Are my data reliable ? Towards an evidential answer

Patrice Buche, Brigitte Charnomordic, Sébastien Destercke

▶ **To cite this version:**

Patrice Buche, Brigitte Charnomordic, Sébastien Destercke. Are my data reliable ? Towards an evidential answer. LFA 2010 - Logique Floue et ses Applications, Nov 2010, Lannion, France. lirmm-00538950

HAL Id: lirmm-00538950

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00538950

Submitted on 6 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mes données sont-elles fiables ? Vers une réponse évidentielle

Are my data reliable ? Towards an evidential answer

P. Buche[1] B. Charnomordic[2] S. Destercke[1]
[1] INRA, CIRAD
[2] INRA

UMR IATE, Campus Supagro, 2 Pl. P. Viala, 34060 Montpellier
UMR MISTEA, Campus Supagro, 2 Pl. P. Viala, 34060 Montpellier

Résumé :

S'il existe de nombreuses méthodes permettant d'intégrer l'information concernant la fiabilité des sources et des données à des modèles d'incertitudes, il existe nettement moins de travaux s'intéressant au problème d'évaluer cette fiabilité. Dans ce papier, nous proposons une méthode permettant d'estimer cette fiabilité, basée sur un ensemble de critères d'évaluation et sur la théorie de l'évidence. L'exemple choisi est tiré d'un cas réel issu de l'application *Sym'Previus*, entrepôt de données en microbiologie prévisionnelle.

Mots-clés :

Fonctions de croyances, fusion d'informations, confiance, sous-ensembles maximaux cohérents.

Abstract:

There are many available methods to integrate data or information source reliability in an uncertainty representation, but there are only a few works focusing on the problem of evaluating this reliability. In this paper, we propose a method to assess data reliability from a set of criteria by the means of evidence theory. The chosen illustrative example comes from real-world data issued from the *Sym'Previus* predictive microbiology oriented data warehouse.

Keywords:

Belief functions, information fusion, confidence, maximal coherent subsets, trust.

1 Introduction

La fiabilité des données a toujours été une préoccupation majeure des scientifiques, imposant le suivi et la description des protocoles expérimentaux, la mise en place de plans d'expérience et de répétitions, et la détection des données aberrantes par des procédures statistiques. Cette notion de fiabilité est importante dans tous les domaines, et dans les sciences de la vie, elle l'est tout particulièrement, en raison de la difficulté de disposer de modèles généraux à toutes les échelles nécessaires.

Les entrepôts de données thématiques ouverts sur le Web offrent des possibilités extrêmement séduisantes d'accroître le nombre de données disponibles pour des traitements suivis d'une prise de décision. Cependant, l'évaluation de la fiabilité des données provenant de ces entrepôts pose de nouvelles questions, qui sont de deux ordres : la prise en compte des sources et celle du contexte. Le développement d'outils génériques permettant une évaluation automatique de la fiabilité des données de ces entrepôts ouverts sur le Web est donc nécessaire en préalable à l'utilisation de ces données dans des processus de décision. Une notion connexe, celle de la pertinence des données pour le processus de décision, est également à étudier.

A l'heure actuelle, des travaux existent sur la notion de *trust* [6], très importante pour le Web sémantique, en particulier sur l'authenticité des sources. Cependant, peu d'auteurs se sont intéressés à la confiance dans le contenu même des sources. Dans [4], Gil et Artz ont dégagé un certain nombre de critères des utilisateurs du Web, intervenant dans l'attribution de degrés de confiance/méfiance (avec un caractère bipolaire) aux données trouvées sur le Web. Au nombre de ces critères figurent la criticalité de l'information, la popularité de la source, la concordance entre sources, la fraicheur de l'information, ... Les méthodes proposées pour la combinaison de ces critères multiples et souvent subjectifs restent simplistes (e.g., moyenne ou maximum). De ce fait, l'interprétation du

résultat de l'estimation de la fiabilité est difficile, ce qui peut être gênant pour des scientifiques.

