Contribution à la recherche de vérité : modèles exploitant des règles d'association extraites d'une base de connaissances

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

[1] LGI2P de l'école des mines d'Alès, Site de Nîmes, Parc G. Besse, F-30 035 Nîmes
{prenom.nom}@mines-ales.fr
[2] UMR Espace-Dev, Université de Montpellier, Rue JF. Breton, Montpellier
isabelle.mougenot@umontpellier.fr

Résumé : Pour contrer les dangers de la désinformation, 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 *déclarations* (i.e. <*sujet, prédicat, valeur*>) diffusées par plusieurs *sources* sur un sujet donné, et tente de déterminer parmi toutes ces déclarations, celle qui constitue un *fait* (une *vérité*). Nous avons récemment montré que la prise en compte d'axiomes fournis par une ontologie de domaine, et en particulier certaines relations transitives, constitue une réelle plus-value et permet d'améliorer la performance des approches existantes. Dans cet article, nous enrichissons nos travaux précédents en étendant l'analyse à d'autres types de relations, pouvant participer à l'identification de règles pour renforcer la confiance dans certaines déclarations et ainsi améliorer l'identification de vérité. A partir d'une base de connaissances, un ensemble de règles est extrait. Un coefficient *propulseur* (*booster*) est alors calculé pour renforcer certaines déclarations. Les premiers résultats de l'évaluation montrent une plus-value par rapport aux approches traditionnelles.

Mots-clés : Détection de vérité, Ontologies, Web sémantique, Détection de règles, Raisonnement.

1 Introduction

Après une certaine période d'euphorie engendrée par l'accès et la diffusion à très grande échelle d'un grand nombre d'informations aussi diverses que peuvent l'être nos différentes activités humaines : relation sociales, activités professionnelles, engagements associatifs et/ou politiques, création artistiques ou photographiques, … 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 et l'adhésion à certaines idéologies qui constitueraient une menace pour nos sociétés. Le site Politifact[1] analyse ainsi depuis plusieurs années les discours des responsables politiques américains afin de déterminer leur part de vérité et de mensonge. 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 trop important pour être traité de façon exhaustive et des approches automatisées sont nécessaires pour les assister dans leur tâche. Pour contrer les dangers de la désinformation, un nouveau domaine de recherche a émergé ces dernières années : la 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 *déclarations*

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

diffusées par plusieurs sources sur un sujet donné, et tente de déterminer parmi toutes ces déclarations, celle qui constitue un *fait* (une *vérité* objective[3]). 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 déclarations 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 déclarations qui traduisent des *faits* (vérités). Nos travaux s'inscrivent dans cette veine et intègrent les ontologies de domaine au processus de détection de vérité afin de renforcer la confiance associée à certaines affirmations. Ainsi, dans (Beretta et al., 2016), nous avons proposé de prendre en compte certaines relations transitives d'une ontologie qui définissent un ordre partiel sur les valeurs pour en confirmer certaines et tenter d'identifier les valeurs *vraies*. Nous considérions alors seulement une portion réduite de l'ontologie, essentiellement au travers de l'exploitation partielle des définitions de classes contenues dans la T-Box. Plus précisément, notre approche se concentrait 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 ressources fournis par des prédicats transitifs supplémentaires (*e.g.* **`partOf`**). Dans cette contribution, nous souhaitons aller plus loin en intégrant une analyse plus large de la A-Box, afin de 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 déclarations. Prenons un 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 déclarations qui proposent une valeur correspondant à l'Espagne ou à des valeurs plus génériques. C'est l'idée que nous explorons dans cette étude.

La section suivante présente le contexte de notre étude, certaines notations et les travaux existants. La section 3 formalise notre approche et détaille l'intégration de règles dans la procédure itérative qui calcule alternativement la fiabilité des sources et la confiance dans les faits pour déterminer les valeurs vraies. La section 4 discute les premiers résultats obtenus et enfin la section 5 conclut cet article et ouvre de nombreuses perspectives de recherche.

2 Positionnement et état de l'art : bases de connaissances et règles d'association

Notre étude s'appuie sur des graphes de connaissances tels que DBpedia (Auer et al., 2007) où les nœuds correspondent à des entités de différents types et les arcs correspondent à des relations, également de différents types, entre ces entités. Nous proposons d'améliorer le processus de recherche de vérité en exploitant les informations contenues dans un tel graphe.

