it possible to clarify the difference existing between *ontologies* and *Knowledge Bases (KB)* (Guarino, 1998). KB is a broader term than ontology and it has several meanings (Kiryakov, 2006). Generally, a KB is a set of structured information represented by a Knowledge Representation formalism, which enables automatic inference. It can include axioms, definitions, facts, statements, and other primitives. The only difference with an ontology is that a KB is not intended to represent a shared or consensual conceptualisation. For this reason in some studies the term *ontology* includes only the intentional aspects of a domain (i.e. *T-Box*). However, an ontology can be seen as a specific sort of KB. Therefore, in this manuscript, the terms ontology and KB will be used interchangeably. The term *ontology* will include both intentional (i.e. *T-Box*) and extensional knowledge (i.e. *A-Box*).

DL-based ontologies become very popular with the advent of Semantic Web. Indeed, as anticipated in this section, the W3C highly contributes to define a number of standard protocols and languages based on DLs to promote Semantic Web and ontologies. The main standards that has been proposed are presented in the next section.

2.3.2 Ontologies in the Semantic Web

In the early 2000s Tim Berners-Lee and its colleagues introduced the Semantic Web (SW) stating that "*The SW is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*" (Berners-Lee, Hendler, & Lassila, 2001). SW has been conceived to overcome limitations regarding the lacks of content structure and content semantics, as well as the absence of a universal data format. In this context, information is ambiguous. Therefore, machines are able to read the provided information, but they are unable to understand it (Sudeepthi, Anuradha, & Babu, 2012). SW want to make Web contents readable and understandable by both human and machines (Al-Feel, Koutb, & Suoror, 2008).

The main aim of Semantic Web

The idea is to define knowledge that is intelligible by autonomous software agents, that can process and manage it effectively. The only way to make the Web content machine-understandable is to explicitly specify the semantics of

Trust in the Semantic Web
Another important concept introduced by Semantic Web is trust, as represented by the last layer of Semantic Web Stack (Horrocks, Parsia, Patel-Schneider, & Hendler, 2005). Indeed, if software agents must be able to operate autonomously, then it is important that they known who to trust. This trust is related to but more complex than the trustworthiness concept that is considered in TD models. Indeed, in TD context, trustworthiness can be defined as the objective probability that a source provide a true information. It is mainly based on its content. Trust in Semantic Web has instead a broader meaning. It is based on the rationale of "Web of trust". Like in human community, trust is based on experience, and by propagating trust among sources. For instance, when one trusts a source s, he also trusts all other sources that are trusted by source s. A source could also consider other source recommendations to trust another one.

each entity (abstract or real) we intend to represent. Specifying the semantics of Web content has important consequences such as the transformation of the Web search (Nuzzolese, Presutti, Gangemi, Musetti, & Ciancarini, 2013).

Semantic Web and entity search

In the usual Web, the search is based on keywords. For instance, when looking for the word "football", a search engine returns all web-pages with "football" in it. In the case of the Semantic Web the search take semantics into account. Therefore, it also returns related contents that describe the entity we are interested in. Considering the previous example, the search engine is able to also retrieve a set of structured information referring to World Cup, professional sport, etc. This kind of search is called entity search. Also important companies, such as Google, started to be interested in Semantic Web due to its potential. Google acquired Freebase that was a large KB whose content was added mainly by users. Freebase was developed by Metaweb company since 2007. Google exploited this large KB to enrich Google Knowledge Graph that officially replaced Freebase in 2015.

Using ontologies to model machine-understandable knowledge

To transform the Web from being machine-readable to being machine-understandable, it is necessary to formally specify rules that indicate how resources can be described. Ontologies are suitable to this task. This is why, ontologies play a prominent role in fulfilling semantic interoperability re-

quired in the Semantic Web (Berners-Lee et al., 2001).

2.3.3 Standards used to represent ontologies

The evolution of the Web towards a semantic dimension has led the W3C to establish several standards which promote common data formats and exchange protocols on the Web. These standards enable the reuse of data and the consequent reduction of data redundancy. In brief, they are a set of guidelines that serve to better specify descriptions of resources and their semantics. First of all, they state that each entity (also called resource) has to be referenced by an IRI (Internationalized Resource Identifier) in order to be unambiguously identified. An IRI is defined as *"a compact sequence of characters that identifies an abstract or physical resource"* (Berners-Lee, 1998). It means that an IRI must contain no space characters and it may refer to an abstract resource such as the concepts "Pablo Picasso" and "painting", as well as to any file that can be retrieved from the Web. Then, they suggest the use of HTTP[11] IRIs in order to facilitate users in accessing the resources on the Web. A HTTP IRI is a string that consists of a scheme, an authority and a path (respecting this order), see Figure 2.4. The scheme indicates the protocol used to access the resource. The authority specifies the server where the resource is located. The path locates the resource within the directory structure of the server. Finally, the standards suggest to provide useful information within an IRI and to include links to other IRIs to enable users to discover related information. Since IRIs are typically long, it is convenient to make use of abbreviated forms to facilitate their readability. A compact IRI includes a namespace and a local name that are separated by a colon. The namespace consists of the protocol, the authority, and in some cases the initial part of the path; instead the local name consists of the rest of the IRI. For example the IRI `http://dbpedia.org/resource/Pablo_Picasso` can

Use of IRIs

Use of HTTP protocol

[11] The acronym means HyperText Transfer Protocol. It refers to a set of rules that regulate communication between a server and a client on the Web.

$$\underbrace{\texttt{http://}}_{\text{schema}}\ \underbrace{\texttt{dbpedia.org/}}_{\text{authority}}\ \underbrace{\texttt{resource/Pablo\_Picasso}}_{\text{path}}$$

Figure 2.4: Example of HTTP IRI.

be abbreviated to `dbp:Pablo_Picasso` where the namespace `dbp` abbreviates `http://dbpedia.org/resource/`. As you can see, usually the local name preserves the substring that is significant to human readers. This abbreviation method will often be used in the rest of this manuscript to facilitate reading.

The most popular ontology languages are RDFS (Resource Description Framework Schema) and OWL (Web Ontology Language) thanks to the Semantic Web and its standardization purposes. These standards represent progressive RDF (Resource Description Framework) enrichments that meet specific needs of expressiveness. RDF establishes a general framework to standardize in an unambiguous way resource descriptions. In addition, the RDFS standard (RDF Schema) introduces some basic elements of knowledge modelling through the notion of class and schema allowing the definition of concepts and the relation among them. It also enables the definition of the interpretation required to perform automatic reasoning. OWL is a language developed to allow a complete ontological representation of knowledge based on the formalisms of description logic. It enables the definition of more complex forms of knowledge. Hereinafter, we present the main features of these languages.

2.3.3.1 RDF

The RDF data-model (Resource Description Framework) (Manola & Miller, 2004a) (Manola & Miller, 2004b) provides a general framework to specify and enrich resources on the Web or knowledge expressed in an ontology. All resources are described through triples; statements with a specific meanings. Specifically, each triple $(subject, predicate, object)$ indicates the value (*object*) assigned to an aspect or a property (*predicate*) characterizing a real-world entity or resource (*subject*). The *subject* can be an IRI or a blank node. Note that blank nodes are treated as simply indicating the existence of a thing, without using an IRI to identify any particular thing. It has been introduced to model resources that is difficult to describe with only binary properties (reification principle). The *object* can be an IRI, a literal or an anonymous node. Finally, the *predicate* is an IRI that indicates the relationship, which exists between the *subject* and the *object* of the triple. Throughout the rest of the manuscript, we will use the term *entity* with reference to resources that occur as subject or object of a triple. An RDF model is a set of these

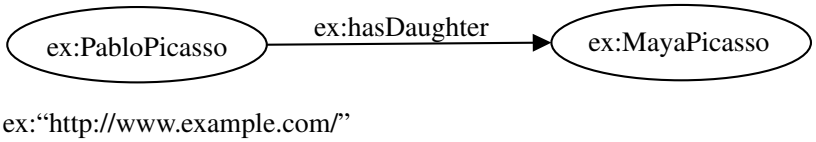


Figure 2.5: Example of RDF graph having only a triple.

triples. It can be represented as an oriented graph, see Figure 2.5. Nodes include all subjects and objects, and arcs include all predicates. Different syntaxes are available to store an RDF data model: RDF, XML, N3, Turtle, or JSON (Beckett, 2004). Moreover, RDF data can be also consumed. Standards such as SPARQL and other RDF languages, e.g. TriQL (Bailey, Bry, Furche, & Schaffert, 2005), have been designed to query it.

2.3.3.2 RDFS

RDFS (Resource Description Framework Schema) (Brickley & Guha, 2004) is an RDF vocabulary, which enriches RDF semantics introducing the notion of concept, i.e. `rdfs:Class`, and the notion of properties (roles) that enable the organization of concepts through hierarchies. In addition it is possible to define domain and range of properties. Considering the RDF triple reported in Figure 2.5, the domain of the predicate *ex:hasDaughter* should specify that *ex:PabloPicasso* must be an instance of the concept *ex:Person* and the range of *ex:hasDaughter* should specify that *ex:MayaPicasso* must be an instance of *ex:Female*. Therefore RDFS makes it possible to model simple ontologies. Moreover, its semantics is inferential. It means that, given a world theory, it is possible to infer new information. An example of RDFS graph is reported in Figure 2.6.

2.3.3.3 OWL

OWL (Web Ontology Language) is a family of languages designed to represent rich and complex knowledge. It is based on the RDF and RDFS languages and introduces complex constructors using DLs. OWL enables the expression of equivalence, union and disjunction between concepts and roles, e.g. the *disjointWith* predicate on Figure 2.7. It also introduces the restrictions, cardinalities and properties of predicates. The standard currently adopted is OWL 2 (Group, 2012). OWL consists of several profiles (Motik et

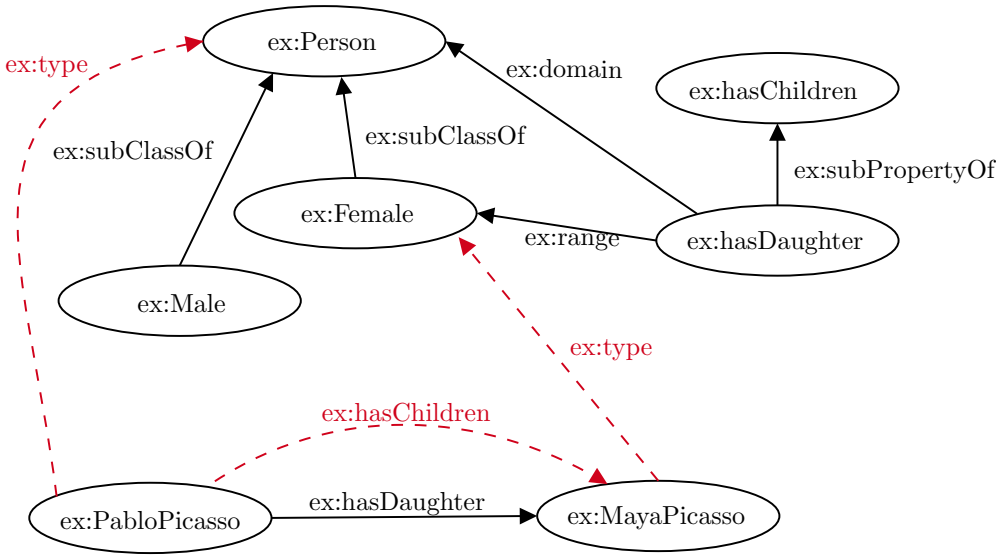


Figure 2.6: Example of RDFS graph. Red dashed lines correspond to some of the statements which can be inferred from the rest of the graph.

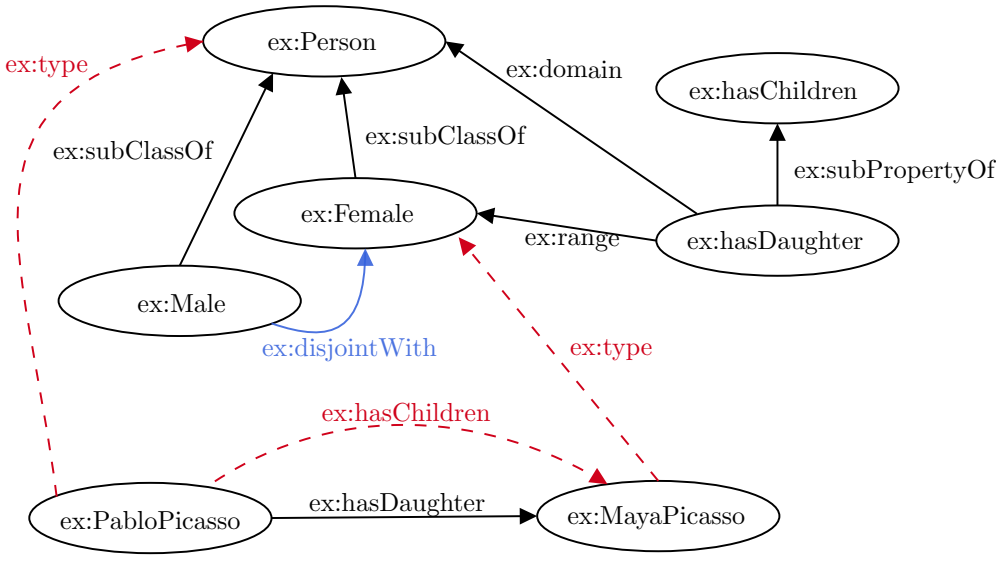


Figure 2.7: Example of OWL graph. Red dashed lines correspond to some of the statements which can be inferred from the rest of the graph. Blue lines indicate the predicates introduced by OWL language.

al., 2008). They are OWL EL (based on the specific DL family languages EL), OWL QL (Query Language) and OWL RL (Rule Language). These profiles are less expressive than the OWL 2 standard and are designed for practical use. They offer different trade-offs between the expressiveness of the ontology on the one hand, and the complexity of reasoning mechanisms on the other.

2.3.4 Open World Assumption

Semantic Web languages such as RDF(S) and OWL operate under the Open World Assumption (OWA). OWA is used in KR to express the fact that knowledge within a system is incomplete. Therefore, if a statement is not contained in the system, then it can be unknown and not necessarily false. Indeed, it may be not explicitly made yet. The opposite of OWA is Closed World Assumption (CWA). CWA is usually used by traditional relational databases. It express the fact that knowledge within a system is complete. Therefore, if a statement is not specified, then the statement is surely false, i.e. negation as failure (a statement is considered false, if it cannot be proved to be true).

2.3.5 Linked Data

Linked Data is an expression that generically refers to the several standards defined by the W3C. These best practices for publishing and connecting structured data on the Web has encouraged several initiatives to create and share linkable data on the Web. The most remarkable one is Linked Open Data (LOD) project. LOD is an open, interlinked collection of datasets containing knowledge on different domains in a machine-understandable form (Bizer, Heath, & Berners-Lee, 2009; Schmachtenberg, Bizer, & Paulheim, 2014). Two popular LOD resources are the following ones.

DBpedia One of the most popular initiative is DBpedia. It is a RDF knowledge base that contains information extracted from different language editions of Wikipedia and its info boxes. DBpedia is broad-coverage, cross-domain, multi-lingual and includes a number of links to other datasets. Because of the many incoming links from other datasets, it has become the

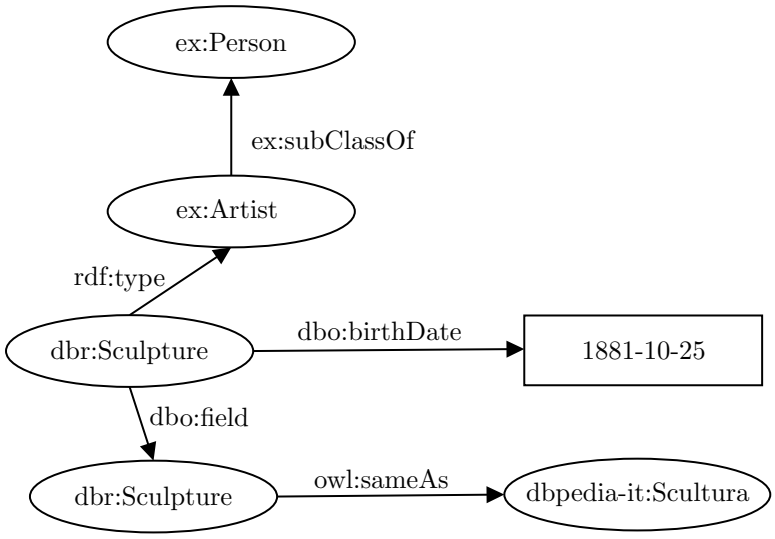


Figure 2.8: Extract of DBpedia. Ellipses surround URI values and rectangles surround literal values.

central hub of the LOD cloud. An extract of DBpedia is reported in Figure 2.8.

Gene Ontology (GO) and gene product annotations. The GO (Ashburner et al., 2000) is a popular ontology related to molecular biology that is used in biomedical and bioinformatics studies. An extract is reported in Figure 2.9. GO defines a structured vocabulary which enables to conceptually annotate gene products on the basis of three different aspects: *molecular functions* they are concerned with, *biological processes* in which they are involved and *cellular components* in which they are located. Each annotation is made in accordance with experimental observations or automatic inferences. For example, the protein shisa-3 can be described by the *molecular function* term protein binding, the *biological process* term multicellular organism development, and the *cellular component* terms endoplasmic reticulum membrane and integral component of membrane.

LOD being ontologies defined on different domain represents a valuable

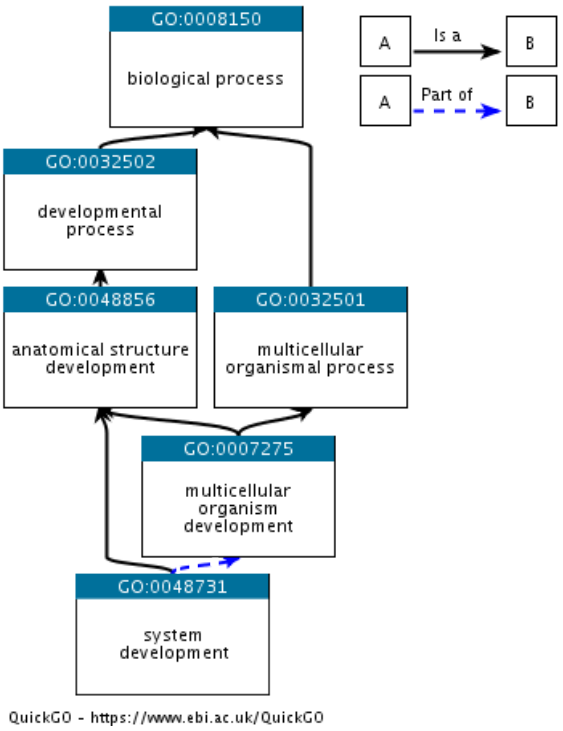


Figure 2.9: Extract of Gene Ontology.

resource. Indeed, they model *a priori* knowledge that is freely accessible avoiding to spend a lot of time in creating new ontologies. Recently, Semantic Web Mining approaches have been proposed. They are models that take advantage of LOD to facilitate data mining and knowledge discovery processes (Buffa, Zucker, Bergeron, & Aouzal, 2016; Ristoski & Paulheim, 2016). In the same perspective, in this thesis, we aim at improving data veracity assessment using *a priori* knowledge contained in LOD to improve TD performance. The idea is that reusable knowledge contained in these ontologies can facilitate understanding of information provided by multiple data sources. Therefore, the evaluation of its veracity can be simplified.

This explains the reason behind discussing data veracity and knowledge representation in the same chapter. More precisely, we have firstly discussed the importance of obtaining reliable information to populate KBs that can be used to support decision-making process. We have specified the meaning of reliable information in this manuscript. An information is considered as reliable if it conforms to reality. Then, we have presented a class of methods,

called Truth Discovery, that are used to identify reliable facts comparing information provided by multiple sources. At this point, the main concepts of knowledge representation have been introduced focusing on ontologies. Indeed, the main contribution of this thesis is to explain how *a priori* knowledge expressed into ontology can be exploited to improve TD performances. In the next chapter, we presented how ontologies can be useful to identify dependencies that may exist between provided values.

Chapter 3

Truth Discovery using partial order of values expressed in ontologies

Contents

This chapter describes how *a priori* knowledge may be exploited to detect important dependencies that may exist among different values in order to improve Truth Discovery (TD) performance. The prior knowledge that is considered specifies partial order relationships between values. This information helps to better identify the actual conflicting values when different ones are provided about the same aspect of a real-world entity. Indeed,

when two values are syntactically different, they are not necessary in conflict. Indeed, one value may semantically support the other. First of all, the importance of taking this kind of prior knowledge into account is explained through an example. Then, the problem setting is formalized discussing the impact of considering this additional knowledge during TD process. Moreover, an adaptation of an existing TD model is developed. Experiments show the validity of the proposed rationale.

Contributions related to this chapter:

Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, and Isabelle Mougenot. (2016). How Can Ontologies Give You Clue for Truth-Discovery? An Exploratory Study. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS '16)*, 12 pages. DOI: `https://doi.org/10.1145/2912845.2912848`

Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, and Isabelle Mougenot. (2018). Truth Selection for Truth Discovery Models Exploiting Ordering Relationship Among Values. *Knowledge-Based Systems*. To appear.

3.1 Exploiting partial order of values within Truth Discovery process

When two values are syntactically different, they are not necessary in conflict. They may semantically support each other. Indeed, a value property of a real-world entity may be expressed with different levels of detail, i.e. different granularities[1]. To the best of our knowledge, the granularity of values has never been considered in TD models. Hereinafter, an example illustrates how this information may be useful for giving insights to discover dependencies that may exist among values and, thus, better identify true claims. In the rest of this chapter, dependent values are interpreted as values that are semantically related because they describe the same concept with different levels of granularity. The example is complemented by a formalisation of the new setting and a model that is proposed to deal with it. First, the representation used to model *a priori* knowledge is described. Then, a method

[1]The granularity is strictly related to Information Content (IC), i.e. informativeness level of a value. The greater the granularity of a value, the higher the IC of this value. Please refer to section 3.1.2.1 for further details on IC.

is explained showing how this additional information is integrated into the traditional TD process enabling the propagation of evidence between values.

3.1.1 Intuition

When asking people the following question: *"Where was Pablo Picasso born?"*, they can reply with different answers depending on their knowledge of the topic and their nature, see Table 3.1. Some people will say *"Málaga"* (the actual birth city of the painter). Other ones will claim other city names, such as *"Madrid"* or *"Paris"*. Instead, other ones will prefer saying more general values such as *"Spain"*. In such a situation, considering that both *"Málaga"* and *"Madrid"* are in *"Spain"*, and only *"Paris"* is in another country is an important information. It suggests that people majority is in accordance with the fact that the painter was born in *"Spain"* even if there is disagreement on which Spanish city. While *"Madrid"* and *"Málaga"* cannot both be true at the same time, both of them implicitly support the more general value *"Spain"*.

This suggestion is made possible because of prior knowledge related to the dependencies between provided values. In this example, the prior knowledge about locations states that *"Madrid"* and *"Málaga"* are both in *"Spain"*, *"Spain"* is in *"Europe"* and so on. This knowledge enables taking advantage of dependencies between values to facilitate the identification of reliable true answers. We assume that sources agree on the prior knowledge we consider since an ontology is defined as a "*shared conceptualization*". Otherwise, if a source has *a priori* knowledge different from the one that is considered, then the meaning of the provided information can be twisted. Potentially, as a consequence, a source can be judge unreliable for a claim it never provides.

Table 3.1: Examples of claims about the birth location of Pablo Picasso.

Source	Object	Predicate	Value
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles

3.1.2 TD-*poset* approach

The first contribution of this thesis aims to study how TD models can be improved by considering prior knowledge in the form of partial order of values, see Figure 3.1. Usually, when different values are provided for the same data item, TD models consider them necessarily conflicting under functional setting. However, as shown by the previous example, this is not always the case; for instance, in a large variety of real-world scenarios, sources can use different degrees of precision while providing their advice about topics for which a dedicated terminology exists, e.g. medicine. Therefore, considering asserted claims to be necessarily disjoint, unrelated and conflicting is not adapted to these situations. Several organizations attempt to create resources that collect dedicated terms of a specific domain, e.g. SNOMED collects medical terms (Spackman, Campbell, & Côté, 1997) and Gene Ontology gathers biological terms used by computational genomic community. Usually, these terms are organized into hierarchies based on taxonomic/subsumptive relations, meronomic/compositional ones and implication/logical ones. Hierarchically structured objects abound in real-world, as a consequence ontologies usually include a portion of knowledge that is hierarchically structured (Joslyn & Hogan, 2010). Hierarchies naturally evoke top-rooted trees as graphical representation. More precisely, hierarchies are correctly represented by directed acyclic graphs, since some objects may have more than one parent. The proper mathematical basis to deal with hierar-

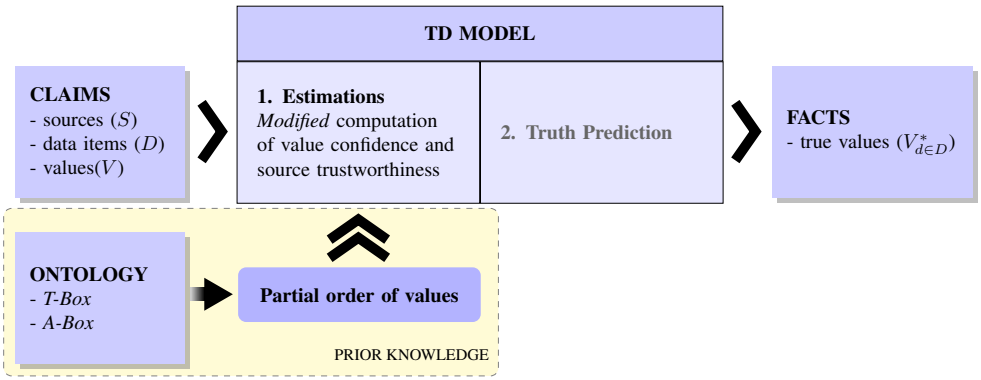


Figure 3.1: Diagram of the overall Truth Discovery (TD) procedure incorporating the partial order of values during the estimation phase to improve TD performance.

chies is order theory. Indeed, the main relations considered to hierarchically structure objects are transitive (Joslyn & Hogan, 2010). This is the motivation that leads us to propose an approach that models value dependencies using partial order and incorporates them into existing truth discovery models to deal with a wider-range of scenarios.

Functional predicate setting

This study focuses on the specific class of TD problems related to functional predicate. The majority of existing models attempt to solve conflicts among claims containing functional predicates assuming that only one value can be true for each data item among the provided ones. However, in cases where different value granularities can be used, different claims related to the same data item d are not independent and, moreover, multiple values may be true at the same time.

Contrary to literature studies (see section 2.2), the problem formulation we tackle here considers a more realistic setting enabling dependencies between values – as we will see it implies important changes in the classical problem formulation, on the assumptions that have to be considered, as well as in the solutions that can be proposed.

3.1.2.1 Partial order of values

Value dependencies are modelled by a partial order $O = (V, \preceq)$ (Davey & Priestley, 1990) that is characterised by a binary relation $\preceq$ defined over V (value set) such that:

- reflexive, i.e. $\forall v \in V : v \preceq v$;
- anti-symmetric, $\forall (v, v') \in V^2 : (v \preceq v' \wedge v' \preceq v) \implies v \equiv v'$;
- transitive, $\forall (v, v', v'') \in V^3 : (v \preceq v' \wedge v' \preceq v'') \implies v \preceq v''$.

This relationship indicates that a value v subsumes another value v' if $v' \preceq v$ or, *vice versa*, v' is subsumed by v. The set V over which a partial order is defined is called partially ordered set (*poset*). Its ordering is considered partial because $\exists (v, v') \in V^2 : v \npreceq v' \wedge v' \npreceq v$. In other words, there are pairs of values that are not ordered.

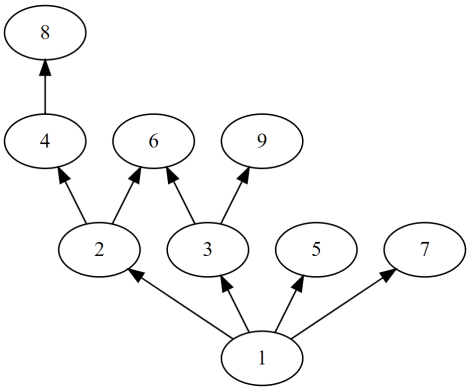
Representing the partial order as a graph

An important characteristic of a partial order of values is that it can always be represented as a Directed Acyclic Graph (DAG). This simplifies its exploitation. Moreover, it allows a graphical representation that facilitates

Figure 3.2: Example of a simple DAG extract(on the left) and its transitive reduction (on the right).

its understanding. A DAG corresponds to a graph $G_O = (V, E)$, where $V = \{v_0, v_1, \cdots, v_m\}$ is the set of values, and $E = \{(x, y) \in V^2 | x \preceq y\}$ is the set of edges specifying the partial ordering that exists between values. Any DAG transitive reduction can easily be obtained using state-of-the-art algorithms (Aho, Garey, & Ullman, 1972). For the reduced graph G_O, the new set of edges is $E' = \{(x, y) \in E | \nexists z \in V \setminus \{x, y\}, x \preceq z \wedge z \preceq y)\}$. For instance, Figure 3.2 denotes an example of partial order represented by a DAG and its transitive reduction.

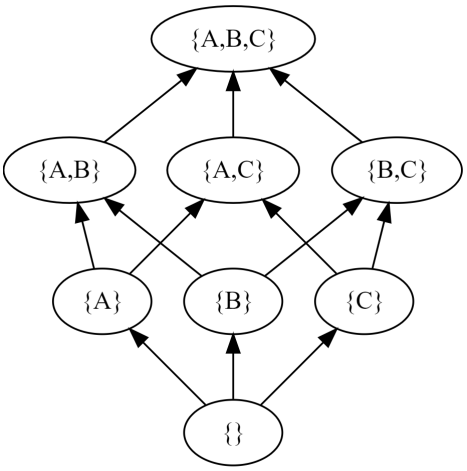
A partial order can be defined on several types of values such as number, sets, categorical data, strings and so on. The semantics of a partial order depends on the binary relation on which it is defined. For instance, considering the numerical values $\{1, \cdots, 8\}$ different partial orders can be obtained based on the relation that is considered. Figure 3.3a reports the partial order that is obtained considering *divides* relation. Indeed, the integer 4, 2 and 1 are divisors of 8, the integer 2 is divisor of 4, and so on. Considering the binary relation "<", i.e. higher than, another partial order is obtained on the same values, see Figure 3.3b. Also sets can be ordered. The *subset* relation can be used to identify a structure among elements of a power set, see Figure 3.4a. For instance, there is a relation between $\{A, B\}$ and $\{A, B, C\}$ because the former set is a subset of the latter set. Moreover, a different partial order can be build on sets of elements considering the *partition* relation. Indeed, the different partitions of a set can be ordered among them. An example of this kind of order is reported in Figure 3.4b.



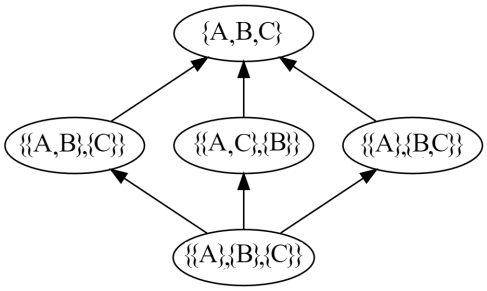
(a) Partial order defined on *divides* relation.

(b) Partial order defined on *higher than* relation.

Figure 3.3: Example of two different partial orders that can be obtained considering the same values included in the following set $\{1, \cdots, 8\}$.



(a) Partial order defined on *subset* relation.



(b) Partial order defined on *partition* relation.

Figure 3.4: Example of two different partial orders that can be obtained considering the set $\{A, B, C\}$.