Dans cet article, nous proposons une méthode plus fine, basée sur la théorie des croyances et adaptée à l'évaluation de la fiabilité de tableaux de données d'expériences scientifiques en Sciences du Vivant, tableaux issus de publications et stockés dans un entrepôt thématique ouvert sur le Web. Dans la section 2, nous proposons tout d'abord des groupes de critères adaptés à cette problématique. La section 3 présente l'intérêt de cette évaluation pour proposer une estimation argumentée de la fiabilité aux utilisateurs d'un entrepôt de données, dans le domaine de la microbiologie prévisionnelle, l'entrepôt *Sym'Previus*. Enfin, la méthode d'évaluation de la fiabilité est décrite dans la section 4.

2 Quelles informations ?

Dans cette section, nous proposons plusieurs groupes de critères pertinents dans l'évaluation de la fiabilité, critères rendus aussi indépendants que possible. Cette hypothèse d'indépendance, même si elle n'est jamais totalement vraie, doit être aussi respectée que possible. Ceci est indispensable pour la combinaison des critères par la théorie des croyances, et facilite le travail de collecte de la connaissance experte. Le tableau 1 résume les critères que nous avons dégagés dans le cadre de la fiabilité de données issues de publications scientifiques.

Un premier sous-groupe de critères se dégage naturellement. C'est celui lié à la fiabilité de la publication d'où proviennent les données. Les critères les plus importants qui apparaissent pour ce sous-groupe sont : le type de source (rapport technique, article de journal, soumis à comité de lecture, ...), le nombre de citations, la réputation du laboratoire ayant réalisé les expériences, et enfin la date de publication, particulièrement importante en Sciences du Vivant. En effet les méthodes expérimentales évoluent très rapidement, et des données *anciennes* sont considérées comme moins fiables, par exemple par les biologistes.

Le second sous-groupe a trait à la qualité de production des données. Il s'agit ici d'examiner la section *Matériel et méthodes* de la publication. Les critères de ce groupe prendront en compte le protocole utilisé et le type d'équipement.

Le troisième et dernier sous-groupe contient des critères statistiques de qualité des données elles-mêmes : le nombre de répétitions, l'intervalle de confiance autour de la mesure, s'il est indiqué. Des procédures de détection des données aberrantes, obtenues par simulation, pourront servir à établir des critères plus sophistiqués.

Une autre notion importante mentionnée dans l'introduction de cet article est celle de la *pertinence* des données. Cette notion intervient dans le traitement automatique de l'information, et permet par exemple d'attribuer un degré de pertinence à des documents obtenus à la suite de requêtes, pertinence par rapport à la question a traiter. Les critères de tri habituellement utilisés sont : la fréquence et la proximité des termes de la requête ou de ceux trouvés dans les documents, la prise en compte des termes rares dans l'index, ... L'utilisation de critères de pertinence pour l'utilisation de données extraites automatiquement du Web dans des problèmes de simulation est également un problème ouvert, que l'on pourrait traiter d'une manière similaire à celle de la fiabilité, sur laquelle nous faisons porter le travail présenté dans cet article.

Les critères du tableau 1 sont conçus pour le traitement des données scientifiques issues de publications, dans le domaine des sciences du vivant. Pour faciliter leur extension à des domaines voisins et la traçabilité des critères pour les utilisateurs, nous proposons de les intégrer dans une ontologie du domaine. Ce point sera détaille dans la section 3.

Source	Production	Statistiques
Type	Protocole	Répétitions
Réputation	Equipement	Intervalle de confiance
Nombre de citations		Procédures de validation des données
Date de publication		

Tableau 1 – Critères de fiabilité

3 Contexte d'utilisation : l'entrepôt *Sym'Previus*

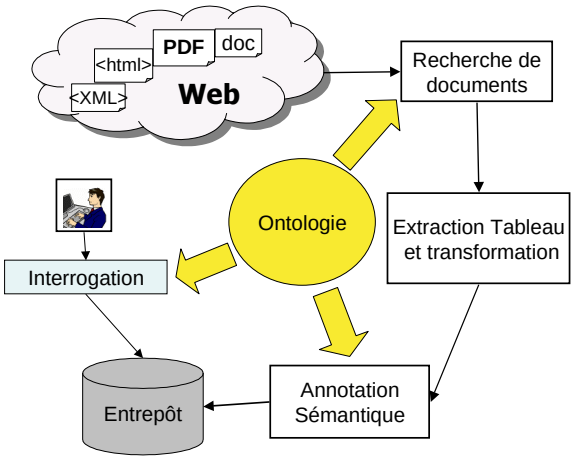

Figure 1 – Grandes étapes de transformation dans @Web.

