Améliorer la recherche de vérité en exploitant la connaissance de domaines via les ontologies et les règles d'association

Valentina Beretta[1], Sylvie Ranwez[1], Sébastien Harispe[1], Isabelle Mougenot[2]

1. LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
6, avenue de Clavières, F-30 319 Alès, France
prenom.nom@mines-ales.fr

2. UMR 228 Espace-Dev, Université de Montpellier
500, rue JF. Breton, F-34 093 Montpellier cedex 5, France
isabelle.mougenot@umontpellier.fr

RÉSUMÉ. Face au danger de la désinformation et de la prolifération de fake news (fausses nouvelles), un domaine de recherche a émergé ces dernières années : la détection de vérité sur le web. 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 différentes sources afin de déterminer celle qui est la plus fiable et digne de confiance. Cette étape est cruciale dans un processus d'extraction de connaissances, par exemple, pour constituer des bases de qualité, sur lesquelles pourront s'appuyer différents traitements *ultérieurs (aide à la décision, recommandation, raisonnement…). Les approches existantes* faisaient jusqu'ici abstraction de la connaissance a priori d'un domaine. Dans cette contribution, nous montrons comment les modèles de connaissance (ontologies de domaine) peuvent avantageusement être exploités pour améliorer les processus de recherche de vérité. Nous insistons principalement sur deux approches : la prise en compte de la hiérarchisation des concepts de l'ontologie et l'identification de motifs dans les connaissances qui permet, en exploitant certaines règles d'association, de renforcer la confiance dans certaines assertions. Chaque approche est validée sur différents jeux de données qui sont rendus disponibles à la communauté, tout comme le code de calcul correspondant aux deux approches.

ABSTRACT. Data veracity is one of the main issues regarding web data. Facing fake news proliferation and disinformation dangers, Truth Discovery models can be used to assess this veracity by estimating value confidence and source trustworthiness through analysis of claims on the same real-world entities provided by different sources. This treatment is crucial within an automated knowledge extraction process, in particular if resulting knowledge bases (KB) are devoted to be used in decision processes. Many studies have been conducted in Truth Discovery domain; however none of them, to our knowledge, take into account the a priori knowledge that may exist regarding a domain (e.g., domain ontologies). This article proposes two ways to reinforce some value confidences and thus source trustworthiness calculus

during this process: the first one considers the concepts' hierarchy and the second one exploits patterns that are extracted from KB using association rule learning techniques. Both approaches are validated and tested using benchmarks, that are freely available as well as the source code.

MOTS-CLÉS : détection de vérité, ontologies, web sémantique, confiance, fiabilité des sources, détection de règles, raisonnement.

KEYWORDS: truth discovery, ontologies, semantic web, value confidence, source trustworthiness, association rule learning, reasoning.

DOI:10.3166/RIA.32.373-405 © 2018 Lavoisier

1. Introduction

Depuis son origine, l'homme laisse des traces de son passage sur Terre. Or, à l'entrée dans le XXI^{e} 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 finding). 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

1. www.politifact.com
2. www.lemonde.fr/verification

provenance. Cette étape est particulièrement importante lorsque l'on souhaite 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.

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 (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 cet article 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 connaissances. 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 l'article.

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, ii) décrire les adaptations des modèles existants nécessaires pour intégrer cette connaissance, iii) proposer une évaluation robuste pour chaque approche.

La section suivante présente le contexte de notre étude, pose la problématique et revient sur les notations de la littérature mises à contribution. On y distinguera les particularités liées à chaque type d'approche : utilisation des liens de l'ontologie ou recherche de motifs. La section 3 formalise le problème de la détection de vérité et donne les détails de chaque solution envisagée. La section 4 détaille la procédure d'évaluation et en particulier la constitution des jeux de test. La section 5 présente les résultats obtenus et les discute, avant la section 6 qui conclut cet article et ouvre de nombreuses perspectives de recherche.

2. État de l'art et positionnement

Par souci de clarté, cette section définit les notations utilisées par la suite. Certaines sont couramment utilisées dans le domaine (Yin et al., 2008 ; Li et al., 2015 ; Berti-Équille et Borge-Holthoefer, 2015), alors que les autres sont introduites pour être utilisées ensuite dans la description de notre approche. Ces notations sont récapitulées dans le tableau 1.

Soit e, une entité 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 $valeur \in V$. Notons que la recherche de vérité envisagée ici ne concerne que des prédicats fonctionnels, c'est-à-dire 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>, et sera appelée assertion[5] tant qu'elle n'est 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 $V^s \subseteq V$.

Tableau 1. Synthèse des notations utilisées

Symbole	Signification
$d \in D$	Une description appartenant à l'ensemble de toutes les descriptions
$s \in S$	Une source appartenant à l'ensemble des sources
$v \in V$	Une valeur appartenant à l'ensemble des valeurs
$V_d \subseteq V$	L'ensemble des valeurs associées par différentes sources à la description d
V^s	L'ensemble des assertions faites par la source s
S^{v_d}	L'ensemble des sources qui proclament une assertion v_d

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

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…). On simplifiera le propos par la suite en ne parlant que de "source".

5 Une assertion pourra donc être notée indifféremment dans la suite sous la forme d'un triplet <entité, prédicat, objet> ou d'une paire (description, valeur) en fonction du contexte.

fonctions : la fiabilité d'une source, que nous noterons t[6], et la confiance dans une assertion que nous noterons c. Ces fonctions sont définies comme suit.

- $t: S \rightarrow [0,1]$, la fiabilité d'une source, représente sa propension à fournir de vraies valeurs. Dans la littérature, le terme poids peut être utilisé (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 ($c(v) \simeq 1$) et sera supposée provenir d'une source fiable. Par ailleurs, une assertion inexacte aura un faible degré de confiance ($c(v) \simeq 0$) 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 (Li et al., 2015) afin de conserver la cohérence de notation dans la suite de l'article.

Définition 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 $v_d^* \in V_d$. 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, $t(s)$, 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 détaillons pas ici l'état de l'art concernant la détection de vérité mais souhaitons en donner une vision synthétique. 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 (Li et al., 2015). Ces approches font l'hypothèse que toutes les sources ont 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.

6. Cette notation fait référence à l'anglais où le terme trustworthiness est employé.

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 et al., 2010). Elles se distinguent par les formulations employées et la procédure itérative utilisée. De plus chaque approche relaxe certaines hypothèses et se concentre sur des aspects particuliers. Par exemple certaines approches prennent l'hypothèse d'une totale indépendance entre les assertions (Li et al., 2015), alors que d'autres utilisent des méthodes de vote complémentaires (Galland et al., 2010). Aucune de ces approches ne considère la connaissance du domaine au cours du processus de détection.

Des approches étendues ont donc été proposées, qui prennent en compte des dépendances possibles entre les assertions exprimées. La plupart de ces approches analysent des dépendances statiques (Blanco, Crescenzi, Merialdo, & Papotti, 2010; Dong, Berti-Equille, Hu, & Srivastava, 2010 ; Dong et al., 2009a ; Pochampally et al., 2014 ; Qi et al., 2013 ; Wang et al., 2015) et une approche est proposée pour prendre en compte la dépendance temporelle (Dong et al., 2009b). Dans cette dernière, les changements de dépendance au cours du temps sont considérés (mises à jour synchrones entre différentes sources). Toutes ces méthodes se basent sur la même intuition que les sources qui partagent les mêmes valeurs fausses sont supposées être interdépendantes. Par exemple, la recopie d'une source à partir d'une autre est estimée (nombreuses redites entre deux sites, par exemple). 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, des similarités entre descriptions, une connaissance antérieure, des techniques de raisonnements, ou encore de l'extraction d'information. TruthFinder (Yin et al., 2008) est une approche qui exploite les dépendances entre les valeurs et qui ajuste son calcul de confiance portant sur une assertion en s'appuyant sur une mesure de similarité. Cette mesure de similarité est par exemple estimée entre des valeurs numériques ou bien textuelles. Dans une autre contribution, le modèle proposé exploite les dépendances entre les descriptions (Meng et al., 2015 ; D. Wang et al., 2015 ; S. Wang et al., 2015). Dans ce cas, les dépendances proviennent des colocations physiques des entités considérées. Dans (Zhao et al., 2012) la distribution des qualités des sources est prise en compte. 3-Estimates introduit la notion de solidité des assertions, c'est-à-dire intègre dans le calcul de fiabilité d'une source, la propension d'une assertion à être associée à une valeur fausse (Galland et al., 2010). Dans (Pasternack et Roth, 2011) des informations complémentaires sont prises en compte, à l'exemple de l'exactitude des extracteurs, de la similarité entre assertions ou encore de l'appartenance à certains groupes d'assertions. Cette dernière est également utilisée dans (Gupta et

al., 2011). L'idée principale consiste à considérer la fiabilité des sources uniquement pour les objets appartenant à un sous-ensemble de sources considérées comme fiables. Enfin, dans (Dong et al., 2015) l'erreur commise par les extracteurs automatiques est prise en compte.

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 (Zhao et al., 2012 ; X. Wang et al., 2015 ; Pochampally et al., 2014). 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). Dans ce cas-là, la confiance est modélisée avec des contraintes. Il est à noter que ces approches sont complètement différentes du contexte d'étude fixé dans la section suivante. En effet, nous considérons dans cet article 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. Deux formes de connaissance a priori sont considérées séparément : tout d'abord l'ordre partiel sur les valeurs est introduit dans la section 2.1 et ensuite les notions de fréquence et de confiance sous-jacente aux règles d'association sont introduites dans la section 2.2.

2.1. Ordre partiel des valeurs et recherche 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, Malaga>
- <Pablo Picasso, bornIn, Europe>
- <Pablo Picasso, bornIn, Granada>

En effet, Granada et Malaga 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 Malaga 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. Cette considération implique des modifications importantes dans la formulation du problème tant au niveau des assomptions considérées qu'au niveau des solutions proposées pour résoudre le problème.