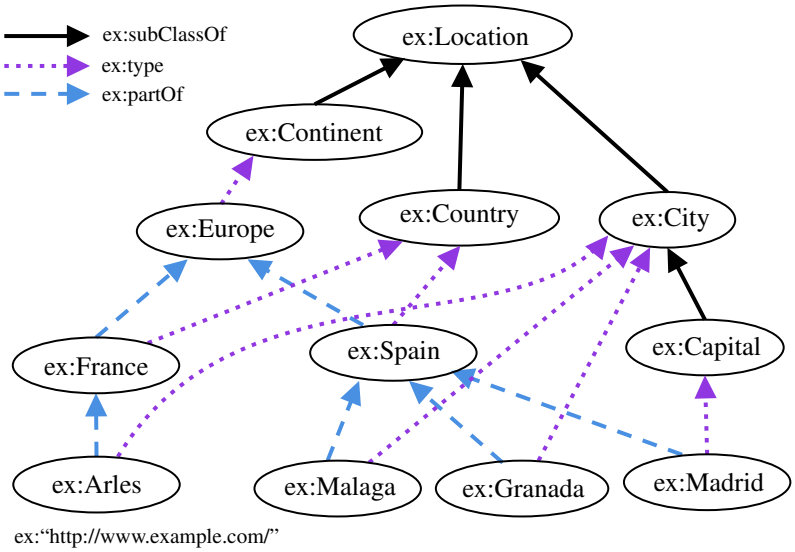


Figure 3.5: Example of a partial ordering of values mixing both taxonomic, *isA* and *part-of* relationships.

In this study, we deal with categorical values on which we consider a partial order. The partial order is assumed to be known *a priori* and defined into an ontology. For each predicate, a partial order can be constructed from the assertions that are contained within an ontology. More precisely, one or more relations can be considered to compose the partial order. The important point is that these relations must preserve, according to the considered predicate, the transitivity property. For instance, the partial order represented in Figure 3.5 is associated with the predicate *birthPlace*. In this case, the transitivity property is well respected considering *subClassOf*, *type* and *partOf*[2]. However, this partial order cannot be used also for the *capitalOf* predicate. Even if the value associated with this predicate has to be a *Location*, in this case the transitivity does not hold anymore for the relationship *partOf*. For example, *Madrid* is *capitalOf Spain*. *Madrid* is *capitalOf* a *Country*, but *Madrid* is not *capitalOf Europe*. Given this observation, we assume only a partial ordering of values which is defined by transitive relationships according to the predicate considered in the claim. After the partial order is composed, we do not consider additional aspects related to the semantics of the relationship defining the partial ordering – the partial ordering can therefore be

[2]To ease the reading the URI prefix are omitted in text.

a taxonomy composed of *subClassOf* relationships, or any directed acyclic graph associated with a more complex semantics.

It is also important to highlight that we compose the partial order considering the CWA. Indeed, when a relationship between two values is missing, we assume that this relation does not exist and thus these values are disjoint. Consider the OWA we cannot consider that two values are independent when a partial order relation does not exist between the two values. Indeed, the relation may be unknown. Values are disjoint/independent if and only if a disjointness axiom is specified. Only in this case the non-existence of a dependency among values is guaranteed. The problem is that disjointness axioms are not often stated (Völker, Fleischhacker, & Stuckenschmidt, 2015). Thus, a lot of values are potentially dependent. In the extreme case in which no disjointness axiom are stated, each value is considered to be dependent from all the others. Since the proposed approach consists of propagating the support[3] given by a value to all its dependent values, it could be not advantageous anymore. For this reason, we prefer to consider CWA in this study.

Important notations used to characterise G_O are formally introduced below to ease the manipulation of the partial order graphs in the next sections. Considering a partial order G_O, we define:

Important notations and functions of a DAG

- a function anc: $V \rightarrow \mathcal{P}(V)$ returning the ancestors of a value x such that $anc(x) = \{y \in V | x \preceq y\}$
- a function $fath$: $V \rightarrow \mathcal{P}(V)$ returning the fathers of a value x such that $fath(x) = \{y \in V | (x \preceq y \wedge \nexists z \in V \setminus \{x, y\}, x \preceq z \wedge z \preceq y)\}$
- a function $desc : V \rightarrow \mathcal{P}(V)$ returning the descendants of a value x such that $desc(x) = \{y \in V | y \preceq x\}$
- a function $chil$: $V \rightarrow \mathcal{P}(V)$ returning the children of a value x such that $chil(x) = \{y \in V | x \in fath(y)\}$
- a set $ROOTS \subseteq V$ containing the root elements of G_O such that $ROOTS(G_O) = \{x \in V | anc(x) = \{x\}\}$
- a set $LEAVES \subseteq V$ containing the leaves of G_O such that $LEAVES(G_O) = \{x \in V | desc(x) = \{x\}\}$.

[3] We do not intend to propagate the disapproval of a value to its independent values.

Given these elements, we can formally define the concept of independent values. Considering a partial order G_O, two values $(x, y) \in V^2$ are independent if and only if $desc(x) \cap desc(y) = \emptyset$. Thus, the term disjoint and independent will be used interchangeably in this chapter. We also define the depth of a node as the maximal length of a path from the node to the tree's root node. Considering G_O, a path from x to y is defined as a non-empty sequence of n different values $\langle v_0, v_1, \cdots, v_{n-1} \rangle$ with $x = v_0$, $y = v_{n-1}$ and for which $\forall i \in [0, n-2]$ $v_{i+1} \in fath(v_i)$. Both number of ancestors and depth of a value are often expected to be directly proportional to its degree of expressiveness, i.e. the level of specificity associated with a value. The more a value is subsumed/the highest a depth of value is, the highest its expressiveness is. On the contrary, the more a value subsumes other values, the lowest its expressiveness is. The specificity of a value can be regarded as the amount of information that it conveys, i.e. its Information Content[4] (IC) (Seco, Veale, & Hayes, 2004). In other words, this quantity represents the degree of abstraction/concreteness of a value w.r.t. an ontology, see section 3.3 in (Harispe, Ranwez, Janaqi, & Montmain, 2015). One of the main IC properties is that the it monotonically decreases from the leaves to the root, i.e. if $x \preceq y$, then $IC(x) \geq IC(y)$ (most often we consider that $IC(root) = 0$ with $root \in ROOTS$). It exists several definitions of IC, each satisfying this property.

Information Content

3.1.2.2 Propagating the evidence associated with values

Modelling value dependencies, the partial ordering specifies the relationships between values indicating which values subsume the others. These relationships have two possible interpretations. Considering $(v, v') \in V^2$ such that $v \preceq v'$ means that:

Meanings of partial order relationship

- v implies v', e.g. $Spain \preceq Europe$ means that claiming that someone was born in *Spain* implicitly implies claiming that this person was born in *Europe*;
- v is possible given v', e.g. $Spain \preceq Europe$ means that claiming that someone was born in *Europe* subsumes the fact that this person was

[4]The IC is usually associated with a concept of an ontology. Since in our approach other partial orders may be considered, we use a different terminology for sake of coherence with the problem setting.

born in one of the country located there including *Spain* (*Spain* is a *plausible* value).

Both interpretations are in accordance with the principles of *Dempster-Shafer theory* (Shafer et al., 1976), also called *Evidential theory*. It is a mathematical framework that deals with uncertainty that is mainly originated by incompleteness, i.e. lack of information, and inconsistency, i.e. conflicting information. An agent is uncertain about a proposition, i.e. claim in our setting, if (s)he does not known its true value. *Evidential theory* aim to discover it assigning beliefs to the true state of a proposition given all available pieces of evidence.

Formally, let Ω denote a finite set of possible states, called the frame of discernment – an exhaustive set of mutually exclusive alternatives, i.e. only one state can be true and the true state is assumed to be in this set. This framework quantifies the available evidence for each possible state defining a mass function $m : 2^{\Omega} \rightarrow [0, 1]$ such that $\sum_{A \subseteq \Omega} m(A) = 1$ with 2^{Ω} being the power set of Ω. Thus, the evidence may be assigned not only to atomic elements, but also to sets of elements. Given our problem setting, the mass function for each data item d is a function m defined from V_d to $[0, 1]$ where V_d is the set of values provided for d. Thus, in this case, the evidence may be assigned not only to the most specific values, i.e. values not subsuming other ones, but also to more general values, namely values subsuming other ones. When $m(v_d) > 0$, v_d is a focal element of m. Moreover, when a value v_d is a focal element ($m(v_d) > 0$) and subsumes several other values, the evidence quantified by its mass function cannot be distributed in any way among those values. Indeed the mass function of v_d represents the fact that v_d but nothing more specific is the true value. This is in accordance with our definition of confidence that collects the evidence provided by sources claiming a certain value. Thus, in our setting, the mass function of each value corresponds to its confidence. Note that if each focal element of m contains only a single element, m is reduced to be a probability distribution. Given a mass function m associated with v_d, two interesting functions can be computed. They are the belief and the plausibility functions that indicate the total mass of information that implies the value v_d and is consistent with v_d, respectively. In other words, the belief represents all the evidence for which v_d is surely true, while plausibility represents the evidence for which

Mass function

v_d could be true, i.e. it is possibly true. They result to be the lower and upper bounds of the interval that measures both the probability of v_d to be true and the ignorance associated with this probability. Thus:

$$Bel(v_d) \leq P(v_d) \leq Pl(v_d) \tag{3.1}$$

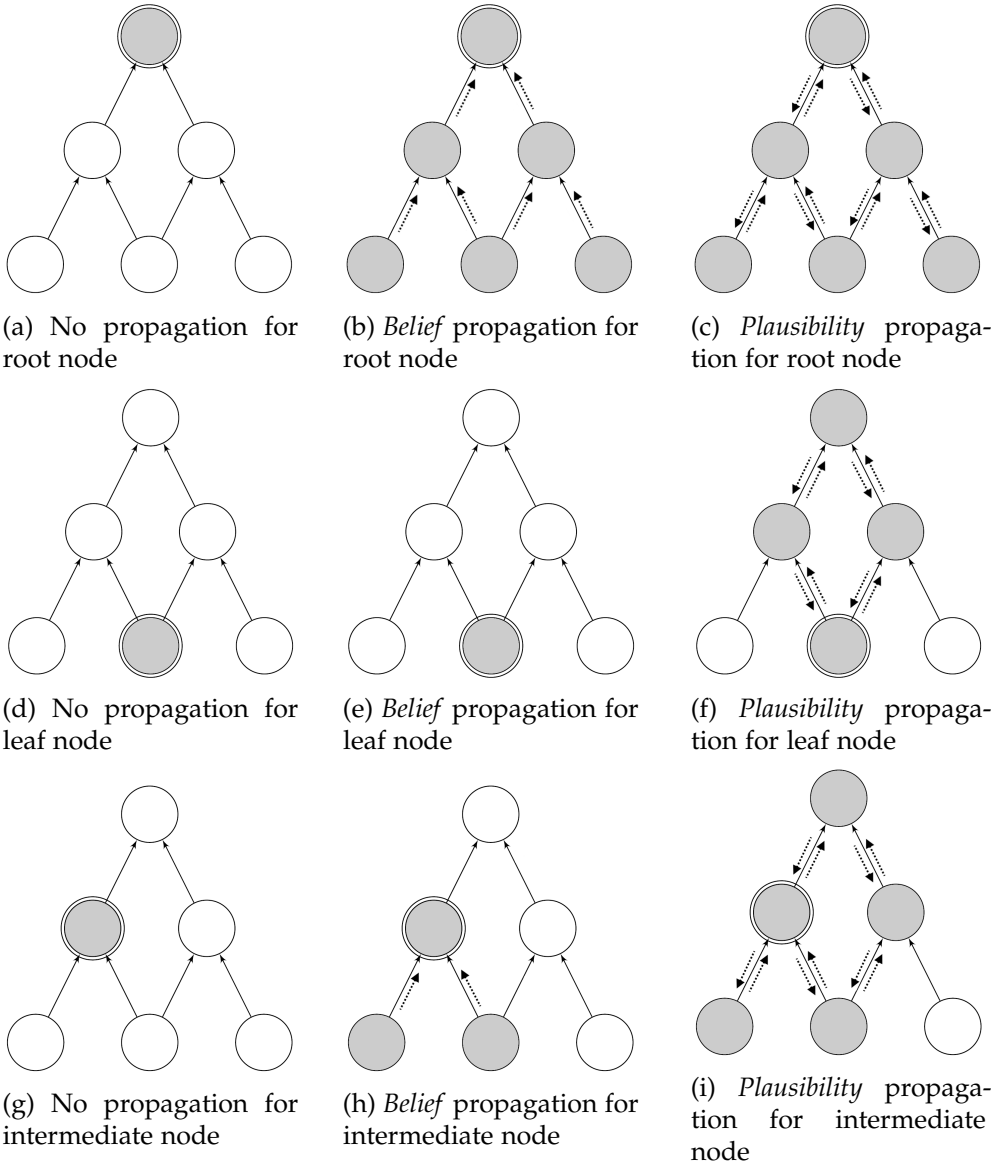
Hereinafter, the classical belief theory notation will be adapted to our setting in order to make the link with it, and to ease the reading.

Belief function

In our setting, considering that the confidence associated with a value can be seen as the quantity returned by its mass function, the belief $Bel : V_d \rightarrow [0, 1]$ of a value v_d is therefore evaluated as:

$$Bel(v_d) = \sum_{v'_d \in desc(v_d)} c(v'_d) \tag{3.2}$$

where $desc(v_d)$ is the set of descendants of v_d according to graph G_O. Summing up the evidence assigned to a value v_d and its descendants (that implicitly support it) this formula is in accordance with the first interpretation of a partial order relationship for which one value implies all its generalization. Indeed, if a source claims v_d for a data item $d \in D$, then it implicitly supports all the claims v'_d that subsume the provided one or includes it, i.e. $\forall v'_d \in \{y \in V_d | v_d \preceq y\}$, $v_d \implies v'_d$. In this case, the evidence is spread in a bottom-up direction: from most specific claims to most general ones. In other words, the evidence of a claim is propagated to all the claims that subsume it considering the partial ordering that exists among values. Thus, the partial order suggests which evidence has to be considered given a certain value. This kind of propagation will be called *belief* propagation in the rest of the thesis. We can distinguish three different situations to compare how the evidence is propagated in practice based on the position in the partial order of the value under examination: (i) the value is the root of the partial order, see Figure 3.6a, (ii) the value is a leaf of the partial order, see Figure 3.6d, and (iii) the value is neither a root nor a leaf, see Figure 3.6g. Considering the *belief* framework and situation (i), when the value is the root, the considered value receives the same evidence, from all its "observed" descendants. see Figure 3.6b. Differently, considering (ii), the value is a leaf, the value we consider does not get any evidence, see Figure 3.6e. In situation (iii), the value is the middle of partial ordering structure, the evidence is given by its descendants, see Figure 3.6h.



(a) No propagation for root node

(b) *Belief* propagation for root node

(c) *Plausibility* propagation for root node

(d) No propagation for leaf node

(e) *Belief* propagation for leaf node

(f) *Plausibility* propagation for leaf node

(g) No propagation for intermediate node

(h) *Belief* propagation for intermediate node

(i) *Plausibility* propagation for intermediate node

Figure 3.6: These graphs show how the evidence is spread considering *belief* and *plausibility* propagation on different types of nodes.

The mass function permits also to calculate the plausibility $Pl_d : V \to [0, 1]$. It is evaluated as follow: *Plausibility function*

$$Pl(v_d) = \sum_{v'_d \in V_d \wedge desc(v'_d) \cap desc(v_d) \neq \emptyset} c(v'_d) \tag{3.3}$$

where $desc(v_d)$ is the set of inclusive descendants of v_d according to the

considered G_O. This formula also permits to spread the evidence. In this case, the bottom-up propagation is completed putting the top-down one before. In this case, additional information is derived also from a value to its specialisations. A source expressing the value $v_d \in V_d$, considers as plausible/possible all claims containing its specialization $v'_d \in V_d$, i.e. values subsumed by v_d. Then, as a consequence of the bottom-up propagation, all values implied by plausible values v'_d are considered plausible as well. This propagation is called *plausibility* propagation. Considering Figure 3.6 and the different situations that can occur, we obtain that in situation (i), when the value is the root, the models behave in the same way that for *belief* propagation. Indeed, the considered value receives the same evidence, from all its "observed" descendants, see Figure 3.6c. Differently, considering (ii), when the value is a leaf, the value obtains information from all more general values, see Figure 3.6f. In situation (iii), the value is the middle of partial ordering structure, the evidence is given not only by its strictly more specific values, i.e. its descendants, but also by the ancestors of its descendants, see Figure 3.6i.

This mathematical framework lends itself to model the proposed rationale. Considering the different interpretations associated with the partial order relationship, two propagation models can be adopted based on, respectively, belief function and plausibility. The former is useful for spreading the evidence from most specific claims to most general ones, while the latter complements the former putting a propagation from most general values to most specific ones before it. For further information and technical details related to the application of belief and plausibility theory to partial ordering please refer to (Harispe, Imoussaten, Trousset, & Montmain, 2015).

3.1.2.3 Implications for the set of true values

In both cases, for each data item $d \in D$, the truth consists of a set of true values and not in a single value anymore[5]. Indeed, considering that a value implicitly supports its generalizations, if a value is considered true, then all its generalization should be considered true as well. Formally, $\forall(v, v') \in$

[5]It is important to underline that the definition of multiple solutions used in multi-truth approaches is different from the definition of true value set discussed herein. We are considering a functional predicate and the solution set is a consequence of using ordered values in the form of a partial order as additional knowledge.

${V_d}^2 : v \preceq v' \wedge tf(v) = true \Rightarrow tf(v') = true$ where tf is the truth function. For instance, considering the usual example, we know that *Málaga* is the actual birthplace of Picasso. Therefore, its generalizations such as *Spain* and *Europe* are considered true as well.

Multiple true values for functional predicate

Given this important consideration, a solution set has to be defined for each data item d, indicated $V^*_{d \in D}$. Constraints that shape its space (i.e. set) can be derived. Like the majority of classical approaches, we assume that the provided claims can be used to identify, not the true value here, but the set of true values associated with a data item. Therefore we assume that the set of true values for a data item d is bounded by the set of possible true values considering observed claims V_d. Based on the considered propagation model, a different set of possible true values is obtained. Given the *belief* propagation model, the set of true values for a data item d will be restricted to:

$$V^*_d \subseteq \cup_{v_d \in V_d} \{v'_d | v_d \preceq v'_d\} \tag{3.4}$$

Indeed this kind of propagation does not permit to say something related to more specific values than the ones stated. Given the *plausibility* propagation model, the true value set will be bounded to:

$$V^*_d \subseteq \cup_{v_d \in V_d} \{v'_d | v_d \preceq v'_d \vee v'_d \preceq v_d\} \tag{3.5}$$

Despite the true value set definition we have just enunciated, we still consider that, in absolute terms, and in accordance with the notion of functional predicate, a single expected true value exits for each data item d, e.g. the exact city of birth. This expected true value can be used to derive the set of true values associated with a data item d:

A single expected true value

$$\forall d \in D, \exists v^*_d \in V^*_d \text{ such as } V^*_d = \{v'_d | v^*_d \preceq v'_d\} \tag{3.6}$$

In other words, all generalizations of the expected true value are included in the set of true values. Nevertheless, without additional knowledge over V^*_d, it is impossible to decide on the true nature of the values v^*_d subsume – in this context, these values are assumed to be *potentially conflicting*. Indeed, the relation $\preceq$ is anti-symmetric, e.g. while saying that someone was born in *Granada* implies that he was born in *Spain*, claiming that the person was born in *Spain* is potentially conflicting with the fact that the person was born in *Granada*. Indeed, the person is not necessarily born in this specific *Spanish*

city even if it could be the case. Moreover, Equation 3.6 indicates that the set of possible true values V_d^* could contain not ordered pairs of values. This is not a problem. Even if these values are not ordered, they values are not conflicting since they have at least one common descendant in V_d^*, i.e. at least the expected true value v_d^*. Formally $\forall(x, y) \in V_d^{*2} : \neg(x \preceq y \vee y \preceq x)$, then $\exists z \in V_d^* : z \preceq x \wedge z \preceq y$. For instance, in Figure 3.5, *Spain* and *Capital* are not conflicting since *Madrid* is both a *Capital* and in *Spain*; however, *Málaga* and *Granada* are examples of conflicting values: they are not ordered and there is no specific value subsumed by both values.
Summarising, considering partial order of values when applying TD models to functional predicate, the solution for a data item d does not consist of a single true value. Instead, it consists of a set of true values V_d^* respecting the properties previously introduced in this section.

3.1.2.4 Applying TD-*poset*: adaptation of an existing model

An existing TD model: Sums

The existing TD model *Sums* (Pasternack & Roth, 2010) has been modified in order to deal with this new problem setting. *Sums* is an iterative procedure in which source trustworthiness and value confidence computations are alternated until convergence. It has already been described in section 2.2.5. Hereinafter, we recall its formulas:

$$t^i(s) = \alpha^i \sum_{v_d \in V_s} c^{i-1}(v_d) \tag{3.7}$$

$$c^i(v_d) = \beta^i \sum_{s \in S_{v_d}} t^i(s) \tag{3.8}$$

With t^i and c^i the estimated source trustworthiness and value confidence respectively at iteration i. Moreover, α^i and β^i are normalization factors that are equal to $1/(max_{s' \in S} \sum_{v_d \in V_{s'}} c^{i-1}(v_d))$ and $1/(max_{v'_d \in V} \sum_{s \in S_{v'_d}} t^i(s))$, respectively. They are required to keep the confidence and trustworthiness scores between 0 and 1. Note that the iterative procedure requires an initialization phase for one of the quantities that has to be estimated. All the confidence values are initialized to the same constant. Then, the trustworthiness of a source $s \in S$ is evaluated summing up all confidences of claims that are provided by s. Similarly, the confidence of a value $v_d \in V_d$ for a given data item $d \in D$ is computed summing up all trustworthiness of sources that claim v_d.

Sums can take advantage of *a priori* knowledge in the form of partial order of values in several ways. First of all, the different propagation processes to spread evidence we have just presented can be applied. Moreover, different formulas can be modified to incorporate the proposed rationale. Even if the additional information is strictly related to values, and therefore it should be integrated into the confidence formula, it can also be integrated into the trustworthiness one. One quantity is estimated using the other one. Moreover, since the algorithm is iterative, a refined estimation of value confidence also leads to a refined estimation of source trustworthiness and *vice versa*. Therefore, the additional information can be integrated into the confidence formula, into the trustworthiness one and in both of them. In the first case a different set of sources is considered for the computation of value confidences. In the second case a different set of claims is considered for the estimation of source trustworthiness. In the third case, a different set of both sources and claims are considered for the evaluation of, respectively, value confidences and source trustworthiness.

Belief-based adaptation

In the case of *belief* propagation framework, evidence associated with a value is spread to its generalizations. This adaptation is called $Sums_{PO}$ because we modified the *Sums* approach taking the partial order of values into account. Thus, to compute the confidence of a value v_d, the confidences of its specializations (that implicitly support it) are also considered. Since the confidence of a value v_d, at iteration i, is computed summing up all trustworthiness of sources providing v_d, when modifying the confidence it suffices to enlarge the set of sources including also the sources providing v'_d (with $v'_d \preceq v_d$). Indeed these sources, providing a specialization of v_d, implicitly support it. Formally, the new set of sources will be denoted by S_{v_d+} and will be composed of either sources providing the claim under examination or sources saying claims that implies v considering the partial ordering of values $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$. Since this model impacts the confidence calculus, it is named $Sums_{PO_C}$.
Differently, when modifying the trustworthiness formula, a new set of claims will be considered. In this $Sums_{PO_T}$ model, the estimation of source trustworthiness is improved by considering also the confidence associated with specializations of v_d, i.e. $v'_d \preceq v_d$. Indeed these values implicitly support v_d, the value used to evaluate the trustworthiness of a source. Formally the new set

of claims considered by $Sums_{PO_T}$ will be $V_{s+} = \{v' \in V | v \in V_s,\ v' \preceq v\}$. In the case where both formulas are modified, i.e. $Sums_{PO_{C+T}}$, the set of sources will be S_{v_d+} and the set of claims will be V_{s+}.

Plausibility-based adaptation

In the case of *plausibility* propagation framework, the model is called $Sums_{PL}$. The new source set will be represented by S_{v++} and will be composed of either sources providing the claim under examination, sources providing claims subsumed by v or sources saying claims induced by those claims subsumed by v, i.e. $S_{v_d++} = \{s \in S_{v'_d} | (v'_d, v''_d) \in {V_d}^2,\ v''_d \preceq v_d \wedge v''_d \preceq v'_d\}$. For instance, the evidence devoted to the claim *"Pablo Picasso was born in Spain"* is propagated to the claim *"Pablo Picasso was born in Madrid"*. The S_{v_d++} set will be used for modifying the formula for computing the confidence level of claims, i.e. $Sums_{PL_C}$. Considering, the modification $Sums_{PL_T}$, of the trustworthiness formula, a new claim set is introduced, represented V_{s++}, where, given the value v, all the plausible claims reported by a source will be employed for evaluating the trustworthiness of this source. The definition of this new set is the following $V_{s++} = \{v' \in V | v \in V_s, v'' \in V,\ v'' \preceq v \wedge v'' \preceq v'\}$. When both formulas are modified, the model is named $Sums_{PL_{C+T}}$. Note that all the different adaptations that are reported in this section are summarised in Table 3.2.

Moreover, note that in the case where no prior knowledge on value ordering is known, the adapted model corresponds to the traditional approach. Indeed if no relationship exists among values, then the S_{v_d+} set will be composed only by the sources claiming the value v and/or the V_{v+} set will be

Table 3.2: The several adaptations that can be obtained applying the proposed approach to *Sums*. Note that the subscripts in the model name specify which formula is modified (*C* indicates the modification of confidence formula and *T* indicates the modification of trustworthiness formula).

Model name	Propagation framework	Old set	New set
$Sums_{PO_C}$	Belief	S_{v_d}	$S_{v_d+} = \{s \in S_{v'_d} \| v'_d \in V_d \wedge v'_d \preceq v_d\}$
$Sums_{PO_T}$	Belief	V_s	$V_{s+} = \{v' \in V \| v \in V_s,\ v' \preceq v\}$
$Sums_{PL_C}$	Plausibility	S_{v_d}	$S_{v_d++} = \{s \in S_{v'_d} \| v'_d, v''_d \in {V_d}^2,\ v''_d \preceq v_d \wedge v''_d \preceq v'_d\}$, with $S_{v_d++} \supseteq S_{v_d+}$
$Sums_{PL_T}$	Plausibility	V_s	$V_{s++} = \{v' \in V \| v \in V_s, v'' \in V,\ v'' \preceq v \wedge v'' \preceq v'\}$, with $V_{s++} \supseteq V_{s+}$

composed only by the explicitly claimed claims. Therefore, the computation will be the same than the *Sums* model.

Necessity of a new method to select the true values

An important consequence of the consideration of a partial ordering of values is the characteristic of resulting value confidence estimations. Therefore, the adaptation of *Sums*, or any other method, also implies modifying the way the true values will be distinguished after the stopping criterion is verified. Existing approaches usually assume that for a specific data item d, elements of V_d are disjoint/independent and they, therefore, recognize the true value of d being that with the highest confidence score. This straightforward procedure cannot be applied using models that consider ordering among values during the estimation phase. Incorporating this information into the model, considering all values associated with a data item, implies that the resulting estimated confidence scores monotonically increase from the leaves to the root, i.e. $\forall v, v' \in V^2$: if $(v \implies v')$, then $c(v) \leq c(v')$. Therefore the most general value (that is implicitly supported by all provided claims) will always have the highest confidence and selecting it as the true value would make no sense. To solve this problem, we propose a post-processing procedure able to select the expected true value for each data item given the estimated confidence scores and the relationships that may exist among values. The next section presents this procedure.

Resulting confidences monotonically increase

3.2 Truth selection algorithm through the use of a partial order among values

Discovering the set of true values implies resolving a problem of sub partial ordering identification w.r.t. the confidence associated with the values. Hereinafter, we present several examples representing situations that may occur.

3.2.1 Intuition

Considering the previous example *"Where was Pablo Picasso born?"*, the provided values in Table 3.1 and the partial order among these values according with Figure 3.5, several situations may arise based on the different value confidence estimations that can be obtained. For instance, two possible value confidence estimations are presented in Figure 3.7. In this figure for each

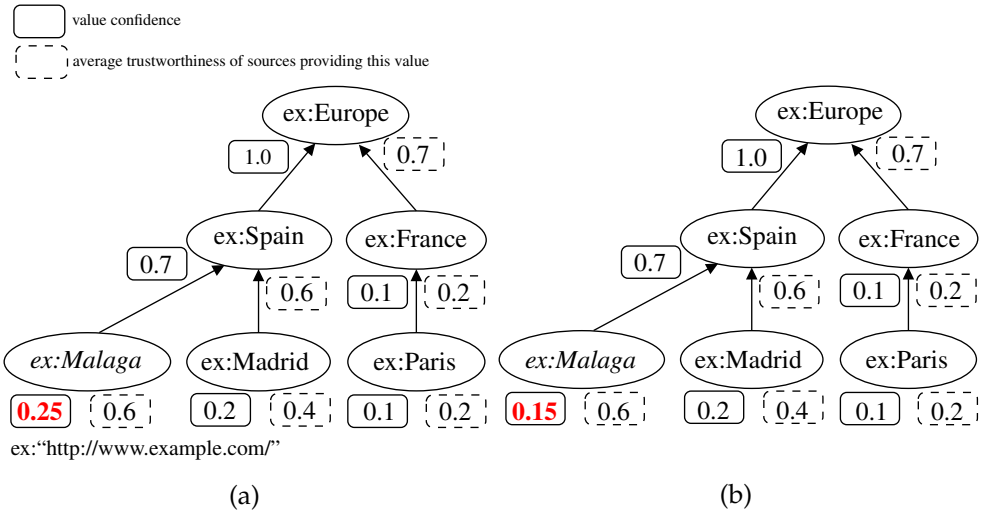


Figure 3.7: Example for representing different scenarios that may occur when *ex:Malaga* is the true expected value. The solid rectangles contain value confidences, and the dashed rectangles indicates the average trustworthiness of sources providing a value.

value, its confidence and its average trustworthiness of sources providing it are reported respectively in the solid and dashed rectangles near the corresponding node. In both cases, the highest confidence is associated with *"Europe"*. Indeed, due to the *Evidential* theory, all sources support it. Selecting this value as the true answer is not informative, indeed, when asking for the birth location of someone, usually the expected answer is a city name, not a continent. When trying to identify the city of birth, different situations can occur, see Figure 3.7a and 3.7b. The two figures differ in the confidence associated to the value *"Málaga"*. In the first case, the true value *"Málaga"* can be identified using a simple procedure that selects the value with the highest confidence at each level of the partial order. This procedure does not work in the second situation. In such a case, it is better to consider also other dimensions such as the average trustworthiness to discriminate different values.

Moreover, when the evidence associated to a value, e.g. city of birth, is too weak, it could be convenient returning its generalization, i.e. country of birth. For instance, considering a confidence score lower than 0.4 not reliable, the returned birth location of the painter should be *Spain* and nothing more specific.

Based on the possible situations that may occur, a parametrisable post-processing procedure has been proposed to face all possible scenarios.

3.2.2 Truth selection algorithm

In this section, we propose a post-processing procedure that selects the true values given the estimations obtained by TD models that relax the assumption related to the disjointness of values.

As shown in Figure 3.8, it involves three steps: selection of the best true value candidates; ranking of selected values; and filtering of ranked values according to defined desirable properties. For instance it may be useful to return a set of solutions that share ordering relationships or, on the contrary, to return a value set composed only of "alternatives" that are not ordered. The choice related to the appropriate features of the solution set depends mainly on the application scenario.

The first step of the process permits to retrieve the most specific true value(s) and all of its ancestors using available information, such as confidence scores and partial order of values. The second step consists of ordering the selected values based on predefined criterion. The third step is required to filter the top-k results. For TD final aim, i.e. identifying the expected true value, k should be equal to 1. However in cases where there is a lot of un-

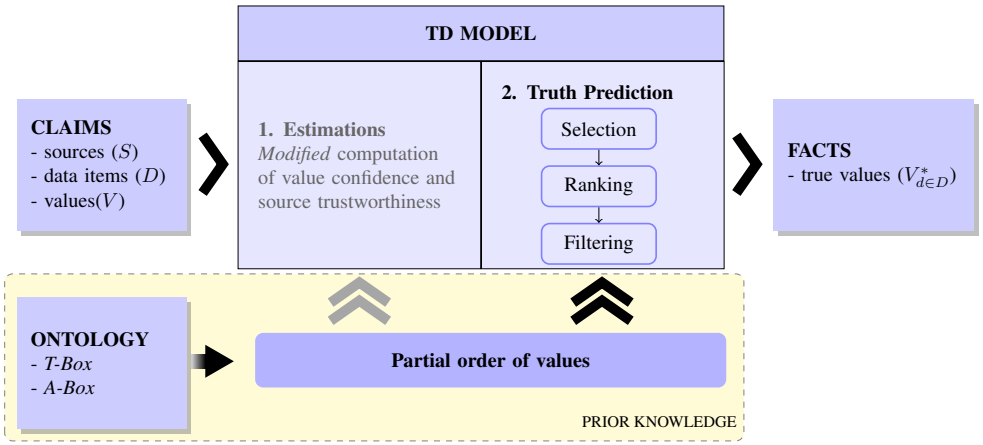


Figure 3.8: Diagram of the overall Truth Discovery (TD) procedure incorporating the partial order of values during the truth prediction phase to improve TD performance.

certainty it may be useful to return a set of values, even if the predicate is functional. Moreover, answers that do not have defined desirable properties are removed from the result list (see "Filtering of top-k true values" paragraph for further details on page 71). These three steps are detailed hereafter.