3.1 Présentation d'@Web

Ce travail sur l'évaluation de la fiabilité des données a été motivé par des besoins rencontrés dans le cadre de la conception d'un entrepôt de données ouvert sur le Web, baptisé @Web [2, 5]. Ses résultats sont actuellement en cours d'implémentation dans les outils liés à @Web. Les grandes étapes de transformation des données provenant du Web sont présentées dans la figure 1. L'intégration dans l'entrepôt des sources de données est réalisée de manière générique. Elle repose sur une ontologie dans laquelle sont rassemblées toutes les connaissances spécifiques d'un domaine d'application. Par conséquent, il suffit de changer d'ontologie pour utiliser @Web sur une nouvelle application. Dans ce papier, nous illustrons ces travaux avec la conception de l'entrepôt de données Sym'Previus en microbiologie prévisionnelle. @Web exploite les tableaux de données présents dans les documents extraits du Web. Dans l'entrepôt Sym'Previus, sont ainsi intégrés des taux et des cinétiques de croissance d'un micro-organisme (par exemple, Listeria monocytogenes) dans un aliment (par exemple, le saumon fumé) qui peuvent être publiés sous forme de tableau dans un article scientifique, un rapport de projet international ou une page Web. La structure des tableaux et le vocabulaire utilisé dans les différentes sources de données étant hétérogènes, @Web les intègre dans l'entrepôt en les indexant de manière semi-automatique avec un vocabulaire standardisé (liste de noms de produits alimentaires, liste de noms de micro-organismes, ...) et un ensemble de relations sémantiques d'intérêt (taux de croissance associé à un couple aliment-micro-organisme, ...) définis dans l'ontologie du domaine d'application. Il est ensuite possible d'interroger l'ensemble des tableaux de données du Web intégrés dans l'entrepôt en utilisant les relations sémantique et le vocabulaire de l'ontologie. Enfin, l'ensemble des critères d'évaluation de la fiabilité d'un tableau de données listés dans la section 2 sont également définis dans l'ontologie. Cela permet d'adapter le choix des critères à chaque domaine d'application tout en préservant la généricité de l'approche.

3.2 Evaluation de la fiabilité des sources de données du Web

Dans l'outil d'aide à la décision *Sym'Previus* (www.symprevius.net), les données stockées dans l'entrepôt sont utilisées pour simuler la

croissance d'un micro-organisme dans un aliment (cf figure 2 et [1]). En préalable aux outils de simulation, il est essentiel de proposer à l'utilisateur une aide dans le choix des sources de données à utiliser. Nous proposons donc une méthode qui permet de calculer une mesure de fiabilité à partir des critères retenus dans la section 2 et qui sera associée à chaque source de données du Web. Elle permettra à terme de renvoyer, en réponse à une requête soumise à l'entrepôt, une liste de sources de données triée par ordre de fiabilité décroissante. Les éléments ayant participé à l'attribution de la fiabilité seront catégorisés et apparaîtront clairement, afin que l'utilisateur puisse décider d'utiliser ou non les données dans le calcul de simulation.

La méthode présentée ici consiste à évaluer la fiabilité à partir de méta-information concernant les tableaux de données. Plus précisément, nous basons notre évaluation sur S sous-groupes $A_1, \ldots, A_S$ distincts de critères, les différents sous-groupes étant considérés indépendants entre eux. Ces sous-groupes correspondent aux critères jugés pertinents (e.g., type de source de données, nombre de citations d'un article, qualité de la méthode employée pour obtenir les données) pour évaluer la fiabilité de la donnée et de sa source. Chaque sous-groupe $A_i = \{a_{i1}, \ldots, a_{iC_i}\}$ contient C_i valeurs différentes.