La dépendance entre les différentes valeurs est exprimée a priori dans une ontologie, sous la forme d'un ordre partiel $O = (\preccurlyeq, V)$ 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 $y \preccurlyeq x$ signifie que y implique x. Par exemple $Espagne \preccurlyeq Europe$ signifie que dire que quelqu'un est né en $Espagne$ implique de dire que cette personne est née en $Europe$. Dans un contexte d'ontologies s'adossant à la famille des logiques de description (Mann 2003), 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 (Hitzler et al., 2009) qui s'appuient sur les logiques de description.

La première approche que nous proposons exploite une portion réduite de l'ontologie, essentiellement constituée des définitions des classes[7] contenues dans la T-Box. Plus précisément, nous nous focalisons sur les ordres partiels des ressources formés par la structuration des classes (e.g., `subClassOf`), le typage des ressources (e.g., `type`) et d'éventuels liens entre les ressources exprimés par des prédicats transitifs supplémentaires (e.g., `partOf`). Nous ne discuterons pas les notions supplémentaires relatives à la sémantique associée aux relations pouvant exister entre les différentes valeurs. 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. Il est à noter que l'ontologie de domaine contient également d'autres types d'information qui peuvent être considérés. Le principal type est sans doute 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 – cf. chapitre 3.3 dans (Sébastien Harispe et al., 2015). 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 $x \preccurlyeq y$, alors $IC(x) \geq$

7. Le terme classe est utilisé ici plutôt que concept, en conformité avec le vocabulaire défini dans RDFS/OWL, langages standard dans le domaine du Web sémantique.

$IC(y)$, ($IC(root) = 0$). 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, 'Malaga' s'avère plus informative (car plus précise, plus spécifique) que 'Europe'. Ainsi, prendre 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.

2.2 Bases de connaissances, règles d'association et recherche 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. Par exemple, le fait qu'une personne soit née en Espagne est fréquemment associé au fait que cette même personne parle espagnol. Ainsi, si l'on recherche le lieu de naissance de Pablo Picasso sachant qu'il parle espagnol, lors de la recherche de vérité le système pourrait renforcer la confiance dans les assertions qui proposent une valeur correspondant à l'Espagne ou à 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 et 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 référer à (Galárraga et al., 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 et 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 :

$$r: B_1 \wedge B_2 \wedge ... \wedge B_n \Rightarrow H \text{ qui est équivalent à } r: \vec{B} \Rightarrow H.$$

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. Pour simplifier l'écriture, une assertion constituée d'un triplet <entité, prédicat,

valeur> sera notée : $prédicat(entité, valeur)$. L'identification des règles par l'analyse de la base de connaissances est réalisée avec AMIE+ (Galárraga et al., 2015). Ces règles ont notamment deux propriétés : i) elles sont connectées (chaque atome est transitivement connecté avec les autres atomes), ii) elles sont fermées, ce qui signifie qu'elles contiennent des variables fermées qui apparaissent au moins deux fois dans la règle.

Plusieurs métriques ont été proposées pour évaluer la qualité d'une règle, dont les plus répandues sont le support et la confiance (Feno, 2007). Notons que la confiance discutée dans cette sous-section concerne les règles d'association et doit être distinguée de la confiance dans les assertions qui a été définie plus haut.

Le support indique la proportion d'entités vérifiant à la fois le corps et la tête de la règle. Dans notre contexte de raisonnement en monde ouvert, nous utilisons la définition de (Galárraga et al., 2015). Pour la règle $r: \vec{B} \Rightarrow H$ où H est composé d'une seule assertion $p(e, v)$, le support est calculé de la façon suivante :

$$supp(r) := supp\left(\vec{B} \Rightarrow p(e, v)\right) := \#(e, v): \exists x_1, \ldots, x_i: \vec{B} \wedge p(e, v) \qquad (1)$$

où $x_1, \ldots, x_i$ représentent les variables contenues dans les atomes de $\vec{B}$ autres que e et v, et $\#(e, v)$ le nombre d'observations du couple (e, v). Ainsi, si l'on considère l'exemple présenté dans le tableau 2, le support de la règle $r: livesIn(e, v) \Rightarrow bornIn(e, v)$ serait égal à 1, étant donné la présence dans la base de connaissances de la paire d'assertions $livesIn(Adam, Paris)$ et $bornIn(Adam, Paris)$ qui la vérifie. Le support peut donc être vu comme la fréquence d'apparition de r dans la base de connaissances KB.

Tableau 2. Extraits de faits de la base de connaissances. Les prédicats correspondant aux en-têtes de colonne et chaque cellule contient un couple (entité, valeur) pour ce prédicat

livesIn	bornIn
(Adam, Paris)	(Adam, Paris)
(Adam, Caen)	(Luca, Rome)
(Luca, Milan)	(Carl, London)
(Bob, Lugano)	

La confiance, quant à elle, indique la proportion d'entités vérifiant la tête, parmi celles qui vérifient le corps. Cette mesure, comprise entre 0 et 1 n'est pas sensible à la taille des données. On peut la calculer de la façon suivante :

$$conf\left(\vec{B} \Rightarrow p(e, v)\right) := \frac{supp\left(\vec{B} \Rightarrow p(e,v)\right)}{supp(\vec{B})} := \frac{\#(e,v): \exists x_1, \ldots, x_i: \vec{B} \wedge p(e,v)}{\#(e,v): \exists x_1, \ldots, x_i: \vec{B}} \qquad (2)$$

Dans notre exemple, $conf\big(livesIn(e,v) \Rightarrow bornIn(e,v)\big) = \frac{1}{4}$. Cette valeur peut être vue comme une estimation de la probabilité de $p(e,v)$ si $\vec{B}$. 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 (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 $d = (entité, prédicat)$, 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. La mesure de PCA_confidence se calcule comme suit :

$$conf_{PCA}\left(\vec{B} \Rightarrow p(e,v)\right) := \frac{supp\left(\vec{B} \Rightarrow p(e,v)\right)}{\#(e,v):\exists x_1,\ldots,x_i,y:\vec{B} \wedge p(e,y)} \quad (3)$$

Dans notre exemple, $conf_{PCA}\big(livesIn(e,v) \Rightarrow bornIn(e,v)\big) = \frac{1}{3}$, puisqu'on rencontre une fois la règle ($livesIn(Adam, Paris)$ et $bornIn(Adam, Paris)$) et trois fois une variable impliquée dans la prémisse de cette règle ($livesIn(Adam, Paris)$, $livesIn(Adam, Caen)$ et $livesIn(Luca, Milan)$).

Dans ce qui suit, nous souhaitons utiliser un coefficient propulseur (booster), calculé à partir de l'identification de règles, qui représentent 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.

Prenons l'exemple très simplifié représenté dans la figure 1. 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. Ce renforcement est apporté par le coefficient propulseur qui 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 considèrerons 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. Le détail concernant ce coefficient booster sera donné dans la section 3.3.

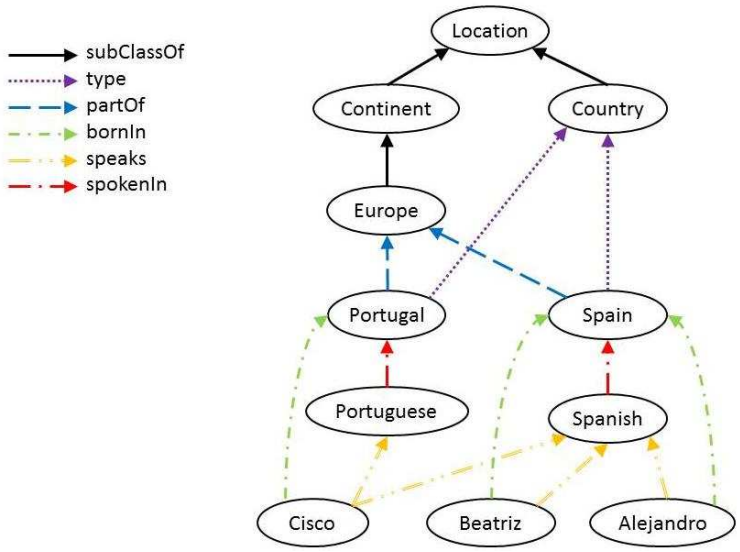

Figure 1. Extrait d'une base de connaissances

3. Formalisation du problème et description de l'approche proposée

Dans un premier temps nous allons reformuler la problématique de façon à ce qu'elle prenne en compte la définition d'une connaissance du domaine ; puis nous détaillerons l'approche que nous avons adoptée pour rechercher la vérité parmi un ensemble d'assertions.

3.1. Reformulation de la problématique

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 suivante :

$$tf : V \rightarrow \{true, false\} \quad (4)$$

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.