3.2.2.1 True value selection

The first part of the post-processing procedure concerns the selection of the promising candidate(s) as the most expected value(s) for each data item. We have defined a selection strategy that takes advantage of the partial order of values and, refining, step by step, the granularity of the true values for each data item. In the rest of this section, an overview of the approach followed by a description of Algorithm 1 on page 69 is provided.

A parametrisable traversing procedure

Starting from the most general value (implicitly supported by all provided values and surely true), the process aims to detect the expected true value(s). Thus, a traversing procedure is applied on the graph that represents the partial order of values. It starts from the root, selects the best alternatives among the children of the considered value, and moves forward through the selected ones. Our assumption is that, at each step, values with the highest confidence should be the most likely to be true. Therefore the choice of the best alternative(s) is done by comparing the confidence scores associated with the children of a value. In the case of functional predicate, the values can be partially ordered by their granularity. Therefore the selection procedure refines, at each step, the level of precision used to describe the expected true value associated with a data item. The semantics of each selected node representing a value expresses the fact that the node is or subsumes the correct solution (i.e. the expected true value). The last selected node should correspond to the most specific true answer that can be identified through the selection process.

The selection process has to address two main undesirable situations that may occur: (1) selection of values having a confidence score that is excessively low to be considered true, and (2) difficulty in discriminating the best alternative(s) among the children of a node since their confidence scores are not significantly different. As a solution, two thresholds have been defined. They have been arbitrarily represented by θ and δ.

The threshold $\theta \in (0, 1]$ specifies the minimum confidence score that is required for a value to be part of the set of true values. Note that the value 0 is not included in the domain of θ. Indeed, considering claims with confidence scores equal to 0 makes no sense because it would mean considering, as truth, values provided by none or totally unreliable sources (all with trustworthiness equal to 0). The confidence score that is compared to θ has to be previously normalized according to each data item, i.e. the confidence score associated with the most general value of each data item has always to be equal to 1. This normalization step is required to avoid the definition of an inconsistent threshold according to the different data items.

Definition of threshold θ

The threshold $\delta \in [0, 1]$ represents the maximum admitted difference between the highest confidence and the confidence of any other selected values. We therefore consider that if the difference between the confidences of two values is less than or equal to δ, then it is hard to make a choice among them. Thus, both values are returned.

Definition of threshold δ

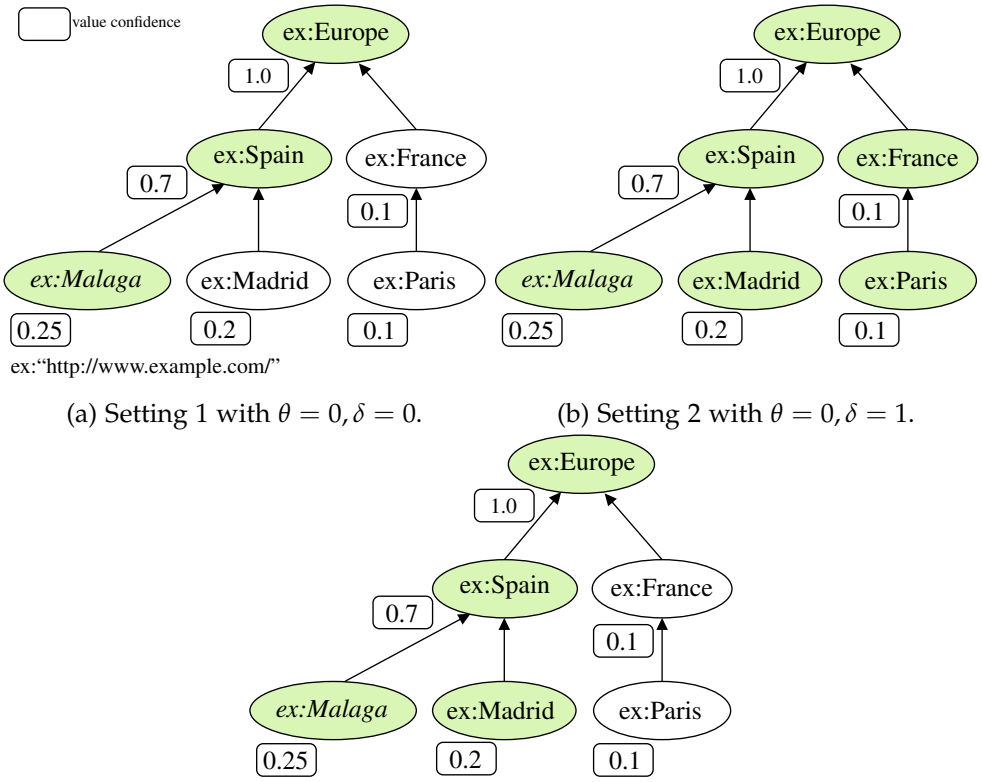
Different parameter settings produce different behaviours of the selection phase ending in the possibility of obtaining different kinds of solution sets. The main ones are summarized in Table 3.3.

Setting 1 ($\theta = \alpha, \delta = 0$) results in a naive greedy algorithm that, at each step, selects values having the highest confidence greater than α without performing any other control. Considering the example of a possible situation resulting from the estimation phase that is reported in Figure 3.7a, the graphical representation of the selection procedure behaviour when $\alpha = 0.0$ is reported in Figure 3.9a. In this figure the set of selected values is highlighted in green.

Setting 2 ($\theta = \alpha, \delta = 1$) is able to return all claimed values and their ancestors

Table 3.3: Interesting settings for the selection procedure.

Setting	θ	δ	**Selection procedure behaviour**
1	α	0	Naive greedy procedure that maximizes the confidence score at each step, until the confidence is higher than α.
2	α	1	All values with confidence higher than α are selected, as well as their ancestors.
3	α	β	At each iteration a value is collected only if the difference of its confidence and the highest confidence at the current step is lower or equal to β. Moreover, the confidence of all values in the returned set is greater than α.



(a) Setting 1 with $\theta = 0, \delta = 0$.

(b) Setting 2 with $\theta = 0, \delta = 1$.

(c) Setting 3 with $\theta = 0, \delta = 0.3$.

Figure 3.9: Example of selection procedure considering different settings. The true expected value is Málaga and the set of values that are returned by the considered setting are highlighted in green.

with confidence higher than α. Indeed, when $\delta = 1$ discriminating values on their confidence is not possible anymore. This setting may seem useless, but it is necessary to obtain a particular set of values at the end of the post-processing procedure. Indeed, in cases where there is a lot of uncertainty, obtaining a set of "promising" alternatives may increase the probability of finding the expected true value. Indeed, the returned set is composed of values that are, as much as possible, fine-grained and semantically different. Therefore, returning all claimed values and their ancestors is useful because, during the ranking phase, it will be possible to position in the first places the most promising alternatives. Applying this setting to the example reported in Figure 3.7a, the returned values are the values that are highlighted in green in Figure 3.9b.

Setting 3 ($\theta = \alpha, \delta = \beta$) is a generalization of the two previous configurations. It selects the set of values with a confidence that is greater than α and that differs, at each step, less than δ from the confidence of the other alternatives. Figure 3.9c reports the values that are selected by the selection procedure using Setting 3 and considering the example reported in Figure 3.7a.

Algorithm 1 reports the pseudo-code of the selection procedure. The algorithm starts performing a transitive reduction of the graph representing the partial ordering (line 2). This ensures that the choice of the best alternative is done among a set of children that do not share ordered relations. Moreover, it avoids useless comparisons of a large number of confidence scores. Then, at each iteration, the algorithm applies a greedy search by maximizing the confidence of the values (lines 5 – 9). It selects all values having confidence higher than or equal to θ whose scores are not significantly different from the highest confidence (line 10). Then, it adds them to the queue if they are not already visited (line 11). The procedure stops when the queue is empty. In order to be in accordance with the definition of true value set (Eq. 3.6), all values that are ancestors of the visited ones will compose the set of possible true values represented by $V^*_{candidates}$. Indeed, due to multiple inheritances some of those values may not have been visited by the greedy

Algorithm of true value selection procedure

Algorithm 1 True value set computation for any $d \in D$ considering a partial order of values represented as a DAG $G_O = (V, E)$, a threshold $\theta \in (0, 1]$, a threshold $\delta \in [0, 1]$, and a function $c : V \to [0, 1]$, i.e. confidence of each value.

1: **procedure** SELECTIONTRUEVALUES($d, G_O, c, \theta, \delta$)
2: $G \leftarrow transitive\_reduction(G_O)$
3: $V^*_{visited} \leftarrow \{\ \}$
4: $queue \leftarrow list(ROOTS(G))$
5: **while** $!(queue.isEmpty())$ **do**
6: $v_d \leftarrow queue.pop()$
7: $V^*_{visited} \leftarrow V^*_{visited} \cup \{v_d\}$
8: $V_{ch} \leftarrow chil(v_d)$
9: $conf_{max} \leftarrow \max_{child \in V_{ch}} (c(child))$
10: $V_{ch^*} = \{v'_d \in V_{ch} : c(v'_d) \geq \theta \wedge (conf_{max} - c(v'_d)) \leq \delta\}$
11: $queue.addAll(V_{ch^*} \setminus V^*_{visited})$
12: **end while**
13: **return** $V^*_{candidates} = \bigcup_{v_d \in V^*_{visited}} anc(v_d)$
14: **end procedure**

procedure (line 12). The fact that confidence estimations monotonically increase according to the partial order guarantees that the confidences related to ancestors of the visited values are higher than or equal to θ.

The termination of Algorithm 1 is ensured by line 6 and line 11. The complexity of the selection of the true value algorithm is related to the number of comparisons required to find the maximum value confidence traversing graph G_O. Therefore, the complexity of the algorithm is $O(E)$ which in turn is bounded by $O(V^2)$. At each step, a number of comparisons equal to the number of children is required. The worst case scenario is verified when the following conditions hold at the same time: (i) graph G_O has depth 2, (ii) its nodes are uniformly distributed between level 2 and 3, (iii) nodes at the same level have the same fathers and the same children and, moreover, (iv) they have equal or not significantly different confidence scores. The conditions (i), (ii) and (iii), related to the topology of the DAG, ensure that the number of comparisons is maximum, and the condition (iv), related to value confidence, guarantees that the procedure traverses all nodes.

All of the configurations of the algorithm input parameters enable us to select a set of possible true values. Since the aim of TD is to find the expected solution, a method that is able to select it is required. The ranking phase described in the next section is devoted to this aim.

3.2.2.2 True value ranking

Given the possible true value set $V^*_{candidates}$ selected in the previous step, we have to define a ranking method in order to select the $k \in \mathbb{N}^+$ most expected values for each data item d where k is a fixed number. In our investigations, k is experimentally set, at the most, at 5. The solution set of most expected true values for a data item d is indicated as $V^*_d \subseteq \mathcal{P}(V^*_{candidates})$.

IC-based ranking

The values in $V^*_{candidates}$ can be ranked based on their IC, see section 3.1.2.1. This method helps to discriminate different values considering their granularity. It is useful for situations in which specific answers are expected/preferred and when there is not much uncertainty on the data item under consideration. Note that in the following experiments IC is a measure computed according to the definition provided by Seco (Seco et al., 2004). Based on the analysis of the partial ordering topology, it takes advantage of the number

of descendants of a value:

$$IC_{Seco}(v_d) = 1 - \frac{log(|desc(v_d)| + 1)}{log(|V|)} \tag{3.9}$$

where $|V|$ is a non-empty set since an ontology is considered to have at least one value, i.e. at least it has one root value.
IC has been proposed as ranking criterion because the user generally expects specific answers. Often general true values, for a data item, are already well known, i.e. it is known that a person was born in a place. If multiple possible true values have the same IC, then a random selection can be performed or another criterion can be used to rank this subset of values.
Alternatively, the values can be ranked based on their source average trustworthiness, denoted WA_{trust}. The rationale is that if a lot of unreliable sources support a false value A, and there are only a few reliable sources that support a true value B, then sources providing A should have lower average trustworthiness scores than sources providing B. This measure is obtained by computing the average trustworthiness associated with sources that explicitly or implicitly claim a value v_d and by weighting it by a normalization factor:

WA_{trust}-based ranking

$$WA_{trust}(v_d) = \left(1 - \frac{1}{\eta + |S_{v_d+}|}\right) avg_{trust}(v_d) \tag{3.10}$$

where the average source trustworthiness is represented by avg_{trust}, S_{v_d+} is the set of sources that implicitly or explicitly provide the value v_d and η is a small number used to avoid that $WA_{trust}(v_d) = 0$ when v_d is provided by only one source. The normalization factor was introduced in order to tune the average according to the number of sources providing the value. Indeed, inspired by a previous study, the higher the number of sources providing a value, the higher our confidence in the computed average should be (Jean, Harispe, Ranwez, Bellot, & Montmain, 2016). Moreover also in this case, when multiple values have the same WA_{trust}, another criterion can be used to rank them.
Once the values are ranked, the next and final step of the post-processing procedure can be performed.

3.2.2.3 Filtering of top-k true values

The filtering phase collects the top-k values in the rank and returns them to end-users. Before performing selection of the top-k values, all the ranked

ones have to be controlled. This is necessary because TD models can be applied to different scenarios: high or low uncertainty situations, high or low risk cases in which making an error is, respectively, very dangerous or not. For instance, if TD models are used to populate a medical knowledge base containing, for each symptom, all possible correlated diseases, then the end-users want to be really careful in accepting a value as true. Therefore, based on the possible application contexts, different properties that the solution set V^* has to respect can be defined. In this way, various true value sets with different characteristics can be identified:

Returning ordered values

- V^*_{ord} that is the solution set containing only values that share partial ordering relationships; formally $\forall(x,y) \in {V^*_{ord}}^2, x \preceq y \vee y \preceq x$. This set is created iteratively selecting and removing the first element of the ranked list returned by previous phase. Each time, this element is added to the solution set only if it is an ancestor or descendant of all elements that are already present in it. This kind of solution can be desirable when there is not much uncertainty (end-users expect to easily find the true answers) or the end-users do not want to deal with potentially different values in a domain where they are not experts.

Returning not ordered values

- V^*_{disj} that is the solution set whose values do not share any partial ordering relationships and are as much as possible very specific and different; formally $\forall(x,y) \in {V^*_{disj}}^2, \neg(x \preceq y) \wedge \neg(y \preceq x) \wedge \nexists\ (w,z) \in {V^*_{candidates}}^2$ such as $w \prec x \vee z \prec y$. This means that all values in the solution set are the most specific among those returned by the selection phase (they have not descendants in the sorted list). In other words, this set of values consists of elements that do not have any of their exclusive descendants in the sorted list. For example, if the ranked list is [*Europe*, *Continent*, *Country*, *City*, *Location*], then considering the partial order in Figure 3.5, the $V^*_{disj} = \{Europe, Country, City\}$. This property can be adopted when there is a lot of uncertainty and especially when the application context allows making errors without dangerous consequences. Indeed, when there is uncertainty, to postpone the selection of true values to the end-users, avoiding to automatically select only a specific value and its ancestor, may be useful. In order to support the end-users final choice, returning a set of values composed of the most promising alternatives is important.

Considering the different parameter settings of the true value selection phase, we observed that obtaining V^*_{ord} is suitable when $\delta = 0$. Indeed, taking the value with the highest confidence at each step, the process ends with the selection of only one specific true value (and its ancestors), see Figure 3.9a. Considering this set of returned values, the first property is often verified without filtering any value out. Note that often very general values are not returned since only the top-k values are selected after the verification of the property. Otherwise, V^*_{disj} is not useful considering $\delta = 0$. Only the single most specific value contained in the set of returned values is selected when this property must hold. Indeed, all of the others share partial ordering relationships. This corresponds to consider that $\delta = 0$, $k = 1$ and a solution set V^*_{ord}. Instead, V^*_{disj} is preferable when $\delta = 1$. Indeed, as explained previously when presenting the different selection procedure settings, in those cases all values having confidence higher than θ are returned by the selection phase (see Figure 3.9b), but for the final aim of TD (finding the truth) it is suitable to only keep the set of "promising" alternatives that correspond to a set of values that are different and specific as much as possible.

3.3 Experiments

In this section we describe all experiments performed to evaluate the validity of the proposed approach. First, we introduce how the synthetic datasets used in the experiments have been generated. Then, we present the methodology that has been applied to evaluate the approach and compare it with existing models. The main aim is not to illustrate that the proposed models are able to outperform existing methods – classical models have initially not been defined to handle the use case studied in this work. We are rather interested by analysing how the proposed adaptations perform depending on the granularity distribution of provided true values, i.e. level of details used to provide a true value (a real world entity can be represented using specific or more abstract values).

3.3.1 Synthetic datasets

Existing publicly available datasets that are usually used in TD domain contain values without partial ordering relationships defined specifically on

value granularities. For instance, the most popular dataset in TD domain is the Author dataset (X. L. Dong et al., 2010) which contains a set of claims where each of them provides a list of authors for a specific book. Full names cannot be express with different level of granularities. At most, a person can provide only the surname of an authors. The same situation is verified for other public datasets, e.g. the Population and the Biography dataset (Pasternack & Roth, 2010). Indeed, the former includes the size of each city provided by several sources, while the second is composed of birth and death dates for a set of people. Therefore, for evaluating models taking advantage of partial ordering of values, we have generated several synthetic datasets based on the well-known knowledge base DBpedia[6] (Auer et al., 2007) and its associated ontology, and the Gene Ontology[7] and the associated annotations of genes (called GOA) proposed by gene ontology community (Ashburner et al., 2000). Precisely, 5 distinct datasets have been generated based on the predicate appearing in their claims: *birthPlace* and *genre* predicates have been selected from DBpedia, and *Cellular Component* (*CC*), *Molecular Function* (*MF*) and *Biological Process* (*BP*) have been selected from Gene Ontology. The detailed procedure which has been used for generating the datasets is described hereinafter.

The main elements required to generate these datasets are:

Required elements for dataset generation

- a ground truth specifying the expected true values for a set of data items;
- a partial order of values;
- a set of sources;
- a set of claims provided by these sources on different data items.

Generation of ground truth and partial order of values

Since similar but different procedures have been conducted for generating the ground truths and the partial order of values for the predicates from the two different ontologies we used, we detail them separately.

DBpedia. Based on this ontology, two datasets have been generated considering `dbo:birthPlace`[8] and `dbo:genre` predicates. For `dbo:birthPlace` predicate, the ground truth has been generated by extracting a set of pair

[6] version 2015-04
[7] version 2016-06-29
[8] The prefix `dbo` stands for `http://dbpedia.org/ontology/`