2.1 Bases de connaissances et recherche de vérité

De façon plus 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 et d'entités (*sujets* et *objets*) sont autorisés. L'objectif ultime de notre approche concerne l'enrichissement de cette base de

[3] Suivant la distinction qui est faite en philosophie entre *vérité de fait* et *vérité de raison*, la définition de la *vérité* considérée ici est la *vérité de fait* : Enoncé qui correspond au réel qu'il décrit (*vérité contingente* selon Leibniz ou *relation de faits* selon Hume).

connaissances afin de constituer un meilleur support pour l'aide à la décision. Chaque élément rajouté à la base doit donc être fiable.

Une paire (*sujet, prédicat)* est appelée une *description*[4] et représente une propriété particulière d'une entité sujet. La valeur associée à cette propriété est représentée par le singleton {*valeur*}. Il est à noter que la recherche de vérité envisagée ici ne concerne que des prédicats fonctionnels, c'est à dire pour lesquels une seule valeur est possible (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[5] peuvent proposer des valeurs différentes et contradictoires pour une même description. Ces propositions, également représentées par des triplets <*sujet, prédicat, valeur*>, sont appelées *déclarations* tant qu'elles ne sont pas validées, i.e. tant que l'on n'a pas identifié la valeur *vraie*. Il est, en effet, nécessaire de déterminer quelle est cette valeur *vraie*, parmi celles proposées, afin de constituer un nouveau *fait* qui pourra être intégré à la base.

Nous ne détaillons pas ici l'état de l'art concernant la détection de vérité. Le lecteur intéressé pourra se reporter à (Berti-Équille & Borge-Holthoefer, 2015) pour un état de l'art approfondi. Dans les approches traditionnelles de détection de vérité, un processus itératif calcule alternativement la *confiance* dans les déclarations et la *fiabilité* des sources afin de déterminer quelle est la valeur vraie la plus probable. La détection de vérité peut également exploiter différentes dépendances : entre les sources (Blanco et al., 2010; Dong et al., 2010; Pochampally et al., 2014; Qi et al., 2013), entre les valeurs (Yin et al., 2008) ou entre les descriptions (Meng et al., 2015; D. Wang et al., 2015; S. Wang et al., 2015). Ces modèles ne fournissent bien souvent qu'un score numérique.

Bien que n'ayant jamais été utilisées dans le domaine de la détection de vérité, il est également possible d'exploiter les cooccurrences par l'identification de règles d'association (Maimon & Rokach, 2005). Cette solution exprime une sémantique et permet une interprétation, c'est pourquoi nous avons choisi d'explorer cette voie. Dans ce qui suit, nous souhaitons utiliser un coefficient propulseur (*booster*), calculé à partir de l'identification de cooccurrences récurrentes entre différents faits afin de renforcer la confiance dans certaines valeurs pendant le processus de détection de vérité.

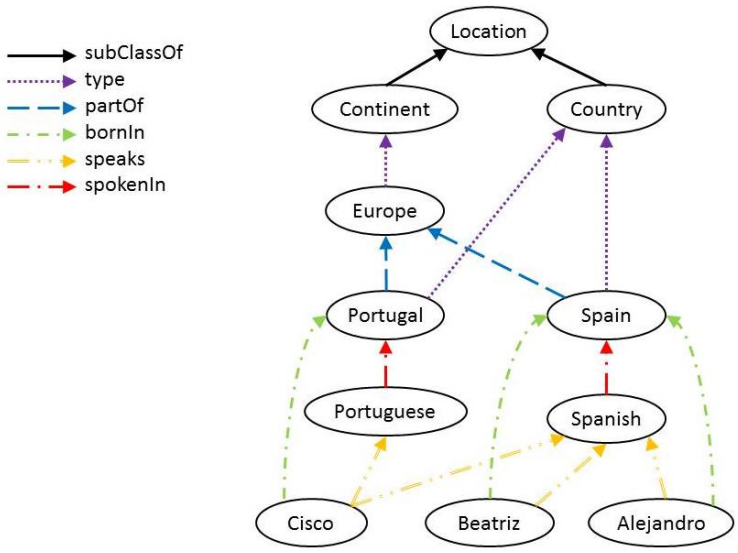

FIGURE 1 – Extrait d'une base de connaissances.