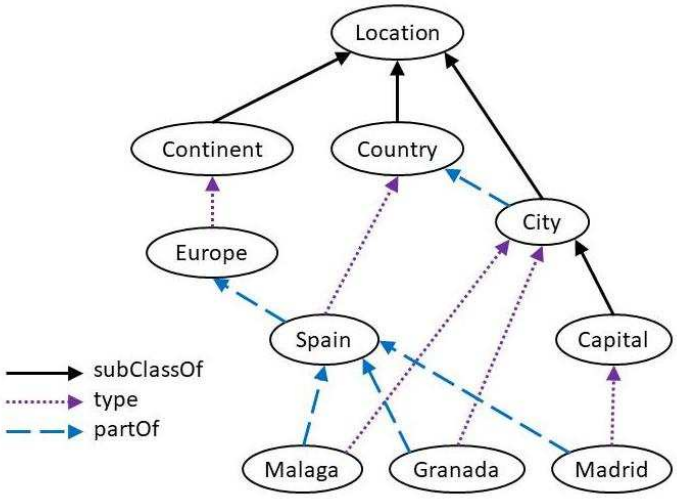

Figure 2. Exemple d'un ordre partiel entre certaines valeurs, qui inclut les relations de spécialisation subClassOf et type, et la relation de composition partOf

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 $\forall v'_d \in d, \{v'_d | v_d \preccurlyeq v'_d\}$, $v_d \Rightarrow v'_d$. 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. Si l'on reprend l'exemple décrit en section 2, les deux assertions <Pablo Picasso, bornIn, Granada> et <Pablo Picasso, bornIn, Malaga> 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 ; formellement $\forall v_d, v'_d \in V_d$, on a $(v_d \preccurlyeq v'_d \wedge tf(v_d)) \Rightarrow tf(v'_d)$. Cette définition signifie qu'un ensemble de valeurs peuvent être considérées comme vraies pour une description $d \in D$ particulière – on note V_d^* l'ensemble des valeurs vraies associées à la description d.

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.

Si l'on observe les contraintes qui définissent l'espace (c'est-à-dire l'ensemble) des valeurs vraies, différentes propriétés générales de $V^*_{d\in D}$ peuvent être exprimées. Ces propriétés sont fondamentales et vont être à la base de la définition de la sémantique du modèle proposé par la suite. Comme dans les approches classiques nous considérons que les assertions fournies permettent à elles seules de dériver, non plus la valeur vraie, mais l'ensemble des valeurs vraies associées à une description. Ainsi, sans connaissance supplémentaire nous considérons que l'ensemble des valeurs vraies associées à une description est inclus dans l'ensemble des valeurs induites par les valeurs de V proposées dans les assertions V_d:

$$V^*_d \subseteq \bigcup_{x\in V_d}\{y|x \preccurlyeq y\} \quad (5)$$

Nous allons cependant toujours considérer que dans l'absolu, et en accord avec la notion de prédicat fonctionnel, une valeur unique permet à elle seule de dériver l'ensemble des valeurs vraies associées à une description :

$$\forall d \in D, \exists x \in V^*_d \text{ tel que } V^*_d = \{y|x \preccurlyeq y\} \quad (6)$$

Cela implique que l'ensemble des valeurs vraies possibles V_d peut contenir des paires de valeurs qui sont incompatibles ou en conflit, c'est-à-dire des paires de valeurs qui ne peuvent pas apparaître toutes les deux dans l'ensemble de valeurs vraies associé à une description. Formellement les valeurs qui sont incompatibles sont représentées par les paires $(x, y) \in V^2$ pour lesquelles $\nexists z \in V$ tel que $\neg(x \preccurlyeq y \vee y \preccurlyeq x) \wedge (z \preccurlyeq x \wedge z \preccurlyeq y)$ – dans la figure 2 Spain et Capital ne sont pas en conflit, Malaga et Granada le sont : ces valeurs ne sont pas ordonnées et il n'existe pas de valeur qui les spécialise toutes les deux. En s'accordant sur ce point, nous considérons de la connaissance non explicitée par l'ordre partiel, par exemple ici, Malaga et Granada font référence à des localisations distinctes – sans territoire partagé. Cependant, et parce que la plupart des techniques de représentation des connaissances basées sur les logiques descriptives considèrent l'assomption d'un monde ouvert, deux valeurs ne peuvent être définies comme entrant en conflit que si elles sont explicitement précisées comme disjointes et si l'on sait que les deux valeurs font en effet référence à deux entités distinctes (assomption du nom unique).

Ainsi pour une description $d \in D$, en fonction des remarques précédentes et d'une valeur vraie $v \in V^*_d$, avec V^*_d inconnu, nous pouvons tout de même inférer de la connaissance sur V^*_d en excluant toutes les valeurs de V_d qui sont en conflit avec v. Néanmoins, sans connaissance supplémentaire sur V^*_d, il est impossible de s'exprimer sur l'ensemble des valeurs qui spécialisent v – dans ce contexte ces valeurs sont considérées comme étant en conflit potentiel. Cette relation entre valeurs n'est pas symétrique : Granada est en conflit potentiel avec Spain, alors que Spain est en accord avec Granada. Dire que quelqu'un est né à Grenade implique de dire qu'il est né en Espagne, alors que le contraire n'est naturellement pas vrai.

De façon plus générale, identifier l'ensemble des valeurs vraies pour une description donnée $d \in D$ revient à identifier l'ensemble $V_d^* \subseteq V_d$ respectant les contraintes (5) et (6) qui maximisent la confiance au regard de la confiance associée aux assertions de V_d qui contiennent les valeurs de V_d^*.

Adopter une telle approche nécessite la définition d'une fonction objectif permettant de calculer la fiabilité associée à une source ; cette fonction étant naturellement définie en tenant compte de l'appréciation de la confiance associée à chaque assertion. Cela nécessite donc de considérer des contraintes ou de la connaissance supplémentaires par rapport aux solutions souhaitées (optimisation de deux critères dépendants). Les approches itératives sont particulièrement adaptées pour amener à la résolution de ce genre de problème. Comme nous l'avons vu lors la définition de la notion de vérité (Equation 4) et lors de la définition des contraintes définies par l'ordre partiel de valeurs, des propriétés intéressantes sur l'ensemble des valeurs vraies peuvent être dérivées pour chaque description. Plus généralement, ces propriétés précisent comment l'information amenée par l'observation des assertions doit être propagée dans l'objectif de distinguer les ensembles de valeurs vraies associées aux descriptions ainsi que la confiance à associer aux sources. Comme nous allons le voir dans la prochaine section, de façon intéressante, la définition de l'espace des valeurs vraies que nous proposons répond au cadre défini par les fonctions de croyance qui sont classiquement utilisées pour traiter des données incertaines et imprécises.

3.2. Utilisation de l'ordre partiel pour la détection de vérité

La modélisation de la solution proposée repose sur les fonctions de croyance introduites dans (Shafer, 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, dans notre cas, peut être vue comme une fonction $m_d: V \rightarrow [0,1]$ qui dépend d'une description $d \in D$ considérée. Cette fonction représente la portion de preuve allouée à une valeur particulière (et non pas plus spécifique). Elle peut être utilisée pour définir la croyance (belief en anglais) qui peut être associée à une valeur donnée.

$$Bel_d(v) = \sum_{v' \preccurlyeq v} m_d(v') \quad (7)$$

Cette formule permet de sommer l'information apportée par l'observation d'une valeur ; elle est ainsi, en totale adéquation avec la définition de l'ensemble des valeurs vraies défini plus haut. 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. La contrainte de place nous empêche de détailler certains aspects techniques de l'approche adoptée et du lien établi avec les fonctions de croyance, mais le lecteur pourra se référer à (Harispe et al., 2015) pour les détails relatifs à l'utilisation de ces fonctions en considération d'un ordre partiel.

À titre illustratif, nous proposons d'adapter le modèle de découverte de vérité Sums, défini dans (Pasternack et Roth, 2010), en y intégrant la nouvelle formulation 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. Les formules utilisées dans la définition originale sont les suivantes :

$$\mathrm{t^i(s)} = \sum_{\mathrm{v_d} \in \mathrm{V^s}} \mathrm{c^{i-1}(v_d)} \tag{8}$$

$$c^i(v_d) = \sum_{\mathrm{s} \in S^{v_d}} t^i(\mathrm{s}) \tag{9}$$

avec t^i l'estimation de la fiabilité associée à une source et c^i la confiance associée à une assertion respectivement à l'itération i. Il faut noter que l'approche itérative requiert une phase d'initialisation pour une des quantités à estimer. Dans nos expérimentations, nous avons choisi d'attribuer une même confiance à toutes les assertions. La fiabilité associée à une source $s \in S$ est ensuite évaluée en sommant les confiances sur les assertions qui lui sont associées. De façon similaire, la confiance associée à une assertion, $c^i(v_d)$, est évaluée en sommant les fiabilités des sources qui expriment cette assertion. A chaque itération une étape de normalisation est appliquée : $t^i(\mathrm{s})$ et $c^i(v_d)$ sont divisés par $max_{s \in S}(t^i(s))$ et $max_{v_d \in V}(c^i(v_d))$ respectivement.

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 S^{v_d} par S^{v_d+}. Nous obtenons donc $c^i(v_d) = \sum_{s \in S^{v_d+}} t^i(\mathrm{s})$ avec S^{v_d+} défini comme l'ensemble des sources qui proclament une assertion donnée et des sources qui proclament des assertions plus spécifiques, c'est-à-dire qui supportent v_d. Autrement dit, $S^{v_d+} = S^{v_d} \cup \{s \in S^{v_d{}'} : v_d{}' \in V \wedge v_d{}' \preccurlyeq v_d\}$. Notez tout de même que la façon de calculer la fiabilité d'une source ne tient pas compte de l'ordre exprimé sur les assertions, pour ne pas intégrer à deux reprises la même information.

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. Pour ce faire, la section 3.4 présentera une stratégie qui permet de sélectionner une valeur vraie pour chaque description qui tient compte de l'ordre partiel sur les valeurs.

3.3. Utilisation de règles pour la détection de vérité

Ici, les règles d'association sont identifiées grâce à AMIE+ (Galárraga et al., 2015) ainsi que les mesures de $support$ et de $PCA - confidence$. Disposant de ces informations, l'approche proposée consiste à adapter les méthodes existantes en

intégrant un coefficient propulseur ($booster$), que nous appelons $boost(d)$, dans la procédure itérative de détection de vérité. Ce facteur a une influence directe sur le calcul de confiance dans une assertion d et donc une influence indirecte sur le calcul de fiabilité des sources.