(*subject*, *object*) associated with the predicate `dbo:birthPlace`. 534723 data items having a unique object for the property `dbo:birthPlace` have been extracted from DBpedia using the SPARQL query 3.1. Indeed, cases in which the same subject has more than one object for this property have been excluded since in some situations these values are even conflicting. For instance the subject `dbr:Francis_Scott_Key`[9] is associated with 4 different values. Among them, the values `dbr:Frederick_County,_Maryland` and `dbr:Carroll_County,_Maryland` are conflicting. Indeed, both are counties of Maryland and it is well known that a person cannot be born in two difference counties.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-
   ↪ ns#>
SELECT ?s, ?o
WHERE{
   ?s dbo:birthPlace ?o .
   {
      SELECT ?s COUNT(?o)
      WHERE{
         ?s rdf:type dbo:Person .
         ?s dbo:birthPlace ?o .
      }
      GROUP BY ?s
      HAVING (COUNT(?o) < 2)
   }
}
```

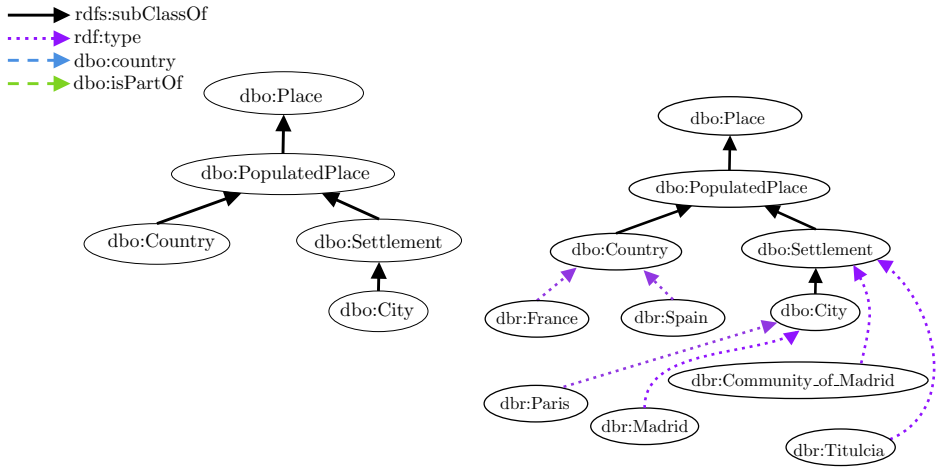
Listing 3.1: SPARQL query

This set of facts defines the set of data items D as well as the true values v_d^* associated with each data item. We assume that all claims are true, indeed, due to the way the dataset is generated, it will not impact our evaluation; considering these claims was mainly motivated by the will to mimic as much as possible a real scenario according to value distribution. For our experiments a subset of 10000 data items having their object values present in the partial

[9]The prefix `dbr` stands for `http://dbpedia.org/resource/`

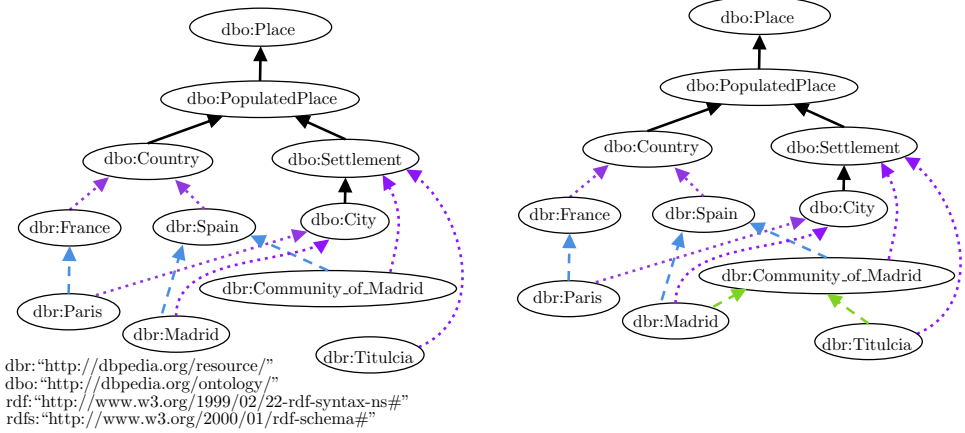
order structure have been filtered out randomly.
The partial order of values has been constructed using DBpedia ontology. First of all, we have selected all the triples that contain `rdfs:subClassOf`[10] as predicate. Among them, we have extracted only the subset of classes that are subsumed by `dbo:Place` class, see Figure 3.10a. Then, we have loaded the triples involving the `rdf:type`[11] predicate making the instances of `dbo:Place`, see Figure 3.10b. Successively, all triples that have `dbo:country` predicate have been added, see Figure 3.10c, as well as the triples with `dbo:isPartof` predicate, see Figure 3.10d. At this point, the partial order has been populated using all triples selected so far without any distinction among the type of relationships considered. Since, `dbo:Thing` is the most abstract concept in DBpedia, we have rooted all the concepts belonging to the graph with it. Then, in order to be sure to obtain a partial order of values, we have checked if the obtained RDF graph respects the properties of a DAG. The examination has shown the presence of cycles that have been induced by incorrectness on part-of triples. Thus, a heuristic has been applied to remove these cycles. For each of them, we have rejected the edge that have as target the node with the highest out-degree. Indeed, the heuristic we have employed hypothesizes that more abstract concepts should have higher out-degree that the less abstract ones. For instance, to remove the cycle existing between the resources `dbr:The_Bronx` and `dbr:New_York_City`, we compared their out-degree that are, respectively, equal to 6 and 65. According to our heuristic the concept `dbr:New_York_City` is more abstract than `dbr:The_Bronx`. Therefore, among the two edges, we eliminated the one whose target was `dbr:The_Bronx`. The validity of this choice is confirmed by the reality. Indeed, the resource `dbr:The_Bronx` represents one of the five boroughs of New York City, and therefore `dbr:The_Bronx` is a part of `dbr:New_York_City`. Analysing the discarded edges, the behaviour supposed by the heuristic has been well respected. After the elimination of the cycles, the partial ordering of values has been obtained.
In order to consolidate the results we obtained, similarly to this procedure, we generated datasets using a different predicate, i.e. `dbo:genre`, and its related partial ordering.
The entire procedure involving the selection of `dbo:birthPlace` and `dbo:genre`

[10] The prefix `rdfs` stands for `http://www.w3.org/2000/01/rdf-schema#`
[11] The prefix `rdf` stands for `http://www.w3.org/1999/02/22-rdf-syntax-ns#`



(a) Extracting only the subset of classes that are subsumed by `dbo:Place`.

(b) Loading triples of type `dbo:Place`.



(c) Loading triples that have predicate `dbo:country`.

(d) Loading triples that have predicate `dbo:isPartof` and the concept `dbo:Thing`.

Figure 3.10: Procedure to construct the partial order associated with *birthPlace* predicate.

triples with the construction of the related partial ordering of values has been implemented using the Semantic Measures Library (Harispe, Ranwez, Janaqi, & Montmain, 2014).

Gene Ontology (GO). Using this ontology we generated three datasets related to the three facets of all gene product properties such as they are classified in GO, i.e. *Cellular Component* (CC), *Molecular Function* (MF) and *Biologi-*

cal Process (BP). These three aspects are treated as three different predicates for the dataset generation procedure. For creating the ground truth for each predicate, we selected a subset of statements reported in the GO annotation file[12] where all descriptions about the functions associated with a gene for a specific aspect are reported. Since product genes may have multiple functions for the same aspect and we deal with functional predicate, we introduce the notion of context_id. The aim is to treat a gene and its associated functions are treated as distinct cases. For each function associated with a gene, a different context_id is considered. Thus, the triple contained in the ground truth are expressed as <gene_id+context_id, predicate, function>. Each of the associated functions is considered as the truth in a specific context. This strategy does not impact the results of our experiments. We are not interested in the semantic meaning of the values, but in the structure among them. Therefore, all the values could be potentially replaced with others. The extraction of the partial ordering of values related to these predicates was simpler than the one for deriving the partial ordering from the DBpedia. Indeed the GO is already a DAG composed of only is-a (if x is-a y, then x is a sub-type of y) and part-of (if y part-of x, then y implies x) relationships, see Figure 2.9. Since all of them are ordering relationships, the construction of the partial order over the values was immediate. We selected all concepts subsumed by the concept *CC*, *MF* and *BP* without any additional operations.

Table 3.4 reports the features associated with the partial ordering structures

Table 3.4: Partial order features: *Cellular Component* (*CC*), *Molecular Function* (*MF*) and *Biological Process* (*BP*) from Gene Ontology and *birthPlace* and *genre* from DBpedia.

Features	*CC*	*MF*	*BP*	*genre*	*birthPlace*
Values	3984	10243	28822	1838	682658
Max depth	12	15	16	8	14
Average depth	5.223	5.610	6.906	3.93	5.424
Average children #	1.451	1.196	1.898	1.041	1.535
Max children #	466	291	451	824	160194
Leaves	3016	8192	14797	1563	663373

[12] goa_human.gaf, v. 2016-07-07

that have been generated for each predicate.

Generation of claims

Considering the ground truth and the partial ordering of value for each predicate, the generation process of the claim set can start. Table 3.5 summarises the features regarding the generation of the claim set that is detailed hereinafter.

Modelling a priori source trustworthiness

First of all, a set of sources with the related source trustworthiness level has been generated. In order to simulate as much as possible a real scenario, we assumed that the majority of the sources are sufficiently reliable and only a few of them are always or never correct. To reproduce this behaviour, for generating pre-defined source trustworthiness, a Gaussian distribution with average and a standard deviation equals to, respectively, 0.6 and 0.4 was used. The trustworthiness of each source is used to decide if the source has to provide a true or false claim. In order to assess that actual trustworthiness score is actually respected for each synthetic datasets, *a posteriori*, we analysed that the rate of true and false values provided by a specific source

Table 3.5: Features of synthetic datasets.

Feature	**Description**
Source Coverage	Each source provides a number of claims that is exponentially distributed.
Source trustworthiness	The trustworthiness distribution is a Gaussian having average and standard deviation equal to, respectively, 0.6 and 0.4. This means that sources are mostly reliable and only a few of them are always or never correct.
# of true claims per source	Each source provide true values according to its trustworthiness level.
# of distinct true values per data item	$1..\|V_d^*\|$ where $V_d^* = \{v \in V : v_d^* \preceq v\}$
Granularity of the provided true value	Each source provides a true value having a granularity that approaches the granularity of the expected true value according to an exponential distribution with high decay rate (EXP), exponential distribution with low decay rate (LOW_E) and a uniform distribution (UNI).
# of distinct false values per data item	1..30 values belonging to $V_d^{false} = \{V_d \setminus V_d^*\} \setminus \{v \| v \preceq v_d^*\}$

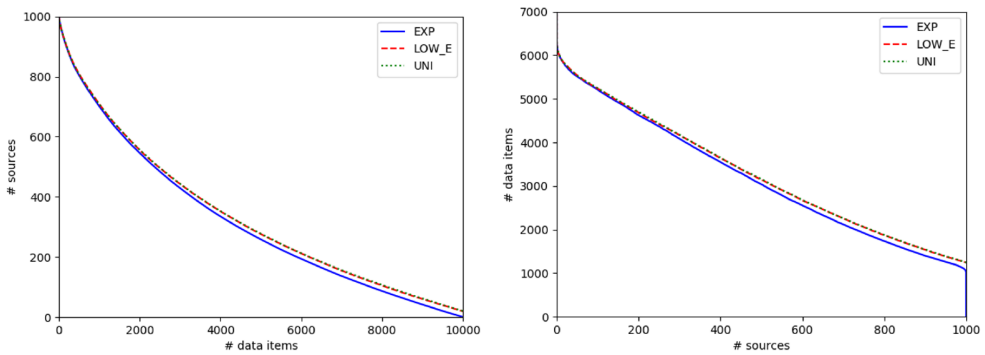
is in accordance with it. (i.e. the pre-assigned trustworthiness level).

Modelling source coverage

A source does not provide a claim for each data item in the ground truth – we consider that sources only express values for some data items. We therefore consider that each source has a different coverage over the entire data item set. In particular, the following behaviour is modelled: a lot of sources provide claims on few data items (in other words, on a few topics), while only few sources claim on a wide range of data items. Considering Web data, this reflects the real-world situation where many websites are specialized on specific topics, while only few websites provide information on a broader range of themes. This rationale represents the long-tail phenomena that is common in many real-world applications. It was mentioned for the first time in the truth-discovery domain by Li et al.(Q. Li, Li, Gao, Su, et al., 2014). The statistic that confirm that this behaviour is respected by the datasets that were generated are reported in Figure 3.11. In Figure 3.11a we observe that approximatively 80% of data items are claimed by less than 500 sources. Figure 3.11b shows that most of sources have provided at least 1000 data items.

Identifying true and false value domains

Each source claims a true or false value for a specific data item according to its trustworthiness.
In case of true claims, the value is selected among the inclusive ancestors of the expected true value specified in the ground truth. In the case of false claims, it is selected from the set of values that are neither inclusive ancestors nor descendants of the expected true one. The descendants represent



(a) # of sources per data item

(b) # of data items per source

Figure 3.11: Statistics of sources-data items for the *Cellular Component* (CC) datasets.

potentially true values on which no other knowledge is given, they cannot therefore be considered all false *a priori*. For this reason, we removed them from false value domain. Formally, given a data item $d \in D$ and its expected true value v_d^* the set of true values is, according to our definition, $V_d^* = \{x | v_d^* \preceq x\}$ and the set of false value is defined by $\{V_d \setminus V_d^*\} \setminus \{x | x \prec v_d^*\}$. In both cases the values are selected according to a similarity measure between the values and v_d^*. To this end, Lin's measure using Sanchez et al. IC has been used – technical aspects related to semantic measures are briefly summarized below; several existing studies contain more details about them (Harispe, Ranwez, et al., 2015; Lin, 1998; Sánchez & Batet, 2011). The semantic similarity of two values u, v that are defined into the partial ordering is computed using $sim_{Lin} : V^2 \rightarrow [0, 1]$:

$$sim_{Lin}(u, v) = \frac{2 \cdot IC(MICA(u, v))}{IC(u) + IC(v)} \tag{3.11}$$

with IC a function used to compute the Information Content of a value by analysing the topology of the partial ordering, refer to the work of Sanchez (Sánchez & Batet, 2011) for the formula used in this stud, and $MICA(u, v)$ the Most Informative Common Ancestor of the two values u, v, i.e. the value that generalizes both u and v which has the higher IC score.[13]

Selecting values for true claims

For the selection of the true values, three different strategies were used to select the granularity of the provided values. This is the main characteristic of the synthetic datasets, indeed based on that, three types of datasets have been generated: EXP, LOW_E and UNI (Figure 3.12) figuring respectively the behaviour of experts, a mix of experts and non-experts, and non-expert users. Therefore EXP simulates cases in which sources are quite sure about the true values, so they tend to claim values similar to the expected one (contained in the ground truth) when they have to provide a true value. As a result, only a limited number of sources claim general true values[14]. UNI reproduces a world where there is a lot of uncertainty, then the sources tend to indiscriminately select the value from the entire set of possible true values,

[13] Semantic similarity computations have been performed using the Semantic Measures Library (Harispe et al., 2014).

[14] According to the partial ordering, the way the true value set is derived, and the selected similarity measure, given a target value of this set, its ancestors with a highest level of specificity are more similar to the target than the ones with a lower level of specificity - specificity refers to the notion of Information Content (Harispe, Ranwez, et al., 2015).

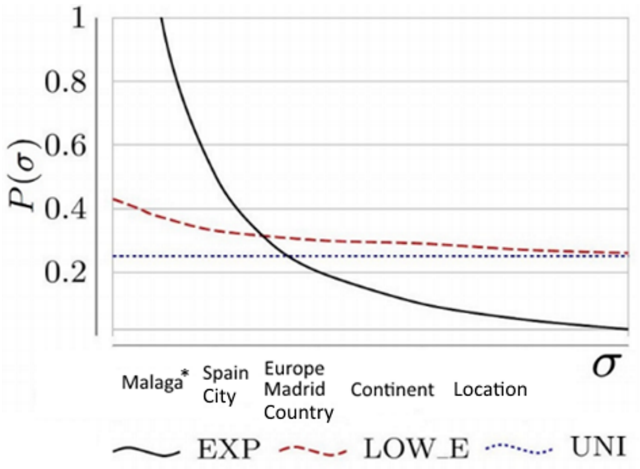


Figure 3.12: Distributions for true value selection.

i.e. uniform distribution among true values. LOW_E is a trade-off between the previous two types. Sources uniformly select the value from the set of possibilities, but there is a slightly higher probability of choosing values similar to the expected one.
For instance, Figure 3.12 reports, on the x-axis, the values of Figure 3.5 ordered according to their similarity measures according to the true value *Málaga*. Considering that v_d^* is *Málaga* and the EXP law, sources will more often provide values such as *Málaga*, *Spain* and *City* than values like *Continent* or *Location*. Otherwise, considering the UNI distribution, the probability of claiming these values will be the same. For each scenario 20 synthetic datasets have been generated.

Selecting values for false claims

For the selection of the false values, only a single strategy has been considered. A source that has to provide a false claim tends to provide a false value that is similar to the expected true value. For instance, considering Figure 3.12, if the true value is *Málaga*, then a source provides the value *France* with a higher probability than the value *Brazil* due to the fact that *France* is an European country. Moreover, sources tend to claim the same false values. Therefore, the probability of a value to be selected as false one increases according to the number of sources that previously claimed it. Therefore, in this case only a single exponential law governs the picking of false values. Considering the similarity measure between the true value v_d^* and the set of possible false values, it means that there is a higher probability to select false

values that are more similar to the truth than values that are more different, i.e. less similar. This happens also in real-world scenario. This setting also permits to reproduce malicious sources that use false value copy to spread false information.

Following these guidelines, the synthetic datasets have been generated to test the proposed model. The experimental settings we considered are presented in the next section.

3.3.2 Experimental setup

In order to provide robust results, considering each predicate, 60 synthetic datasets have been generated (20 for each different granularity distribution used to select the true values). Several experiments have been conducted on them. A deeper investigation has been done for the belief propagation framework than for the plausibility. The reason behind this choice is related to the preliminary results obtained during the experimental campaign. They show that the *belief* propagation outperform *plausibility* propagation. These results are reported in Appendix A.

Estimation phase settings

To obtain estimation of trustworthiness and confidence, we tested $Sums_{PO_C}$ approach. We initialized each value confidence to the value 0.5 in order to start the estimation phase. The stopping criterion used for the iterative procedure is the same as the one employed in original paper of *Sums* (Pasternack & Roth, 2010); the procedure was stopped after 20 iterations.

Truth prediction settings

Once the estimation were obtained, the post-processing procedure were applied. Table 3.6 reports the experimental settings that were tested. The name associated with each setting indicates the delta value. When $\delta = 0$ the approach is called TSbC (Truth Selection of the Best Child). Indeed, the selec-

Table 3.6: Set of experiments performed for each predicate using the belief propagation framework.

Setting Name	θ	δ	**Rank**		**Filter**
			1^{st}	2^{nd}	
TSbC$_{trust}$	0,..., 0.5	0	WA_{trust}	IC_{Seco}	V^*_{ord}
TSbC$_{IC}$	0,..., 0.5	0	IC_{Seco}	WA_{trust}	V^*_{ord}
TSaC$_{trust}$	0,..., 0.5	1	WA_{trust}	IC_{Seco}	V^*_{disj}
TSaC$_{IC}$	0,..., 0.5	1	IC_{Seco}	WA_{trust}	V^*_{disj}

tion algorithm chooses, at each step, the value with the highest confidence. In other words, it selects the best node among the children of the considered one. Otherwise, when $\delta = 1$ the approach is called TSaC (Truth Selection of all Children). Indeed, using this configuration the algorithm selects, at each step, all the children of the considered nodes. Moreover, the subscript of the setting name specifies the first ranking criterion used, i.e. TSbC$_{IC}$ means that IC is used for the ranking phase as first criterion to order the values. For all the experiments, different threshold θ were used: 0, 0.1, 0.2, 0.3, 0.4, 0.5. Given the considerations at the end of the section 3.2.2.3, that when δ is equal to 0, we test only the property of the solution set indicating that its values share ordering relationships. Indeed, the selection procedure in this case chooses, at each step, only a single values having the highest confidence. Therefore only a single most specific value and its ancestors can be returned. No alternatives to the most specific value will be selected. When δ is equal to 1, we test only the property indicating that the values in the solution set do not share a partial order. The procedure may select more than one branch. In this situation, if we force the returned true values to share an ordering relationship, we oblige the algorithm to select only one path. Thus, the main advantage of this configuration, i.e. to propose a set of alternatives, is wasted.

The algorithms have been implemented in Python 3.4. The source code and datasets associated with the proposed approach are open-sourced and published on the Web at `https://github.com/lgi2p/TDSelection`. Otherwise, experiments related to existing models were performed using the DAFNA-EA[15] implementation (Waguih & Berti-Equille, 2014). This API provides the source code for the main existing models.

3.3.3 Evaluation methodology

The evaluation of the model we proposed was carried out using both traditional and hierarchical performance measures of classification problems.

Precision and Recall

Among traditional metrics, precision and recall were mainly used to compare our approach with the existing models that do not consider the partial order. Our positive class consists of all pairs $(d \in D, v_d^* \in V_d)$ where v_d^* is the

[15] `http://www.github.com/daqcri/DAFNA-EA`

expected true value contained in the ground truth for the data item d, and the negative class is composed of all pairs $(d \in D, v_d \in V_d \setminus \{v_d^*\})$. Therefore, the precision is the proportion of pairs (d, v_d^*) returned by the approach among all the pairs it returns. The recall is the proportion of pairs (d, v_d^*) returned by the approach among all pairs contained in the ground truth.

Hierarchical Evaluation Measures

The Hierarchical Evaluation Measures (HEM) were used to analyse the behaviour obtained by different parameter settings of our approach (Kosmopoulos, Partalas, Gaussier, Paliouras, & Androutsopoulos, 2015). Indeed, hierarchical metrics distinguish the severity of different errors taking the hierarchy of classes into account. Reasonably if *Málaga* is the true value, then an approach that returns *France* should be less penalized than another that returns *Brazil*. Indeed *France* and *Málaga* are located in the same continent, i.e. Europe, while *Brazil* is located in the American continent. A detailed study related to hierarchical measures was presented in (Kosmopoulos et al., 2015). They distinguish the main dimensions that characterize hierarchical classification problems and suggest, for each possible combination, which are the best evaluation metrics to use. They recommend $F_{LCA}, P_{LCA}, R_{LCA}$ and $MGIA$ when dealing with single-label problem and DAG hierarchy. This situation corresponds to our initial problem settings: for each data item there is a single expected true value and our partial order among values is represented using a DAG. F_{LCA}, P_{LCA} and R_{LCA} are set-based measures. They use hierarchical relations to augment the sets of returned and true values and to compute precision and recall. Since adding ancestors over-penalize errors that occur to nodes with many of them, $F_{LCA}, P_{LCA}, R_{LCA}$ use the notion of the Lowest Common Ancestor to limit this undesirable effect. $MGIA$ is a pair-based metric that uses graph distance measures to compare returned and true values. Its limitation is that it does not change with depth. For further details related to the computation of these measure please refer to (Kosmopoulos et al., 2015). Now, we briefly describe the main characteristics of these hierarchical measures through an illustrative example. This enable the reader to better understand the result discussion in the next section. Considering the DAG in Figure 3.5 and *Málaga* as the true value, the HEMs related to several returned values are reported in Table 3.7. As shown, if the returned value is more general than the expected one, then P_{LCA} is not affected, while R_{LCA} decreases when increasing the distance from the

Table 3.7: Example of hierarchical evaluation measures (HEM) considering the Directed Acyclic Graph (DAG) in Figure 3.5 and *Málaga* as the true value.

Returned value	P_{LCA}	R_{LCA}	F_{LCA}	$MGIA$
Málaga	1	1	1	1
Spain	1	0.5	0.7	0.9
Country	1	0.3	0.5	0.8
Madrid	0.5	0.5	0.5	0.8
France	0.5	0.3	0.4	0.7

expected value. Otherwise, if the returned value is an error (neither the expected value nor more general one), then P_{LCA} and R_{LCA} decrease according to the position of the returned value in the partial order. $MGIA$ indicates the distance among the returned value and the expected one without considering if one value is more general or specific than the other.

3.4 Results and discussion

All of the experimental settings described in section 3.3 were tested. Here, the results are presented and discussed. Note that a robust analysis was conducted given the artificial nature of the synthetic datasets.

$TSaC_{trust}$ and $TSbC_{IC}$ are the most effective settings

Results show that our approach enables successfully addressing the problem of selecting true values. Recall that our study considers a setting where value confidence estimations according to the partial order of values monotonically increases. The most effective configuration settings of our selection procedure were $TSaC_{trust}$ and $TSbC_{IC}$ as shown in Figure 3.13 and in Figure 3.14. These settings coupled with the $Sums_{PO_C}$ model were able to outperform, in terms of recall, existing TD methods on the different datasets and predicates that were used for the experiments, see Figure 3.15, Figure 3.16 and Figure 3.17. Note that in these experiments we compared our post-processing strategies considering $k = 1$ with the other models. Indeed, the general aim of TD is to return a single answer for each data item.

Hereinafter we detail the comparison of the proposed approach with existing TD models and we study different configuration settings of the post-processing procedure analysing its behaviour considering different k, δ and θ values and different dataset types.
Both $TSaC_{trust}$ and $TSbC_{IC}$ obtained good performance, but $TSaC_{trust}$ was

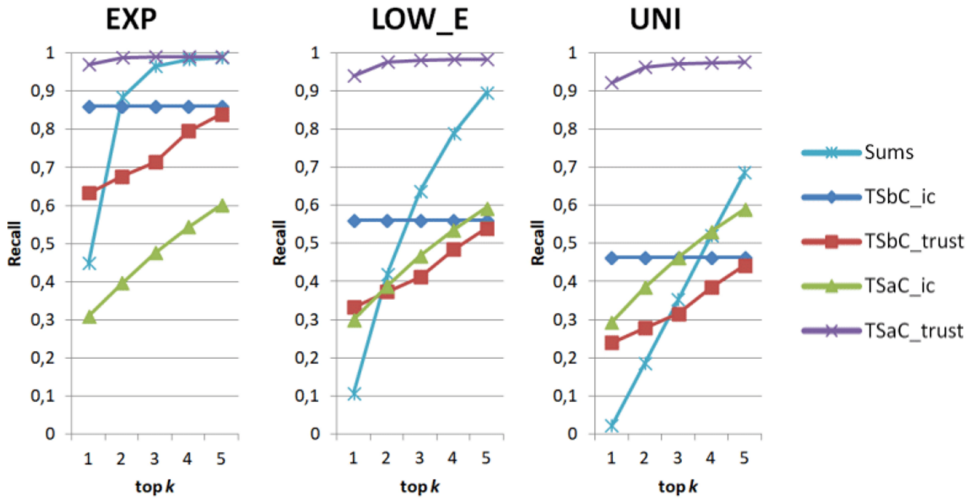


Figure 3.13: Recall obtained by applying our approach and the proposed models (with $\theta = 0$) on the synthetic datasets *genre* with respect to the dataset type and number of returned values.

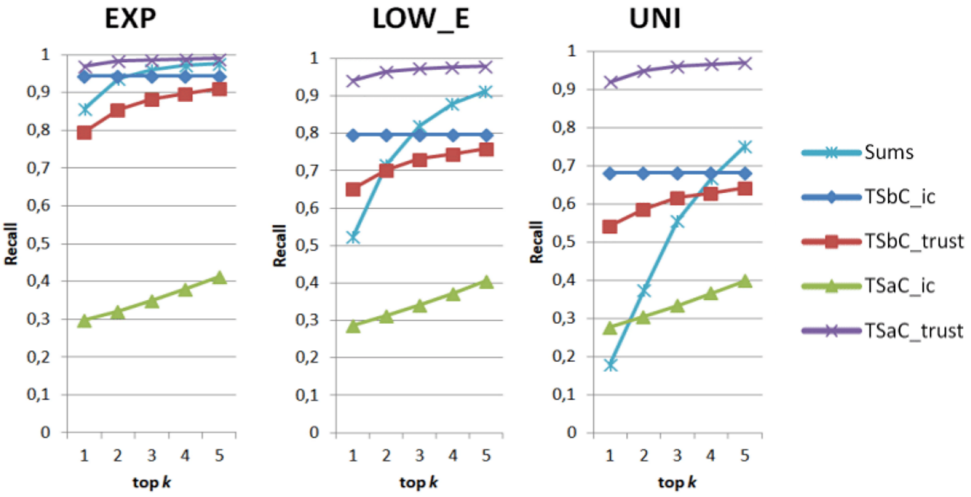


Figure 3.14: Recall obtained by applying our approach and the proposed models (with $\theta = 0$) on the synthetic datasets *Molecular Function - MF* with respect to the dataset type and number of returned values.

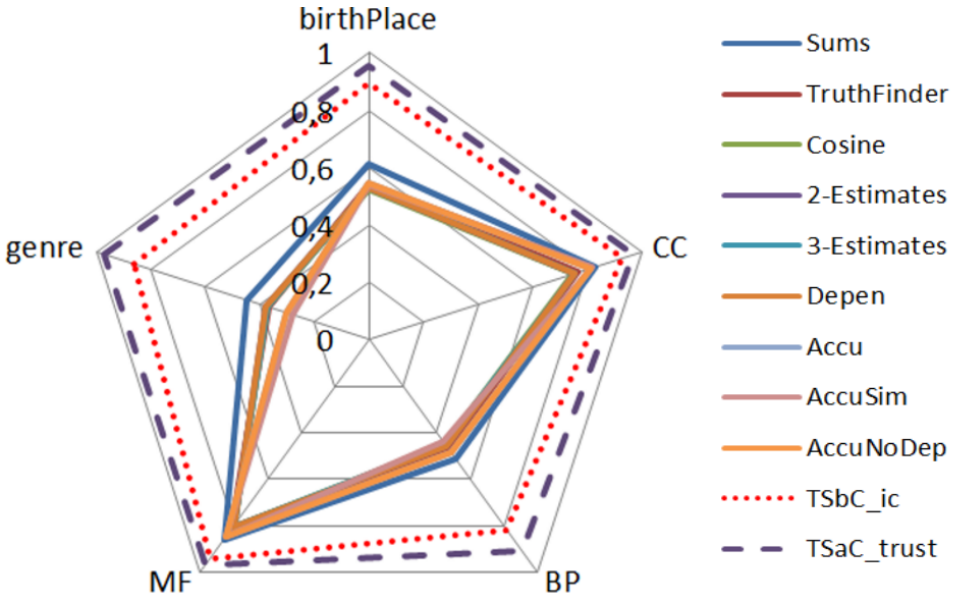


Figure 3.15: Recall obtained by applying our approaches TSbC$_{IC}$ (dotted line) and TSaC$_{trust}$ (dashed line), both with $k = 1$ and $\theta = 0$, and the models provided by DAFNA API (solid lines) on the EXP synthetic datasets.

TSaC$_{trust}$ performance analysis

the most robust approach independently of the predicate and dataset type, as shown in Figure 3.15, Figure 3.16 and Figure 3.17. It resulted to be only slightly influenced by source disagreement increase (UNI dataset case). Indeed, TSaC$_{trust}$ aimed to analyse and compare the trustworthiness of sources providing the most specific values that do not share partial order relationships. This was done selecting and returning all provided values higher than θ, i.e. $\delta = 1$, then ranking the values according to the weighted average trustworthiness of sources claiming them. Finally, filtering the first k values that did not share ordering relationships. Following this post-processing procedure, TSaC$_{trust}$ performance was not affected when the number of sources providing true general values increased (UNI dataset). Precisely, analysing the recall obtained by the different models from EXP to UNI dataset types, we observed that, when increasing the number of sources that provided general true values, TSaC$_{trust}$ had a recall drop equal to 0.073 against a recall drop around 0.528 obtained by existing TD models. Indeed, the average recall, over the different predicates, obtained by TSaC$_{trust}$ was 0.954, 0.912 and 0.881 respectively for EXP, LOW_E and UNI dataset types. The average recall achieved by existing TD models was 0.595, 0.243 and 0.067 respectively

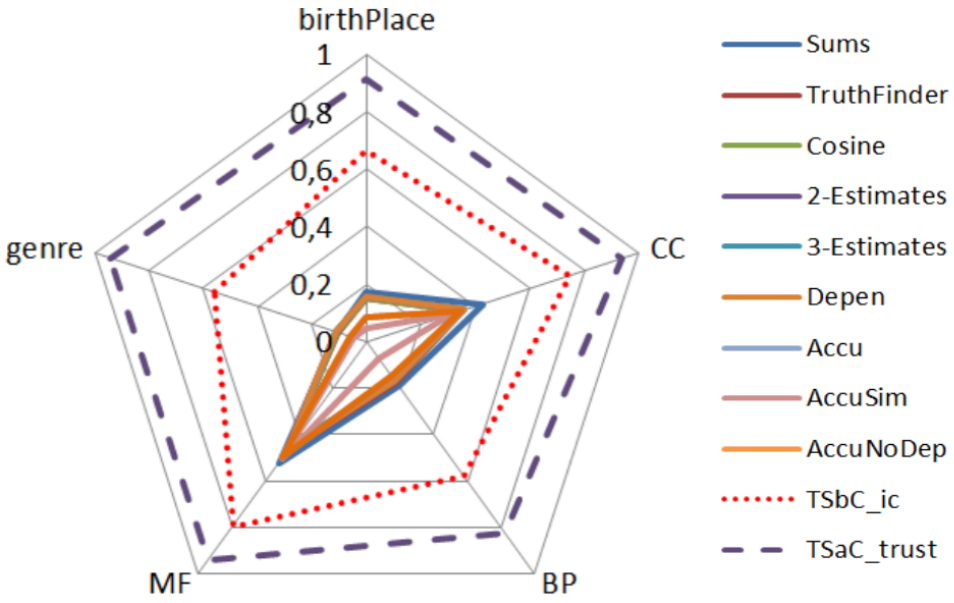


Figure 3.16: Recall obtained by applying our approaches TSbC$_{IC}$ (dotted line) and TSaC$_{trust}$ (dashed line), both with $k = 1$ and $\theta = 0$, and the models provided by DAFNA API (solid lines) on the LOW_E synthetic datasets.

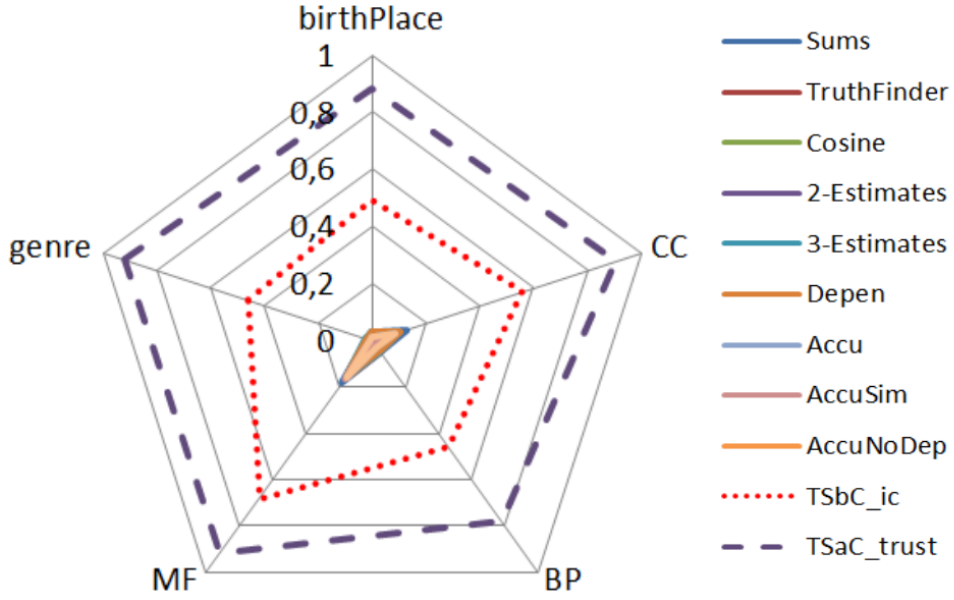


Figure 3.17: Recall obtained by applying our approaches TSbC$_{IC}$ (dotted line) and TSaC$_{trust}$ (dashed line), both with $k = 1$ and $\theta = 0$, and the models provided by DAFNA API (solid lines) on the UNI synthetic datasets.

for EXP, LOW_E and UNI dataset types.

TSbC$_{IC}$ performance analysis

On the contrary, TSbC$_{IC}$ performance was more influenced by source disagreement increase than TSaC$_{trust}$ performance. TSbC$_{IC}$ is the post-processing strategy that employed the greedy algorithm to select the true value, i.e. at each step the selection phase chooses the values with the highest confidence. Then, it ordered them according to their IC. Finally, it kept only values that shared a partial order. Therefore, it used as selection criterion, at each step, the value confidence. When sources provided more general true values, the information associated with these claims were propagated to less values. Thus the confidence estimations were less informative in the last steps of the procedure. However, also TSbC$_{IC}$ outperformed existing methods obtaining recall levels that were equal to 0.889, 0.670 and 0.531 for EXP, LOW_E and UNI dataset respectively (with a recall drop of 0.358).

Performance analysis considering different predicates

Observing Figure 3.15, Figure 3.16 and Figure 3.17 we analysed the predicates for which our approaches, TSaC$_{trust}$ and TSbC$_{IC}$, obtained slightly lower performances. Even in these cases our models stills outperformed existing ones.

Considering TSaC$_{trust}$, the worst recall performance were achieved for *birthPlace* and *BP* predicate. Analysing the features shown in Table 3.4 related to the different predicate partial orders, it is clear that this configuration setting was influenced by the average number of children in the partial order. Indeed *birthPlace* and *BP* are the two predicates with the highest children average number. Moreover, the ranking obtained ordering the predicates according to their recall corresponded to the ranking obtained ordering the predicate according to the children average number in decreasing order.

Otherwise, when considering TSbC$_{IC}$ approach the worst performance in terms of recall were obtained considering *genre* and *BP* predicate. We found out that TSbC$_{IC}$ performance depended both on the children average number and the average depth of expected solutions according to the maximum depth. Indeed, at each step of TSbC$_{IC}$ the probability of error is related to the number of alternatives among which the procedure can select a value. Moreover, it is also related to the percentage of the partial order that the selection procedure has to traverse in order to reach the expected solutions according to the maximum depth. The probability of error increased when the part of the graph to traverse augmented. For instance *genre* predicate had the lowest children average number, but it obtained performance lower

than *MF*, *CC* and *birthPlace* predicate. This is because its expected values had a depth that required to traverse a bigger part of the partial order than in the other cases.

Comparison of different truth selection procedure configurations

Before comparing the proposed approaches with the existing models, several experiments were conducted to understand the best parametrization for the post-processing procedure among the different settings reported in Table 3.6. First of all, we compared the different post-processing strategies we proposed, evaluating the recall at different levels of k. The results are reported in Figure 3.13 and in Figure 3.14. Note that we show the results for the predicates *genre* and *MF*, but a similar behaviour was obtained with all the others.