3.3 Détermination des opinions expertes associées aux critères

Afin de pouvoir calculer la mesure de fiabilité sur les tableaux de données, il est nécessaire de déterminer les opinions des experts sur les critères pris en considération. Par exemple, si l'on considère le sous-groupe de critères $A_1 =$ *type de la source de données* (ici le sous-groupe est réduit à un critère), qui peut avoir plusieurs valeurs ($a_{11} =$ *article de journal*, $a_{12} =$ *rapport gouvernemental*, $a_{13} =$ *rapport de projet international*, $a_{14} =$ *autre rapport*), une opinion experte doit être définie pour chaque valeur prise par le sous-groupe. Cette opinion est exprimée sur un ensemble de modalités (voir par exemple le tableau 2), auxquelles l'on pourrait ajouter la modalité *sans opinion*. Par exemple, l'expert associe au couple *(type de la source de données, article de journal)* ($A_1 = a_{11}$) l'opinion *très fiable*.

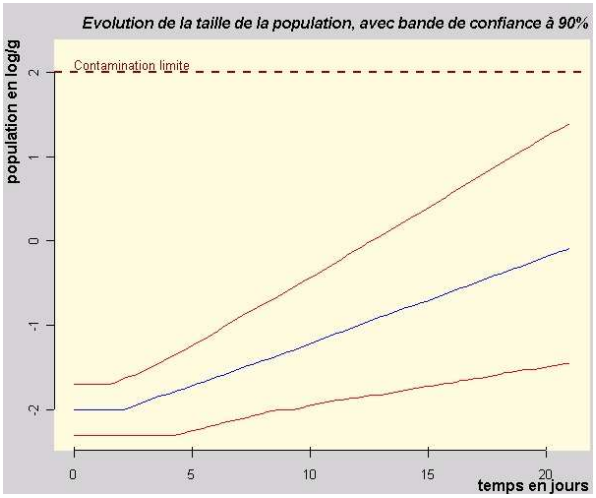

Figure 2 – Simulation de la croissance de Listeria monocytogenes dans le saumon fumé calculée à partir des données de l'entrepôt.

Dans cet article, nous choisissons de représenter les modalités utilisées dans les opinions expertes au moyen de sous-ensembles flous sur un domaine ordonné $\Theta = \{\theta_1, \ldots, \theta_N\}$ ($\theta_i < \theta_j$ ssi $i < j$) représentant la fiabilité. Cette étape de traduction (ou de modélisation) d'une opinion linguistique en ensemble flou sur Θ nous offre un degré de liberté qui peut permettre d'adapter le système afin que ses réponses soient conformes à ce qui est attendu (pour peu qu'un ensemble d'entrainement soit disponible).

θ_1 correspond donc à une donnée totalement non fiable, et θ_N à une donnée totalement fiable (cf figure 3). Le tableau 2 rappelle les cinq modalités utilisées dans notre illustration. Les résultats sont évidemment valables pour un nombre quelconque de modalités.

4 Evaluation de fiabilité

Après les brefs rappels nécessaires, cette section expose l'aspect technique de la méthode proposée pour évaluer la fiabilité de données à

très peu fiable	peu fiable	moyennement fiable	fiable	très fiable

Tableau 2 – Modalités choisies pour la fiabilité

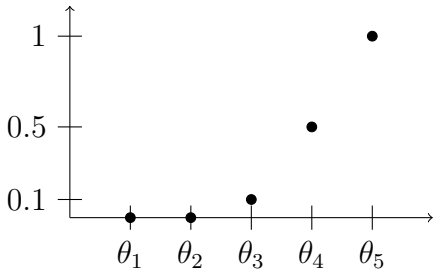

Figure 3 – Ensemble flou définissant la modalité *très fiable* sur le domaine Θ avec 5 éléments.

partir de différents sous-ensembles de critères jugés indépendants entre eux. Etant donné $\Theta = \theta_1, \ldots, \theta_N$, nous noterons $I_{a,b} = \{\theta_a, \ldots, \theta_b\}$ un ensemble tel que $a \leq b$ et $\forall\ c$ t.q. $a \leq c \leq b$, $\theta_c \in I_{a,b}$.