En utilisant les notations introduites dans la section 2, une adaptation de la méthode Sums (Pasternack & Roth 2010), peut être modélisée comme suit pour tenir compte de l'information amenée par les règles :

$$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{v_d' \in V^{s'}} c^{i-1}(v_d')\right)} \sum_{v_d \in V^s} c^{i-1}(v_d) \quad (10)$$

$$c^i(v_d) = \frac{1}{norm_{v_d}}\left((1-\gamma) confidence_{basic}(v_d) + \gamma . boost(v_d)\right) \quad (11)$$

avec $\gamma \in [0,1]$ une poids qui représente l'influence relative accordée aux sources et à la base de connaissances ; $confidence_{basic}$ une fonction de V dans $[0,1]$ qui représente la confiance donnée par les sources à une assertion ; et $boost$ une fonction de V dans $[0,1]$ qui représente la confiance dans une assertion provenant de l'application des règles identifiées grâce à l'analyse de la base de connaissances.

Le paramètre γ dépend du contexte et sera fixé en fonction de la stratégie choisie. Dans l'évaluation qui sera présentée dans la section 5, plusieurs valeurs seront considérées et leur impact sera discuté.

Le calcul de $confidence_{basic}(d)$ est réalisé comme dans la méthode Sums :

$$confidence_{basic}(v_d) = \frac{\sum_{s \in S^{v_d}} t^i(s)}{max_{v_d' \in V} \sum_{s' \in S^{v_d'}} t^i(s')} \quad (12)$$

où l'on retrouve au numérateur, la somme des fiabilités associées à toutes les sources qui émettent une assertion et au dénominateur, un facteur de normalisation c'est-à-dire la confiance maximale associée à une assertion.

Le facteur propulseur, booster, cherche à synthétiser les informations données par toutes les règles obtenues pour chaque assertion. Par exemple pour $bornIn$, à partir du graphe de la Figure 1, on peut obtenir les règles suivantes :

- $speaks(x,z) \wedge officialLanguage(y,z) \Rightarrow bornIn(x,y)$
- $speaks(x, Spanish) \wedge officialLanguage(Spain, Spanish) \Rightarrow bornIn(x, Spain)$.

Comme nous l'avons montré dans la section 2.2, chaque règle peut être évaluée par différentes métriques.

Le support et la confiance représentent deux caractéristiques d'une même règle. Dans notre cas il est important de considérer ces deux mesures. En effet, dans certains cas, une règle r pourra avoir une mesure de $conf_{PCA}(r) = 1$ et une mesure de support $supp(r) = 2$ alors qu'une autre règle r' pourra avoir également une

mesure $conf_{PCA}(r') = 1$ mais un support $supp(r') = 100$. Dans ce cas-là, on préfèrera se baser sur la règle r' qui a été observée un plus grand nombre de fois que la règle r. Dans le même ordre d'idée, choisir une règle uniquement parce qu'elle a une mesure de confiance bien supérieure aux autres règles n'a de sens que si elle a été observée un grand nombre de fois. Nous avons donc choisi une fonction d'agrégation qui permet de considérer simultanément ces mesures dans un même indicateur. Nous nous sommes notamment basés sur le modèle d'agrégation proposé dans (Jean et al., 2016). Ce modèle a été adapté à notre contexte et résulte dans la formulation suivante. Soit une règle r, son support $supp(r)$ et sa confiance $conf_{PCA}(r)$, le score obtenu par agrégation de ces différentes caractéristiques est donné par :

$$\text{score(r)} = \left(1 - \frac{1}{\text{supp(r)}}\right) \text{conf}_{\text{PCA}}(\text{r}) \tag{13}$$

Ainsi en pondérant la mesure de confiance dans une règle par les occurrences de cette règle, on accorde plus de confiance dans les règles qui sont les plus fréquentes (avec un support élevé).

Les scores des différentes règles qui concernent une même assertion peuvent ainsi être agrégés. En effet, notre objectif reste bien de renforcer la confiance dans certaines assertions en utilisant tous les motifs identifiés.

Soit une assertion $v_d = p(d, o)$, une base de connaissances KB et un ensemble de règles $R = \{r: B_1 \wedge \ldots \wedge B_n \Rightarrow p'(x, y)\}$ extraites à partir de KB, nous considérons que le facteur propulseur doit être fonction du pourcentage de règles qui sont vérifiées par l'assertion considérée. Pour chaque assertion, l'ensemble des règles R_{v_d} à considérer (règles éligibles) est un sous-ensemble des règles extraites : $R_{v_d} \subseteq R$. Ces règles doivent répondre à certaines contraintes : contenir le prédicat p dans la tête de la règle et avoir un corps composé uniquement d'atomes valides (c'est-à-dire contenus dans KB). Formellement $R_{v_d} = \{r: B_1 \wedge \ldots \wedge B_n \Rightarrow p'(x, y) \in R | (p' = p) \wedge T(B_1 \wedge \ldots \wedge B_n) = 1\}$ avec $T(B) = 1$ (resp. $= 0$) une fonction qui indique que le corps de la règle est vérifié (resp. n'est pas vérifié). Le facteur propulseur peut alors être défini comme suit.

$$\text{boost(d)} = \left(1 - \frac{1}{1 + \sum_{\text{r} \in \text{R}_{\text{v}_\text{d}}} \text{score(r)}}\right) \frac{\sum_{\text{r} \in \text{R}^{\text{T}}_{\text{v}_\text{d}}} \text{score(r)}}{\sum_{\text{r} \in \text{R}_{\text{v}_\text{d}}} \text{score(r)}} \tag{14}$$

où $\text{R}^{\text{T}}_{v_d} = \{r: B_1 \wedge \ldots \wedge B_n \Rightarrow p'(x, y) \in R_{v_d}: T(r) = 1\}$ et où $T(r) = 1$ (resp. $= 0$) représente le fait que la règle r soit vérifiée (resp. fausse). Autrement dit, l'ensemble des règles éligibles est composé des règles qui sont vérifiées au moins par une instanciation de leur corps, c'est-à-dire contenant le même sujet pour le prédicat considéré. Ce facteur propulseur est compris entre 0 et 1.

Prenons l'exemple du tableau 3. Si l'on considère les deux règles suivantes :

- $r^1: speaks(x, z) \wedge officialLanguage\ (y, z) \Rightarrow bornIn(x, y)$
- $r^2: resident(x, France) \Rightarrow bornIn(x, France)$

où le score de r^1 est de 0.55 et celui de r^2 est de 0.75. Le coefficient propulseur pour l'assertion proposée par la source s_1 est de 0.245 car les deux règles sont éligibles, mais seule r^1 est vérifiée ce qui donne : (1-1/(1+0.55+0.75)) * (0.55/(0.55+0.75)). Par contre, pour l'assertion proposée par la source s_2 le score sera de 0 car même si les deux règles sont éligibles, aucune des deux n'est vérifiée.

Tableau 3. Éléments d'une base de connaissances

Sources/origine	Assertions
s_1	$bornIn(Picasso, Spain)$
s_2	$bornIn(Picasso, UK)$
KB	$officialLanguage(Spain, Spanish)$
KB	$speaks(Picasso, Spanish)$
KB	$resident(Picasso, France)$

3.4. Sélection de valeurs vraies

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 de facto être associées à un fort degré de confiance. Conformément à l'ordre défini sur les valeurs, la confiance associée à ces valeurs génériques sera renforcée par la confiance attribuée aux valeurs qu'elles subsument. Dans cette hypothèse, 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, Malaga> 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 forcément d'un grand intérêt.

Pour pallier ce problème, 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: θ et δ.

Le paramètre θ permet de spécifier un seuil de confiance minimal en deçà duquel la valeur ne sera pas considérée comme candidate possible à la valeur vraie. Il est important de noter que le score de confiance qui doit être comparé à θ doit être au préalable normalisé en fonction de chaque description, c'est-à-dire le score maximal de confiance associé à chaque description doit être égal à 1. Cette normalisation permet d'éviter les seuils inconsistants en fonction des descriptions.

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.

La prise en compte de ces différents paramètres induit des comportements différents lors de la sélection et, par conséquent, peut conduire à différents ensembles de solution. Par exemple, si l'on positionne $\theta = 0$ et $\delta = 0$ on reproduit le cas d'un algorithme glouton basique, qui sélectionne à chaque itération les valeurs qui ont la confiance maximale supérieure à θ, sans contrôle supplémentaire. D'autre part, si on fixe $\theta = 0$ et $\delta = 1$, l'ensemble des valeurs ayant une confiance supérieure à θ est retourné comme résultat. Si cette configuration peut sembler inutile, elle permet toutefois d'obtenir un ensemble de valeurs possibles à la fin de la procédure de post traitement. Cet ensemble est composé d'alternatives intéressantes qui sont tout à la fois spécifiques (avec un fort IC), mais sémantiquement différentes. Nous augmentons ainsi la probabilité de trouver une valeur vraie en augmentant le nombre de concepts différents considérés. Cette stratégie permet, entre autres, de traiter les cas empreints d'une forte incertitude. L'idée consiste alors à retourner comme résultat toutes les valeurs et leurs ancêtres, et ensuite utiliser la phase d'ordonnancement pour positionner en premières places les alternatives les plus prometteuses en fonction des propriétés attendues.