We observe that the best results were obtained by the TSaC_{trust} for any k value. It took advantage of the fact that it returned a set of alternatives as different as possible from each other and, at the same time, as specific as possible. Usually TSbC_{IC} also outperformed the baseline model (*Sums*), but for higher values of k it was worse than *Sums*. This is because we forced the result of TSbC_{IC} to share ordered relationships, while in the case of *Sums*, k values with the highest confidence were returned (no additional filter was applied on these values). Note that the recall of TSbC_{IC} did not improve when increasing the value of k. This means that a situation in which a returned value is more specific than the expected one never occurs. This is in accordance with the policy we adopted to generate the synthetic datasets. Given the expected value, we cannot say anything about its descendants. Each of them may be a true specification of the expected truth or not. Consequently, we removed all of the descendants from the set of possible true and false values. In other words, no sources provide a claim that contains one of the descendants of the expected value associated with the considered data item. Otherwise, in all the other configurations, increasing the number of values returned (k) enhanced the recall.
The TSaC_{IC} and TSbC_{trust} configurations were for the majority of cases worse than those of the baseline approaches. TSaC_{IC} consists of the selection strategy with $\delta = 1$, i.e. all provided values having confidence higher than θ are selected, and the use of IC as first ranking criterion. It obtained quite low performance because IC_{Seco} was not a good discriminator among values that did not share ordering relationships. Indeed it is based on the number of de-

scendant values and it may happen in situations in which x is the expected value and y has the same father as x. If x has descendants, while y has none, y will be preferred by the ranking based on the IC_{Seco} even if it is not a true value. Thus, the WA_{trust} ranking is more suitable in these cases.
Otherwise TSbC$_{trust}$ is a post-processing strategy with $\delta = 0$, i.e. at each step of the selection process only one value is selected, with the use of source average as ranking criterion. Obtaining low recall for this model means that WA_{trust} was not a good discriminator to rank the values sharing partial order relationships returned by the selection phase.
Moreover, Figure 3.13 and Figure 3.14 show that when disagreement among sources providing true values increased these two latter approaches (TSaC$_{IC}$ and TSbC$_{trust}$) could be useful. The recall they obtained for $k = 1$ was higher than the recall of *Sums* model. Therefore in case of high level of disagreement also a no optimal procedure can be advantageous.
As expected, in all the cases, the precision always decrease when increasing k. Moreover, comparing the different settings of the proposed approach, we observed that the ranking based on their precision performances was the same that the one obtained according to their recall. Therefore, we omit these repetitive results.
Our further analysis focused on models TSaC$_{trust}$ and TSbC$_{IC}$ they were the models among the proposed ones that achieved the best performances. We examined the impact of different threshold values, setting $k = 1$, according to the hierarchical evaluation metrics: $F_{LCA}, P_{LCA}, R_{LCA}$ and $MGIA$. The results are reported in Table 3.8. Considering TSbC$_{IC}$, we noticed that, when slightly increasing θ, $MGIA$ increased in the majority of the cases. This occurred because there are expected values (supported by few reliable sources) with a confidence lower than false ones (supported by many unreliable sources), even though the former have a higher WA_{trust} than the latter. Thus, using TSbC$_{IC}$ and $\theta = 0$, these values were selected as true values. Increasing θ allows the procedure to avoid a part of these errors. Indeed, eliminating the values with confidence score very low enables the procedure to return, with high probability, the father of the expected value. However, further increasing the threshold caused a loss of $MGIA$ because the returned values result to be very general. This does not happen with TSaC$_{trust}$ since this kind of errors are already overcome considering WA_{trust} as first ranking criterion.

Table 3.8: Hierarchical Evaluation Measures (HEM) obtained for the different predicates with respect to the model and the threshold θ considered.

		Model											
		TSbC$_{IC}$ ($\delta = 0$)						TSaC$_{TRUST}$ ($\delta = 1$)					
Predicate	**HEM**	$\boldsymbol{\theta}$						$\boldsymbol{\theta}$					
		0	0.1	0.2	0.3	0.4	0.5	0	0.1	0.2	0.3	0.4	0.5
CC	F_{LCA}	**0.836**	0.826	0.770	0.694	0.617	0.561	**0.958**	0.890	0.783	0.690	0.613	0.560
	P_{LCA}	0.824	0.874	0.943	0.986	**0.991**	0.988	0.959	0.954	0.967	0.985	**0.989**	**0.989**
	R_{LCA}	**0.862**	0.812	0.693	0.568	0.469	0.407	**0.959**	0.861	0.701	0.563	0.465	0.406
	MGIA	0.879	**0.910**	0.907	0.890	0.855	0.818	**0.963**	0.945	0.917	0.887	0.851	0.816
MF	F_{LCA}	**0.878**	0.865	0.800	0.697	0.637	0.572	**0.962**	0.914	0.807	0.695	0.636	0.572
	P_{LCA}	0.870	0.907	0.960	0.990	**0.994**	**0.994**	0.964	0.965	0.971	0.989	**0.994**	**0.994**
	R_{LCA}	**0.898**	0.850	0.729	0.568	0.492	0.414	**0.963**	0.893	0.734	0.567	0.491	0.414
	MGIA	0.909	**0.937**	0.926	0.892	0.862	0.824	**0.966**	0.958	0.928	0.890	0.860	0.824
BP	F_{LCA}	**0.745**	0.689	0.620	0.540	0.484	0.438	**0.881**	0.725	0.607	0.527	0.477	0.436
	P_{LCA}	0.732	0.859	0.957	**0.979**	0.976	0.968	0.886	0.935	0.963	**0.976**	0.974	0.967
	R_{LCA}	**0.783**	0.624	0.494	0.391	0.335	0.293	**0.882**	0.642	0.481	0.379	0.329	0.291
	MGIA	0.792	**0.853**	0.836	0.774	0.707	0.641	**0.881**	0.865	0.815	0.754	0.696	0.635
birthPlace	F_{LCA}	**0.791**	0.773	0.709	0.640	0.587	0.532	**0.946**	0.855	0.713	0.627	0.576	0.530
	P_{LCA}	0.788	0.841	0.936	0.988	**0.993**	0.990	0.948	0.941	0.953	0.988	**0.991**	0.989
	R_{LCA}	**0.800**	0.744	0.601	0.483	0.424	0.372	**0.946**	0.813	0.602	0.469	0.414	0.369
	MGIA	0.909	**0.912**	0.897	0.877	0.845	0.807	**0.968**	0.948	0.900	0.869	0.838	0.805
genre	F_{LCA}	**0.784**	0.775	0.708	0.657	0.617	0.571	**0.963**	0.930	0.729	0.657	0.617	0.571
	P_{LCA}	0.781	0.791	0.855	0.979	0.995	**0.997**	0.966	0.952	0.878	0.980	0.994	**0.997**
	R_{LCA}	**0.793**	0.774	0.641	0.505	0.454	0.409	**0.962**	0.920	0.660	0.505	0.454	0.409
	MGIA	0.903	**0.904**	0.889	0.887	0.867	0.833	**0.974**	0.967	0.897	0.887	0.867	0.833

Moreover, we observed that, in the majority of cases, when increasing θ the R_{LCA} always decreased, while the P_{LCA} always increased. Precisely, the highest R_{LCA} for both TSaC$_{trust}$ and TSbC$_{IC}$ was obtained with $\theta = 0$. The highest P_{LCA} was obtained for both approaches with different θ values depending on the predicate as shown in Table 3.8.

Summarising, the most effective configuration settings were TSaC$_{trust}$ and TSbC$_{IC}$. They were both able to obtain better performance than existing TD models. We noted that increasing the number of values returned for each data item allow increasing the performance. Nevertheless this can be applied only in the case where a group of experts can select the true values among the ones proposed by the proposed approach for each data item. Otherwise, we have to force the parametrization $k = 1$. Regarding the threshold θ, a high θ value is recommended when the application scenario does not permit to assume many risks. In this case it is important to have a high precision. In other words, obtaining a general true value rather than a potentially false one is preferred. Therefore, the different parameter settings of the proposed post-processing procedure allow dealing with different application scenarios taking their requirements into account.

The results obtained by the experiments presented in this chapter show that using *a priori* knowledge efficiently improve TD performances. These important results drive us to further exploit knowledge expressed into ontologies. While in this chapter we exploited a priori knowledge in the form of partial order of values to identify dependencies among values, in the next chapter we describe an approach that identifies dependencies that may exist among data items.

Chapter 4

Truth Discovery based on recurrent patterns derived from an ontology

Contents

In this chapter we describe a novel approach to enrich traditional TD models by incorporating information associated with recurrent patterns extracted from ontology. These patterns have been identified in the form of rules using a state-of-the-art rule mining system. They have been used to increase confidences of claims they support. Precisely, each rule contributes to confidence estimation according to its quality. In order to take different quality aspects into account, a function that aggregates existing rule quality metrics has also been defined. The proposed approach has been evaluated through an extensive evaluation; interestingly, the results show that TD framework can benefit from information expressed by rules. Datasets and source code proposed in this study are open-sourced and freely accessible online[1].

[1] `https://github.com/lgi2p/TDwithRULES`

Contributions related to this chapter:

Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, and Isabelle Mougenot. (2018). Combining Truth Discovery and RDF Knowledge Bases to their mutual advantage. In *Proceedings of the 17th International Semantic Web Conference (ISWC '18)*, 16 pages.

4.1 Incorporating recurrent patterns into Truth Discovery framework

This section starts by presenting an example that shows how recurrent patterns can be helpful to facilitate the TD task. Before detailing the proposed model, all key elements of rules and their quality metrics are introduced. Indeed, rules have been chosen, among a set of alternatives methods, as a model to represent recurring behaviours detected in an existing ontology. Then, the proposed model is finally described and evaluated.

4.1.1 Intuition

In the example introduced in chapter 3, *a priori* knowledge related to relationships existing among collected answers, see Table 3.1, was used to reinforce confidence in certain replies when they are supported by the other provided ones. For instance, the fact that *"Málaga"* and *"Madrid"* are both Spanish cities increased the confidence of the claim *"Pablo Picasso was born in Spain"*. The validity of this answer could be further strengthen by other kind of *a priori* knowledge. For instance, given all *facts* contained in an ontology, whose extract is represented in Figure 4.1, the following consideration can be derived. Usually, if someone speaks the official language of a country, then there is a high probability that this person was born in that country. Given this observation and knowing that *"Pablo Picasso speaks Spanish"* should contribute to increase the confidence of the claim *"Pablo Picasso was born in Spain"*. Indeed, the official language of *Spain* is *Spanish*. Therefore, general patterns that frequently occur can be derived from available knowledge when it is not limited to value dependencies, but it is also related to other relationships such as other people and their information. Most often people having similar characteristics should have similar answers for the considered aspect of interest, i.e. the birth location in our case. This additional

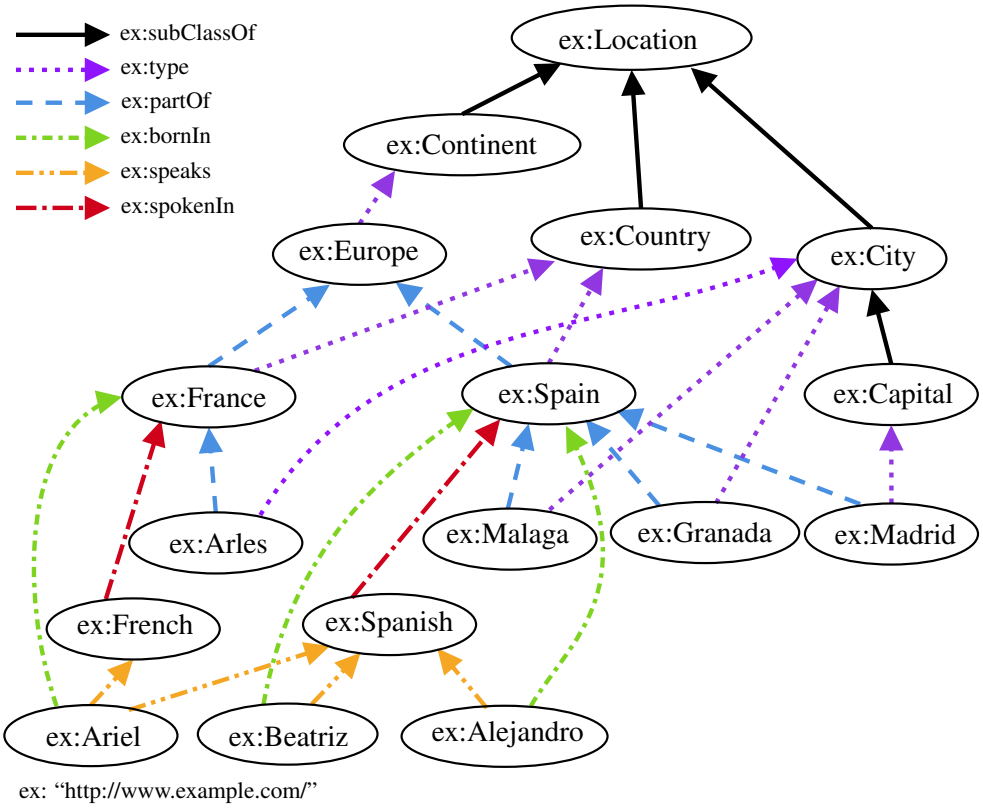


Figure 4.1: Extract of some *facts* from an ontology.

information can be used to reinforce confidence in answers that match the replies associated with similar people. In this way, true answer identification can be improved by taking advantage of dependencies that may exist among different subject entities.

4.1.2 Recurrent pattern detection

In this thesis, we propose to take advantage of recurring patterns to better discriminate true claims from a set of conflicting ones. These patterns can be identified using several methods that have been proposed in Artificial Intelligence (AI). Indeed, AI has extensively studied the process of inferring recurrent patterns that explain observations of particular instances, i.e. induction. Numerous models have been proposed to address the problem of induction in several research fields, such as association rule mining, link mining, decision tree techniques. In the rest of this section, we briefly in-

troduce these domains to then focus on rule mining techniques. Rules have been adopted to model recurring patterns in this thesis.

Link Mining (LM) consists of a set of techniques that gain insight from graph structure analysis. Among the different tasks addressed by this discipline, such as object clustering, sub-graph discovery (Getoor & Diehl, 2005), we are mainly interested in link prediction. Approaches in this field exploit recurrent pattern detection in order to identify missing links, i.e. missing relationships between entities. They return a list of potentially new links ordered according to their likelihood of existence (Lao, Mitchell, & Cohen, 2011). These patterns also incorporate the dependencies among objects since these models consider graph structures (Jensen, 1999). Several probabilistic models, embedding techniques as well as factorization ones have been proposed to efficiently discover new or missing relationships in data, exploiting both latent and graph features (Nickel, Murphy, Tresp, & Gabrilovich, 2016). Link prediction models can be used to enhance ontologies. Ontologies can be represented as graphs of interrelated heterogeneous objects where nodes are subject and object entities, and edges are predicates linking them. For instance, a model addressing knowledge base completion uses this approach (Nickel et al., 2016). In order to apply these models in RDF KB context, the information regarding the type of each URI must be available. Otherwise, frequency analysis cannot be performed since each URI is never duplicated in a RDF graph (Chi, Yang, & Muntz, 2004; Kuramochi & Karypis, 2001). Given our problem setting, the set of links returned by these models should represent the set of claims that an existing ontology endorses. Moreover, the confidence that could be associated with each link by these models could represent the degree of support provided by the ontology to a claim. However, in this thesis, we prefer to derive this degree using rule mining techniques. Indeed, their results, i.e. rules, are easily interpretable (Agrawal, Imieliński, & Swami, 1993). Even though also decision trees are methods that generate rules, we decided not to adopt them for various reasons. For instance, they generate rules that are usually less compact than the rules produced by rule mining algorithms. Indeed, they learn rules all at once, and not sequentially as done by rule mining approaches. Internal nodes of decision trees do not produce any rules (although tentative rules can be derived) leading to more complex rules. Moreover, a theoretical difference exists between decision

Identifying recurrent patterns using rule mining

trees and association rules. While decision trees are classification techniques that aim to predict the class attribute given a set of labelled examples, association rule mining tries to discover associations that exist among items with no particular focus on a target one (Ordonez & Zhao, 2011). Therefore, their rules are different in their meaning. While decision trees identify regularities that enable distinguishing instances of different classes, rule mining techniques identify regularities in data without any expectation.

4.1.2.1 Rule mining

Rule mining (RM) aims to extract interesting correlations, frequent patterns or causal structures that may exist among sets of items. More precisely, given structured data, such as a KB, RM intends to learn logical rules in an unsupervised manner. Generally, a rule specifies which attribute value conditions need to occur in order to observe other attribute value conditions. Hereinafter, we formally introduce rules and related concepts. Formally, given a KB K, a rule r is an implication such that *Rules*

$$B_1 \wedge B_2 \wedge \cdots \wedge B_n \rightarrow H_1 \wedge H_2 \wedge \cdots \wedge H_m \tag{4.1}$$

Usually, it is abbreviated $\widehat{B} \rightarrow \widehat{H}$ where the set $\widehat{B}$ is called *body*, or antecedent, and $\widehat{H}$ is called *head*, or consequent. Note that the sets $\widehat{B}, \widehat{H} \subset K$ and $\widehat{B} \cap \widehat{H} = \emptyset$. Both sets are composed of atoms. Each atom, denoted $p(s, o)$, represents a relation between s and o that can be variables or constants.[2] *Atoms*

Given our problem setting, we are particularly interested in Horn rules. Considering Datalog-style, a Horn rule is characterized by a single atom in the head (Nebot & Berlanga, 2012), i.e. $\widehat{B} \rightarrow p(s, o)$. When the head atom corresponds to a claim, the confidence of this claim can be reinforced. To identify the set of claims that are inferred by a rule, its body atoms need to be instantiated[3]. Indeed a rule can infer the head atom when its body holds. A body $\widehat{B}$ holds under an instantiation σ in K, if each atom in $\widehat{B}$ holds (Galárraga & Suchanek, 2014). In turn, an atom a holds under σ in a KB K if $\sigma(a) \in K$. Thus the instantiated head atom that is obtained by instantiating the body can be seen as a claim that is supported by a certain rule. Indeed, each instantiated atom $p(s, o)$ can also be represented as an RDF triple *Horn rules*

[2]Note that, in this section, the symbol s represents the *subject* of a relation. It does not represent a *source* unless otherwise stated.

[3]The instantiation of an atom is a substitution of its variables with constants.

$\langle subject, predicate, object \rangle$ where all atom variables have been replaced by URIs. Each RDF triple can be seen in turn as a claim v_d where d represents the pair $(subject, predicate)$ and v is the *object* occurring in the RDF triple.
RM has been extensively investigated in information system community (Fürnkranz & Kliegr, 2015), but only few studies have focused on RDF data (Barati, Bai, & Liu, 2017). This kind of data has introduced new challenges related to data dimension, lack of information and presence of errors. Therefore traditional techniques have to be adapted to these new settings (Quboa & Saraee, 2013). Initially, rule mining on RDF data have been used to perform exploratory analysis identifying patterns in the given dataset (Böhm et al., 2010), i.e. descriptive task. Lately, it has been used but not limited to infer new knowledge generalizing rules mined from the given dataset, i.e. predictive task (L. Galárraga et al., 2015). Other tasks in Semantic Web domain that can benefit from rule mining methods are, for instance, ontology learning, ontology alignment, canonicalization of open KBs and error correction, all related to data integration task (Galárraga, 2014; Galárraga, 2015). Initial studies related to ontology learning and reasoning in Semantic Web were based on inductive logic programming (ILP), which merges statistics and reasoning techniques (Muggleton, 1995). Using logic programming to represent hypotheses, examples and background knowledge, these models learn from the set of positive and negative examples any regularities and describe them in the form of logical hypothesis. Examples of these methods are FOIL (Quinlan, 1990), WARMER (Goethals & Bussche, 2002), ALEPH (Muggleton, 1995), Sherlock (Schoenmackers, Etzioni, Weld, & Davis, 2010) and Quick-FOIL (Zeng, Patel, & Page, 2014). These models adopt different strategies to generate rules. For instance, FOIL performs a hill-climbing search limiting the search space through constraints, while WARMER uses language bias models to restrict it. Otherwise, ALEPH is based on inverse entailment in order to refine rules. Then, Sherlock uses probabilistic graphical models to infer new facts, given a target relation, through first-order Horn rules. Quick-FOIL is an improvement of FOIL and applies a new scoring function for the search phase. It also employs a new pruning strategy, but it still requires to explicitly provide negative examples as all the other approaches. This is one of the main problem of ILP approaches (L. A. Galárraga, Teflioudi, Hose, & Suchanek, 2013). To overcome it, models able to infer rules from only positive example have been proposed (Muggleton, 1995, 1996; Schoen-

Rule mining in Semantic Web

mackers et al., 2010). Even if these new models resolve the aforementioned limitation, they still assume that the ontological knowledge bases they use do not contain factual errors. Using RDF KBs, we cannot ensure the absence of errors. Moreover, Open World Assumption (OWA) holds for RDF KBs. Indeed, OWA is considered when dealing with incomplete information. It permits distinguishing between unknown and false information. A triple that does not appear in KB is not systematically false as considered under the Closed World Assumption (CWA). Also the scalability issue of ILP methods does not permit an easy application of them in the context of Semantic Web where large RDF KBs exist. Therefore, as already anticipated, these approaches have to be adapted to the new settings (Quboa & Saraee, 2013).

Open World Assumption vs. Closed World Assumption

An example of such an adaptation is AMIE (L. Galárraga et al., 2015). To face OWA, its authors have introduced the Partial Completeness Assumption (PCA): if a KB contains some object values for a given pair (*subject*, *predicate*), it is assumed that all object values associated with it are known. This assumption enables generating counter-examples that are necessary for rule mining models, but do not appear in RDF KBs which often contain only positive facts. Using this new assumption, a new confidence measure, called $conf_{PCA}$, has been introduced to better discriminate different rules, see section 4.1.2.2. The rules inferred by AMIE are connected and closed. A rule is connected if each of its atom share at least a variable or an entity with at least another atom. A rule is closed if all its variables appear at least twice. These rules are generated following the procedure below. Initially AMIE considers all possible head atoms as rules of size 1. Then, if a rule meets certain criteria, such as the $conf_{PCA}$ higher than a threshold, it is returned as output by the system. Successively, if it does not exceed the maximum number of atoms, then it is also refined. A pruning phase follows in order to eliminate not relevant rules.

Rule mining with RDF data

A similar procedure is followed by another model presented in (D'Amato, Staab, Tettamanzi, Minh, & Gandon, 2016). It differs from AMIE in the fact that it exploits terminological axioms and deductive reasoning to prune rules that are inconsistent with the ontology. Moreover different operator are used for extending rules.

Otherwise, RDF2Rules uses a completely different strategy to generate rules introducing the concept of frequent predicate cycles (Z. Wang & Li, 2015).

They are interesting frequent patterns in a KB. Precisely, each cycle is a set of predicate paths[4] starting with the same entity variable and ending with the same entity variable. This model generates rules from the set of frequent predicate cycles that are identified in a KB. Note that from a cycle with k predicates, k rules can be generated (Z. Wang & Li, 2015). At the end, rules are evaluated and pruned when necessary.
Alternative assumptions and metrics have been proposed to extract rules under OWA also for other purposes such as ontology learning (Lehmann & Völker, 2014; Nickel et al., 2016; Tanon, Stepanova, Razniewski, Mirza, & Weikum, 2017). In this study, we use AMIE+ because it is a well evaluated state-of-the-art method and its source code is freely available online.

4.1.2.2 Rule quality metrics

Any rule, independently from the system used to extract it, can be evaluated by several quality metrics; among them the most well-recognized measures are *support* and *confidence* (Agrawal et al., 1993; Maimon & Rokach, 2005; Ventura & Luna, 2016). *Support* represents the number of instantiations for which a rule is verified, i.e. the frequency of a rule in a KB, while confidence is the percentage of instantiations of a rule among the instantiations of its body in the KB. In other words, the confidence count the number of correct predictions obtained by applying a rule. Hereinafter we present how these metrics are computed using the formal definition adopted by the authors of AMIE (L. Galárraga et al., 2015) for sake of coherence and clarity. In this thesis, we do not propose a comparison among different quality metrics because it is out of the scope of this study. Here the primary aim is to evaluate the potential of integrating knowledge extracted from an RDF KB into a TD process. However, since we are aware that robust metrics could have an impact in TD results, we plan to investigate such a comparison in future studies.
Considering a Horn rule $r : \widehat{B} \rightarrow H$ where $H = p(s, o)$, its *support* is defined as the number of different *(s,o)* pairs of the atom head that appear in the KB when instantiating the rule. It is evaluated as:

Rule support

$$supp(\widehat{B} \rightarrow p(s, o)) := \#(s, o) : \exists z_1, \ldots, z_n : \widehat{B} \wedge p(s, o) \tag{4.2}$$

[4] A path is a sequence of entity variables and predicates.

where $z_1, \ldots, z_n$ are the variables contained in the atoms of the rule body $\widehat{B}$ apart from s and o, and $\#(s,o)$ is the number of different pairs s and o. Otherwise, its *confidence* is computed using the following formula:

Rule confidence

$$conf(\widehat{B} \rightarrow p(s,o)) := \frac{supp(\widehat{B} \rightarrow p(s,o))}{\#(s,o) : \exists z_1, \ldots, z_n : \widehat{B}} \tag{4.3}$$

Considering this formula, all predictions, provided by a rules, that do not appear in the KB are considered wrong. This is in accordance with CWA. To deal with OWA, Galarraga et al. defined a new *confidence*, called $conf_{PCA}$ (L. Galárraga et al., 2015). It considers the PCA assumption making possible to distinguish between false and unknown facts considering PCA. In this setting, if a predicate related to a particular subject, never appears in the KB, then it can neither be considered as true nor false. This new *confidence* based on PCA is evaluated as follows:

Rule confidence based on partial completeness assumption

$$conf_{PCA}(\widehat{B} \rightarrow p(s,o)) := \frac{supp(\widehat{B} \rightarrow p(s,o))}{\#(s,o) : \exists z_1, \ldots, z_n, y : \widehat{B} \wedge p(s,y)} \tag{4.4}$$

Considering PCA, $conf_{PCA}$ normalizes the *support* by the set of true and false facts that does not include the unknown ones. Indeed, this time a prediction is considered false only if there is at least one occurrence of any object for the predicate and subject appearing in the prediction.

Example. Given a KB K, reported in Table 4.1, and the following rule:

- $r_1 : speaks(x,z) \wedge officialLang(y,z) \rightarrow bornIn(x,y)$

Its *support* is 1. Indeed only one distinct *(subject,object)* pair results from the instantiations of the rule that appears in the KB.

Table 4.1: Illustrative set of triples.

predicate	*subject*	*object*	*predicate*	*subject*	*object*
officialLang	(Spain,	Spanish)	bornIn	(Dalì,	Spain)
officialLang	(French,	France)	bornIn	(Gauguin,	France)
speaks	(Dalì,	Spanish)	residentIn	(Picasso,	Paris)
speaks	(Dalì,	French)	residentIn	(Giotto,	Florence)
speaks	(Monet,	French)	cityOf	(Florence,	France)
speaks	(Picasso,	Spanish)	cityOf	(Paris,	France)
speaks	(Giotto,	Italian)			

Its *confidence* is 1/4 because there is one positive example for r_1 (the prediction *bornIn(Dalì, Spain)* appears in K) and three negatives examples (the predictions *bornIn(Dalì, France)*, *bornIn(Monet, France)* and *bornIn(Picasso, Spain)*). Its *PCA confidence* is 1/2 because there is the same positive example of before (the prediction *bornIn(Dalì, Spain)* appears in K) but only one negative example (the predictions *bornIn(Dalì, France)*. The predictions *bornIn(Monet, France)* and *bornIn(Picasso, Spain)* are not anymore considered as negative examples since in K there is no information about neither where *Monet* and *Picasso* were born nor where they were not born.

In the next section, presenting the proposed approach, we describe how these quality measures are combined into a single one. Indeed, considering several quality aspects is important because each rule will contribute according to its quality in the computation of the overall evidence that supports a certain claim.

4.1.3 TDR approach: Truth Discovery using Rules

The second contribution of this thesis aims at studying how extracted rules can be integrated into truth discovery models to improve their performance, see Figure 4.2. This figure also shows that the partial order of values is taken into account as well. The overall idea is that, when recurrent patterns

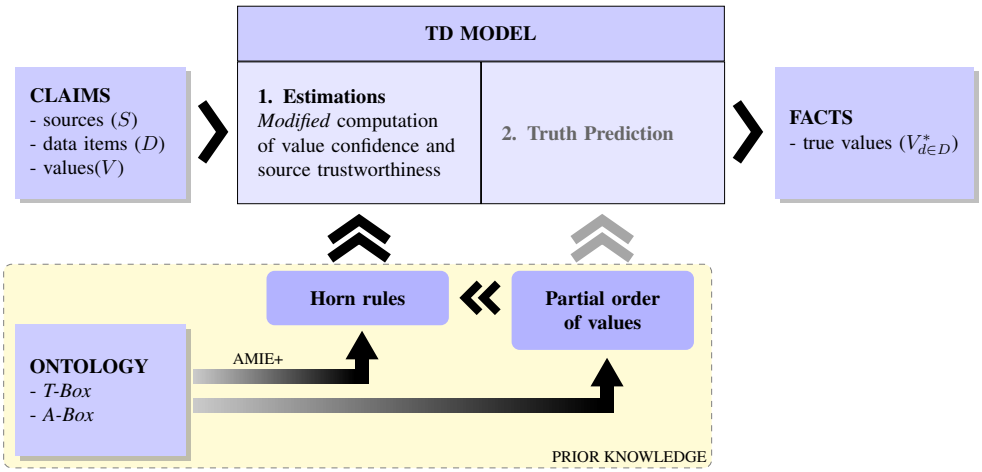


Figure 4.2: Diagram of the overall Truth Discovery (TD) procedure incorporating the recurrent patterns during the estimation phase to improve TD performance.

occur and they concern a particular data item, all confidence of those claims concerning these patterns may be increased. To this end, we defined the concepts of *eligible* and *approving* rules to identify the most useful rules that have to be considered when evaluating the confidence of a claim. Then we describe how information associated with these rules is quantified to further introduce the new confidence estimation formulas used by the proposed TD framework.

4.1.3.1 Eligible and approving rules

Considering the entire set of extracted rules (denoted R) may not be useful to improve value confidence. For instance, some rules could not be related to a given data item. Therefore, for each claim v_d, where $d = (subject, predicate)$, only *eligible* rules are used as potential evidence to improve its confidence estimation. They are defined in the following way.

Definition 4.1 (Eligible Rule) *Given a KB K, a set of rules $R = \{r : \widehat{B} \rightarrow H\}$ extracted from K, where $H = p(s, o)$, and a claim v_d, where d=(subject, predicate), a rule $r \in R$ is an **eligible rule** when its body holds (all of its body atoms appear in K when instantiating all rule variables) with respect to the data item subject. Moreover, its head predicate has to correspond to the one in the claim under examination, i.e. $(\sigma(\widehat{B}) \in K) \wedge (H = p(subject, o))$.* *Eligible rules*

In our context, the eligibility of a rule depends on the subject and the predicate that compose a data item d. Thus, all claims related to the same data item $d = (subject, predicate)$ have the same set of eligible rules denoted $R_d = \{r \in R \mid (\sigma(\widehat{B} \in K) \wedge (H = predicate(subject, o))\}$.
Once eligible rules with respect to a claim v_d are collected, the proposed approach checks how many of these rules endorse (approve) v_d, i.e. how many rules support v_d.

Definition 4.2 (Approving Rule) *Given a KB K, a set of eligible rules $R_d = \{r : \widehat{B} \rightarrow H\}$, where $H = predicate(subject, o)$, and a claim v_d, where d=(subject, predicate), a rule $r \in R_d$ is an **approving rule** when the value predicted by r corresponds to the claimed value v_d, i.e. $(\sigma(\widehat{B}) \in K) \wedge (H = predicate(subject, v_d))$.* *Approving rules*

The set of approving rules for v_d is represented by $R_d^v \subseteq R_d$ where d indicates that the rules are eligible for a certain data item d and v indicates that the

rules predict/support value v. Formally we have $R_d^v = \{r \in R_d \mid (\sigma(\widehat{B}) \in K) \wedge (H = predicate(subject, v_d))\}$.

Example. Given a KB K, reported in Table 4.1, and the following rules:

- $r_1 : speaks(x, z) \wedge officialLang(y, z) \rightarrow bornIn(x, y)$
