Incertain et inconnu, deux facettes de la cotation

Adrien Revault d'Allonnes, Herman Akdag, Bernadette Bouchon-Meunier

Incertain et inconnu, deux facettes de la cotation

Adrien Revault d'Allonnes, Herman Akdag, and Bernadette Bouchon-Meunier

Laboratoire d'Informatique de Paris VI – LIP6
4 place Jussieu, 75005 Paris
{Adrien.Allonnes, Herman.Akdag, Bernadette.Bouchon-Meunier}@lip6.fr

Résumé : La génération automatique de connaissances s'assortit généralement d'une mesure de confiance. Les systèmes d'apprentissage évaluent leurs performances en fonction des standards et de leurs pertinences. Les outils de recherche d'informations classent leurs résultats selon diverses stratégies, en fonction du contexte d'utilisation.

Cette nécessité émane autant des algorithmes, des modèles que des faits eux-mêmes. Cependant, la majorité des degrés de confiance affectés à des informations le sont de manière globale.

Nous percevons la cotation comme la projection de différentes dimensions d'incertitude ou d'imperfections sur la donnée elle-même.

Pour être utile, la cotation doit être compréhensible. Nous nous proposons donc de focaliser notre attention sur la représentation de la cotation. Pour la favoriser, nous proposons de distinguer la cotation indéterminée de son évaluation impossible.

Mots-clés : Cotation d'information, incertitude, logique multivalente, évaluation impossible.

1 Introduction

Le traitement automatique de l'information se généralise avec l'accès ubiquitaire à des masses de données. Pour pouvoir servir dans des situations délicates comme l'aide à la décision ou la production de données sensibles, les informations émanant de ces traitements doivent s'accompagner d'un coefficient reflétant la confiance qui peut leur être délivrée.

Cette confiance pourra, selon le cas de figure, intégrer l'incertitude de l'information (le fait décrit est sûr), sa qualité (complétude, précision, ...), la crédibilité de son mode de production ou tout autre combinaison de critères d'intérêt. La projection de l'évaluation de ces critères sur l'information considérée est ce que nous appelons **la cotation de l'information**.

Pour être utile, la cotation d'une information doit être interprétable par son destinataire. Elle doit donc représenter des notions pertinentes et cohérentes. Nous allons voir que la recherche sur la mesure de l'incertitude propose des outils lisibles de qualification de faits.

Ce qui nous intéresse particulièrement ici est la distinction que nous faisons entre un fait dont la cotation ne peut être calculée d'un autre dont la cotation n'est ni positive ni négative. Cette notion, initialement confuse dans le modèle que nous employons, a fini par disparaître. Nous proposons donc de la réintégrer, en la clarifiant.

2 Représentation de la cotation

L'incertitude des informations produites et manipulées par les systèmes automatiques est un sujet largement étudié par diverses communautés scientifiques. Les théories de manipulation de l'incertitude proprement dite (Bouchon-Meunier, 2007), les méthodes statistiques (National Institute of Standards and Technology, 2003) et les travaux étudiant la discipline plus spécifique de la qualité des données et de l'information (Batini & Scannapieco, 2006) proposent nombre d'outils pour sa modélisation et sa manipulation.

Parmi ces travaux, on retiendra plus particulièrement ceux qui s'évertuent à mettre en avant la lisibilité de leurs qualificatifs. Zadeh (2002), par exemple, étend ses travaux sur les variables linguistiques (Zadeh, 1975a,b) et développe la possibilité de la manipulation d'opérations sur des concepts et relations formulés avec des mots et de raisonnement en langage naturel.

Notre but étant moins de raisonner que de représenter la cotation, nous parlerons ici plus particulièrement des méthodes symboliques, également issues des logiques non-classiques. Celles-ci permettent, en effet, d'exprimer l'incertitude liée à une information selon une échelle discrète et lisible, associée à des termes expressifs.

2.1 La logique multivaluée pour exprimer la cotation

La manipulation de degrés lisibles, par opposition à des statistiques numériques, permet à un utilisateur, novice ou non, d'assimiler rapidement le doute formulé, et d'agir en conséquence. Pour être efficace, ces degrés doivent néanmoins répondre à certaines obligations, autant d'un point de vue théorique que sémantique. Parmi ces prérequis, on retiendra :

1. la comparabilité des degrés
2. leur combinaison systématique et interprétable
3. leur correspondance aux intuitions qu'ils prétendent représenter

La logique multivaluée propose d'utiliser des degrés de vérité d'un ensemble $\mathcal{L}_M = \{\tau_0, \ldots, \tau_{M-1}\}$ (où τ_0 est interprété comme faux et τ_{M-1} comme vrai), totalement ordonné et assure donc le premier point. Ces degrés de vérité sont associés à des termes (cf. tableau 1), favorisant ainsi la lisibilité.