C'est pourquoi, étant donné un ensemble de valeurs vraies retourné par la procédure de sélection, une méthode d'ordonnancement doit être spécifiée, afin d'identifier la valeur vraie 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, lorsque $\delta = 0$. Cette approche est appelée 'Sélection des valeurs vraies par les meilleurs enfants' ou SVmE_{IC}. Un autre mode de sélection consiste à ordonner les valeurs en utilisant la moyenne des fiabilités de leurs sources. Cette approche est appelée 'Sélection des valeurs vraies en retournant tous les enfants' ou SVtE_{trust}. L'hypothèse retenue dans cette méthode d'ordonnancement est la suivante : si de très nombreuses sources non fiables soutiennent une valeur fausse A (augmentant ainsi son score de confiance – Sums ne procède pas à une normalisation basée sur le nombre de sources qui proposent une valeur ce qui peut fausser les résultats) et que peu de sources fiables soutiennent une valeur vraie B, alors les sources qui ont proposé B doivent avoir une meilleure moyenne de score de fiabilité.

4. Évaluation de la méthode

Notre objectif étant d'adapter des méthodes existantes afin de prendre en compte la connaissance du domaine et la relation d'ordre entre les valeurs, nous avons été amenés à créer un nouveau jeu de tests car aucun de ceux proposés dans la littérature ne faisait l'hypothèse de relations possibles entre les valeurs associées à des descriptions.

4.1. Constitution du jeu de test

En effet, l'un des jeux de données les plus populaires dans ce domaine, à savoir celui des Auteurs présenté dans (Dong et al., 2010) contient une liste d'auteurs pour un ensemble de livres. Il est clair qu'on ne peut pas, sans traitements particuliers, considérer de relation d'ordre partiel sur ces auteurs identifiés par leurs noms propres. La même constatation s'impose pour les jeux de données proposés par (Pasternack & Roth 2010) qui concerne pour l'un la population (la taille de chaque ville) et pour l'autre des données biographiques (date de naissance et de décès de personnes).[8]

Nous avons alors créé un jeu de test qui regroupe i) un ensemble de descriptions pour lesquelles ii) les valeurs vraies sont connues, ainsi que iii) un ensemble de sources et iv) un ensemble d'assertions associées à chaque source.

Nous avons pour cela collecté un ensemble de faits de DBpedia (Auer et al., 2007) considérés comme étant tous vrais (postulat). Nous nous sommes focalisés pour cette extraction sur le prédicat `dbpedia-owl:birthPlace` (version 2015-04) et nous avons choisi les faits pour lesquels il n'y avait pas de doublon. Nous avons

8. Nous précisons cependant qu'il est possible de considérer des ordres sur des valeurs qui ne sont pas de facto structurables, en générant pour cela des structurations artificielles à partir de sous-ensembles des valeurs proposées par les sources. Ce traitement dépasse le cadre étudié dans ce papier et ne sera donc pas discuté ici ; il souligne cependant que l'approche n'est pas nécessairement utilisable dans le seul contexte du traitement de données structurées a priori.

ensuite généré des sources avec un degré de fiabilité associé. Nous avons utilisé une distribution gaussienne avec une moyenne et un écart type respectivement de 0.6 et de 0.4 pour simuler autant que possible un scénario réaliste. En effet, cette configuration permet d'affecter à la majorité des sources un score de fiabilité moyen et à quelques sources seulement un score de fiabilité soit très élevé soit très faible. Nous avons ensuite respecté les règles suivantes :

– Une source ne propose pas des assertions pour l'ensemble des descriptions. C'est le cas dans la réalité, où par exemple la majorité des sites web se concentrent sur quelques sujets spécifiques alors que très peu couvrent des sujets très vastes (Wikipedia, par exemple). Ainsi chaque source a une couverture partielle d'un domaine et de nombreuses sources vont donc proposer des assertions sur quelques descriptions et peu de sources vont couvrir un vaste ensemble d'assertions.

– Une source propose une assertion vraie en fonction de son degré de fiabilité et peut choisir comme valeur, un ancêtre de la valeur identifiée comme étant vraie (nous utilisons pour cela une mesure de similarité, grâce à la SML[9], afin de nous limiter dans la liste des ancêtres). Trois types de jeux de données[10] ont été définis : EXP, LOW_E et UNI qui diffèrent par la stratégie de sélection des valeurs vraies qui seront associées aux sources (cf. figure 3) ; étant donné une description $d \in D$, la valeur vraie de base v_d^* est utilisée pour dériver l'ensemble V_d^* qui rassemble toutes les valeurs vraies qui correspondent aux différentes stratégies correspondant aux jeux de données cités. Une mesure de similarité sémantique est utilisée pour déterminer pour chaque valeur de cet ensemble sa proximité avec la valeur v_d^*. Cette mesure est utilisée pour ordonner les valeurs de V_d^*. Étant donné cet ordre sur les valeurs, dans le cas du jeu de données EXP, les sources tendent à proposer des valeurs très proches de la valeur vraie de base. Ainsi on autorisera peu de sources à proposer des valeurs très génériques pour une description particulière. Pour la génération du jeu de données UNI, les valeurs proposées par les sources sont sélectionnées indépendamment de leur similarité avec la valeur vraie de base. Dans le cas de LOW E, la majorité des valeurs sont sélectionnées comme dans le jeu UNI (c'est-à-dire avec une même probabilité d'être choisies), mais dans quelques cas, on favorise des valeurs qui sont très proches de la valeur vraie. Ainsi, par exemple, étant donné $v_d^* = Malaga$, dans le cas de EXP, beaucoup de sources vont proposer des valeurs telles que $Malaga, Spain$ et peu de sources vont proposer $Europe$ ou $Place$, tandis que dans le cas de UNI le nombre de sources proposant ces valeurs sera approximativement le même.

– Une source propose une assertion fausse, en fonction de son degré de non-fiabilité ($1 - degré\ de\ fiabilité$) et choisit à cet effet des valeurs qui n'ont aucun lien de généralisation/spécialisation avec la valeur vraie donnée (pour ce faire une mesure de similarité est également utilisée). En effet, les ancêtres d'une valeur vraie sont également vrais et leurs descendants constituent des valeurs vraies potentielles. Elles ne peuvent donc pas être considérées comme fausses a priori. Il est à noter que

9. Semantic Measure Library (Harispe et al., 2013)
10. 20 jeux de données pour chacun des trois types cités.

seule une loi exponentielle est utilisée pour choisir les valeurs fausses. L'idée est qu'en considérant une source peu fiable, il y ait une plus grande probabilité de sélectionner une valeur fausse qui est proche de la valeur attendue. C'est-à-dire, si la valeur attendue est Malaga, il sera plus probable que la source non fiable donne comme fausse valeur une autre ville d'Europe plutôt que Pékin (ville en Chine). Cette configuration permet de reproduire le comportement de sources malveillantes qui utilisent la copie de valeurs fausses pour disséminer de fausses informations. De plus, pour mieux simuler un contexte réel, une valeur fausse déjà proposée a une plus grande probabilité d'être sélectionnée que les autres (celles qui n'ont jamais été proposées).

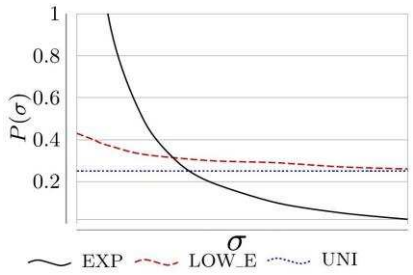

Figure 3. Illustration des distributions utilisées pour la sélection des valeurs en considérant en abscisse les scores de proximité sémantique et en ordonnée la *probabilité qu'une valeur associée à un score donné soit sélectionnée*

Basés sur ces règles, différents jeux de données ont été générés (20 ensembles de données pour chaque type considéré). À noter que chaque jeux de données est générés à partir de 1 000 sources et 10 000 descriptions). Pour chaque jeu de données, nous avons testé les approches en évaluant le pourcentage de valeurs vraies et de valeurs fausses proposées par chaque source – estimé par les approches, au regard du degré de fiabilité qui avait été associé au préalable à chaque source lors de la génération des données de test. Ainsi, nous considérons que plus une approche permet de distinguer les valeurs vraies et fausses proposées par une source indépendamment des paramètres utilisés pour générer ces valeurs, plus celle-ci est performante et robuste.

4.2. Cadre expérimental

Nous avons implémenté quatre modèles différents à partir de la méthode Sums. La méthode Sums dite traditionnelle (M_1) est celle qui est proposée par les auteurs de (Pasternack & Roth 2010). La méthode M_2 consiste à intégrer à Sums la prise en compte des règles identifiées (après analyse de la A-Box) lors du calcul de la confiance dans une assertion (comme décrit ci-dessus). La méthode M_3 est la méthode qui consiste à tenir compte uniquement des relations transitives (une partie

de la T-Box) en plus de la méthode Sums et enfin la méthode M_4 consiste à tenir compte à la fois des relations définies dans l'ontologie et des règles identifiées par l'analyse de la A-Box dans le calcul de la confiance associée aux assertions. Le tableau 4 synthétise ces différents modèles et les équations associées respectivement au calcul de la confiance dans les assertions et au calcul de la fiabilité des sources dans le processus itératif de recherche de vérité.