- $r_2 : residentIn(x, w) \wedge cityOf(w, y) \rightarrow bornIn(x, y)$

When considering the following claims on the birth location of some painters $\langle Picasso, bornIn, Spain\rangle$, $\langle Picasso,\ bornIn,\ Málaga\rangle$ and $\langle Giotto, bornIn,\ Italy\rangle$, the set of eligible rules for data item $d_A = (Picasso, bornIn)$ is $R_{d_A} = \{r_1, r_2\}$. Indeed, the predicate in the head corresponds to claimed one and when replacing all occurrences of variable x with *Picasso* in r_1's and r_2's body, they are both verified. Otherwise, when $d_B = (Monet, bornIn)$ the set of eligible rules is $R_{d_B} = \{r_2\}$ because, even if head and claim predicate are the same considering both rules, substituting x variable with *Monet* the body of r_1 is not verified.
The set of approving rules for the first, second and third claims are respectively $R_{d_A}^{Spain} = \{r_1\}$, $R_{d_A}^{Málaga} = \emptyset$ and $R_{d_B}^{Italy} = \{r_2\}$.

Before explaining how additional information related to eligible and approving rules is quantified and then incorporated into TD framework, we describe a function used to integrate two quality aspects we are interested in, for each rule. It permits better weighing each rule contribution during the evaluation of a claim.

4.1.3.2 Combining rule's quality measures

Support and conf_{PCA} represent different aspects of a rule, see section 4.1.2.2. We propose an aggregate function to combine them into a single quality metric since, in our context, taking both aspects into account is important. Indeed, it may happen that two rules r_1 and r_2 have the same *confidence*, but different *supports*. For instance, if $conf_{PCA}(r_1) = conf_{PCA}(r_2) = 0.8$, $supp(r_1) = 5$ and $supp(r_2) = 500$, then r_2 deserves a higher level of *credibility* than r_1 since r_2 has been more observed than r_1.
To address this problem, a function $score : R \rightarrow [0, 1]$ is defined. It is based on Empirical Bayes (EB) methods (Robbins, 1956). EB adjusts estimations when resulting from few examples that may happen by chance. The estima-

Using Empirical Bayes to better aggregate rule quality metrics

tions are modified according to available examples and prior expectations. When many examples are available, estimation adjustments are small. On the contrary, when there are only few examples, the adjustments are more important. They are corrected according to the average value that is expected by *a priori* knowledge. Given a family of the prior distribution that represents available data, EB is able to directly estimate its hyper parameters from the data. Then, it updates the prior representing our belief with new evidence. In other word, the estimation that can be computed from the new examples is modulated with respect to prior expectation. The new estimation corresponds to the expected value of a random variable following the updated distribution. In our case, a more robust conf_{PCA}, i.e. the proportion of positive examples among all the considered ones, has to be estimated. The prior expectation on our data can be modelled using a *Beta* distribution that is characterized by parameters α and β. Once the model estimates them, it uses this distribution as prior to modulate each individual estimate. This estimation will results to be equal to the expected value of the updated distribution $Beta(\alpha + X, \beta + (N - X))$ where X is the number of new positive examples and N is the total number of new ones. The new expected value is $(\alpha + X)/(\alpha + \beta + N)$. This value is returned by the aggregation function. Summarizing, given the hyper parameters α_S and β_S, the value returned by *score* for a rule $r : \widehat{B} \rightarrow p(x, y)$ is computed as follows:

Rule score function

$$score(r) = \frac{\alpha_S + supp(r)}{\alpha_S + \beta_S + \sum_j supp(\widehat{B} \rightarrow p(x, j))} \tag{4.5}$$

where $supp(r)$ is the *support* of r and $\sum_j supp(\widehat{B} \rightarrow p(x, j))$ is the number of triples containing data item (x, p). The returned score appears to be similar to $conf_{PCA}$, but it takes the cardinality of the examples into account.
Once this score is estimated for each rule, the proposed approach sums up all this new information that is integrated in the value confidence estimation formula.

4.1.3.3 Assessing rule's viewpoint on a claim confidence

All the evidence provided by rules for a claim v_d is summarized in a *boosting factor* that can be seen as the confidence that is assigned by these rules to v_d. Precisely, it represents the proportion of eligible rules that confirm a given claim v_d. In other words, the percentage of approving rules out of

Claim boosting factor

the entire set of eligible rules is evaluated, i.e. $|R_d^v|/|R_d|$. It is returned by a function $boost : D \times V \rightarrow [0,1]$. As anticipated, the proposed model weights each rule differently according to its quality *score*. The higher is the *score* of a rule, the strongest should be its impact on computing the *boosting factor*. Intuitively, given a claim v_d where $d = (subject, predicate)$ and a set of rules R extracted from a KB K, the proposed model evaluates the *boosting factor* in the following way:

Claim boosting factor based on recurrent patterns

$$boost(d, v_d) \approx \frac{\sum\limits_{r \in R_d^v} score(r)}{\sum\limits_{r \in R_d} score(r)} \tag{4.6}$$

where R_d^v is the set of approving rules, R_d is the set of eligible rules and $score : R \rightarrow [0,1]$ represents the quality score associated with a rule (as detailed in section 4.1.3.2). Since the *boosting factor* consists in evaluating a proportion, EB is used also in this case to obtain a better estimation that is less prone to be result of chance. As explained in section 4.1.3.2, when applying EB, initially the parameters α_b and β_b of a *Beta* distribution are estimated from available data using methods of moments. Then this prior is updated based on evidence associated with a specific v_d. Thus, *boosting factor* corresponds to the expected value of the updated prior that is equal to:

$$boost(d, v_d) = \frac{\alpha_b + \sum\limits_{r \in R_d^v} score(r)}{\alpha_b + \beta_b + \sum\limits_{r \in R_d} score(r)} \tag{4.7}$$

where α_b and β_b are the hyper parameters of the Beta distribution that represents the available examples. Since AMIE does not consider any *a priori* knowledge such as the partial order of values to extract rules, we decided to use it to further exploit rule information and to compute a more refined boosting factor. Precisely, considering a partial order $\mathcal{V} = (V, \preceq)$, when a rule r explicitly predicts a value v, we assume that it implicitly supports all more general values v' such that $v \preceq v'$. In other words, the evidence provided as support by a rule to a value is propagated to all its generalization:

Claim boosting factor based on recurrent patterns and partial order

$$boost_{PO}(d, v_d) = \frac{\alpha_b + \sum\limits_{r \in R_d^{v+}} score(r)}{\alpha_b + \beta_b + \sum\limits_{r \in R_d} score(r)} \tag{4.8}$$

Therefore, in this case the boosting factor $boost_{PO}(d, v_d)$ (the subscript underlines the fact that the Partial Order among values is considered) indicates the

percentage of approving rules (for both the value under examination and all of its more specific values) out of all eligible rules. Therefore, the set R_d^v in Eq. 4.7 is replaced by the set $R_d^{v+} = \{r \in R_d \mid \widehat{B} \wedge H = p(x, v'), v' \preceq v\}$.

4.1.3.4 Applying TDR to existing model: *Sums*$_{RULES}$

All elements required to integrate information given by recurrent patterns into TD models have been defined. Therefore, we can proceed describing the adaptation of an existing model. Since the *boosting factor* is related to a claim, only the confidence formula has been updated. As proof of concept, in this study, we modified *Sums* (Pasternack & Roth, 2010) whose estimation formulas are:

$$t^i(s) = \frac{1}{\max_{s' \in S} \sum_{v'_d \in V_{s'}} c^{i-1}(v'_d)} \sum_{v_d \in V_s} c^{i-1}(v_d) \tag{4.9}$$

$$c^i(v_d) = \frac{1}{\max_{v'_d \in V} \sum_{s' \in S_{v'_d}} t^i(s')} \sum_{s \in S_{v_d}} t^i(s) \tag{4.10}$$

We modified Eq. 4.10 proposing the new $Sums_{RULES}$. It integrates the additional information given by rules into the confidence formulas as follows:

Rule-based adaptation

$$c^i_{RULES}(v_d) = \frac{1}{norm_{v_d}} \Big[(1-\gamma)c^i(v_d) + \gamma\ boost(d, v_d)\Big] \tag{4.11}$$

where $\gamma \in [0, 1]$ is a weight that calibrates the influence that is assigned to information provided by sources and to information contained in an external KB during the value confidence estimation. For sake of coherence, when using $boost_{PO}$ we considered the partial order also for the computation of the confidence formula applying the belief propagation as proposed in chapter 3. We refer to this model as $Sums_{PO}$ and its confidence formula as $c^i_{PO}(v_d)$. The confidence of v_d is therefore computed considering all trustworthinesses of the sources that associate with a data item d the value v under examination, or a more specific one than v. Indeed as stressed before when claiming a value, we also consider that a source implicitly supports all its generalizations. Similarly, the model that integrates both $boost_{PO}$ and rules is indicated as $Sums_{RULES\&PO}$ and is defined as follows:

Rule and partial order-based adaptation

$$c^i_{RULES\&PO}(v_d) = \frac{1}{norm_{v_d}} \Big[(1-\gamma)c^i_{PO}(v_d) + \gamma\ boost_{PO}(d, v_d)\Big] \tag{4.12}$$

Note that, while *Sums* and $Sums_{RULES}$ return a true value for each data item selecting the value with the highest confidence, $Sums_{RULES\&PO}$ and $Sums_{PO}$ required the truth prediction procedure proposed in chapter 3 to select the most informative true values.

4.2 Experiments

Estimation phase settings

In order to obtain an extended overview of the proposed approach, several experiments were carried out on synthetic datasets. Their aim is to determine the improvement obtained by $Sums_{RULES}$ (Eq. 4.11) and $Sums_{RULES\&PO}$ (Eq. 4.12, $\gamma > 0$) with respect to their respective baseline, i.e. *Sums* (Pasternack & Roth, 2010) (Eq. 4.10) and $Sums_{PO}$ (Eq. 4.12, $\gamma = 0$) considering different scenarios. In both cases, the baseline corresponds to set $\gamma = 0$ in the new confidence formula of the proposed models. Note that to analyse the effect of incorporating rules only $TSbC_{IC}$ has been considered as post-processing procedure. Indeed, since rules are defined but not limited to specific values, their cannot improve the $TSaC_{TRUST}$ procedure whose rationale is to return different and specific values as much as possible. A comparison with existing models is also presented. Also in this case, we initialized all value confidences at 0.5 and we used as stopping criteria the maximum number of iterations (fixed at 20).

Using AMIE to extract rules

The rules used in the following experiments, as well as their *support* and $conf_{PCA}$ were extracted from DBpedia by AMIE. In order to filter out the most useless rules, we selected 62 rules for the predicate *birthPlace* and 47 rules for the predicate *genre*. Examples of these rules are reported in Table 4.2.

The synthetic datasets were used to evaluate the proposed model on different scenarios depending on the granularity of the provided true values.

Table 4.2: Examples of rules extracted by AMIE from DBpedia for `db-owl:birthplace` predicate.

@prefix db: <http://dbpedia.org/resource/>.
@prefix db-owl: <http://dbpedia.org/ontology/>.

?a `db-owl:deathPlace` ?b	→	?a `db-owl:birthPlace` ?b
?a `db-owl:country` ?b	→	?a `db-owl:birthPlace` ?b
?a `db-owl:deathPlace` ?b ∧ ?b `db-owl:language` `db:English_language`	→	?a `db-owl:birthPlace` ?b

These scenarios simulate to deal with experts or non-expert users. For instance, When dealing with experts, they usually provide specific true values (EXP datasets). Otherwise, when dealing with non-expert users, they also provide general values that remain true (UNI datasets).

Expected, general and erroneous values

To evaluate the performance in this setting, we measured the expected values rate/recall (returned values that correspond to expected ones), true but more general values rate (returned values that are more general than the expected ones) and erroneous values rate (values that are neither expected nor general) obtained by different model settings. Note that, during the analysis of datasets, we noted that only 25-30% of their data items had at least one eligible rule. Therefore, the performance could be further improved when eligible rules are associated with all data items.

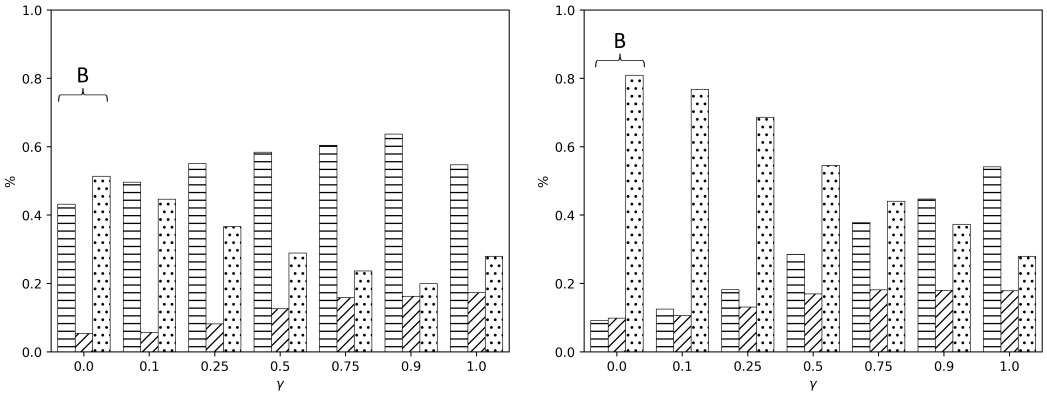
4.3 Results and discussion

The results, summarized respectively in Figure 4.3 and 4.4, and Figure 4.5 and 4.6, show that the proposed approaches enable TD models to benefit from the use of *a priori* knowledge given by an external and reliable ontology. Indeed, usually the number of correct *facts* that are identified by the proposed models increases compared to the baseline. Intuitively, since the number of correct *facts* increases, a new KB that is populated with the true claims identified by the improved TD will have a higher quality.

Performance analysis for birthPlace dataset

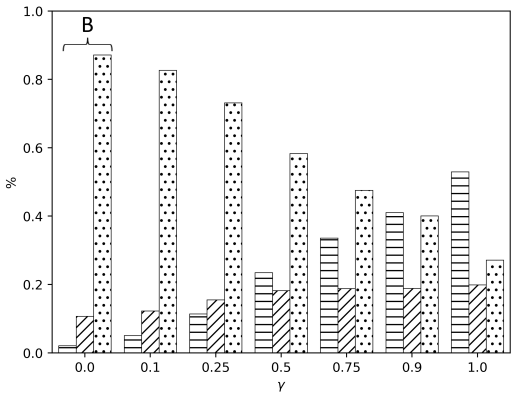
Considering *birthPlace* predicate, the improvement obtained by considering both $Sums_{RULES}$ and $Sums_{RULES\&PO}$ was always greater for UNI datasets than for EXP or LOW_E datasets. Since identifying true values in UNI setting was harder than in the other cases (the highest disagreement among sources on the true values is modeled by UNI), the baseline obtained the lowest recall. Using additional information tackles the high level of disagreement among sources and thus enables full exploitation of the higher scope for improvement that was available in the case of UNI setting.

Considering $Sums_{RULES}$ the best recall was obtained with different γ values. For UNI datasets, the optimal configuration was when $\gamma = 1$. In such a case, it was considered that no information provided by sources was useful and that only rules should be used to solve conflicts among claims (when rules are available). This was true only for the extreme situation represented by



(a) *genre* - EXP

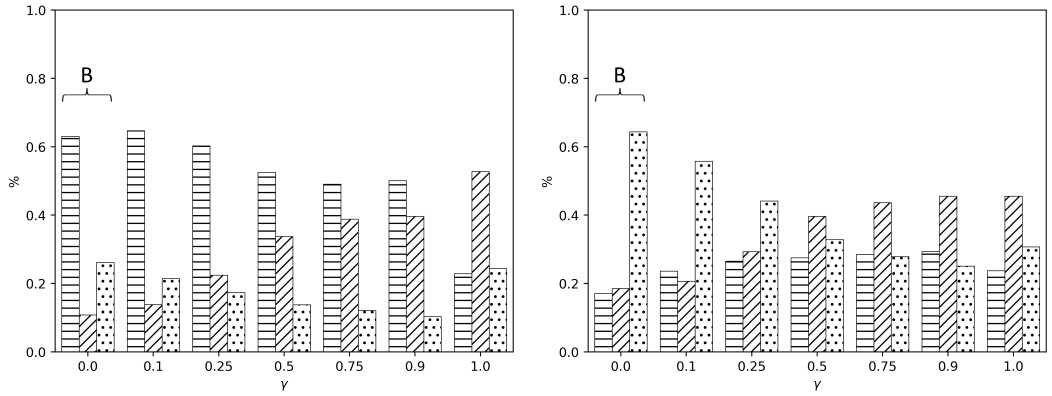
(b) *genre* - LOW



(c) *genre* - UNI

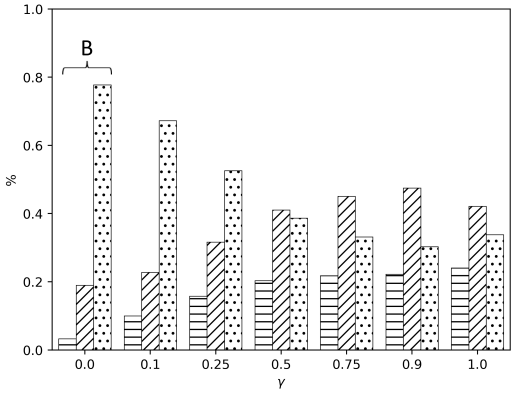
Figure 4.3: Expected (horizontal line bars), true but more general (diagonal line bars) and erroneous values (dotted bars) obtained by $Sums_{RULES}$ on different *genre* datasets with several γ's values. Letter B indicates the bars showing the performance obtained by the baseline.

UNI datasets where disagreement among sources was so high that the recall obtained by baseline model remained under 10%. Indeed, in the other cases it was advantageous to take both source trustworthiness and rule information into account. For EXP datasets, the optimal γ value was 0.1, while for LOW_E it was 0.9. Low γ values were preferred in EXP settings because in this case sources that provide true values are quite sure about the expected one, and it is thus less useful to consider the rules' viewpoints. Moreover, this setting was the only situation where considering external knowledge was damaging in terms of recall. Nevertheless, the error rate obtained by



(a) *birthPlace* - EXP

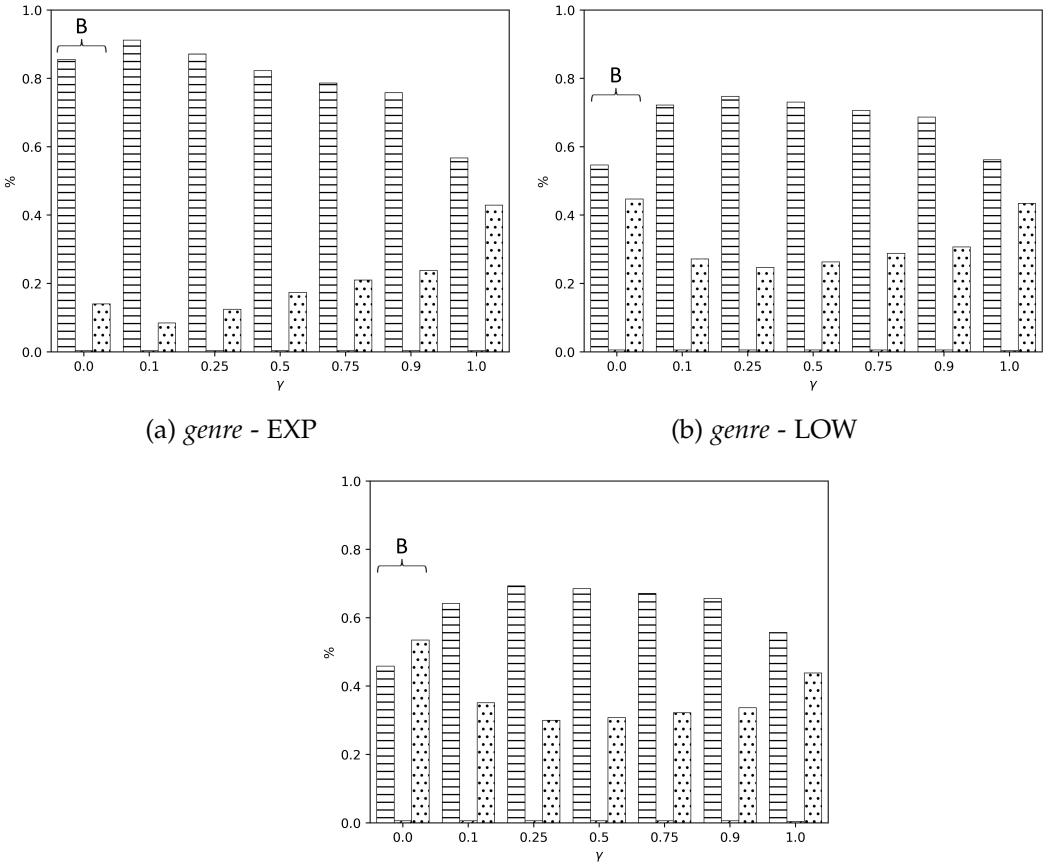
(b) *birthPlace* - LOW



(c) *birthPlace* - UNI

Figure 4.4: Expected (horizontal line bars), true but more general (diagonal line bars) and erroneous values (dotted bars) obtained by $Sums_{RULES}$ on different *birthPlace* datasets with several γ's values. Letter B indicates the bars showing the performance obtained by the baseline.

$Sums_{RULES}$ when $0 < \gamma < 1$ was always lower than the error rate achieved when $\gamma = 0$. This is explained by the fact that the average IC of values inferred by rules extracted for the *birthPlace* predicate is around 0.53. This means that they often infer values that are general. Many returned values, selected with the highest value confidence criteria, were therefore more general than the expected one but not erroneous. In other words, the rules associated with the *birthPlace* predicate were more effective for discovering the country of birth than the expected location. However using rules were useful.



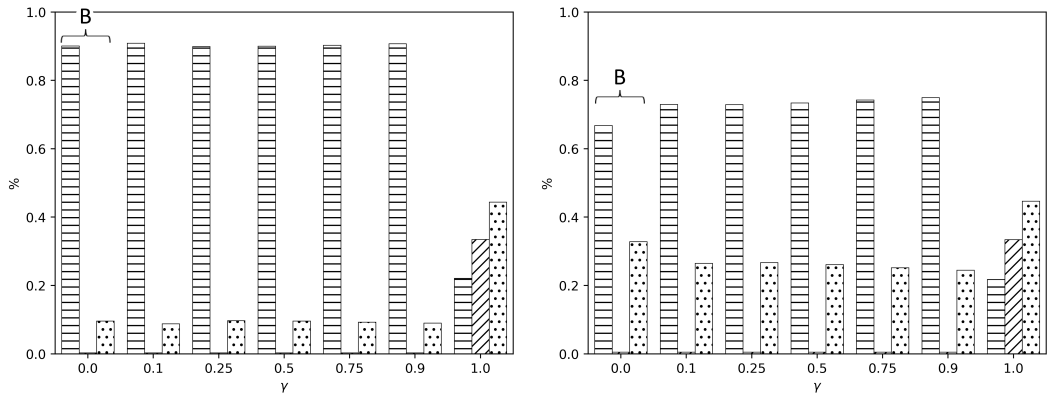
(a) *genre* - EXP

(b) *genre* - LOW

(c) *genre* - UNI

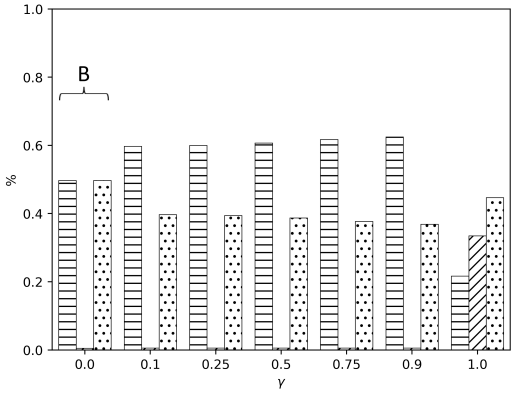
Figure 4.5: Expected (horizontal line bars), true but more general (diagonal line bars) and erroneous values (dotted bars) obtained by $Sums_{RULES\&PO}$ on different *genre* datasets with several γ's values. Letter B indicates the bars showing the performance obtained by the baseline.

The limitation related to rules that support general values was in part overcome by considering $Sums_{RULES\&PO}$, which also takes the partial order of values into account. In this case rules can improve the selection of the correct value during the first steps of the selection procedure. They were able to handle and dominate the false general values supported by many sources. The selection process was then continued with the fine-grained values evaluated based only on source trustworthiness information since no evidence provided by rules was available. For $Sums_{RULES\&PO}$ tested on EXP datasets, low γ values were preferred, while on LOW_E and UNI datasets high γ



(a) *birthPlace* - EXP

(b) *birthPlace* - LOW



(c) *birthPlace* - UNI

Figure 4.6: Expected (horizontal line bars), true but more general (diagonal line bars) and erroneous values (dotted bars) obtained by $Sums_{RULES\&PO}$ on different *birthPlace* datasets with several γ's values. Letter B indicates the bars showing the performance obtained by the baseline.

values led to the best performance.

Performance analysis for genre predicate

Considering *genre* predicate we observe a similar behaviour on both models. Interestingly we notice that the enhancement in terms of performance were higher for *genre* predicate than for *birthPlace*. This is due, once again, to the different IC associated with the values that can be inferred by rules extracted for the two predicates. As already said, the rules associated with *birthPlace* often infer general values. Instead, the IC of values inferred by rules associated with *genre* is higher, i.e. 0.95. In other words, the rules as-

sociated with *genre* were more effective to discover the expected values than the rules associated with *birthPlace*. The different IC also explains why for $Sums_{RULES\&PO}$ low γ values (around 0.25) were preferred to obtain the best performance. Rules predicting specific values cannot dominate the low confidence associated with specific values (not affected by belief propagation). Indeed these rules have the opportunity to impact also the last steps of the selection procedure where the value confidence is low.

$Sums_{RULES\&PO}$ is the most effective model

The best overall recall was obtained by $Sums_{RULES\&PO}$ for both predicates, see Figure 4.3 and Figure 4.4, and Figure 4.5 and Figure 4.6. This model considers the two kinds of *a priori* knowledge: extracted rules and partial order of values.

The evaluations just presented were conducted on synthetic datasets. In the next chapter, we propose to test the proposed approaches on real-world datasets. We describe a real-world scenario where TD models can be applied. We then report the results obtained by the proposed approaches and existing models in this real-world setting.

Chapter 5

Truth Discovery on real-world datasets

Contents

In this chapter, we describe the behaviour of the proposed models when they are applied on real-world data. We propose to use Truth Discovery (TD) models for serving Knowledge Base Population. More precisely, the TD models proposed in the previous chapters are exploited to identify new facts about actual values of missing entity properties in DBpedia. Considering this application scenario, we discuss advantages and disadvantages of the proposed models.

5.1 Application context and its specificities

Recently, several initiatives have been developed to automatically populate large Knowledge Bases (KBs), such as Yago, DBpedia, etc., with Web data. The increasing interest in the creation and enrichment of these KBs is due to the fact that they can be used for several tasks. For instance, they can be exploited for question-answering or used as background knowledge for En-

tity Linking, and other Word Sense Disambiguation-related tasks. Although the current size of these KBs is quite important, they are still far from being complete (Buche, Dervin, Haemmerle, & Thomopoulos, 2005). For instance, a study found that the properties *place of birth* and *nationality* are missing, respectively, for 71% and 75% of entities of type person in Freebase[1] (X. Dong et al., 2014). This issue can limit the great potential of these KBs. Increasing their completeness is therefore important. The research areas of Knowledge Base Completion (KBC) and Knowledge Base Population (KBP) address this problem. Both of these fields aim at increasing KB completeness, but using different methodologies.

Knowledge Base Completion

KBC models automatically infer missing facts based only on existing facts in a KB. No external text collections are used. These approaches can be divided into: methods based on Markov random fields that make inferences using first-order logic (Jiang, Lowd, & Dou, 2012) or probabilistic soft logic (Pujara, Miao, Getoor, & Cohen, 2013); embedding strategies that identify new facts to add into a KB analysing latent factors (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013; Nickel, Tresp, & Kriegel, 2011); path ranking methods consisting of algorithms that search new facts (i.e. new links between entities) using random walks (X. Dong et al., 2014; Lao et al., 2011).

Knowledge Base Population

KBP instead aims to discover facts about entities from a large collection of texts in order to augment KBs. It consists mainly of two tasks: (i) entity discovery and linking, and (ii) Slot-Filling. The task (i) links entity mentions in text to entities in KBs. The task (ii) adds information about one or more properties of an entity into the KBs in the form of triples. The Slot-Filling task thus help to increase the completeness of KBs when missing values are known. Indeed, this task requires as input to specify both entities and properties whose values have to be identified from a large collection of texts. The majority of successful approaches in this field are based on distant supervision. This technique is used to create training data. More precisely, training examples are automatically generated by labeling relation mentions that appear in a new corpus according to relations and instances that are already listed in an external knowledge base. Then distance supervision-based models train a multi-class classifier for each slot type based on the generated

[1] Freebase has been integrated into Knowledge Google Graph since 2015.

training data and hand-crafted features. Alternative methods have been proposed. For instance, methods based on open-domain information extraction and manually defined rules (Soderland, Gilmer, Bart, Etzioni, & Weld, 2013). Other models are based on unsupervised techniques. For instance, a study proposed to use the ensemble weak minority clustering to discover patterns that identify relevant relations between entities of a certain type (Ageno, Comas, Naderi, Rodríguez, & Turmo, 2013). However, this kind of models obtain low performance compared to the others.

TD-based Slot-Filling

Here, we propose to use TD models to serve Slot-Filling purpose exploiting Web data as text collection. This idea resulted from the necessity of having structured data as input of TD models. Considering a real-world scenario where it is difficult to have structured content, it is mandatory to define a pre-processing phase that extract the structured claims from Web. We can identify two main advantages for using TD in this context. First, the proposed approach is based on Web data. Thus, a text collection is not required anymore for slot filling purpose. Second, no training phase is necessary. Indeed, TD models are unsupervised techniques. Obviously, we are aware that a more rigorous evaluation is necessary to compare the proposed framework with traditional Slot-Filling systems. Here, we focus on the comparison among the different TD models when applied on the same real-world scenario. Indeed, the primary aim is to evaluate the impact of considering *a priori* knowledge when using TD models in a real-world scenario, which has its own characteristics. A previous study proposed to use TD in KBP setting (H. Li et al., 2014). Their study differs from our setting because they focus on the Slot-Filling validation task. They compare the values proposed by the different Slot-Filling systems using the TD rationale in order to identify the true values. In other words, they proposed a multi-dimensional TD model that incorporates the information associated with multiple Slot-Filling systems and their reliability into TD formalization. We consider instead TD models to compare the information extracted by different text collections, i.e. different websites. It is also important to note that several challenges, such as TAC Slot filling task and ISWC Challenge 2017, have been launched by several conferences to attract an increasing number of researchers in this domain. They provide datasets that can be used for KBP purpose. However, these datasets focus on properties such as addresses, phone numbers,

websites; they do not correspond to attribute values that are ordered (even if ad-hoc orderings could be defined for each type of attribute values, e.g. all specific German addresses could be defined as German addresses, which specializes EU addresses... and so on). In our study, we wanted to evaluate refined TD approaches that can take advantage of value orderings on value granularities and rules; this would have been difficult with existing datasets. Building specific datasets allows avoiding the issue of dealing with ad-hoc value orderings. This is why, we build new real-world datasets. We consider the collected datasets as a further contribution of this thesis. However, we already planned to extend the evaluation of proposed approaches analysing their behaviour on the datasets that have been proposed by the challenges defining ad-hoc orders.

Hereinafter, we present the use-case study we considered to compare the behaviour of the different TD models when applied on collected Web data. Indeed, as shown by a previous study, it does not exist a TD model that always outperforms the other (Waguih & Berti-Equille, 2014). It depends on the considered dataset. Retrieved Web data may have intrinsic characteristics that could result in different algorithm behaviours.

5.2 Truth Discovery-based Knowledge Base Population

The goal here is to test the proposed TD models in a realistic scenario. In this perspective, we decided to exploit Web data in order to identify missing true values in existing RDF KBs for a list of pairs *(subject, property)*. Each pair correspond to a data item. Indeed, each pair represents an aspect (i.e. *property*) of a real-world entity (i.e. *subject*), whose value is missing in the considered KB. Considering the context of the Web, the different website domains are regarded as information sources. Since multiple sources may provide conflicting values for the same data items, TD helps us to discriminate the true values from the false ones.

TD models required structured claims as input in order to be applied. However, the majority of Web content is unstructured, i.e. text. Therefore it is necessary to perform a pre-processing step that extracts structured claims

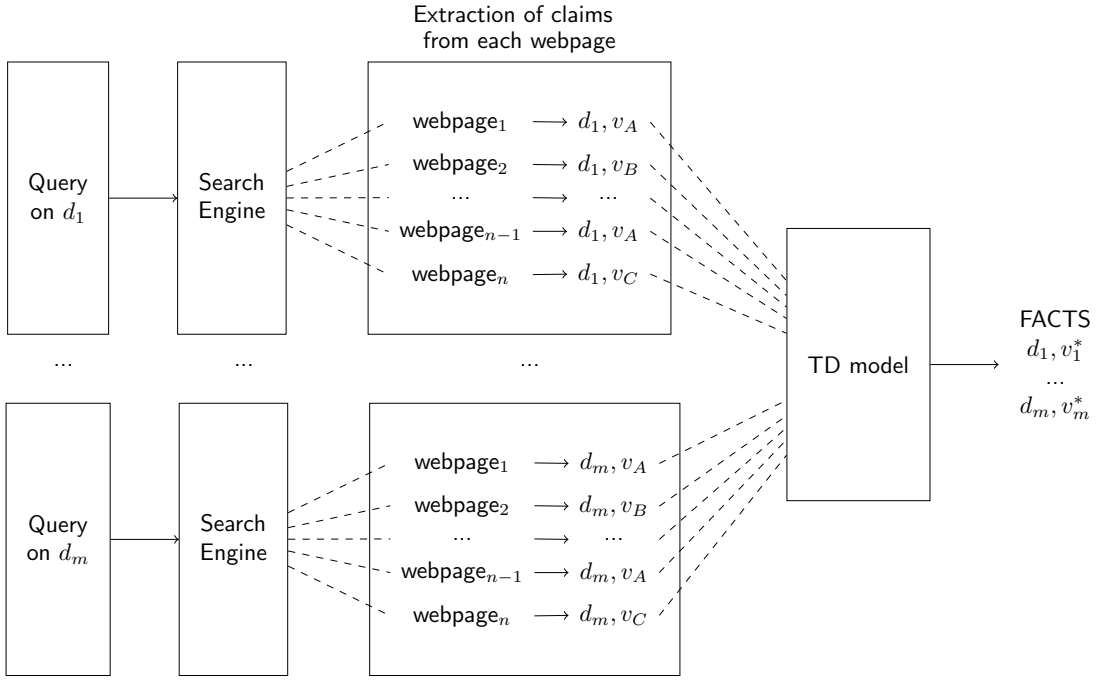


Figure 5.1: Procedure that uses TD models to exploit Web data to identify new facts on certain data items.

Obtaining structured claims from Web data

from Web data. The entire procedure we suggested to increase KB completeness is reported in Figure 5.1. As it shows, the execution of a TD model is preceded by some steps that aim to collect a set of claims for a set of data items we are interested in. Several queries are submitted to a search engine in order to retrieve relevant web pages for the considered data items. Then, claims are extracted from the content of the web pages and used as input for the TD model. Hereinafter, we will detail all the steps related to the preprocessing phase. First of all, for each data item whose value is missing in the KB, a query is generated. Each query is composed of two keywords. The first keyword corresponds to the label that represents the full name of the considered DBpedia instance, i.e. the subject of the data item. The second keywords is the natural language expression that indicates the considered property, i.e. the predicate of the data item. For instance, if we are interested in discovering the birth location of Pablo Picasso, i.e. data item d = *(Picasso, bornIn)*, then the corresponding query will be *"Pablo Picasso" AND "was born"*. In the query, each keyword is included in the quotation marks.

Query formulation

Moreover these keywords are linked by the AND operator. This kind of query forces a search engine to return only web pages in which there is at least one occurrence of both exact keywords appearing in the query. Once the query has been created, it is submitted to a search engine.

Websites as sources of information

A set of web pages is thus returned. The domain names of these web pages are considered as information sources. A web page is referred by a complete web address that contains a domain name, as well as other components needed to locate the specific page within a website. We consider as information sources the domain name in order to increase the probability that information sources cover more than one data item. For instance, if a claim is extracted from `https://en.wikipedia.org/wiki/Pablo_Picasso`, then the source that we consider for this claim is `https://en.wikipedia.org`. While a web page often focuses on a specific subject, the corresponding website may also concern other topics. For example, the website Wikipedia covers many subjects, but its several web pages focus on different themes.

Claim extraction procedures

At this point, given a web page that has been returned by a query, a claim can be extracted. Two naive information extraction procedures were defined[2] : procedure A and procedure B. They differ from each other in the set of conditions that must hold to extract a triple of terms as a claim. First of all, given a web page, the two procedures have to detect all the occurrences of the three components of a claim: its subject, its predicate (that together define the data item) and its value. This part is common for both procedures. In order to maintain as simple as possible the procedures, we replaced the entity matching phase that is normally performed by KBP models with a simple string matching phase[3]. Given a web page and a string representing the full name of the considered data item subject, we check if this string appears in the web page. The same control is done for checking if there is at least one occurrence of the string representing the data item property in a natural language form. Then, for the identification of potential values that can be provided by a web page, we use DBpedia Spotlight[4] (Mendes, Jakob,