Le cadre formel de ces logiques offre des opérateurs de combinaison (Seridi & Akdag, 2001) permettant d'assurer la cohérence des résultats. Dans Revault d'Allonnes *et al.*

Crédibilité	Degré
Totalement improbable	τ_0
Plutôt improbable	τ_1
Possible	τ_2
Plutôt probable	τ_3
Extrêmement probable	τ_4

TAB. 1 – Un exemple de degrés de vérité dans $\mathcal{L}_5$

(2007), nous avons proposé un formalisme de combinaison de degrés de vérité qui intégrait la fiabilité de la source et permettait de modéliser différentes stratégies de l'utilisateur.

Notre propos ici est de pallier à un des manques de la logique multivaluée : la différence entre l'inconnu et l'incertain.

2.2 Représenter l'inconnu : un degré incertain

Dans Besombes & Revault d'Allonnes (2008) nous proposions un procédé de calcul de la cotation. Bien qu'il ne s'y agisse pas de logique multivaluée, ces travaux manipulaient des degrés discrets, lisibles. Parmi eux en figurait un essentiel : « *La confiance ne peut être estimée* ».

Ce degré est capital dans l'expression de l'incertitude. Il témoigne de la différence entre une confiance moyenne, ni positive ni négative, et l'impossibilité d'évaluer un critère. On voit, d'ailleurs, que nombre de théories de l'incertitude l'intègrent. Il est équivalent, en logique possibiliste, à une possibilité totale et une nécessité nulle. De même, en théorie des sous-ensembles flous, lorsque l'incertitude est totale la seule appartenance certaine est à l'ensemble de discernement (cf. fig. 1). Quand rien n'est connu sur un événement, toutes les alternatives sont équiprobables.

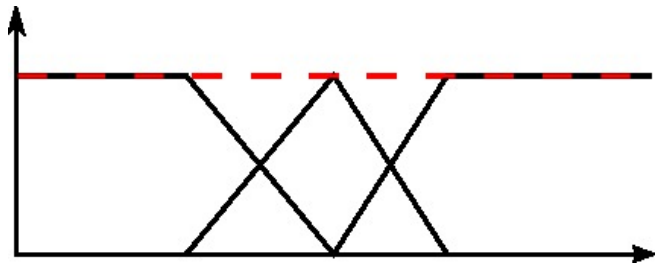

FIG. 1 – Incertitude totale sur une partition floue (*en pointillés*)

Pourtant, la logique multivaluée ne permet pas de représenter cet état de fait. Chez Łukasiewicz (1920) et Post (1921), le degré ajouté à la logique classique cumulait ces rôles. Dubois (2010) nous explique comment dès l'origine cette confusion entre valeur de vérité

d'une proposition et interprétation épistémique a parasité la théorie. Par la suite, il nous montre que toutes les logiques multivaluées, continues ou discrètes, pâtissent de cette confusion résolue par la distinction entre nécessité et possibilité.

Bien qu'à notre avis la modélisation de l'incertitude en théorie des sous-ensembles flous reste pertinente, nous ne pouvons que constater que dans $\mathcal{L}_3$, la distinction entre l'inconnu et l'incertain est ténue. De même, les extensions ultérieures de la logique symbolique confondent généralement l'inconnu et l'indéterminé. Nous proposons donc un nouveau degré de vérité, 'tau inconnu', noté $\tau_?$. Comme nous le présentions dans Besombes & Revault d'Allonnes (2008), ce degré représentant une connaissance nulle, il est de nature différente des autres éléments de $\mathcal{L}_.$ Tau inconnu aura donc les propriétés suivantes :

1. $\neg\tau_? = \tau_?$
2. $\forall\tau_\alpha \in \mathcal{L}_M, \tau_? \odot \tau_\alpha = \tau_\alpha$
3. $\tau_? \odot \tau_\alpha = \tau_? \Leftrightarrow \tau_\alpha = \tau_?$[1]

De toute évidence, tau inconnu ne préserve pas l'ordre total de $\mathcal{L}_M$. Il est également important de noter que $\tau_?$ est un élément neutre de tous les opérateurs. On démontre également aisément que malgré son irrégularité, $\tau_?$ préserve les principales propriétés des degrés de vérité, notamment celles de la logique classique :

1. Involution de la négation : $\neg\neg\tau_? = \tau_?$
2. Lois de de Morgan : $\neg(\tau_? \wedge \tau_\alpha) = \neg\tau_? \vee \neg\tau_\alpha$ et $\neg(\tau_? \vee \tau_\alpha) = \neg\tau_? \wedge \neg\tau_\alpha$

Les seuls axiomes de l'algèbre de Boole que $\tau_?$ ne préserve pas sont :