4.1 Bref rappel

Fonctions de croyance Une fonction de croyance est une fonction m des sous-ensembles d'un espace Θ dans l'intervalle unité $[0, 1]$ t.q. $\sum_{E\subseteq\Theta} m(E) = 1$, $m(E) \geq 0$ et $m(\emptyset) = 0$. Les ensembles E ayant une masse positive sont appelés ensembles focaux. Nous noterons $\mathcal{F}_m$ les ensembles focaux d'une fonction de croyance m. A partir de cette fonction, deux fonctions d'ensembles, les mesures de plausibilité et de crédibilité, sont définies comme dans [7] :

$$Bel(A) = \sum_{E\neq\emptyset, E\subseteq A} m(E); \quad Pl(A) = \sum_{E, E\cap A\neq\emptyset} m(E)$$

où la fonction de crédibilité mesure la quantité d'information qui étaye forcément A, et la fonction de plausibilité la quantité d'information qui pourrait étayer A.

Agrégation de fonctions de croyances venant de sources distinctes Etant donné S fonctions de croyances $m_1, \ldots, m_S$ définies sur un espace Θ, leur agrégation ou fusion en une fonction de croyance unique m lorsqu'elles sont jugées distinctes (indépendantes) peut s'écrire, $\forall E \subseteq \Theta$

$$m(E) = \sum_{\substack{E_i\in\mathcal{F}_i \\ \oplus_{i=1}^{S}(E_i)=E}} \prod_{i=1}^{S} m_i(E_i), \qquad (1)$$

avec $\mathcal{F}_i$ les ensembles focaux de m_i, et $\oplus_{i=1}^{S}(E_i) = E$ une combinaison d'opérateurs ensemblistes. Nous retrouvons la combinaison conjonctive si $\oplus = \cap$ et la combinaison disjonctive si $\oplus = \cup$.

Combinaison par sous-ensembles maximaux cohérents (SMC) sur espaces ordonnés Soit $K = \{1, \ldots, k\}$ ensembles I_{a_k,b_k}. Utiliser la méthode des SMC sur ces intervalles revient à prendre l'intersection au sein des ensembles $\overline{K_j} \subset N$ t.q. $\cap_{i\in K_j} I_{a_i,b_i} \neq \emptyset$ et qui sont maximaux avec cette propriété, pour ensuite faire l'union des résultats (i.e. $\cup_j \cap_{i\in K_j} I_{a_i,b_i}$). Nous noterons l'opérateur SMC par $\oplus_{SMC}$. Dans le cas d'espaces ordonnés (le cas ici), l'algorithme 1 détaillé dans [3] fournit un moyen facile et efficace de reconnaître les sous-ensembles maximaux cohérents. Son résultat est illustré par la figure 4, où quatre intervalles (à valeurs réelles) I_1, I_2, I_3, I_4 sont fusionnés. Les deux sous-ensembles maximaux cohérents sont (I_1, I_2) et (I_2, I_3, I_4), le résultat de leur fusion par l'opérateur SMC est donc $(I_1 \cap I_2) \cup (I_2 \cap I_3 \cap I_4)$.

La combinaison par sous-ensembles maximaux cohérents permet donc de gagner un maximum de précision (conjonction) tout en gérant le conflit via la réconciliation (disjonction) des sous-ensembles maximaux. Notons que la conjonction (resp. la disjonction) est retrouvée lorsque tous les ensembles sont consistants (resp. conflictuels) entre-eux, i.e., ont une intersection non-vide (resp. une intersection vide 2 à 2).

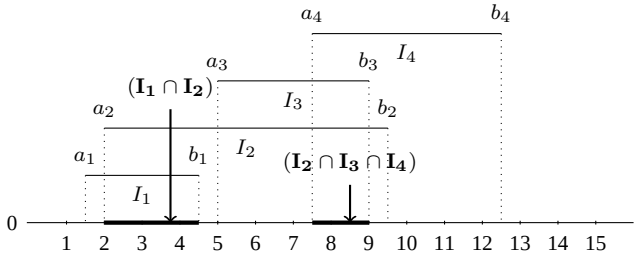

Figure 4 – Méthode des sous-ensembles maximaux cohérents : illustration

De plus, cette combinaison a l'avantage d'autoriser une explication des résultats, compréhensible par les utilisateurs, comme nous le verrons par la suite.