[2]We are aware that advanced extraction techniques have been proposed in the literature. However, we decided to define naive extraction procedures that do not require any supervision and any training phase to avoid relying on additional external resources, such as labelled data.

[3]Problems related to synonyms, polysemy are therefore not addressed in this study.

[4]http://www.dbpedia-spotlight.org/ .

García-Silva, & Bizer, 2011). The primary purpose of DBpedia Spotlight is to interlink unstructured content and DBpedia. Indeed, it is a tool that annotates mentions of all DBpedia resources in natural language texts. More precisely, it performs the following steps to reach its goal:

- spotting, it consist of identifying substrings that may be entity mentions;
- candidate selection, it proposes a set of candidate meanings for the substrings identified in the previous phase;
- disambiguation, it selects the most likely candidate meanings for each substring;
- filtering, it adjusts the annotations according to user needs (i.e. needs of annotating only instances of a certain type).

The only limitation of this annotation system is that it recognizes only resources defined in DBpedia. In our experiment this is not a problem because we will try to enhance the completeness of DBpedia itself. Once all occurrences of the three elements of a claim have been identified, a claim can be extracted from a web page if and only if some constraints hold. The two procedures consider two different sets of constraints. Procedure A selects a value as the claimed one if and only if the value co-occurs in the same sentence of the term representing the considered property of a data item. Moreover, if multiple values are identified, then the value that has the lowest distance from the considered property is selected. If there are multiple pairs (property, value) with these characteristics, the value of the first pair occurring in the text is selected as the value provided by this web page for a data item. Procedure B instead requires one more constraints to be verified. A value can be selected only if it appears after the first occurrence of the subject full name in the text of the web page.

Procedure A

Procedure B

In the next section, we specify the practical aspects regarding the real-world dataset collection. We report some examples to show the different claims that the different protocols extract. Then, we describe characteristics of the collected real-world datasets. Finally, we report the results we obtained over these datasets.

5.3 Experiments

As use-case, we decided to identify the missing people birth location exploiting the proposed TD approaches. Considering DBpedia, this task consists of identifying the value associated with the `dbo:birthPlace`[5] predicate of a subset of DBpedia instances of type `dbo:Person`. This missing value is therefore a location. Since results of experiments on synthetic data showed that the most interesting results are obtained by considering both extracted rules and partial order of values, we compared the results obtained in this case with the results obtained by existing TD methods[6] (Waguih & Berti-Equille, 2014). Note that for the estimation phase we used $\text{Sums}_{RULES\&PO}$ formula. Indeed, setting $\gamma = 0$ corresponds to the proposed approach that does not consider rules. For the truth prediction phase we used TSbC_{IC} algorithm since, as already explained in chapter 4 on page 110, rules have no impact on TSaC_{TRUST}. The evaluation protocol for these experiments consisted in counting the number of values returned by a model that are equal to the expected ones. In this setting, the number of general values returned were not analysed since the main aim of TD models, as well as Slot-Filling, is to return the expected values, not their generalizations.

Estimation phase settings

Truth prediction settings

5.3.1 Dataset collection

In order to apply the proposed TD approaches to serve Slot-Filling purpose, it is necessary to collect a set of claims provided by multiple sources. To gather these claims, it is required to have a list of data items whose birth location is unknown. As list of data items, we randomly selected a subset of 564 DBpedia instances of type `dbo:Person` having at least one eligible rule (considering the rules generated by AMIE to infer values for the predicate *birthPlace*). Moreover, in order to evaluate the proposed models, the selected data items must have their values for the DBpedia `dbo:birthPlace` property. The value reported in DBpedia acts as our gold standard. In this way, we can evaluate the models comparing the value returned by TD models with the one in the ground truth. Note that, for the gold standard, we used a subset of 480 data items. Indeed, the actual true values of 84 data items were not

Ground truth

[5]The prefix `dbo` stands for `http://dbpedia.org/ontology/`

[6]For these models we used the implementation available at `http://www.github.com/daqcri/DAFNA-EA`.

in the partial order we considered. However, as usually done for existing TD benchmark, TD models base their computations on all data items for better estimating value confidences and source trustworthiness, while for the evaluation phase they consider only a subset of these data items. As *a priori* knowledge, we considered the same the partial order and rules used for synthetic datasets .

A priori knowledge

Given the list of data items previously selected, the real-world datasets could be collected. For each data item we retrieved a set of web-pages (up to 50) containing at least one occurrence of the subject full name and the expression "born", i.e. usually used to introduce the birth location of a person[7]. Given a web-page and its data item we used procedure A and procedure B for extracting a claim, respectively, for dataA and dataB. The set of figures from Figure 5.2 to Figure 5.4 report examples of web pages that were returned as results for the query *"Pablo Picasso" AND "born"*. In these figures, the terms that may be selected as part of a claim are underlined by a dashed line, surrounded by a solid rectangle or surrounded by a dotted rectangle.

Claim collection

Claim extraction examples

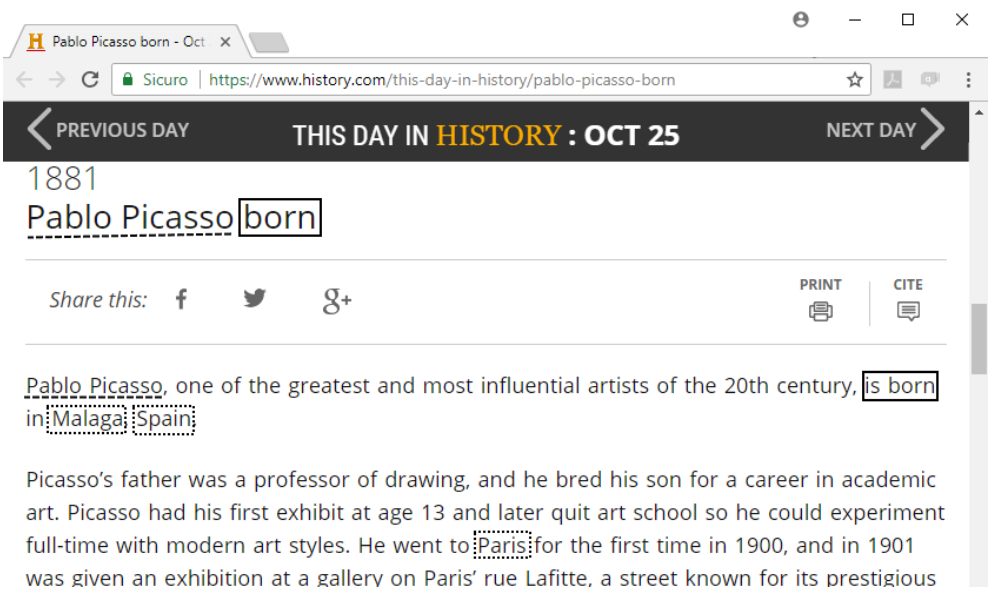


Figure 5.2: Example of a web page containing information on Pablo Picasso birth location. Dotted, solid and dashed lines indicate the occurrence of a subject, predicate and value respectively.

[7]We did not adopt query expansion techniques in order to keep this phase as simple as possible.

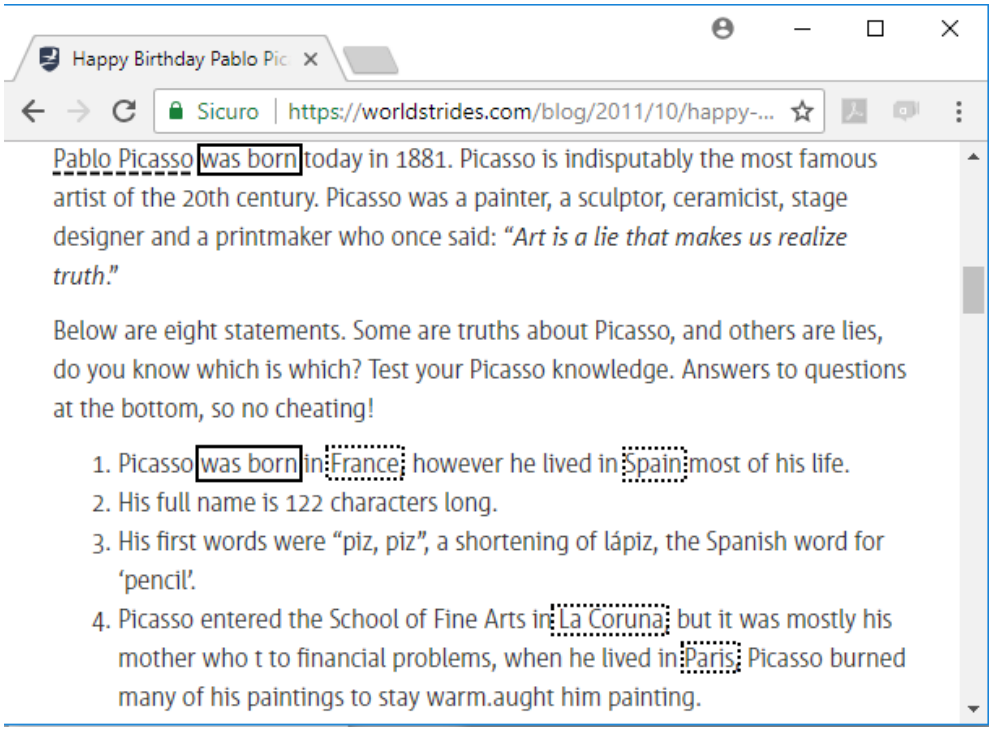


Figure 5.3: Example of a web page containing information on Pablo Picasso birth location. Dotted, solid and dashed lines indicate the occurrence of a subject, predicate and value respectively.

Based on the style that it is used to highlight a term, this term could be, respectively, the subject, the property and the value of the extracted claim. Once all these kinds of terms have been identified in a web page, one of the procedures can be applied. Considering Figure 5.2, DBpedia spotlight identified more than one location: *Málaga*, *Spain* and *Paris*. *Paris* is immediately discarded because it does not co-occur in the same sentence of the term *born*. Then, between *Málaga* and *Spain*, both procedures will return *(Picasso, bornIn, Málaga)* as claim. Indeed, *Málaga* is the nearest one to the term *born*. The rationale is that, when referring to the same person, the city is usually followed by the country of birth in natural language sentences. Moreover, the higher the distance from the term *born*, the higher the probability that the location is referring to another predicate will be. This claim is also returned by procedure B because the painter full name appears once before the occurrence of the pair *(property, value)*. A similar situation is reported in Figure 5.3 where both procedures extract the claim *(Picasso, bornIn, France)*.

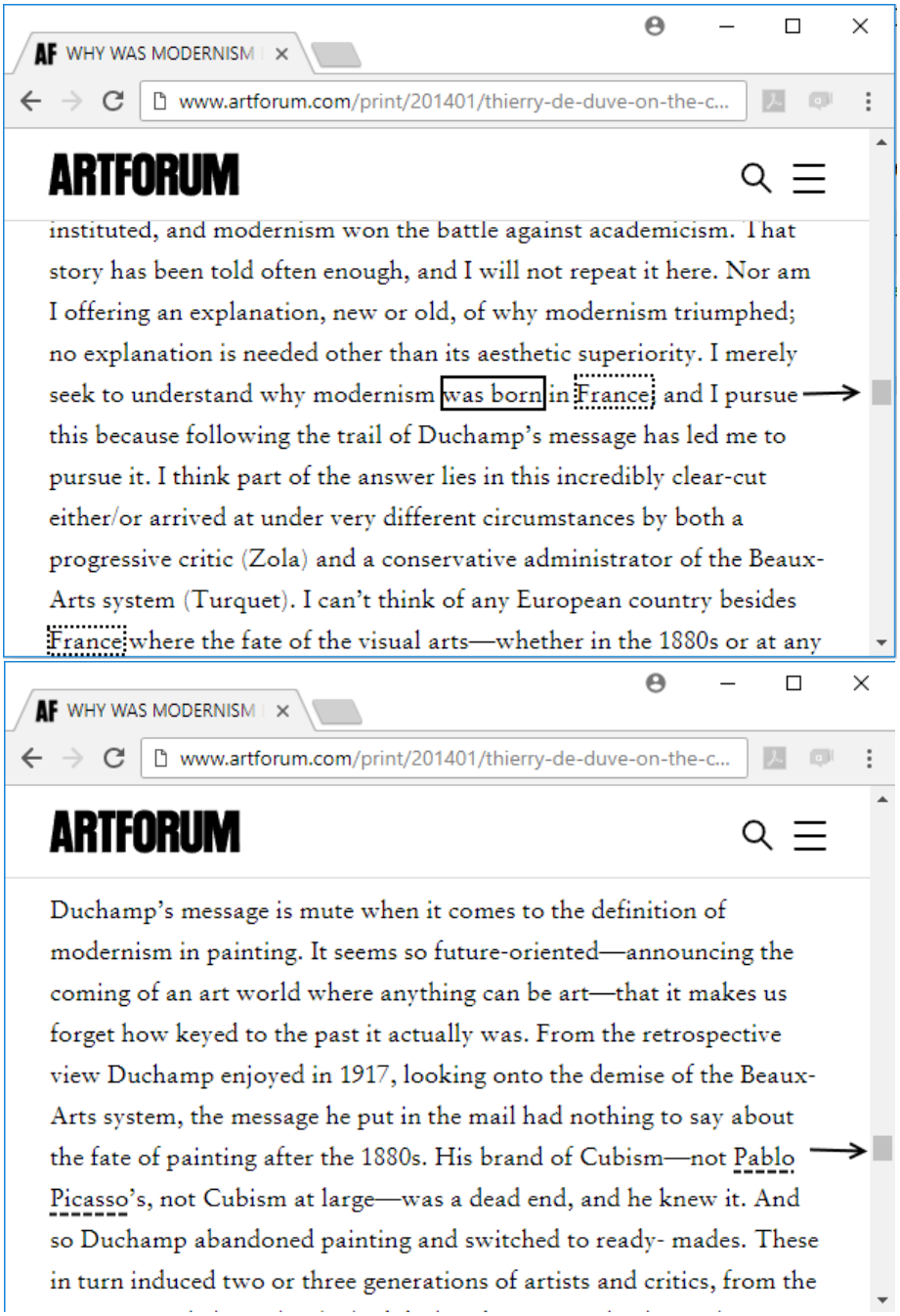


Figure 5.4: Example of a web page containing information on Pablo Picasso birth location. The two screenshots represent different parts of the same web page as highlighted by the scroll bar. Dotted, solid and dashed lines indicate the occurrence of a subject, predicate and value respectively.

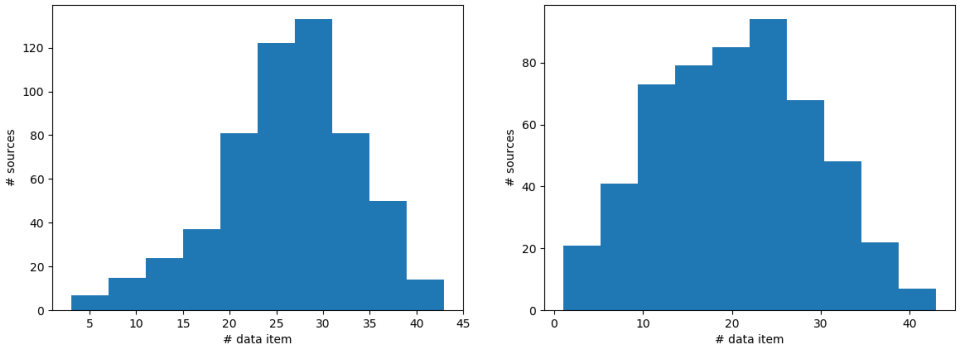
Considering this example, it is simple to understand that the source does not state this claim as surely true. Indeed, it appears in a list of true/false questions about Picasso. However, considering very naive extraction systems that do not consider information about the context of the claim, this claim is extracted as a statement provided by this website. In this kind of situation, comparing information provided by several sources (as done by TD models) can be useful. The last example is reported in Figure 5.4. Here, the candidate claim to be extracted is *(Picasso, bornIn, France)*. Protocol A extracts this claim since its constraints are verified. Protocol B instead does not extract any claim from this website because the first occurrence of the full name *Pablo Picasso* is after the sentence containing the term *born* and *France*, see the second screenshot reported in Figure 5.4.

DataA and DataB characteristics

Table 5.1 reports some characteristics of the datasets we collected. The fact that procedure A is less strict results in obtaining more noisy data than procedure B. Indeed, beside obtaining a higher number of sources and claims, as well as a higher number of sources per data item (see Figure 5.5a and Figure 5.5b), it obtains a higher number of different values for each data item. Unfortunately, both datasets suffer from the power law phenomena. In both cases, this phenomena is over-expressed as shown in Figure 5.6a and Figure 5.6b. Detailed numbers are reported in Table 5.2. It shows the percentage of sources having a coverage higher than a certain number of data items. Only 24% of sources provide values for more than one data item. This percentage decreased up to 1.4% considering sources providing values for more than 20 data items.

Table 5.1: Features of dataA and dataB.

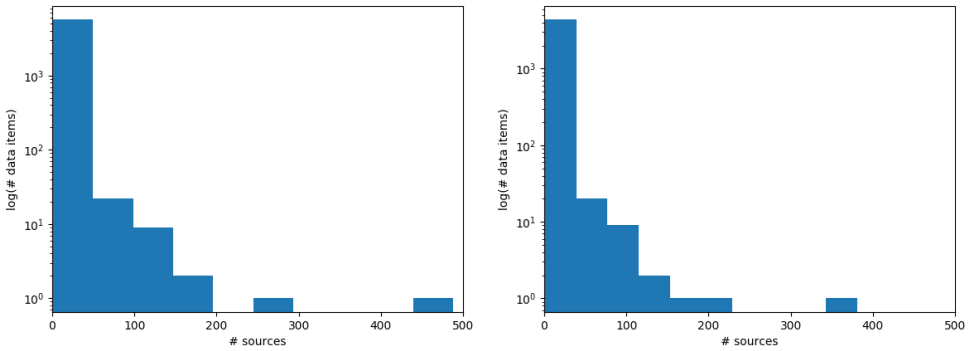
Features	dataA	dataB
# data items	564	538
# sources	5692	4396
# values	2843	2101
average # sources per data item	25.88	20.27
max # source	43	43
average # data items per source	2.56	2.48
max # data items	488	381
average # of values per data item	13.55	9.8



(a) DataA (b) DataB

Figure 5.5: Distribution of number of sources per data item for the two datasets.

Moreover, we noted that values that are more specific than the expected value (contained in the ground truth) were provided in the collected claims. In these cases, we manually checked if these specifications were true. For 20 instances that we manually checked, 10 were found true specifications. Extraction procedures, source code and obtained datasets are available online at https://github.com/lgi2p/TDwithRULES.



(a) DataA (b) DataB

Figure 5.6: Distribution of number of data items per source for the two datasets.

Table 5.2: Coverage analysis.

% of sources that cover a number of data items higher than...	**dataA**	**dataB**
>1	24.0 %	23.5 %
>2	12.9 %	12.5 %
>3	8.9 %	8.3 %
>5	5.4 %	5.3 %
>10	2.9 %	2.9 %
>20	1.4 %	1.4 %
>30	0.9 %	1.0 %
>50	0.6 %	0.6 %

5.3.2 Results and discussion

Table 5.3 shows the results obtained by the best configuration of parameters for the different proposed models. We can observe that in both datasets DataA and DataB we improved the performance of respectively 18% and 14% with respect to the baseline, i.e. *Sums* – the approach we decided to modify. Therefore the main finding of the experiments conducted on synthetic datasets were confirmed. Considering *a priori* knowledge is useful in TD settings. Especially, when both kind of knowledge were considered. The difference we noted was in the best parameterization of θ and γ. In real-world setting, the best θ was usually 0.05. This means that the minimum value confidence required to be part of the true value set was 0.05. This is in accordance with our expectations. Indeed, the only reason for which in synthetic datasets the best θ was 0 was the method we used to generate the datasets. All values that were more specific than the expected values were associated neither with true nor false claims. Moreover the best γ was 0.3. This may depend on the different distribution of confidence estimations when considering only the source viewpoint. The configuration of synthetic datasets results to obtain higher value confidence estimations that in case of real-world data. Therefore, to avoid that the importance given to KB information dominates the information provided by sources, lower γ values were preferred.

Considering a priori knowledge results in an improvement of performances

When comparing the proposed approach to existing TD models, it did not outperform the other approaches, see Table 5.4. The results reported in this table show that existing approaches have almost the same behaviour on both

Table 5.3: Recall obtained using *Sums* and its modifications on DataA and DataB.

Model	DataA	DataB
Sums	0.448	0.473
$Sums_{PO}$ ($\gamma = 0.0, \theta = 0.05$)	0.517	0.566
$Sums_{RULES\&PO}$ ($\gamma = 0.3, \theta = 0.0$)	0.527	0.548
$Sums_{RULES\&PO}$ ($\gamma = 0.3, \theta = 0.05$)	0.565	0.590
$Sums_{RULES\&PO}$ +post-proc. ($\gamma = 0.3, \theta = 0.1$)	**0.631**	**0.614**

Table 5.4: Recall obtained using existing Truth Discovery models on DataA and DataB.

Existing Model	DataA	DataB
Voting	0.640	0.625
TruthFinder	**0.646**	0.622
2-Estimates	0.631	0.635
3-Estimates	0.008	0.612
Cosine	0.636	0.635
AccuCopy	0.638	0.640
Accu	0.638	**0.660**
Depen	0.431	0.494
AccuSim	0.413	0.448
SimpleLCA	0.631	**0.660**
GuessLCA	0.644	0.646

datasets. The only exception is given by 3-Estimates that obtained very poor performances on dataA. We hypothesized that the estimations of claim difficulty were distorted because, when considering dataA, the number of different values that has been provided for the data items is higher than when considering dataB. Note that our study focused on modifying *Sums* which is considered to be one of the most well studied model, but not necessary the most effective one.

Limitations of Sums

After investigating the errors, we found out that it was mainly due to a limitation of *Sums*: it rewards sources having high coverage and, meanwhile, penalizes the ones having low coverage. Indeed *Sums* computes the trustworthiness of a source summing up all the confidence of the claims it provides. Thus the higher the number of claims a source provides, the higher the trustworthiness of this source will be. The problem is that *Sums* does not distinguish between sources providing always true values, but having different coverage. While Wikipedia.org is correctly considered as a high reliable source, a fan club website, specialized on its favourite actor, is incorrectly considered unreliable. Indeed even if the information it provides is correct, since it covers only one data item, its trustworthiness will be lower than the one of Wikipedia.org (source having a high coverage). In real-world datasets there are very few sources having high coverage, while the majority

of them have a low coverage – power law phenomenon. In this scenario the sources having high coverage dominate the specialized ones. Therefore, no extraction errors from high coverage sources are allowed. Indeed if an incorrect value is extracted from Wikipedia.org (for instance when the sentence refers to another person), this will be incorrectly considered as the true one. Since this cannot be guaranteed (the extraction procedures we defined are voluntary naive), we propose a post-processing procedure that alleviates this problem. Before selecting the true value, this procedure sets equal to 0 all the confidence of values that are provided by only a single source. We assume that it is highly improbable that the true value is provided only once. This solution, indicated as $Sums_{RULES\&PO}$ + post-proc., allows comparable results considering DataA and DataB. Indeed using this post-processing procedure we are able to avoid some of the extraction errors (occurring more with the first extraction procedures, the most naive, i.e. it has one less constraint), but we are not able to avoid to assign lower trustworthiness levels to specialized sources.

The importance of considering power law phenomena

Given these observations, in real-world settings considering the power law phenomenon is very important. The results show that *Sums* is not valuable for this kind of situation. Nevertheless, using additional information (partial order and extracted rules) improved the results with respect to the baseline approach, and this is promising for the principles introduced in this study. As shown in Table 5.3, the improvement considering this information was of 18% for DataA and of 19% for DataB. Through this study we also show that TD models can be used to improve correctness and granularity of values in DBpedia. Indeed, using the proposed approach, claims on data items can easily be collected on the Web. When more specific values than the one contained in DBpedia are found, they can be verified using TD model.

Thus in this chapter, we showed the potential of using TD model for serving KBP. Experiments also show the potential of using *a priori* knowledge to improve existing approaches. *Sums* obtained very poor performance on real-world data compared to the other existing models. Although, when incorporating *a priori* knowledge into *Sums*, comparable performances were obtained. Several limitations of *Sums* were highlighted by the experiments. They give us some interesting ideas for further enhancing the proposed framework. Interesting future directions are presented in the next chapter.

Chapter 6

Conclusions

Contents

This chapter discusses the conclusions of the research presented in this thesis. Firstly the research contributions are outlined. Then, an analysis of their limitations enable us to define a list of potential future works.

6.1 Thesis contributions

6.1.1 Incorporating semantic dependencies among values to improve value confidence estimation

Literature analysis showed that some researchers exploit value dependencies to increase or decrease confidence in a certain claim based on the other claims. In their studies, they measure the support between values using edit distance when dealing with string, and numerical interval when dealing with numerical values. For instance, they assume that the lower the edit distance between values, the higher the support between them is. In this research instead we proposed to derive the support provided by a value to another value analysing the semantic dependencies that may exist between them. More precisely, we considered that a value supports all its generalizations. We model these dependencies through a partial order of values extracted from an ontology. This partial order is exploited by the confidence estimation phase. Indeed, it indicates the set of values that need to be considered during confidence estimation, i.e. all specifications of the value under examination. Indeed, these specifications implicitly support it. We also identified an important theoretical consequence of considering semantic dependencies among data items when dealing with functional predicates. Now, multiple values can be true: the expected true value and its generalizations. Indeed, the expected true value supports its generalizations.

6.1.2 Considering semantic dependencies among values during the truth estimation phase

We also proposed a truth prediction phase that aims to identify the most expected true value among the set of values. It uses both the partial order and the estimations obtained by the previous phase. Indeed, the confidence estimations monotonically increase with respect to the partial order of values. Thus, the value with the highest confidence is always the most general value.

This value is supported by all the others. Even if the most general values is surely true, it is not very informative. Therefore we define a parametrisable procedure that identifies the expected true value handling different scenarios. Experiments we conducted on synthetic and real-world datasets showed that the TD performance improved with respect to the baseline model when considering semantic dependencies among values.

6.1.3 Incorporating dependencies among data items to improve value confidence estimation

While previous studies exploited spatial and temporal dependencies among data items to identify similar ones, we consider *a priori* knowledge expressed into an ontology. We identify similar data items based on the predicates (and the corresponding values) they have in common. Indeed, similar data items should have similar values. Also in this case, we modified the value confidence formula to integrate the support provided by the *a priori* knowledge we consider to a certain claim. This support is based on the analysis of the properties associated with similar data items (to the data item under examination). We use rules mined from an ontology to quantify this support. Indeed, a body of a rule indicates the set of properties and corresponding values that often occur together with the property and value in the head of this rule. Thus, if the body of the rule holds, the confidence in the value predicted by this rule can be increased. Also in this case, experiments showed that considering *a priori* knowledge in the form of rules is worthwhile to improve the TD performances.

6.1.4 Use-case study on real-world data

In order to test the proposed approaches on real-world data, we decided to exploit Web data. Since TD models require structured claims as input, we needed to specify a pre-processing procedure that was able to extract these structured claims from free text. Considering the text as input makes our problem setting equivalent to slot-filling task. Therefore, we proposed the use of this procedure that coupled a naive extraction process with TD models to serve Slot-Filling purpose. We compared the different performances obtained, on the collected real-world datasets, by the proposed TD models and existing TD methods.

6.1.5 Synthetic datasets, real-world datasets and source code.

We agree on the importance of sharing all artefacts we used in the experiments. Synthetic datasets, real-world datasets and source codes implementing the proposed approaches are therefore open source, documented and freely available in accordance with scientific standards in Computer Science at `https://github.com/lgi2p/TDSelection` and `https://github.com/lgi2p/TDwithRULES`.

6.2 Limitations

Considering *a priori* knowledge during the estimation of value confidences is worthwhile. The research reported in this manuscript shows that the proposed approaches result to be effective. Indeed, the performances increase when they are compared to the baseline model. Although we identify here their main limitations.

- The problem settings considered in this study are limited to the analysis of TD when dealing with functional and static predicates. This means that we focus on properties having a single true value that does not change over the time. Although, in the real-world, a lot of aspects associated with entities are non-functional and dynamic.
- The automatic construction of the partial order of values does not consider the Open World Assumption. If a relationship between two entities does not exist, we do not distinguish if this relation actually does not exist or if it is just unknown.
- The confidences associated with sources that are hubs are usually overestimated. This is due to the fact that the power law phenomenon is over-expressed in the real-world scenario and the model we adapt to integrate *a priori* knowledge, i.e. *Sums*, does not deal with this kind of situations.
- The best configuration of parameter γ is established based on empirical evaluation and not automatically estimated. γ regulates the weight that the additional knowledge related to data item dependencies should have in the confidence estimation formula. Considering

the parameters that regulate the truth prediction phase (θ and δ), we have the same limitation.

- The *a priori* knowledge used in the experiments is assumed to be reliable. Errors can be introduced into both estimation phase and truth prediction phase without assessing the quality, in the term of reliability, of the considered *a priori* knowledge. Moreover, the *a priori* knowledge is extracted from a single KB. This can limit the potential of the proposed approaches due to incompleteness problem of KBs.
- The source codes that is shared with the scientific community has to be run from command line. Even if we provided all the instructions necessary to run the scripts, the use of the command line could be a deterrent for some end-users.

6.3 Perspectives

The following sections give the direction for future studies, based both on the limitations analysed before and the extensions that can add value to the proposed approaches.

6.3.1 Application to multi-truth scenario

In the real-world, a lot of entity properties are non-functional, i.e. multiple expected true values exist. For instance, the authors of a book or the children of a person are often more than one. It is of primary importance to modify the proposed models in order to deal with this kind of situations. This is not straightforward. Indeed, considering partial order implies that each value among the multiple true values can be represented with a different level of granularity. Thus, it may occur that a subset of these values is exhaustively expressed by a general concept. In this case, important considerations need to be done.

6.3.2 Static "vs." dynamic properties

Several true value properties are dynamic, i.e. they change over the time. For instance, the claim "The president of America is Donald Trump" is currently true, but at some point in the future it will be false. Therefore an important

further step would be to take temporal dimension into account when evaluating the veracity of claims. Obviously, also the *a priori* knowledge may evolve over the time. Therefore, methods that update it are necessary as well.

6.3.3 Considering OWA when generating partial order of values

Since Semantic Web is based on OWA, it is important to take this assumption into account when extracting *a priori* knowledge from its resources. In the future, we plan to modify the generation of the partial order of values making a distinction between a relationship that does not exist and a relationship that is unknown. The idea is to use the disjointness information among concepts to detect when a relationship does not surely exist. In all the other cases, a relationship could just be unknown. When this is the case, a weak support can be propagated between values that share this relationship.

6.3.4 Over expression of power law in real-world scenario

Since the magnitude of this problem, it is really important to propose models that take it into account. Therefore, we intend to incorporate *a priori* knowledge into models able to overcome this issue during the confidence estimation.

6.3.5 Extracting *a priori* knowledge from multiple ontologies

Although we consider different ontologies, such as DBpedia and Gene Ontology, in the experiments that were conducted, we always considered to use *a priori* knowledge of a single ontology, i.e. either DBpedia or Gene Ontology. Considering multiple ontologies for extracting *a priori* knowledge could increase the probability of identifying dependencies among values and among data items. The higher the number of dependencies that are identified, the higher the impact of the proposed approaches will be. The several studies that has been conducted into the ontology alignment field can be used to facilitate this improvement (David, Guillet, Gras, & Briand, 2006). Moreover, assessing the quality of each ontology that it is considered is important. Indeed, it could enable us to automatically set the value of parameter γ. Ideally, high quality ontologies should have an higher impact than low quality ones.

6.3.6 Graphical User Interface

The importance of implement Graphical User Interface (GUI) resides in the fact that it enables to simplify the use of the source code for the end-users. A GUI could encourage less technical researchers to use the proposed models. Through a GUI it will be possible to set the parameters and to easily select the input files. Moreover, it will be also possible to consult the resulting estimations as well as the facts that are identified.

At the end of the writing of this manuscript, I am convinced that, during these three years, the research conducted on the data veracity problem has been challenging but satisfying. As highlighted by the numerous perspectives resulting from this research, it is necessary to further study this problem. As far as I can, in the coming years, I will look into these perspectives to further contribute to this domain which I believe is truly important for future society.

Appendices

Appendix A

Empirical analysis: supplementary results

This appendix provides additional results for several experiments and discussions which are presented in the manuscript.

A.1 Synthetic data: study of source trustworthiness estimations

Supplementary results of chapter 3 are provided hereinafter. During preliminary experiments we evaluated the source trustworthiness estimation error rate obtained by the proposed approaches. Indeed, using synthetic datasets for the experiments, the actual trustworthiness score associated with a source was known. It was established during the initial phase of dataset generation procedure to later decide if a source claims a true or false value for a given data item. Therefore, it was possible to evaluate the error ε_{MODEL} using the following formula:

$$\varepsilon_{MODEL} = \frac{1}{|S|} \sum_{s \in S} abs(t_{MODEL}(s) - t_{EXP}(s)) \tag{A.1}$$

where $t_{MODEL}(s)$ represents the trustworthiness of source s estimated by the model under examination and $t_{EXP}(s)$ represents the actual trustworthiness of s (that is established *a priori*). For each model, the error rate was calculated as the average error rate obtained on 20 different datasets for each considered dataset type.

Table A.1: Source trustworthiness estimation error rate obtained applying *Sums*, $Sums_{PO}$ and $Sums_{PL}$ on *birthPlace* and *genre* datasets generated from DBpedia.

Predicate	Dataset type	Model				
		-	Belief			Plaus
		Sums	$Sums_{PO_C}$	$Sums_{PO_{C+T}}$	$Sums_{PO_T}$	$Sums_{PL_C}$
birthPlace	EXP	0.206	**0.172**	0.177	**0.172**	0.189
	LOW_E	0.251	**0.173**	0.179	0.174	0.197
	UNI	0.269	**0.172**	0.177	**0.172**	0.200
genre	EXP	0.234	**0.173**	0.175	**0.173**	0.187
	LOW_E	0.280	**0.173**	0.176	**0.173**	0.194
	UNI	0.283	**0.172**	0.175	**0.172**	0.196

A.1.1 Results

Experiments were conducted for each synthetic dataset using *Sums* and its improvement $Sums_{PO}$ and $Sums_{PL}$. For the *belief* propagation framework $Sums_{PO}$, the adaptation consists of modifying the confidence formula $Sums_{PO_C}$, the trustworthiness one $Sums_{PO_T}$ or both of them $Sums_{PO_{C+T}}$. For the *plausibility* propagation framework $Sums_{PL}$, the adaptation consists only of modifying the confidence formula $Sums_{PL_C}$. Indeed, since the improvements we obtained were lower than the ones achieved considering the belief propagation framework, we decided to not further perform experiments in this direction.