1. Lois d'identité : $\tau_M \wedge \tau_? = \tau_M$ et $\tau_0 \vee \tau_? = \tau_0$
2. Lois d'inversion : $\neg\tau_? \odot \tau_? = \tau_?$[1]
3. Lois d'absorbtion : $\tau_? \vee (\tau_? \wedge \tau_\alpha) = \tau_\alpha$ et $\tau_? \wedge (\tau_? \vee \tau_\alpha) = \tau_\alpha$

Si le degré de vérité que nous proposons ne vérifie pas toutes les propriétés des autres éléments de $\mathcal{L}_M$, il n'en reste pas moins qu'il permet de représenter un aspect capital de la cotation qui manquait cruellement à la logique multivaluée. Bien que la comparabilité des degrés de vérité était la première des contraintes que nous voulions imposer à notre méthode de représentation de la cotation, nous pensons que le gain d'expressivité de l'ajout de tau inconnu compense la perte.

En effet, tau inconnu n'a jamais besoin d'être comparé à un autre degré. Si, à l'issue d'un processus de production d'une information, aucun des critères d'évaluation de la confiance n'a pu être établi, il semble raisonnable que la cote finale ne soit pas une confiance moyennement indéterminée. En revanche, dès qu'une des étapes du procédé permet de formuler une hypothèse sur la cotation, les évolutions successives resteront déterminées.

De ce fait, si le score final est moyen, l'utilisation de la cotation pour ordonner les informations reprendra tout son sens. À l'inverse, une information cotée en sortie de système $\tau_?$ sera nécessairement marquée comme information à vérifier.

Afin de préserver les propriétés mathématiques de la logique symbolique, nous proposons donc de représenter la cotation sur $\mathcal{L}_M \cup \{\tau_?\}$.

[1] Où $\odot$ représente un opérateur quelconque sur les degrés de vérités.

3 Conclusion

Nous concevons la cotation comme un coefficient de confiance associé à une information. Ce coefficient témoigne à la fois de l'incertitude du fait décrit par l'information, de la fiabilité de la source et du processus l'ayant produite, de la qualité du modèle et d'autres facteurs éventuellement mesurables qualifiant l'information. Afin de favoriser l'exploitation de la cotation par les utilisateurs, nous souhaitons que ce score soit exprimé de manière interprétable et justifiable.

Nous avons proposé ailleurs un processus d'évaluation de la cotation. Nous cherchons maintenant à la représenter efficacement. Pour ce faire, nous nous servons de l'attirail offert par les logiques symboliques. Confronté à une limitation de notre formalisme de choix, nous proposons d'assouplir ses contraintes afin de représenter la nuance entre l'impossibilité d'estimer la cotation et une valeur neutre. Nous vérifions dans quelle mesure nos assouplissements remettent en question le formalisme et concluons que le gain en expressivité pour la cotation est suffisamment important pour les justifier.

Références

BATINI C. & SCANNAPIECO M. (2006). *Data Quality*. Springer.

BESOMBES J. & REVAULT D'ALLONNES A. (2008). An Extension of STANAG2022 for Information Scoring. In *Proceedings of the 11th International Conference on Information Fusion*, p. 1635–1641, Koeln, Germany.

BOUCHON-MEUNIER B. (2007). *La logique floue*. Presse Universitaire de France.

DUBOIS D. (2010). *Degrees of Truth, Ill-Known Sets and Contradiction*, In B. BOUCHON-MEUNIER, L. MAGDALENA, M. OJEDA-ACIEGO, J.-L. VERDEGAY & R. R. YAGER, Eds., *Foundations of Reasoning under Uncertainty*, volume 249 of *Studies in Fuzziness and Soft Computing*, p. 65–83. Springer.

ŁUKASIEWICZ J. (1920). O logice trójwartościowej (On three-valued logic). *Ruch Filozoficzny*, **5**, 170–171.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (2003). *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH.

POST E. L. (1921). Introduction to a General Theory of Elementary Propositions. *American Journal of Mathematics*, **43**, 163–185.

REVAULT D'ALLONNES A., AKDAG H. & POIREL O. (2007). Trust-moderated information-likelihood. A multi-valued logics approach. In *Computation and Logic in the Real World, CiE 2007*, Sienna, Italy.

SERIDI H. & AKDAG H. (2001). Approximate Reasoning for Processing Uncertainty. *Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII)*, **5**(2), 110–118.

ZADEH L. A. (1975a). The Concept of a Linguistic Variable and its Application to Approximate Reasoning – I. *Information Sciences*, **8**(3 & 4), 199–249 & 301–357.

ZADEH L. A. (1975b). The concept of a linguistic variable and its application to approximate reasoning – III. *Information Sciences*, **9**(1), 43–80.

ZADEH L. A. (2002). From computing with numbers to computing with words – From manipulation of measurements to manipulation of perceptions. *International Journal of Applied Mathematics and Computer Science*, **12**(3), 307–324.