Algorithme 1 : Sous-ensembles maximaux cohérents sur des intervalles

Input : k ensembles $I_{a,b}$
Output : Liste de sous-ensembles maximaux cohérents $\overline{K}_j$
Liste = $\emptyset$; j=1 ; $\overline{K} = \emptyset$;
Ordonnancer de manière croissante $\{a_i, i = 1, \ldots, k\} \cup \{b_i, i = 1, \ldots, k\}$ (en cas d'ex-aequo, ranger a_i avant b_i) ;
Les renommer $\{c_i, i = 1, \ldots, 2k\}$ avec $type(i) = a$ si $c_i = a_m$ et $type(i) = b$ si $c_i = b_m$;
for $i = 1, \ldots, 2k-1$ **do**
  if $type(i) = a$ **then**
    Ajouter source m à K t.q. $c_i = a_m$;
    if $type(i+1) = b$ **then**
      Ajouter $\overline{K}$ à Liste ($\overline{K}_j = \overline{K}$) ;
      $j = j + 1$;
  else
    Enlever source j de $\overline{K}$ t.q. $c_i = b_m$;

4.2 Méthode

Nous considérons qu'à chaque valeur $a_{ij} \in A_i$ est associée une fonction de croyance $m_{a_{ij}} : 2^{|\Theta|} \rightarrow [0, 1]$ induite par l'ensemble flou [1] (par exemple, celui de la figure 3) correspondant à l'opinion donnée par l'expert lorsque $A_i = a_{ij}$. Vu qu'un expert est consistant avec lui-même (ou peut-être supposé comme tel), les éléments focaux de $m_{a_{ij}}$ correspondront à des intervalles.

Exemple 1. Reprenons le cas ébauché dans la section 3, et ajoutons au critère A_1 le critère $A_2 = \{a_{21}, a_{22}\}$ concernant le fait que l'expérience ait été répétée ou pas (a_{21} = Répétée). Supposons que le tableau de données auquel on s'intéresse provienne d'un article soumis à comité de lecture ($A_1 = a_{11}$) mais que l'expérience n'ait pas été répétée ($A_2 = a_{22}$). Une donnée venant d'un article soumis à comité de lecture est jugée "très fiable", mais l'absence de répétition indique une donnée "peu fiable".

Les fonctions de croyances correspondant à ces deux avis sont, sur l'échelle donnée dans le tableau 2 :

	$\{\theta_3, \theta_4, \theta_5\}$	$\{\theta_4, \theta_5\}$	$\{\theta_5\}$
$m_{a_{11}}$	0.1	0.4	0.5

	$\{\theta_1, \theta_2, \theta_3\}$	$\{\theta_2, \theta_3\}$	$\{\theta_2\}$
$m_{a_{22}}$	0.1	0.4	0.5

Etant donné les valeurs $A_i = a_{ij_i}$ prises par les différents critères pour une donnée particulière (l'indice j_i représentant le numéro de la modalité associée au critère A_i et affectée à cette donnée), nous proposons de construire une fonction de croyance m_g reflétant la fiabilité globale d'une donnée en fusionnant $m_{a_{1j_1}}, \ldots, m_{a_{Sj_S}}$ par l'équation (1), en prenant $\oplus = \oplus_{SMC}$. L'utilisation de (1) est justifiée par le fait que les sous-ensembles de critères A_i sont jugés indépendants, tandis que l'utilisation de $\oplus_{SMC}$ comme opérateur ensembliste permet de gérer facilement les conflits (assez probables pour la plupart des données).

Le résultat de la fusion sur l'exemple 1 est donné dans le tableau 3.

Notons que le résultat de l'aggrégation peut comporter des éléments focaux qui ne sont pas des intervalles (par exemple, $\{\theta_2, \theta_5\}$). De tels ensembles proviennent de contradictions dans

1. Si $0 \leq \alpha_1 \leq \ldots \leq \alpha_M = 1$ sont les valeurs d'un ensemble flou μ, alors, pour $i = 1, \ldots, M$, $m(\{\theta | \mu(\theta) \geq \alpha_i\}) = \alpha_i - \alpha_{i-1}$.