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.

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

[5] 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".

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 la déclaration <Picasso, bornIn, Spain> doit être renforcée, ainsi que les déclarations qui contiennent des valeurs plus génériques[6]. Ce renforcement est apporté par le coefficient propulseur qui représente le degré de soutien (la caution) apporté pour cette déclaration par les informations contenues dans KB. Ainsi le postulat de base du processus de recherche de vérité présenté en introduction sera modifié et nous considèrerons désormais que les *faits* (*vérités*) sont des *déclarations* 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 déclaration 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.

2.2 Détection de règles d'association : principes et mise en œuvre

Comme mentionné dans la synthèse sur les règles d'association présentée dans (Maimon & Rokach, 2005), il est difficile d'avoir une vue exhaustive des travaux dans ce domaine. Pour des applications en lien avec le Web sémantique, on peut 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 & Saraee, 2013).

Dans la suite, la notation utilisée pour les règles sera celle de Datalog. 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 :

$R: B_1 \wedge B_2 \wedge \ldots \wedge B_n \Rightarrow H$ 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 *déclaration* constituée d'un prédicat défini et de sujet et objet qui peuvent être des variables. Pour simplifier l'écriture, une déclaration constituée d'un triplet <*s*, *p*, *v*> sera notée : $p(s, v)$. 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, i.e. elles contiennent des variables fermées, c'est-à-dire qui apparaissent au minimum 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).

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 déclaration $p(s, v)$, le support est calculé de la façon suivante :

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

[6] Ce dernier point ne sera pas discuté ici puisqu'il rejoint la contribution de (Beretta et al., 2016).

où $x_1, \dots, x_i$ représentent les variables contenues dans les atomes de $\vec{B}$ exceptées s et v. Si l'on considère l'exemple présenté dans la TABLE 1, le *support* de la règle $r: livesIn(s,v) \Rightarrow bornIn(s,v)$ serait égal à 1, étant donné qu'on ne trouve dans la base de connaissances que la paire de déclarations $livesIn(Adam, Paris)$ et $bornIn(Adam, Paris)$ qui la vérifie.

TABLE 1 – *Extraits de faits de la base de connaissances. Les prédicats sont notés en tête 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 :

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

Dans notre exemple, $conf\left(livesIn(s,v) \Rightarrow bornIn(s,v)\right) = \frac{1}{4}$. Cette mesure de confiance a été définie dans un contexte de raisonnement en monde fermé, où l'on considère comme fausses les déclarations 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. C'est pour cela que 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 un triplet qui concerne une description, i.e. une paire (*sujet*, *prédicat*), alors toutes les valeurs possibles pour cette description sont connues. Autrement dit, si une description n'apparait jamais dans la base, elle n'est considérée ni comme étant vraie, ni comme étant fausse. La mesure de *PCA confidence* se calcule comme suit :

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

Dans notre exemple, $conf_{PCA}\left(livesIn(s,v) \Rightarrow bornIn(s,v)\right) = \frac{1}{3}$, 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 $livesIn(Luca, Milan)$) .

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, Harispe,

Ranwez, Bellot, & Montmain, 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 :

$$score(R) = \left(1 - \frac{1}{supp(R)}\right) conf_{PCA}(R) \quad (2)$$

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).

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 déclaration d et une influence indirecte sur le calcul de fiabilité des sources.

TABLE 2 – *Synthèse des notations utilisées dans notre approche.*

symbole	*signification*
$d \in D$	Une déclaration appartenant à l'ensemble de toutes les déclarations
$s \in S$	Une source appartenant à l'ensemble des sources
D^s	L'ensemble des déclarations faites par la source s
S^d	L'ensemble des sources qui proclament une déclaration d
$t^i(s)$	Calcul de la fiabilité (*trustworthiness*) d'une source à l'étape i
$c^i(d)$	Calcul de la confiance dans une déclaration à l'étape i

En utilisant les notations synthétisées dans la TABLE 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 :

$$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} C^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d) \quad (5)$$

$$c^i(d) = \frac{1}{norm_d}\left((1-\gamma) confidence_{basic}(d) + \gamma . boost(d)\right) \quad (6)$$

avec $\gamma \in [0,1]$ un poids qui représente l'influence relative accordée aux sources et à la base de connaissances ; $confidence_{basic}$ une fonction de D dans [0,1] qui représente la confiance donnée par les sources à une déclaration ; et $boost$ une fonction de D dans [0,1] qui représente la confiance dans une déclaration 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 4, plusieurs valeurs seront considérées et leur impact sera discuté.