Hereinafter, only results obtained on the *birthPlace* datasets are described. However, similar outcomes were achieved with the majority of the other datasets (different predicates and/or ontologies), see Table A.1 and Table A.2 for predicates extracted from DBpedia and GO respectively.

Considering the results reported in Table A.1 for *birthPlace* predicate, we observed that using knowledge in the form of a partial ordering of values coupled with the belief propagation framework led to reduce the average error rate compared to *Sums*, i.e. the baseline approach, for all synthetic datasets, no matter which formula was modified.
Moreover, the proposed models result to be more robust than *Sums* independently of the propagation model adapted to spread evidence and the adapted formula. Indeed, considering Figure A.1, the results of the experi-

Table A.2: Source trustworthiness estimation error rate obtained applying *Sums*, $Sums_{PO}$ and $Sums_{PL}$ on *CC*, *MF* and *BP* datasets generated from GO.

Predicate	**Dataset type**	**Model**				
		-	**Belief**			**Plaus**
		Sums	$Sums_{PO_C}$	$Sums_{PO_{C+T}}$	$Sums_{PO_T}$	$Sums_{PL_C}$
CC	EXP	0.186	**0.173**	0.176	**0.173**	0.189
	LOW_E	0.207	**0.172**	0.178	**0.172**	0.196
	UNI	0.212	**0.173**	0.178	**0.173**	0.199
MF	EXP	0.178	**0.173**	0.177	**0.173**	0.192
	LOW_E	0.190	**0.173**	0.178	**0.173**	0.200
	UNI	0.191	**0.173**	0.178	**0.173**	0.202

ments illustrate that granularity distribution of true values does not decrease the performance of the proposed model $Sums_{PO}$ while impacting the performance of *Sums*.

Precisely, when only one of the formulas changes ($Sums_{PO_C}$ and $Sums_{PO_T}$), the same outcomes were obtained. Differently, when both of them were modified ($Sums_{PO_{C+T}}$), the results were deteriorated. This means that counting two times the same information was not useful and the performance of the

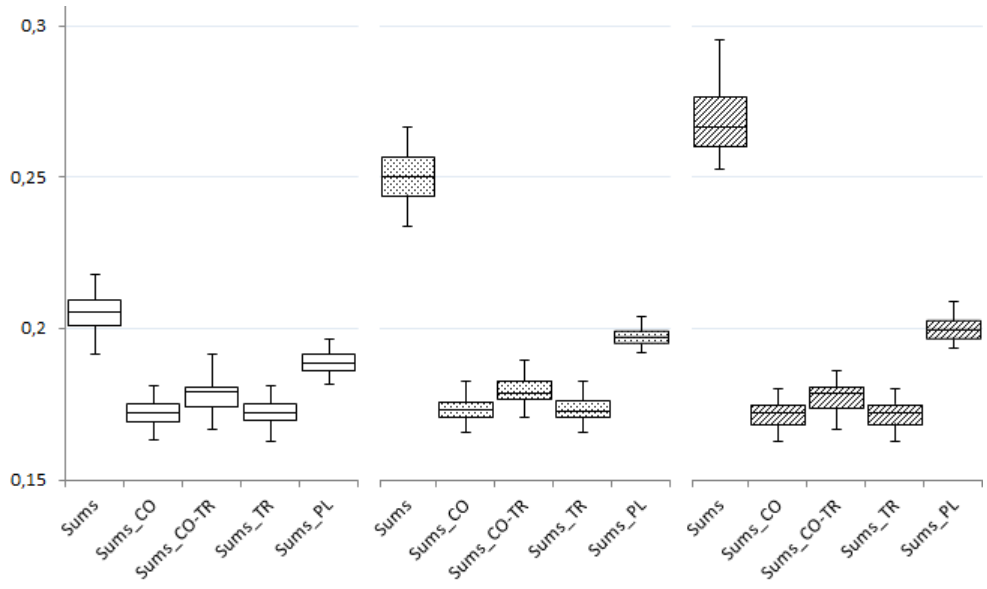


Figure A.1: Error rate for *birthPlace* predicate with respect to dataset type (EXP = white boxes, LOW_E = dotted boxes and UNI = diagonal line boxes) and the applied model ($Sums_{PO_C}$, $Sums_{PO_{C+T}}$, $Sums_{PO_C}$ and $Sums_{PL_C}$).

model was damaged. A reduction in the error rate was obtained also using the rationale of plausibility for propagating evidence.
Evaluating carefully the results we noticed that the belief-based adaptations outperform the plausibility-based approach. In the case of *birthPlace* datasets, $Sums_{PO_C}$ and $Sums_{PO_T}$ allowed to reduce the average error rate of 29.0% w.r.t. *Sums* approach, while $Sums_{PL_C}$ was able to reduce by 19.4% the error rate compared to *Sums*. These quantities have been obtained averaging the error rate computed over the different kinds of datasets, listed in Table A.1, for each propagation approach. This outcome is a consequence of the different number of values that receive the same evidence in the two different approaches used for spreading information. In the belief-based models, the evidence of a value is provided only to more general values. Differently, in the case of plausibility, given a certain value, its evidence is propagated to all values more general than its descendants, as well as the descendants itself. Therefore, in this case, the spreading process provides the same information to a bigger number of other values compared to the belief propagation model. Thus, the evidence results to be less informative than in belief-based models and it does not represent an advantage to better estimate value confidence anymore.

Moreover, the greater improvement in terms of error rate was achieved when the experiments were conducted on UNI datasets. Indeed, this is the scenario that can benefit most from the partial order of values. In this kind of datasets the true values provided by sources are selected independently from their similarity with the true value in the ground truth. Therefore, in this case, usually a broader number of different values is provided for the same data item and the disagreement among sources is higher than in the other datasets. The correct estimation of source trustworthiness and value confidence results to be more difficult than in the previous cases. Results clearly show that *Sums* adaptations using prior knowledge about values dependencies can compensate this added complexity. Indeed they can take advantage of the partial order among values, when it exists, to link them and to propagate their evidence. Using the UNI dataset coupled with the $Sums_{PO_C}$ or $Sums_{PO_T}$ models, we observed a 36.1% error rate decreasing compared to traditional *Sums* approach. The magnitude of error is maintained even in the case of $Sums_{PO_{C+T}}$. Even in this case the improvement obtained

with the $Sums_{PO_{PL}}$ is lower (25.7%). See Table A.1 for further details.
Otherwise, the EXP dataset is the scenario that can less benefit from the proposed approach. Indeed, the advantage given by evidence derived by correlated values is not very worthy. Since, in this case, the sources tend to be in agreement with each other, the evidence propagation can be employed only for few claims. In the EXP dataset the provided true values are usually similar to the expected one. Thus, often the cardinality of the set of claimed true values tend to be small – limited diversity in terms of claimed true values. It means that the majority of the true values provided for a specific data item will be the same. As a result the quantity estimations cannot be better refined in order to improve the overall performance. Precisely, considering EXP datasets and belief propagation framework, analysing the $Sums_{PO_C}$ and $Sums_{PO_T}$ models we observed a similar error rate reduction of 16.5% with respect to *Sums* approach. In the case of $Sums_{PO_{C+T}}$ we obtain 14.1% error rate reduction. Slightly lower it is the gain in terms of error rate with the plausibility using $Sums_{PL_C}$. It is equal to 8.3%.
As expected, the LOW_E datasets, where the true value selection is governed by a distribution sharing both exponential and uniform features, obtain performance that are in the middle among the results achieved by UNI and EXP datasets.

Moreover, by analysing all cases in which the error of trustworthiness estimation produced by the adapted models was worse than the one made by traditional one, we observed that source trustworthinesses were underestimated. This is due to the behaviour of the adaptations. *Per* design, they tend to be more careful in assigning trust to sources claiming specific values. This respects the rationale for which the general values are more easily true than specific ones. Since we assign trustworthiness score only based on the confidence of provided claims; if a claim contains a general value, then the probability than the source claims the truth will be considered to be higher. Increasing the level of specificity, the probability that the value is true decreases. Therefore, proposed models penalize in terms of source trustworthiness those sources that tend to provide specific values that are not supported by other claims.

As anticipated, the same results and behaviour are obtained also by the experiments performed considering the other predicates and the related partial

ordering of values, see Table A.1 and Table A.2. The only exception is given by the plausibility-based adaption applied to the *CC*, *MF* and *BP* datasets. In these experiments we obtained a worse error rate than in the original *Sums* approach. This is due to two main factors: (i) the way in which false values are selected during the generation of datasets; the majority of them tends to be similar to the set of possible true values, i.e. the false values are chosen among the set of most similar values in the taxonomy without considering inclusive ancestors and descendants of the true value reported in the ground truth; (ii) in CC and MF datasets there are more values that are not leaves and that have few ancestors than in birthPlace and genre cases; indeed, there are 3148 and 3823 values having three ancestors maximum in, respectively, CC and MF datasets against the 0 and 10 values contained in the birthPlace and genre ground truths.

In these conditions, the probability to spread evidence from true values to false ones increases a lot. Especially when the true value in the ground truth corresponds to a value near to the root of the taxonomy. In this case, the false values are selected among the other values, not sharing any order relationship with the true one, that are near to the root. Since the probability that two values near to the root have at least a leaf in common is higher than two values far from the root, there is a high probability of propagating evidence from true values to false ones.

Therefore, the results of our experiments suggest that way of propagating evidence of the plausibility framework generates a lot of noise decreasing the performance of the models.

References

Ageno, A., Comas, P. R., Naderi, A. M., Rodríguez, H., & Turmo, J. (2013). The talp participation at tac-kbp 2013. *Proceedings of the 6th Text Analysis Conference*.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207–216.

Aho, A. V., Garey, M. R., & Ullman, J. D. (1972). The transitive reduction of a directed graph. *SIAM Journal on Computing*, *1*(2), 131–137.

Al-Feel, H. T., Koutb, M., & Suoror, H. (2008). Semantic web on scope: A new architectural model for the semantic web. *Journal of Computer Science*, *4*(7), 613–324.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–36.

Anokhin, P., & Motro, A. (2001). Data integration: Inconsistency detection and resolution based on source properties. *Proceedings of the International Workshop on Foundations of Models for Information Integration*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25–29.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The Semantic Web*.

Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., & Nardi, D. (2003). *The description logic handbook: Theory, implementation and applica-*

tions. Cambridge university press.

Bailey, J., Bry, F., Furche, T., & Schaffert, S. (2005). Web and semantic web query languages: A survey. *Proceedings of the 1st international conference on Reasoning Web*, 35–133.

Barati, M., Bai, Q., & Liu, Q. (2017). Mining semantic association rules from rdf data. *Knowledge-Based Systems*, *133*, 183–196.

Beckett, D. (Ed.). (2004). *RDF/XML Syntax Specification (Revised).* World Wide Web Consortium.

Berners-Lee, T. (Ed.). (1998). Uniform resource identifiers (uri): Generic syntax [Computer software manual]. (RFC 2396)

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

Berti-Équille, L., & Borge-Holthoefer, J. (2015). *Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics*. Morgan & Claypool Publishers.

Berti-Equille, L., Sarma, A. D., Xin, Dong, Marian, A., & Srivastava, D. (2009). Sailing the information ocean with awareness of currents: Discovery and application of source dependence. *Proceedings of the Biennial Conference on Innovative Data Systems Research*, 1–6.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, *5*(3), 1–22.

Blanco, L., Crescenzi, V., Merialdo, P., & Papotti, P. (2010). Probabilistic Models to Reconcile Complex Data from Inaccurate Data Sources. *Proceedings of the 22nd International Conference on Advanced Information Systems Engineering*, 83–97.

Bleiholder, J., & Naumann, F. (2009). Data fusion. *ACM Computing Surveys (CSUR)*, 1–41.

Böhm, C., Naumann, F., Abedjan, Z., Fenz, D., Grütze, T., Hefenbrock, D., Pohl, M., & Sonnabend, D. (2010). Profiling linked open data with prolod. *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on*, 175–178.

Boley, H. (2000). Relationships between logic programming and rdf. *Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence*, 201–218.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Ad-*

vances in neural information processing systems, 2787–2795.

Brickley, D., & Guha, R. V. (2004). *RDF vocabulary description language 1.0: RDF schema* (W3C Recommendation). W3C. (http://www.w3.org/TR/2004/REC-rdf-schema-20040210/)

Buche, P., Dervin, C., Haemmerle, O., & Thomopoulos, R. (2005). Fuzzy querying of incomplete, imprecise, and heterogeneously structured data in the relational model using ontologies and rules. *IEEE Transactions on Fuzzy Systems*, *13*(3), 373-383.

Buffa, M., Zucker, C. F., Bergeron, T., & Aouzal, H. (2016). Semantic web technologies for improving remote visits of museums, using a mobile robot. *Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference*.

Chi, Y., Yang, Y., & Muntz, R. R. (2004). Hybridtreeminer: An efficient algorithm for mining frequent rooted trees and free trees using canonical forms. *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*, 11–20.

D'Amato, C., Staab, S., Tettamanzi, A. G., Minh, T. D., & Gandon, F. (2016). Ontology enrichment by discovering multi-relational association rules from ontological knowledge bases. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, 333–338.

Davey, B. A., & Priestley, H. A. (1990). *Introduction to lattices and order*. Cambridge university press.

David, J. (2007). Association rule ontology matching approach. *International Journal on Semantic Web and Information Systems*, *3*(2), 27–49.

David, J., Guillet, F., Gras, R., & Briand, H. (2006). Conceptual hierarchies matching: an approach based on discovery of implication rules between concepts. *Proceedings of the 17th biennial European Conference on Artificial Intelligence*, *6*, 357–361.

De Bruijn, J., Martin-Recuerda, F., Manov, D., & Ehrig, M. (2004). D4. 2.1 state-of-the-art-survey on ontology merging and aligning v1. *SEKT Project deliverable D*, *4*, 2–1.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, 601–610.

Dong, X. L., Berti-Equille, L., Hu, Y., & Srivastava, D. (2010). Global detection of complex copying relationships between sources. *Proceedings of the VLDB Endowment*, *3*(1), 1358–1369.

Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009a). Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, *2*(1), 550–561.

Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009b). Truth Discovery and Copying Detection in a Dynamic World. *Proceedings of the VLDB Endowment*, *2*(1), 562–573.

Dong, X. L., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., & Zhang, W. (2014). From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, *7*(10), 881–892.

Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S., & Zhang, W. (2015). Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, *8*(9), 938–949.

Dong, X. L., & Naumann, F. (2009). Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, *2*(2), 1654–1655.

Eiermann, B., Rahmner, P. B., Korkmaz, S., Landberg, C., Lilja, B., Shemeikka, T., Veg, A., Wettermark, B., & Gustafsson, L. L. (2010). Knowledge bases for clinical decision support in drug prescribing–development, quality assurance, management, integration, implementation and evaluation of clinical value. In *Decision support systems.* InTech.

Euzenat, J., & Shvaiko, P. (2013). *Ontology matching, second edition*. Springer.

Feldman, F. (1970). Leibniz and leibniz'law. *The Philosophical Review*, *79*(4), 510–522.

Fürnkranz, J., & Kliegr, T. (2015). A brief overview of rule learning. *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, 54–69.

Galárraga, L. (2014). Applications of rule mining in knowledge bases. *Proceedings of the 7th Workshop on Ph.D Students*, 45–49.

Galárraga, L. (2015). Interactive rule mining in knowledge bases. *Actes des 31^e Conférence sur la Gestion de Données, Île de Porquerolles*.

Galárraga, L., & Suchanek, F. M. (2014). Towards a numeric rule mining language. *Proceedings of Automated Knowledge Base Construction workshop*.

Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. M. (2015). Fast rule mining in ontological knowledge bases with amie+. *The VLDB Journal*, *24*(6), 707–730.

Galárraga, L. A., Teflioudi, C., Hose, K., & Suchanek, F. (2013). Amie: association rule mining under incomplete evidence in ontological knowledge bases. *Proceedings of the 22nd international conference on World Wide Web*, 413–422.

Galland, A., Abiteboul, S., Marian, A., & Senellart, P. (2010). Corroborating information from disagreeing views. *Proceedings of the 3rd ACM international conference on Web Search and Data Mining*, 131–140.

Gao, J., Li, Q., Zhao, B., Fan, W., & Han, J. (2015). Truth discovery and crowdsourcing aggregation: A unified perspective. *Proceedings of the VLDB Endowment*, *8*(12), 2048–2049.

Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *SIGKDD Explor. Newsl.*, *7*(2), 3–12.

Goethals, B., & Bussche, J. V. d. (2002). Relational association rules: Getting warmer. *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*, 125–139.

Group, W. O. W. (2012). *OWL 2 web ontology language. document overview (second edition)* [W3C Recommendation].

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

Guarino, N. (1998). *Formal ontology in information systems: Proceedings of the first international conference (fois'98), june 6-8, trento, italy* (Vol. 46). IOS press.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In *Handbook on ontologies* (pp. 1–17). Springer.

Guzman-Arenas, A., Cuevas, A.-D., & Jimenez, A. (2011). The centroid or consensus of a set of objects with qualitative attributes. *Expert Systems with Applications*, *38*(5), 4908 - 4919.

Harispe, S. (2014). *Knowledge-based semantic measures: From theory to applications* [PhD Thesis]. Université de Montpellier.

Harispe, S., Imoussaten, A., Trousset, F., & Montmain, J. (2015). On the consideration of a bring-to-mind model for computing the information content of concepts defined into ontologies. *Proceedings of the 2015 IEEE International Conference on Fuzzy Systems*, 1-8.

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2014). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, *30*(5), 740–742.

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, *8*(1), 1–254.

Horrocks, I., Parsia, B., Patel-Schneider, P., & Hendler, J. (2005). Semantic web architecture: Stack or two towers? *Proceedings of the 3rd International Workshop on Principles and Practice of Semantic Web Reasoning*, 37–41.

Jean, P.-A., Harispe, S., Ranwez, S., Bellot, P., & Montmain, J. (2016). Uncertainty detection in natural language: a probabilistic model. *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, WIMS 2016, Nîmes, France, June 13-15*, 10:1–10:10.

Jensen, D. (1999). Statistical challenges to inductive inference in linked data. *Proceedings of the 2nd International Conference on Artificial Intelligence and Statistics*.

Jiang, S., Lowd, D., & Dou, D. (2012). Learning to refine an automatically extracted knowledge base using markov logic. *Proceedings of IEEE 12th International Conference on the Data Mining*, 912–917.

Joslyn, C., & Hogan, E. (2010). Order metrics for semantic knowledge systems. *Proceedings of the 5th International Conference on Hybrid Artificial Intelligence Systems*, 399–409.

Kalfoglou, Y., & Schorlemmer, M. (2003). Ontology mapping: The state of the art. *Knowl. Eng. Rev.*, *18*(1), 1–31.

Kiryakov, A. (2006). Ontologies for knowledge management. *Semantic Web Technologies: trends and research in ontology-based systems*, 115–138.

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, *46*, 668–677.

Knap, T., Michelfeit, J., & Necaskỳ, M. (2012). Linked open data aggregation: Conflict resolution and aggregate quality. *Proceedings of the IEEE 36th Annual Conference Workshops on Computer Software and Applications*, 106–111.

Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., & Androutsopoulos, I. (2015). Evaluation measures for hierarchical classification: a

unified view and novel approaches. *Data Mining and Knowledge Discovery*, *29*(3), 820–865.

Krotofil, M., Larsen, J., & Gollmann, D. (2015). The process matters: Ensuring data veracity in cyber-physical systems. *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, 133–144.

Kuramochi, M., & Karypis, G. (2001). Frequent subgraph discovery. *Data Mining, 2001. ICDM 2001, Proceedings IEEE international conference on*, 313–320.

Lao, N., Mitchell, T., & Cohen, W. W. (2011). Random walk inference and learning in a large scale knowledge base. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 529–539.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lehmann, J., & Völker, J. (Eds.). (2014). *Perspectives on ontology learning* (Vol. 18). AKA Heidelberg.

Li, C., Sheng, V. S., Jiang, L., & Li, H. (2016). Noise filtering to improve data and model quality for crowdsourcing. *Knowledge-Based Systems*, *107*(Supplement C), 96 - 103.

Li, H., Zhao, B., & Fuxman, A. (2014). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. *Proceedings of the 23rd international conference on World Wide Web*, 165–176.

Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., & Han, J. (2014). A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, *8*(4), 425–436.

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. (2014). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. *Proceedings of the 2014 ACM SIGMOD international conference on Management of Data*, 1187–1198.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., & Han, J. (2015). A survey on truth discovery. *SIGKDD Explor. Newsl.*, *17*(2), 1–16.

Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., & Han, J. (2016). Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 1986–1999.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference of Machine Learning*, *98*, 296–304.

Ma, F., Li, Y., Li, Q., Qiu, M., Gao, J., Zhi, S., Su, L., Zhao, B., Ji, H., & Han, J. (2015). Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 745–754.

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook* (Vol. 2). Springer.

Manola, F., & Miller, E. (2004a). *RDF primer*. W3C Recommendation 10 February 2004.

Manola, F., & Miller, E. (Eds.). (2004b). *Rdf primer.* World Wide Web Consortium.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. *Proceedings of the 7th International Conference on Semantic Systems*, 1–8.

Meng, C., Jiang, W., Li, Y., Gao, J., Su, L., Ding, H., & Cheng, Y. (2015). Truth discovery on crowd sensing of correlated entities. *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, 169–182.

Minsky, M. (1974). A framework for representing knowledge.

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., & Lutz, C. (2008). *Owl 2 web ontology language: Profiles* (Vol. 27).

Muggleton, S. (1995). Inverse entailment and progol. *New generation computing*, *13*(3-4), 245–286.

Muggleton, S. (1996). Learning from positive data. *International Conference on Inductive Logic Programming*, 358–376.

Mukherjee, S., Weikum, G., & Danescu-Niculescu-Mizil, C. (2014). People on drugs: credibility of user statements in health communities. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, 65–74.

Nakashole, N., & Mitchell, T. M. (2014). Language-aware truth assessment of fact candidates. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *1*, 1009–1019.

Nebot, V., & Berlanga, R. (2012). Finding association rules in semantic web data. *Knowledge-Based System*, *25*(1), 51–62.

Nenov, Y., Piro, R., Motik, B., Horrocks, I., Wu, Z., & Banerjee, J. (2015). Rdfox: A highly-scalable rdf store. *Proceedings of the 14th International*

Semantic Web Conference, 3–20.

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11–33.

Nickel, M., Tresp, V., & Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. *ICML*, *11*, 809–816.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175.

Nuzzolese, A. G., Presutti, V., Gangemi, A., Musetti, A., & Ciancarini, P. (2013). Aemoo: exploring knowledge on the web. *Proceedings of Web Science 2013 (co-located with ECRC 2013)*, 272–275.

Ordonez, C., & Zhao, K. (2011). Evaluating association rules and decision trees to predict multiple target attributes. *Intelligent Data Analysis*, *15*(2), 173–192.

Ouyang, R. W., Kaplan, L. M., Toniolo, A., Srivastava, M., & Norman, T. J. (2016). Aggregating crowdsourced quantitative claims: Additive and multiplicative models. *IEEE Transactions on Knowledge and Data Engineering*, *28*(7), 1621–1634.

Ouyang, R. W., Srivastava, M., Toniolo, A., & Norman, T. J. (2016). Truth discovery in crowdsourced detection of spatial events. *IEEE Transactions on Knowledge and Data Engineering*, *28*(4), 1047–1060.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin UK.

Pasternack, J., & Roth, D. (2010). Knowing what to believe (when you already know something). *Proceedings of the 23rd International Conference on Computational Linguistics*, 877–885.

Pasternack, J., & Roth, D. (2011). Making better informed trust decisions with generalized fact-finding. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2324–2329.

Pasternack, J., & Roth, D. (2013). Latent credibility analysis. *Proceedings of the 22nd international conference on World Wide Web*, 1009–1020.

Perez, C. (2010). Technological revolutions and techno-economic paradigms. *Cambridge journal of economics*, *34*(1), 185–202.

Plous, S. (1993). *The psychology of judgment and decision making*. Mcgraw-Hill Book Company.

Pochampally, R., Das Sarma, A., Dong, X. L., Meliou, A., & Srivastava, D.

(2014). Fusing data with correlations. *Proceedings of the 2014 ACM SIGMOD international conference on Management of Data*, 433–444.

Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. *Proceedings of the 12th International Semantic Web Conference*, 542–557.

Qi, G.-J., Aggarwal, C. C., Han, J., & Huang, T. (2013). Mining collective intelligence in diverse groups. *Proceedings of the 22nd International Conference on World Wide Web, WWW '13*, 1041–1052.

Quboa, Q. K., & Saraee, M. (2013). A state-of-the-art survey on semantic web mining. *Intelligent Information Management*, *5*(01), 1–10.

Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine learning*, *5*(3), 239–266.

Richards, I. A., & Ogden, C. K. (1923). *The meaning of meaning: a study of the influence of language upon thought and of the science of symbolism*. Routledge and Kegan Paul Ltd., London, tenth edition.

Ristoski, P., & Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, *36*, 1–22.

Robbins, H. (1956). An empirical bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 157–163.

Rocha, O. R., Zucker, C. F., & Giboin, A. (2018). Extraction of relevant resources and questions from dbpedia to automatically generate quizzes on specific domains. *Proceedings of the International Conference on Intelligent Tutoring Systems*, 380–385.

Samadi, M., Talukdar, P., Veloso, M., & Blum, M. (2016). Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 222–228.

Sánchez, D., & Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of biomedical informatics*, *44*(5), 749–759.

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. *Proceedings of the 13th International Semantic Web Conference*, 245–260.

Schoenmackers, S., Etzioni, O., Weld, D. S., & Davis, J. (2010). Learning first-

order horn clauses from web text. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1088–1098.

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. *Proceedings of the 16th European Conference on Artificial Intelligence*, 1089–1090.

Shafer, G., et al. (1976). *A mathematical theory of evidence* (Vol. 1). Princeton university press Princeton.

Smyth, P., Fayyad, U. M., Burl, M. C., Perona, P., & Baldi, P. (1995). Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, 1085–1092.

Soderland, S., Gilmer, J., Bart, R., Etzioni, O., & Weld, D. S. (2013). Open information extraction to kbp relations in 3 hours. *Proceedings of the 6th Text Analysis Conference*.

Sowa, J. (1984). *Conceptual structures: Information processing in mind and machine*. Addison-Wesley.

Spackman, K. A., Campbell, K. E., & Côté, R. A. (1997). Snomed rt: a reference terminology for health care. *Proceedings of the American Medical Informatics Association Annual fall Symposium*, 640.

Staab, S., & Studer, R. (2009). *Handbook on ontologies - second edition*. Springer.

Su, L., Li, Q., Hu, S., Wang, S., Gao, J., Liu, H., Abdelzaher, T. F., Han, J., Liu, X., Gao, Y., et al. (2014). Generalized decision aggregation in distributed sensing systems. *Proceedings of the 2014 IEEE 35th Real-Time Systems Symposium*, 1–10.

Sudeepthi, G., Anuradha, G., & Babu, M. S. P. (2012). A survey on semantic web search engine. *International Journal of Computer Science Issues*, *9*(2), 241–245.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. *Doubleday*.

Tanon, T. P., Stepanova, D., Razniewski, S., Mirza, P., & Weikum, G. (2017). Completeness-aware rule learning from knowledge graphs. *Proceedings of the 16th International Semantic Web Conference*, 507–525.

Ventura, S., & Luna, J. M. (2016). Quality measures in pattern mining. In (pp. 27–44). Springer, Cham.

Völker, J., Fleischhacker, D., & Stuckenschmidt, H. (2015). Automatic acquisition of class disjointness. *Web Semant.*, *35*(P2), 124–139.

Waguih, D. A., & Berti-Equille, L. (2014). Truth discovery algorithms: An experimental evaluation. *CoRR*, *abs/1409.6428*.

Wan, M., Chen, X., Kaplan, L., Han, J., Gao, J., & Zhao, B. (2016). From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wang, D., Abdelzaher, T., & Kaplan, L. (2015). *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann.

Wang, D., Abdelzaher, T., Kaplan, L., Ganti, R., Hu, S., & Liu, H. (2013). Exploitation of physical constraints for reliable social sensing. *Proceedings of the 2013 IEEE 34th Real-Time Systems Symposium*, 212–223.

Wang, D., Kaplan, L., Le, H., & Abdelzaher, T. (2012). On truth discovery in social sensing: A maximum likelihood estimation approach. *Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, 233–244.

Wang, S., Su, L., Li, S., Hu, S., Amin, T., Wang, H., Yao, S., Kaplan, L., & Abdelzaher, T. (2015). Scalable social sensing of interdependent phenomena. *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, 202–213.

Wang, S., Wang, D., Su, L., Kaplan, L., & Abdelzaher, T. F. (2014). Towards cyber-physical systems in social spaces: The data reliability challenge. *Proceedings of the 2014 IEEE 35th Real-Time Systems Symposium*, 74-85.

Wang, X., Sheng, Q. Z., Fang, X. S., Yao, L., Xu, X., & Li, X. (2015). An integrated bayesian approach for effective multi-truth discovery. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 493–502.

Wang, X., Sheng, Q. Z., Yao, L., Li, X., Fang, X. S., Xu, X., & Benatallah, B. (2016). Empowering truth discovery with multi-truth prediction. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 881–890.

Wang, Z., & Li, J. (2015). Rdf2rules: Learning rules from RDF knowledge bases by mining frequent predicate cycles. *CoRR*, *abs/1512.07734*.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 2035–2043.

Yin, X., Han, J., & Yu, P. S. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796-808.

Yin, X., & Tan, W. (2011). Semi-supervised truth discovery. *Proceedings of the 20th international conference on World Wide Web*, 217–226.

Yu, D., Huang, H., Cassidy, T., Ji, H., Wang, C., Zhi, S., Han, J., Voss, C., & Magdon-Ismail, M. (2014). The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. *Proceedings of the 25th International Conference on Computational Linguistics*, 1567–1578.

Zeng, Q., Patel, J. M., & Page, D. (2014). Quickfoil: Scalable inductive logic programming. *Proceedings of the VLDB Endowment*, *8*(3), 197–208.

Zhao, B., & Han, J. (2012). A probabilistic model for estimating real-valued truth from conflicting sources. *Proceedings of the 10th International Workshop on Quality in DataBases*.

Zhao, B., Rubinstein, B. I., Gemmell, J., & Han, J. (2012). A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, *5*(6), 550–561.