Sous-ensemble	Masse globale de croyance
$\{\theta_3\}$	0.05
$\{\theta_2, \ldots, \theta_5\}$	0.21
$\{\theta_1, \ldots, \theta_5\}$	0.04
$\{\theta_2, \theta_4, \theta_5\}$	0.20
$\{\theta_1, \theta_2, \theta_3, \theta_5\}$	0.05
$\{\theta_2, \theta_5\}$	0.25
$\{\theta_2, \theta_3, \theta_5\}$	0.2

Tableau 3 – Exemple de sous-ensembles et masses de croyance résultant de la fusion

les informations relatives aux données, et traduisent le fait qu'une donnée peut être soit fiable, soit non-fiable, les informations qu'on possède sur cette dernière ne permettant pas de trancher. Sur le principe, la présence de tels ensembles focaux n'est pas gênante, puisqu'ils sont disjonctifs (une seule des valeurs est la "vraie") et que l'agrégation par SMC vise justement à résoudre au mieux les inconsistances dans l'information. Notons également que des informations fortement conflictuelles génèreront, après agrégation, une information assez imprécise (càd des ensembles focaux larges).

4.3 Résumé et ordonnancement

Afin de résumer l'information contenue dans m_g, nous proposons de calculer les espérances inférieure $\underline{\mathbb{E}}_g(f_\Theta)$ et supérieure $\overline{\mathbb{E}}_g(f_\Theta)$ de la fonction $f_\Theta : \Theta \rightarrow \mathbb{R}$ telle que $f_\Theta(\theta_i) = i$ (chaque élément recevant son rang comme valeur). L'espérance inférieur peut se calculer comme suit :

$$\underline{\mathbb{E}}_g(f_\Theta) = \sum_{A \subseteq \Theta} m(A) \min_{\theta \in A} f_\Theta(\theta),$$

l'espérance supérieure étant obtenue en replaçant le minimum par un maximum. Supposons maintenant qu'un ensemble $D = \{e_1, \ldots, e_d\}$ de d tableaux de données soient disponibles, et que la fiabilité sur chacune d'elles soit décrite par des fonctions de croyances $m_{g_1}, \ldots, m_{g_d}$. Une utilisation pertinente de ces résultats est de chercher à ordonnancer les données par fiabilité décroissante pour les présenter à l'utilisateur. Outre l'intérêt immédiat de ce tri, il peut également permettre de valider les résultats obtenus, sous réserve que, pour chacune des données, une estimation (e.g., donnée par un expert) soit disponible.

Une fois les fiabilités estimées, les masses de croyance obtenues sur les différents sous-ensembles ne permettent pas un classement immédiat. Cependant, la fiabilité d'une donnée i étant résumée par un intervalle $[\underline{\mathbb{E}}_{g_i}(f_\Theta), \overline{\mathbb{E}}_{g_i}(f_\Theta)]$, il apparaît naturel d'utiliser un ordre partiel pour arranger les données par groupe , i.e., construire une partition $\{D_1, \ldots, D_O\}$ ordonnée de $\{e_1, \ldots, e_d\}$, où D_1 correspond aux données les plus fiables, D_O aux moins fiables. Nous proposons d'utiliser la relation d'ordre partielle $\leq_\mathbb{E}$ telle que $e_i \leq_\mathbb{E} e_j$ ssi $\underline{\mathbb{E}}_{g_i}(f_\Theta) \leq \underline{\mathbb{E}}_{g_j}(f_\Theta)$ et $\overline{\mathbb{E}}_{g_i}(f_\Theta) \leq \overline{\mathbb{E}}_{g_j}(f_\Theta)$ ($e_i <_\mathbb{E} e_j$ quand au moins une des deux inégalités est stricte). Pour un sous ensemble quelconque $F \subseteq \{e_1, \ldots, e_d\}$, nous noterons $opt(\mathbb{E}, F)$ l'ensemble des données optimales, c'est-à-dire non dominées au sens de $\leq_\mathbb{E}$:

$$opt(\mathbb{E}, F) = \{e_i \in F | \ \nexists e_j \in F, \ t.q. \ e_i <_\mathbb{E} e_j\}.$$

Nous pouvons maintenant proposer une partition $\{D_1, \ldots, D_O\}$ définie récursivement de la façon suivante :

$$D_i = opt(\mathbb{E}, (\{e_1, \ldots, e_d\} \setminus \bigcup_{J=0}^{i-1} D_j)), \quad (2)$$

avec $D_0 = \emptyset$. Cette méthode correspond à conserver à chaque itération l'ensemble des données optimales, et à supprimer ces dernières de l'ensemble considéré à l'itération suivante. Elle est résumée par l'algorithme 2. Notons que l'ordonnancement pourrait se faire dans l'ordre inverse, i.e. placer dans D_O toutes les données qui n'en dominent pas d'autres, les supprimer de $\{e_1, \ldots, e_d\}$, et itérer l'opération jusqu'à obtenir une partition, les données à la fiabilité mal connue se retrouveraient alors dans le bas du classement, alors que l'ordre induit par la partition (2) les place en haut du classement.