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

$$confidence_{basic}(d) = \frac{\sum_{s \in S^d} t^i(s)}{max_{d' \in D} \sum_{s' \in S^{d'}} t^i(s')} \quad (7)$$

où l'on retrouve au numérateur la somme des fiabilités associées à toutes les sources qui émettent une déclaration et un facteur de normalisation au dénominateur, i.e. la confiance maximale associée à une déclaration.

Le facteur propulseur, *booster*, cherche à synthétiser les informations données par toutes les règles obtenues pour chaque déclaration. 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 un score unique. Les scores des différentes règles qui concernent une même déclaration peuvent ainsi être agrégés. En effet, notre objectif reste bien de renforcer la confiance dans certaines déclarations en utilisant tous les motifs identifiés.

Soit une déclaration $d = p(s,o)$, 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 la déclaration considérée. Pour chaque déclaration, l'ensemble des règles R_d à considérer (règles éligibles) est un sous-ensemble des règles extraites : $R_d \subset R$. 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 (i.e. contenus dans KB). Formellement $R_d = \{r: B_1 \wedge \ldots \wedge B_n \Rightarrow p'(x,y) \in R | (p' = p) \wedge T(B_1 \wedge \ldots \wedge B_n) = 1\}$ 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.

$$\text{boost(d)} = \left(1 - \frac{1}{1+\sum_{\text{r}\in \text{R}_\text{d}} \text{score(r)}}\right) \frac{\sum_{\text{r}\in \text{R}_\text{d}^\text{T}} \text{score(r)}}{\sum_{\text{r}\in \text{R}_\text{d}} \text{score(r)}} \quad (8)$$

où $\text{R}_\text{d}^\text{T} = \{r: B_1 \wedge \ldots \wedge B_{|r|} \Rightarrow p'(x,y) \in R_d : T(r) = 1\}$ 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, i.e. contenant le même sujet pour le prédicat considéré.

Ce facteur propulseur est compris entre 0 et 1.

Prenons l'exemple de la *TABLE 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 d_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 d_2 le score sera de 0 car même si les deux règles sont éligibles, aucune n'est vérifiée.

TABLE 3 – *Eléments d'une base de connaissances.*

Sources/origine	*Déclarations*
s_1	$d_1 = bornIn(Picasso, Spain)$
s_2	$d_2 = bornIn(Picasso, UK)$
KB	$d_3 = officialLanguage(Spain, Spanish)$
KB	$d_4 = speaks(Picasso, Spanish)$
KB	$d_5 = resident(Picasso, France)$

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 les règles identifiées (après analyse de la A-Box) lors du calcul de la confiance dans une déclaration (comme décrit ci-dessus). La méthode M_3 est la méthode présentée dans (Beretta et al., 2016) qui consiste à tenir compte uniquement des relations transitives 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 confiance associée aux déclarations. La *TABLE 4* synthétise ces différents modèles et les équations associées respectivement au calcul de la confiance dans les déclarations et au calcul de la fiabilité des sources dans le processus itératif de recherche de vérité.

TABLE 4 – *Récapitulatif des différents modèles utilisés pour la recherche de vérité.*