Tableau 4. Récapitulatif des différents modèles utilisés pour la recherche de vérité

M_1 – Sums traditionnel
$$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{v'_d \in V^{s'}} c^{i-1}(v'_d)\right)} \sum_{v_d \in V^s} c^{i-1}(v_d)$$
$$c^i(v_d) = confidence_{basic}(v_d) = \frac{1}{max_{v'_d \in D}\left(\sum_{s' \in S^{v'_d}} t^{i(s')}\right)} \sum_{s \in S^{v_d}} t^i(s)$$
M_2 – Sums traditionnel + prise en compte des règles d'association
$$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{v'_d \in V^{s'}} c^{i-1}(v'_d)\right)} \sum_{v_d \in V^s} c^{i-1}(v_d)$$
$$c^i(v_d) = \frac{1}{norm_{v_d}}[(1-\gamma) confidence_{basic}(v_d) + \gamma . boost(v_d)]$$
M_3 – Sums traditionnel + propagation en fonction des relations d'ordre
$$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{v'_d \in V^{s'}} c^{i-1}(v'_d)\right)} \sum_{v_d \in V^s} c^{i-1}(v_d)$$
$$c^i(v_d) = adaptedConfidence(v_d) = \frac{1}{max_{v'_d \in D}\left(\sum_{s' \in S^{v'_{d^+}}} t^i(s')\right)} \sum_{s \in S^{v_{d^+}}} t^i(s)$$ avec $S^{v_d+} = S^{v_d} \cup \{s \in S^{v'_d} : v'_d \in V, v'_d \preccurlyeq v_d\}$
M_4 – Sums traditionnel + propagation en fonction des relations d'ordre + Règles d'association
$$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{v'_d \in V^{s'}} c^{i-1}(v'_d)\right)} \sum_{v_d \in V^s} c^{i-1}(v_d)$$
$$c^i(v_d) = \frac{1}{norm_{v_d}}[(1-\gamma) adaptedConfidence(v_d) + \gamma . boostProp(v_d)]$$ avec $boostProp(v_d) = \left(1 - \frac{1}{\sum_{r \in R_{v_{d^+}}} score(r)}\right) \frac{\sum_{r \in R^T_{v_{d^+}}} score(r)}{\sum_{r \in R_{v_{d^+}}} score(r)}$ et $R_{v_d+} = R_{v_d} \cup \{R_{v'_d} \in R : v'_d \in V, v'_d \preccurlyeq v_d\}$

Il est à noter que $boostProp(v_d)$ est un coefficient calculé à partir du coefficient boost qui provient de l'application des règles d'association mais auquel on applique une propagation. En effet, soit une règle dont la tête est égale à $H = p(s, o)$, une telle règle se vérifie et donc confirme toutes les règles plus génériques au regard de l'ontologie de domaine (c'est-à-dire les règles qui impliquent des concepts plus génériques que ceux de la règle considérée). Autrement dit, le facteur de boost correspondant doit être propagé à tous les ancêtres. Ce facteur permet d'assurer la monotonie de la fonction de confiance associée aux assertions. En effet, la confiance $c(v_d)$ dans une assertion v_d telle que $v'_d \preccurlyeq v_d$ doit être supérieure ou égale à la confiance $c(v'_d)$ associée à l'assertion v'_d.

4.3. Méthodologie d'évaluation

Pour chaque expérimentation, la valeur initiale de la confiance a été fixée arbitrairement à 0,5. Le critère d'arrêt de l'itération est le même que dans (Pasternack et Roth, 2010) : le nombre maximal d'itérations est fixé (ici à 20). L'algorithme a été implémenté en Python et les tests ont été réalisés sur un PC Intel Core 2 Duo processor (2.93GHz/8.00GB). À noter que les différents jeux de données utilisés ainsi que le code de calcul sont rendus disponibles à la communauté à l'adresse https://github.com/lgi2p/TDwithRULES. Nous ne pouvons pas baser nos évaluations sur les mesures de rappel comme c'est souvent le cas dans la littérature. En effet, contrairement aux autres approches, nous sélectionnons un ensemble de valeurs comme pouvant être vraies. La probabilité que la valeur exacte appartienne à cet ensemble est donc de facto supérieure. Nous avons donc analysé la proportion de valeurs vraies retournées par les différentes méthodes et qui correspondent à des valeurs attendues - pour chaque description (entité, prédicat), la valeur attendue est contenue dans un corpus de référence.

5. Résultats

Sur chaque jeu de données, tous les modèles présentés dans la section précédente ont été appliqués. Tous les résultats synthétisés dans les figures et les tableaux ont été obtenus en calculant la moyenne sur les 20 jeux de test de chaque type.

Nous avons tout d'abord analysé l'impact de la prise en compte de l'ordre partiel sur les valeurs dans un modèle de détection de vérité (M_3) en fonction de modèles de détection de vérité existants, c'est-à-dire Sums ou M_1. Comme en témoignent les résultats présentés dans la figure 4, la prise en compte de cette connaissance améliore les performances de M_1 (Sums) en terme de rappel, lorsque le calcul de confiance est effectué avec les paramètres suivants pour la procédure de sélection : $\delta = 0$ ou 1 (respectivement SVmE $_{IC}$ et SVtE $_{trust}$) et θ plus petit que 0.2. Dans tous les cas, les meilleures performances sont obtenues avec SVtE $_{trust}$ quand $\theta = 0$. Dans le cas du jeu de test birthPlace, la prise en compte de l'ordre partiel dans une adaptation du modèle Sums conduit à une amélioration des performances par rapport aux approches existantes pour chaque type de jeu de données (Exp,

Low_E et Uni). Les meilleurs résultats sont obtenus avec Uni. En effet, lors de la constitution de ce jeu de données, les valeurs vraies proposées par les sources sont choisies indépendamment de leur similarité avec la valeur vraie. Ainsi, un plus grand nombre de valeurs différentes sont proposées pour la même description. Cela signifie que les sources sont plus en désaccord sur une description que dans les autres jeux de test. C'est pourquoi l'estimation de la confiance dans les valeurs est plus aisée que dans les autres jeux de test. Les résultats montrent clairement que le modèle M_3, en utilisant les connaissances a priori sur l'ordre de valeurs, peut compenser cette complexité supplémentaire. En effet, ce modèle peut utiliser l'ordre partiel parmi les valeurs pour croiser l'information associée aux observations distinctes.

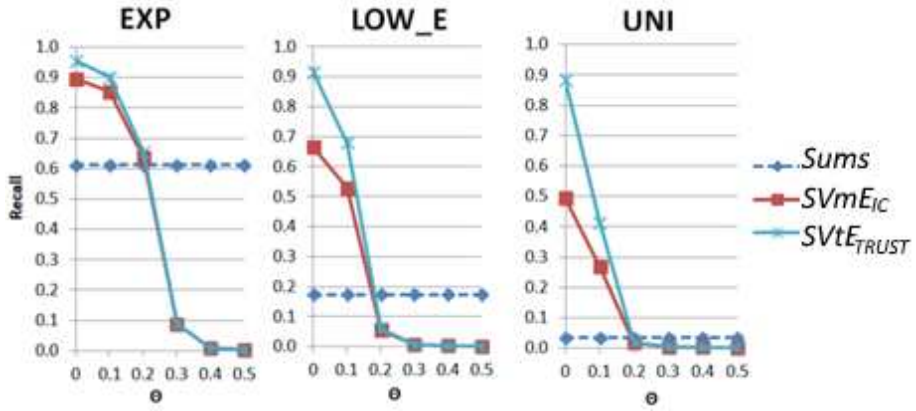

Figure 4. Synthèse des résultats. Ici le rappel (taux de valeurs attendues obtenues) est présenté en fonction du jeu de test et de la méthode de calcul : Sums (M_1) ou Sums+ ordre partiel(M_3) couplé avec deux différentes configurations de l'algorithme de sélection des valeurs vraies ($SVmE_{IC}$ et $SVtE_{trust}$) *et de* θ

Ainsi nous avons analysé les performances obtenues en introduisant de l'information fournie par une connaissance a priori sous la forme de règles. Dans nos expérimentations, nous avons sélectionné uniquement les règles détectées par AMIE+ qui ont une couverture supérieure à 0.012 pour la tête afin de limiter les erreurs qui pourraient être introduites par des règles peu significatives, i.e. supportées par trop peu d'exemples. Nous restreignons ainsi le nombre de règles considéré à 62. Dans chaque cas, l'identification de vérité est réalisée par un processus itératif où les calculs de confiance dans les assertions et de fiabilité des sources sont ceux présentés dans les tableaux 5 et 6 (M_2 et M_4).

Ici, nous avons analysé la proportion de valeurs vraies retournées par les différentes méthodes qui correspondent à des valeurs attendues (pour chaque description (sujet, prédicat), la valeur attendue est contenue dans un corpus de référence) – n_{vrai}, c'est-à-dire le rappel. Nous avons aussi analysé la proportion de

valeurs plus générales que celles attendues – n_{gen}, pour mieux comprendre les effets des règles. Et enfin le taux d'erreur est indiqué (valeurs proposées qui sont totalement différentes et décorrélées de la valeur attendue) – n_{faux}.

Étant donné les résultats obtenus précédemment, nous fixons $\theta = 0$, à la fois pour les approches SVmE $_{IC}$ et SVtE $_{trust}$ dans le cas du modèle M_4. Nous nous focalisons donc désormais sur l'analyse des comportements influencés par le paramètre γ.

En premier lieu, nous pouvons constater que l'introduction des règles d'association dans le contexte de SVtE $_{trust}$, c'est-à-dire quand δ = 1, n'améliore pas les résultats. Cela s'explique par le fait que les règles concernent souvent des concepts très génériques et cela n'impacte donc pas la confiance dans les valeurs plus spécifiques. Étant donné que la procédure de sélection, dans le cas où δ = 1, renvoie les valeurs différentes les plus spécifiques et avec une confiance supérieure à θ, l'information apportée par les règles n'a pas d'impact sur les résultats. Nous ne détaillerons donc pas ces résultats dans ce papier. Par contre, dans le cas où l'on a δ = 0 cette information additionnelle apportée par les règles est d'intérêt. En effet, elle permet d'augmenter la probabilité de choisir le 'bon' chemin, lors du parcours du graphe des valeurs candidates lors de la sélection de la valeur sélectionnée par l'approche automatisée. On note aussi que le rappel augmente significativement.