Algorithme 2 : Ordonnancement des données par fiabilité

Input : $\{e_1, \ldots, e_d\}, m_{g_1}, \ldots, m_{g_d}$
Output : Partition $\{D_1, \ldots, D_O\}$
F = D=$\{e_1, \ldots, e_d\}$; j=1;
for $i = 1, \ldots, d$ **do**
  Calculer $[\underline{\mathbb{E}}_{g_i}(f_\Theta), \overline{\mathbb{E}}_{g_i}(f_\Theta)]$
while $F \neq \emptyset$ **do**
  foreach $e_i \in F$ **do**
    if $\nexists e_j \in F$ *t.q.* $e_i \leq_{\mathbb{E}} e_j$ **then**
      Mettre e_i dans D_j
  $F = F \setminus D_j$;
  $j = j + 1$

5 Conclusion et perspectives

Nous nous sommes intéressés à l'estimation de la fiabilité d'un tableau de données extrait automatiquement d'une publication scientifique. Nous avons proposé un certain nombre de critères organisés en sous-groupes, ce qui permet de les structurer dans une ontologie de domaine. L'intérêt pour cette problématique est né d'un besoin pratique d'affecter un indice de fiabilité à des données d'un entrepôt ouvert sur le Web. Tout en conservant une démarche très générique, nous nous sommes attachés au cas pratique de l'entrepôt *Sym'Previus*, qui accueille des données de microbiologie prévisionnelle, en vue de les intégrer dans des modèles de simulation.

Nous avons élaboré une méthode basée sur la théorie des croyances, avec un opérateur de combinaison basé sur les sous-ensembles maximaux cohérents, qui présente l'avantage de gérer le conflit entre croyances via la réconciliation, et de fournir une explication du résultat via la présentation à l'utilisateur des SMC. Nous avons également proposé des outils pratiques pour résumer et ordonancer l'information fournie par le résultat de ces combinaisons, dans le souci d'offrir un réponse compréhensible par l'utilisateur. Ce travail ouvre plusieurs perspectives : généralisation à d'autres domaines, étude de la notion de pertinence, prise en compte de sources multiples.

Références

[1] J.-C. Augustin, V. Zuliani, Cornu, and G. M. Growth rate and growth probability of listeria monocytogenes in dairy, meat and seafood products in suboptimal conditions. *Journal of Applied Microbiology*, (99) :1019–1042, 2005.

[2] P. Buche, J. Dibie-Barthélemy, and H. Chebil. Flexible sparql querying of web data tables driven by an ontology. In *FQAS*, volume 5822 of *Lecture Notes in Computer Science*, pages 345–357, 2009.

[3] D. Dubois, H. Fargier, and H. Prade. Multi-source information fusion : a way to cope with incoherences. In Cepadues, editor, *Proc. of French Days on Fuzzy Logic and Applications (LFA)*, pages 123–130, La rochelle, 2000. Cepadues.

[4] Y. Gil and D. Artz. Towards content trust of web resources. In *WWW'06 : Proceedings of the 15th international conference on World Wide Web*, pages 565–574, New York, NY, USA, 2006.

[5] G. Hignette, P. Buche, J. Dibie-Barthélemy, and O. Haemmerlé. Fuzzy annotation of web data tables driven by a domain ontology. In *ESWC*, volume 5554 of *Lecture Notes in Computer Science*, pages 638–653, 2009.

[6] H. Prade. A qualitative bipolar argumentative view of trust. In *SUM '07 : Proceedings of the 1st international conference on Scalable Uncertainty Management*, pages 268–276, Berlin, Heidelberg, 2007. Springer-Verlag.

[7] G. Shafer. *A mathematical Theory of Evidence*. Princeton University Press, New Jersey, 1976.