M_1 – *Sums* traditionnel
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = confidence_{basic}(d) = \frac{\sum_{s \in S^d} t^i(s)}{max_{d' \in D}\left(\sum_{s' \in S^{d'}} t^i(s')\right)}$
M_2 – *Sums* traditionnel + prise en compte des règles d'association
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = \frac{1}{norm_d}[(1-\gamma) confidence_{basic}(d) + \gamma . boost(d)]$
M_3 – *Sums* traditionnel + propagation en fonction des relations transitives de l'ontologie
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = adaptedConfidence(d) = \frac{\sum_{s \in S^{d+}} t^i(s)}{max_{d' \in D} \sum_{s' \in S^{d'}} t^i(s')}$ $avec\ S^{d+} = S^d \cup \{s \in S^{d'} : d' \in D \wedge d' \preccurlyeq d\}$
M_4 – *Sums* traditionnel + propagation en fonction des relations transitives + Règles
$t^i(s) = \frac{1}{max_{s' \in S}\left(\sum_{d' \in D^{s'}} c^{i-1}(d')\right)} \sum_{d \in D^s} c^{i-1}(d)$
$c^i(d) = \frac{1}{norm_d}[(1-\gamma) adaptedConfidence(d) + \gamma . boostProp(d)]$ avec $boostProp(d) = \left(1 - \frac{1}{\sum_{r \in R_{d+}} score(r)}\right) \frac{\sum_{r \in R^T_{d+}} score(r)}{\sum_{r \in R_{d+}} score(r)}$ et $R_{d+} = R_d \cup \{R_{d'} \in R : d' \in D \wedge d' \preccurlyeq d\}$

Il est à noter que $boostProp(d)$ 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 (i.e. les règles qui impliquent des concepts plus génériques que ceux de la règle considéré). 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 déclarations. En effet, la confiance $c(d)$ dans une déclaration d telle que $d' \preccurlyeq d$ doit être supérieure ou égale à la confiance $c(d')$ associée à la déclaration d'.

4 Discussion des résultats

Les expérimentations qui suivent ont été réalisées sur le corpus exploité dans (Beretta et al., 2016), disponible à l'adresse https://doi.org/10.6084/m9.figshare.4616071. Ce jeu de test a été réalisé à partir de l'extraction du prédicat **`dbpedia-owl:birthPlace`** dans DBpedia (version 2015-04) qui permet de disposer du lieu de naissance des personnes connues. Trois jeux de test ont été générés (EXP, LOW_E et UNI) qui diffèrent par la stratégie de sélection des valeurs vraies (plus ou moins spécifiques) – cf. (Beretta et al., 2016) pour plus de détail.

Dans nos expérimentations, nous avons sélectionné uniquement les règles détectées par AMIE+ qui ont une couverture supérieure à 0.012 pour la tête. 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 déclarations et de fiabilité des sources sont ceux présentés dans la ***TABLE 4***. La sélection de la valeur vraie est ensuite réalisée par un algorithme glouton – cf. (Beretta et al., 2016).

TABLE 5 – *Synthèse des résultats obtenus avec les modèles de détection de vérité M_2 et M_4 appliqués aux trois types de corpus. Différentes valeurs de γ ont été testées. Les valeurs indiquées en rouge indiquent les plus mauvais résultats et les résultats en gras les meilleurs.*

		Jeu de données EXP			***Jeu de données LOW_E***			***Jeu de données UNI***		
	γ	n_{vrai}	n_{gen}	n_{faux}	n_{vrai}	n_{gen}	n_{faux}	n_{vrai}	n_{gen}	n_{faux}
M_2 *Sums* + Règles	0*	*0,6267*	*0,1136*	*0,2596*	*0,1726*	*0,1896*	*0,6378*	*0,0335*	*0,1953*	*0,7712*
	0,25	**0,6278**	0,1433	0,2289	0,2085	0,2168	0,5746	0,0737	0,2316	0,6947
	0,50	0,6036	0,1919	0,2045	0,2200	0,2663	0,5137	0,0983	0,2814	0,6203
	0,75	0,5811	0,2404	0,1785	0,2310	0,3125	0,4565	0,1167	0,3284	0,5549
	0,90	0,5944	0,2567	**0,1489**	**0,2507**	0,3439	0,4054	0,1274	0,3695	0,5031
	1,00	0,1578	**0,3205**	0,5217	0,1279	**0,4751**	**0,3970**	**0,1293**	**0,4388**	**0,4318**
M_4 *Sums* adapté + Règles	0*	*0,8955*	***0,0033***	*0,1012*	*0,6645*	*0,0048*	*0,3307*	*0,4938*	*0,0055*	*0,5007*
	0,25	0,8995	0,0033	0,0971	0,6927	0,0049	0,3024	0,5371	0,0057	0,4572
	0,50	0,8983	0,0033	0,0983	0,6969	0,0049	0,2981	0,5454	0,0057	0,4489
	0,75	0,8997	0,0033	0,0970	0,7040	0,0050	0,2911	0,5547	0,0057	0,4396
	0,90	**0,9041**	0,0034	**0,0925**	**0,7142**	0,0050	**0,2808**	**0,5666**	0,0058	**0,4276**
	1,00	0,1018	**0,1545**	0,7437	0,1006	**0,1540**	0,7454	0,1002	**0,1546**	0,7453