Tableau 5. Synthèse des résultats obtenus avec les modèles de détection de vérité M_2 appliqués aux trois types de corpus. Différentes valeurs de γ ont été testées. Les valeurs en rouge/souligné indiquent les plus mauvais résultats et les résultats en gras les meilleurs

Sums + Règles		γ					
		0*	*0.25*	*0.50*	*0.75*	*0.9*	*1.0*
Jeu de données EXP	n_{vrai}	*0,6267*	**0,6278**	0,6036	0,5811	0,5944	0,1578
	n_{gen}	*0,1136*	0,1433	0,1919	0,2404	0,2567	**0,3205**
	n_{faux}	*0,2596*	0,2289	0,2045	0,1785	**0,1489**	0,5217
Jeu de données LOW_E	n_{vrai}	*0,1726*	0,2085	0,2200	0,2310	**0,2507**	0,1279
	n_{gen}	*0,1896*	0,2168	0,2663	0,3125	0,3439	**0,4751**
	n_{faux}	*0,6378*	0,5746	0,5137	0,4565	0,4054	**0,3970**
Jeu de données UNI	n_{vrai}	*0,0335*	0,0737	0,0983	0,1167	0,1274	**0,1293**
	n_{gen}	*0,1953*	0,2316	0,2814	0,3284	0,3695	**0,4388**
	n_{faux}	*0,7712*	0,6947	0,6203	0,5549	0,5031	**0,4318**

L'ensemble des résultats pour les modèles M_2 et M_4 sont présentés dans les tableaux 5 et 6. Notons que le modèle M_1 correspondant à la méthode Sums est équivalent au modèle M_2 pour lequel la valeur de γ est égale à zéro (première ligne

de M_2). De même, le modèle M_3 est équivalent au modèle M_4 pour lequel $\gamma = 0$. En effet, dans ce cas les règles ne sont pas prises en compte dans le calcul. Nous avons déjà montré que la prise en compte de la propagation d'information en fonction de l'ontologie du domaine (M_3) apportait une plus-value par rapport à l'approche classique (M_1) (Beretta et al., 2016). Les tableaux 5 et 6 confirment que la prise en compte d'une relation d'ordre partiel entre les valeurs a un impact significatif sur le taux d'erreur. En effet, celui-ci est sensiblement meilleur (plus faible) pour tous les jeux de données sur lesquels on utilise une adaptation prenant en compte l'ontologie de domaine.

Tableau 6. Synthèse des résultats obtenus avec les modèles de détection de vérité M_4 appliqués aux trois types de corpus. Différentes valeurs de γ ont été testées. Les valeurs en rouge/souligné indiquent les plus mauvais résultats et les résultats en gras les meilleurs

Sums adapté + Règles		γ					
		0*	0.25	0.50	0.75	0.9	1.0
Jeu de données EXP	n_{vrai}	*0,8955*	0,8995	0,8983	0,8997	**0,9041**	0,1018
	n_{gen}	*0,0033*	0,0033	0,0033	0,0033	0,0034	**0,1545**
	n_{faux}	*0,1012*	0,0971	0,0983	0,0970	**0,0925**	0,7437
Jeu de données LOW_E	n_{vrai}	*0,6645*	0,6927	0,6969	0,7040	**0,7142**	0,1006
	n_{gen}	*0,0048*	0,0049	0,0049	0,0050	0,0050	**0,1540**
	n_{faux}	*0,3307*	0,3024	0,2981	0,2911	**0,2808**	0,7454
Jeu de données UNI	n_{vrai}	*0,4938*	0,5371	0,5454	0,5547	**0,5666**	0,1002
	n_{gen}	*0,0055*	0,0057	0,0057	0,0057	0,0058	**0,1546**
	n_{faux}	*0,5007*	0,4572	0,4489	0,4396	**0,4276**	0,7453

Considérer uniquement l'influence des règles lors du processus de recherche de vérité consisterait à choisir comme valeur $\gamma = 1$. Dans la grande majorité des cas, cette configuration offre les résultats les moins bons si l'on considère le nombre de valeurs vraies attendues et le taux d'erreur. Ce résultat s'explique facilement. Les règles reposent uniquement sur une analyse statistique de KB et peuvent parfois ne pas être valides pour toutes les entités. Par contre, cette configuration fournit toujours le meilleur taux en termes de valeur générique. Ce constat confirme l'intuition suivante : l'application des règles tend à favoriser une connaissance générique. En effet, plus le recouvrement de la tête d'une règle est élevé, plus le nombre d'instances pour lesquelles elle est valide est grand. Pour les valeurs plus génériques, il est donc d'autant plus facile de trouver des règles qui seront vérifiées.

Cependant, même si l'application des règles d'association favorise les valeurs plus génériques lors de la recherche de vérité, il est intéressant de les prendre en compte. Elles permettent en effet, en accordant plus de confiance dans les valeurs génériques, de favoriser certaines branches lors de l'exploration de l'arbre des

valeurs par l'algorithme glouton de sélection des valeurs vraies. Elles permettent donc d'éviter certaines erreurs lors de l'amorce du processus de sélection des valeurs vraies.

Si l'on considère le modèle M_2 (Sums+Règles), on remarque une amélioration des résultats par rapport à la méthode de base, même si cette amélioration est moins flagrante qu'avec les autres modèles (M_3 – Sums adapté et M_4 – Sums adapté+règles). Cette amélioration s'explique par la propagation des confiances sur les valeurs (en respectant l'ordre partiel donné par l'ontologie de domaine). Cette propagation dans le cas des modèles M_3 et M_4 concerne l'ensemble des valeurs alors que les règles ne concernent, quant à elles, qu'une sous-partie de ces valeurs.

Tous modèles confondus, les meilleurs résultats sont obtenus avec la méthode M_4 et un coefficient $\gamma = 0.9$, et ce sur tous les types de corpus testés. Cette configuration permet d'obtenir le plus grand nombre de valeurs vraies attendues et de diminuer le taux d'erreur. Sur ces deux critères, le gain est d'autant plus grand que la disparité (contradictions) entre les sources est grande.

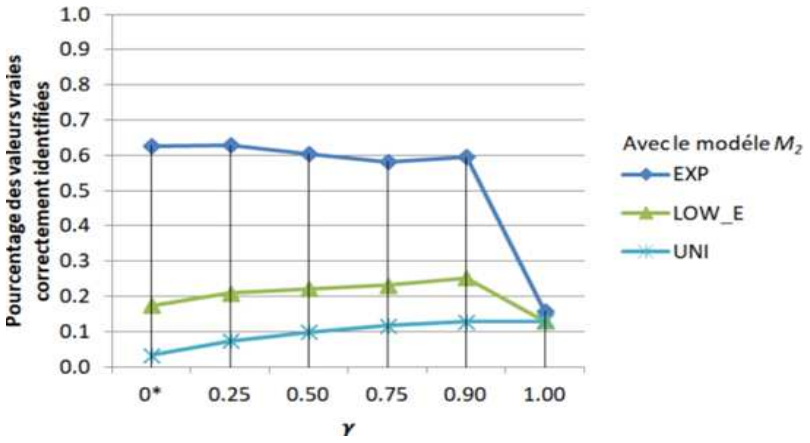

Figure 5. Synthèse des résultats pour M_2: rappel (taux de valeurs attendues *obtenues) en fonction du jeu de test, de la méthode utilisée et de* γ

L'importance de la valeur de γ est également à souligner. On le voit bien sur les résultats obtenus avec le modèle M_2 (pour lequel on ne tient pas compte de la propagation sur les valeurs proposées). Pour le jeu de données EXP dans lequel les sources ont tendance à être plus en accord sur la valeur proposée et cette valeur étant très spécifique, les meilleurs résultats en terme de précision sont obtenus avec un coefficient $\gamma = 0.25$. On voit donc que si l'on est dans un cas où les sources sont relativement fiables sur un sujet donné (site web spécialisé, par exemple), il est préférable d'accorder plus de confiance aux assertions qu'elles proclament. Par contre, pour les deux autres types de jeu de données (LOW_E et UNI) dans lesquels les désaccords entre les sources sont nombreux et vont en croissant, il est préférable

de s'appuyer sur les règles d'association identifiées ($\gamma = 0.9$ dans le premier cas et $\gamma = 1$ dans le cas de UNI). La comparaison des performances obtenues au travers des différents modèles est illustrée dans les figures 5 et 6.

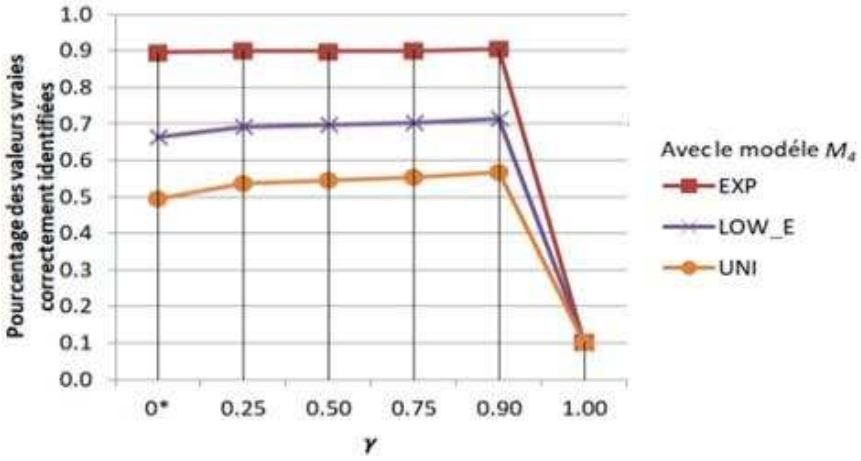

Figure 6. Synthèse des résultats pour M_4 : rappel (taux de valeurs attendues *obtenues) en fonction du jeu de test, de la méthode utilisée et de* γ.

6. Conclusion et perspectives

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.