Nous avons évalué la précision de notre approche. L'ensemble des résultats est présenté dans la ***TABLE 5***. Nous avons analysé la proportion de valeurs vraies retournées par les différentes méthodes et qui correspondent à des valeurs attendues (pour chaque description (*sujet*, *prédicat*), la valeur attendue est contenue dans un corpus de référence) – n_{vrai}. La proportion de valeurs plus générales que celles attendues est également considérée – n_{gen}. Et enfin le taux d'erreur est indiqué (valeurs proposées qui sont totalement différentes et décorrélées de la valeur attendue) – n_{faux}. 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 dans (Beretta et al., 2016) 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).

A l'inverse, 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 lequel 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.

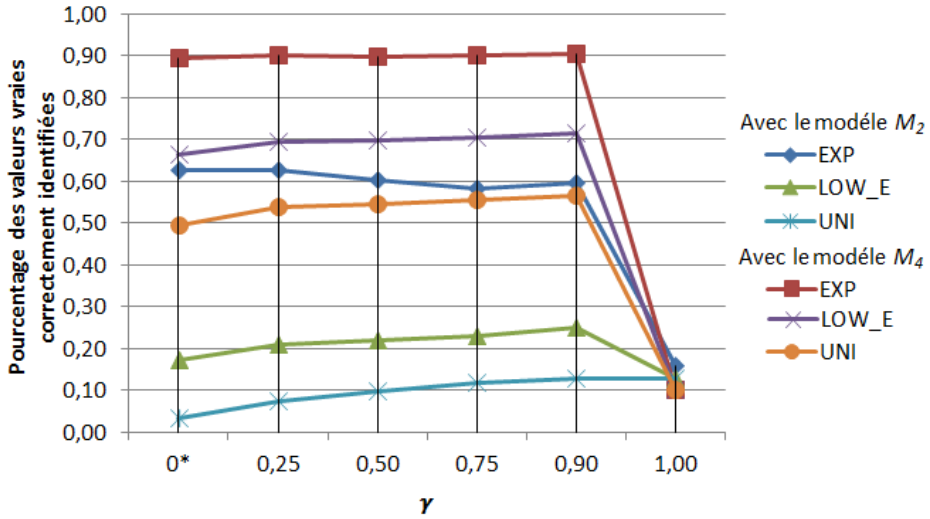

FIGURE 2 – Synthèse des résultats : précision (taux de valeurs attendues obtenues) en fonction du jeu de test, de la méthode utilisée et de γ.

Si l'on considère maintenant 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, ce gain est d'autant plus grand que la disparité (contradictions) entre les sources est grande.

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 obtenus en terme de précision sont 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 donnés (site Web spécialisé, par exemple), il est préférable d'accorder plus de confiance aux déclarations 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 la **FIGURE 2**.

5 Conclusion et perspectives

A 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. Dans une contribution précédente, nous avons montré comment certaines relations de cette ontologie permettait d'améliorer les approches existantes. Dans cette présente contribution, nous montrons 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. 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). Ces développements sont en cours. Nous souhaitons également effectuer des tests sur d'autres jeux de données avec des prédicats plus ou moins spécialisés par rapport à un domaine. La propagation vers les concepts plus généraux (selon la propagation des *croyances*) produit des améliorations. Il est aussi souhaitable d'effectuer une propagation vers les concepts plus spécifiques suivant le mode de propagation des *possibilités* en théorie des croyances (modèle en cours de définition). Enfin, nous souhaitons également intégrer ces modules logiciels dans une chaîne réelle d'enrichissement de bases de connaissances basée sur une extraction à partir de textes.

Références

AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., & IVES, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L.

Nixon, … P. Cudré-Mauroux (Eds.), *The Semantic Web, Lecture Note in Computer Science* (Vol. 4825, pp. 722–735). Springer Berlin Heidelberg.

BERETTA, V., HARISPE, S., RANWEZ, S., & MOUGENOT, I. (2016). Utilisation d'ontologies pour la quête de vérité : une étude expérimentale. In *Actes des 27es Journées francophones d'Ingénierie des Connaissances IC2016*. Montpellier, France.

BERTI-ÉQUILLE, L., & BORGE-HOLTHOEFER, J. (2015). *Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics* (Synthesis). Morgan & Claypool Publishers.

BLANCO, L., CRESCENZI, V., MERIALDO, P., & PAPOTTI, P. (2010). Probabilistic models to reconcile complex data from inaccurate data sources. In B. Pernici (Ed.), *Advanced Information Systems Engineering: 22nd International Conference CAiSE 2010 Proceedings* (pp. 83–97). Hammamet, Tunisia: Springer-Verlag.

DONG, X. L., BERTI-EQUILLE, L., HU, Y., & SRIVASTAVA, D. (2010). Global detection of complex copying relationships between sources. In E. Bertino, P. Atzeni, K. L. Tan, Y. Chen, & Y. C. Tay (Eds.), *Proceedings of the VLDB Endowment* (Vol. 3, pp. 1358–1369). VLDB Endowment.

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

GALARRAGA, L., TEFLIOUDI, C., HOSE, K., & SUCHANEK, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE. *The VLDB Journal The International Journal on Very Large Data Bases*, *24*(6), 707–730.

GALLAND, A., ABITEBOUL, S., MARIAN, A., & SENELLART, P. (2010). Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10* (pp. 131–140). New York, New York, USA: ACM Press.

JEAN, P.-A., HARISPE, S., RANWEZ, S., BELLOT, P., & MONTMAIN, J. (2016). Uncertainty Detection in Natural Language: A Probabilistic Model. In *6th International Conference on Web Intelligence, Mining and Semantics, WIMS'16* (p. 10:1-10:10). Nîmes, France: ACM Press, New York (USA).

LI, Y., GAO, J., MENG, C., LI, Q., SU, L., ZHAO, B., … HAN, J. (2015). A Survey on Truth Discovery. *ACM SIGKDD Explorations Newsletter*, *17*(2), 1–16.

MAIMON, O., & ROKACH, L. (2005). *Data Mining and Knowledge Discovery Handbook*. (O. Maimon & L. Rokach, Eds.). Springer US.

MENG, C., JIANG, W., LI, Y., GAO, J., SU, L., DING, H., & CHENG, Y. (2015). Truth Discovery on Crowd Sensing of Correlated Entities. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, SenSys '15* (pp. 169–182). Seoul, South Korea: ACM Press, New York (USA).

PASTERNACK, J., & ROTH, D. (2010). Knowing What to Believe (when you already know something). In *23rd International Conference on Computational Linguistics, COLING'10* (pp. 877–885). Stroudsburg, PA, USA: Association for Computational Linguistics.

POCHAMPALLY, R., DAS SARMA, A., DONG, X. L., MELIOU, A., & SRIVASTAVA, D. (2014). Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14* (pp. 433–444). New York, New York, USA: ACM Press.

QI, G.-J., AGGARWAL, C. C., HAN, J., & HUANG, T. (2013). Mining collective intelligence in diverse groups. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13* (pp. 1041–1052). New York, New York, USA: ACM Press.

QUBOA, Q., & SARAEE, M. (2013). A state-of-the-art survey on semantic web mining. *Intelligent Information Management*, *5*(1), 10–17.

WANG, D., ABDELZAHER, T., & KAPLAN, L. (2015). *Social Sensing: Building Reliable Systems on Unreliable Data*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

WANG, S., SU, L., LI, S., HU, S., AMIN, T., WANG, H., … ABDELZAHER, T. (2015). Scalable Social Sensing of Interdependent Phenomena. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks, IPSN '15* (pp. 202–213). Seattle, Washington: ACM, New York, NY, USA.

WANG, Z., & LI, J. (2015). RDF2Rules: learning rules from RDF knowledge bases by mining frequent predicate cycles. *arXiv Preprint 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), 796–808.