Cet article 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 termes 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 cet article). Une évaluation au travers de 60 jeux de données de trois types distincts a été menée. Les résultats

montrent qu'une adaptation des méthodes traditionnelles qui intègre la prise en compte d'une structuration entre les valeurs, au travers d'une ontologie de domaine, conduit à de meilleurs résultats. Par ailleurs, cette approche est plus robuste, car moins sensible à la nature des jeux de données utilisés. En effet, certains jeux contenaient une proportion de valeurs vraies variable, pour refléter les cas où de nombreuses sources émettent des assertions potentiellement contradictoires sur certaines entités. Nous avons montré comment les relations d'ordre associées à cette ontologie permettaient d'améliorer l'approche existante.

Nous montrons également que la A-Box associée a également une grande influence et peut améliorer le processus de recherche de vérité. En considérant l'ordre partiel qui existe entre les valeurs proposées par différentes sources, l'utilisation de règles d'association collectées après l'analyse de la A-Box, permet de favoriser certaines valeurs plus génériques et ainsi d'améliorer la stratégie de sélection des valeurs vraies. En fonction du contexte, une bonne paramétrisation permettra d'obtenir de meilleurs résultats que les approches classiques.

Cette étude préliminaire souligne l'apport que constitue la prise en compte de l'ordre défini entre les concepts d'une ontologie et de la connaissance exprimée par les règles dans la détection de vérité et ouvre de nombreuses perspectives. Nous souhaitons compléter cette étude en considérant d'autres méthodes de référence comme AverageLog, Investment et PooledInvestment (Pasternack & Roth, 2010), et Cosine et 2-Estimated (Galland et al., 2010) afin de vérifier la flexibilité de l'approche. Nous souhaitons également effectuer des tests sur d'autres jeux de données avec des prédicats propres à un domaine et plus ou moins spécialisés. Par ailleurs, la procédure de propagation peut être modifiée. Notre approche ne considère, à l'heure actuelle qu'une propagation ascendante, inspirée par la propagation des croyances. Cette propagation peut être améliorée en y intégrant une propagation descendante, telle que la propagation des vraisemblances en théorie des croyances (plausibilité). La confiance d'une assertion sera alors dépendante de l'observation des assertions plus génériques et plus spécifiques. Ensuite, nous analyserons d'autres caractéristiques qui peuvent être intégrées à la détection de vérité. En effet, nous n'avons considéré ici que l'ordre partiel défini sur les valeurs, mais nous n'avons pas tenu compte de la sémantique associée aux concepts de l'ontologie qui pourrait être utilisée pour propager l'évidence que constitue une valeur pour les autres valeurs. Enfin, nous souhaitons également évaluer les approches proposées dans une chaîne de traitements réelle d'enrichissement de bases de connaissances à partir de textes.

Bibliographie

Auer S. et al., (2007). DBpedia: A Nucleus for a Web of Open Data. In K. Aberer et al., eds. The Semantic Web, Lecture Note in Computer Science. Springer Berlin Heidelberg, pp. 722–735.

Beretta V. et al., (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. Nîmes, France, pp. 15:1-15:12.

Berti-Équille L. & Borge-Holthoefer J. (2015). *Veracity of Data : From Truth Discovery* Computation Algorithms to Models of Misinformation Dynamics, ser. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, pp.1-155.

Blanco L. et al., (2010). Probabilistic Models to Reconcile Complex Data from Inaccurate Data Sources. In Proceedings of the 22nd International Conference on Advanced Information Systems Engineering, Hammamet, Tunisia, pp.83–97.

Boley H. (2000). Relationships between logic programming and RDF. In Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence, Melbourne, Australia, Melbourne, Australia,, pp. 201-218.

Dong X.L. et al., (2010). Global detection of complex copying relationships between sources. In Proceedings of the VLDB Endowment, 3(1-2), pp.1358–1369.

Dong X.L. et al., (2015). Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. In Proceedings of the VLDB Endowment, 8(9), pp. 938–949.

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

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

Feno D.R. (2007). Mesures de qualité des règles d*'association : normalisation et* caractérisation des bases. Université de la Réunion, France.

Galárraga L. et al., (2015). Fast rule mining in ontological knowledge bases with AMIE+. The VLDB Journal, 24(6), pp.707–730.

Galland A. et al., (2010). Corroborating Information from Disagreeing Views.In Proceedings of the third ACM international conference on Web search and data mining, New York City, NY, USA, pp.131–140.

Gupta M., Sun Y. & Han J. (2011). Trust analysis with clustering. In Proceedings of the 20th international conference companion on World wide web, pp. 53-54.

Harispe S. et al., (2015). On the consideration of a bring-to-mind model for computing the Information Content of concepts defined into ontologies. In Proceedings of IEEE International Conference on Fuzzy Systems, Istanbul, Turkey, pp. 1-8.

Harispe S. et al., (2015). Semantic Similarity from Natural Language and Ontology Analysis. Synthesis Lectures on Human Language Technologies, 8(1), pp.1–254.

Harispe S. et al., (2013). SML: semantic measure library. Available at: http://www.semantic-measures-library.org/sml/.

Hitzler P. et al., (2009). OWL 2 Web Ontology Language Primer. W3C recommendation, pp.1–123.

Jean P.-A. et al., (2016). Uncertainty Detection in Natural Language: A Probabilistic Model. In Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics. Nîmes, France, pp. 10:1-10:10.

Li Y. et al., (2015). A Survey on Truth Discovery. ACM SIGKDD Explorations Newsletter, 17(2), pp.1–16.

Maimon O. & Rokach L. (2005). Data Mining and Knowledge Discovery Handbook. O. Maimon & L. Rokach (eds.), Springer US Publisher, pp.1-1285.

Mann C.J.H. (2003). The Description Logic Handbook – Theory, Implementation and Applications, Kybernetes, 32(8-9).

Meng C. et al., (2015). Truth Discovery on Crowd Sensing of Correlated Entities. In Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, Seoul, Republic of Korea, pp.169–182.

Nenov Y. et al., (2015). RDFox: A Highly-Scalable RDF Store. In Proceedings of the 14th International Semantic Web Conference, Bethlehem, Pennsylvania, pp. 3-20.

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

Pasternack J. & Roth D. (2011). Making better informed trust decisions with generalized fact-finding. In Proceedings of the22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, 3, pp.2324–2329.

Pochampally R. et al., (2014). Fusing data with correlations. In Proceedings of the 2014 ACM Special Interest Group on Management of Data, Snowbird, USA, pp.433–444.

Qi G.-J. et al., (2013). Mining collective intelligence in diverse groups. In Proceedings of the 22nd international conference on World Wide Web, pp. 1041–1052.

Quboa Q.K. & Saraee M. (2013). A State-of-the-Art Survey on Semantic Web Mining. Intelligent Information Management, Rio de Janeiro, Brazil, 5, pp.10–17.

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

Shafer G. (1976). A Mathematical Theory of Evidence, Princeton: Princeton University Press.

Wang D., Abdelzaher T. & Kaplan L. (2015). Social Sensing: Building Reliable Systems on Unreliable Data, Morgan Kaufmann Publishers, San Francisco, CA, USA, pp. 1-232.

Wang S. et al., (2015). Scalable Social Sensing of Interdependent Phenomena. In Proceedings of the 14th International Conference on Information Processing in Sensor Networks, Seattle, USA, pp.202–213.

Wang X. et al., (2015). An Integrated Bayesian Approach for Effective Multi-Truth Discovery. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, Melbourne, Australia, pp. 493–502.

Wang Z. & Li J. (2015). RDF2Rules: Learning Rules from RDF Knowledge Bases by Mining Frequent Predicate Cycles. arXiv:1512.07734.

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), pp.796–808.

Zhao B. et al., (2012). A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. In Proceedings of the VLDB Endowment, 5(6), pp.550–561.

BON DE COMMANDE D'ABONNEMENT 2018

2018 SUBSCRIPTION FORM

Renvoyer à / *Return to*: Lavoisier SAS, Abonnements Revues
14, rue de Provigny – 94236 Cachan cedex – France
tel : (33) 01-47-40-67-00 – Fax : (33) 01-47-40-67-02 – abonne.ria@lavoisier.fr

REVUE D'INTELLIGENCE ARTIFICIELLE		
RIA – VOLUME 32/2018	**6 N°/AN** *(3 issues/year)*	
Tarif d'abonnement	TTC FRANCE	HT ÉTRANGER (*)
Version imprimée *incluant la version on line*	415 €	478 €
Version on line + archives	378 €	378 €
ABONNEMENT AUX 4 REVUES RSTI	**6 TSI + 6 RIA + 6 ISI + 3DN = 21 N°/AN**	
Tarif d'abonnement	TTC FRANCE	HT ÉTRANGER (*)
Version imprimée *incluant la version on line*	1 325 €	1 494 €
Version on line + archives	1 206 €	1 206 €

CONDITIONS D'ABONNEMENT / *CONDITIONS OF SUBSCRIPTION*

Les abonnements sont enregistrés à réception de leur règlement et sont acceptés pour l'année civile uniquement. / *Subscriptions are entered upon receipt of payment and are accepted for a calendar year only.*

(*) Pour les tarifs TTC étranger, merci de nous contacter / *Other countries rates are available on our web site: http://www.revuesonline.com or on request (revues.abo@lavoisier.fr)*

Nom / *Name* . . . . . .

Organisation / *Organization* . . . . . .

Adresse / *Address* . . . . . .

. . . . . .

Code postal – Ville / *ZIP – City* . . . . . .

Pays / *State* . . . . . .

☐ Règlement par chèque joint à l'ordre de Lavoisier / *Cheque enclosed payable to Lavoisier*

☐ Règlement par carte VISA /*Payment by VISA card*

N°carte / *Card No* ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Date d'expiration / *Expiry Date*

3 derniers chiffres du cryptogramme au dos de votre carte ☐☐☐

The last 3 digits of the cryptogram on the reverse of your card

Date et signature / *Date and signature*