

Évaluation sémantique d'informations symboliques : la cotation

Adrien Revault d'Allonnes

▶ **To cite this version:**

Adrien Revault d'Allonnes. Évaluation sémantique d'informations symboliques : la cotation. Intelligence artificielle [cs.AI]. Laboratoire d'informatique de Paris 6 [LIP6]; Paris 6, 2011. Français. NNT : . tel-01559975

HAL Id: tel-01559975

https://univ-paris8.hal.science/tel-01559975

Submitted on 11 Jul 2017

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.

Public Domain

UPMC SORBONNE UNIVERSITÉS

École Doctorale d'Informatique, Télécommunication et Électronique de Paris

Évaluation sémantique d'informations symboliques : la cotation

THÈSE

présentée et soutenue publiquement le 11 juillet 2011

pour l'accession au grade de

Docteur de l'Université Pierre et Marie Curie – Paris 6

(spécialité informatique)

par

Adrien Revault d'Allonnes

Composition du jury

Rapporteurs :	Salem Benferhat Laurence Cholvy
Examinateurs :	Herman Akdag Bernadette Bouchon-Meunier Philippe Capet Marie-Jeanne Lesot Olivier Poirel
Invité :	Michel Goya

Office National d'Étude et de Recherche Aérospatiales – Laboratoire d'Informatique de Paris 6

ONERA THE FRENCH AEROSPACE LAB

LIP6

Résumé

La confiance en une information est une mesure de la foi qu'on peut lui apporter, c'est-à-dire qu'elle indique combien on peut la croire. Si la confiance constitue un sujet de recherche très étudié, dans la plupart des cas sa mesure témoigne plus de la qualité de la production de l'information sur laquelle elle porte qu'elle ne révèle si l'on peut s'y fier. Comme on ne se satisfait pas du fait qu'une nouvelle émane d'une source sûre pour la croire, l'évaluation automatique de la confiance nécessite un modèle plus riche, capable de décrire pour quelles raisons son objet est ou n'est pas crédible. C'est à cette problématique que nous nous sommes attaché dans nos travaux.

Après avoir étudié la perception d'un mode de représentation de la confiance, nous proposons d'aborder séparément son expression et le processus gouvernant son établissement, de distinguer la cote de la cotation. Nous étudions les prérequis à la définition de la cote et en déduisons les dimensions incontournables à sa construction. Nous en proposons un découpage permettant de regrouper les différents critères d'évaluation selon leur objet et leur influence et assurant leur indépendance et leur non-redondance. Nous nous attachons également à assurer la lisibilité des mesures participant à l'évaluation de la cote en proposant de les apprécier sur des échelles discrètes explicitées d'étiquettes linguistiques.

Une fois cette sélection des dimensions effectuée, nous nous posons la question de leur combinaison pour modéliser le processus d'établissement de la confiance. Afin de répondre à ce problème, nous proposons une philosophie de l'intégration des dimensions à la confiance : nous façonnons une architecture à la cotation. Nous fournissons à cette architecture une représentation sous la forme d'une chaîne de cotation, celle-ci mettant en avant l'ordre de prise en compte des dimensions d'intérêt et leur influence sur la hausse ou la baisse de la cote. Nous montrons également comment la souplesse de notre modèle permet de représenter différentes postures de crédulité de l'utilisateur, adaptabilité essentielle à la modélisation de principes subjectifs.

Suite à ces définitions, nous proposons une formalisation théorique du procédé de cotation ainsi que de la cote, son mode d'expression. Exploitant l'expressivité de la logique multivaluée, nous choisissons ce formalisme pour exprimer nos propositions. Afin de préserver la nuance importante entre l'impossibilité de mesurer et la mesure neutre, nous l'étendons en lui adjoignant un nouveau degré de vérité. Dans ce cadre de logique symbolique étendue, nous définissons les opérateurs de combinaison permettant de représenter l'ensemble de nos propositions et formalisons la modélisation de la crédulité.

Nous considérons ensuite la mise en pratique de notre modèle dans l'extraction et la cotation d'informations symboliques. Dans un premier temps nous examinons la transposition de la cotation au problème de l'extraction de connaissances à partir de textes. Nous détaillons successivement la cotation de l'extraction d'informations, puis celle de leur fusion, en examinant pour chacune la transposition des dimensions qui la constituent. Nous implémentons, ensuite, un démonstrateur pour la mise en œuvre de nos propositions. Enfin, nous appliquons modèle et démonstrateur à un cas réel d'extraction et de cotation de réseau social.

Mots-clés: Cotation, confiance, qualité de l'information, fusion d'informations, logique multivaluée, extraction d'informations, extraction de réseaux sociaux

Abstract

Confidence in information should represent how far one can believe it, how much faith to put in it. Trust is a thriving field of study yet, in general, it tends to measure quality of the process responsible for producing the information rather than advise on whether to believe it or not. In the same way that hearing a fact from a trustworthy source is insufficient to fully believe it, automatic evaluation of trust in an information requires a rich model capable of explicitly puting forward why what it qualifies should or should not be believed. This is the problem we have tackled in our work.

From a careful study of an existing representation of confidence, we choose to split the problem in two : the encoding of trust, i.e. how it is represented, and the rules governing its appraisal, i.e. how it is evaluated. We derive the quintessential dimensions participating in the building of trust from the prerequisites imposed on the definition of its encoding. We offer a categorisation of these dimensions which gathers the evaluated criteria according to their object and influence and thus ensures their independence and non-redundancy. We also take great care of ensuring the readability of the measures involved in the assessment by proposing their expression along discrete scales made explicit through the use of linguistic labels.

After these dimensions have been selected, we can address the problem of their combination to model the trust-building process. We solve this problem by proposing a philosophy of integration fo the dimensions, that is, we shape the architecture of information scoring. We provide this architecture with a representation as a scoring-chain which highlights the order in which dimensions are considered and the influence they have on the increase or decrease of the confidence evaluation. We also show how the flexibility of our model can be used to represent different user gullibility-postures, an essential adaptability for the modeling of subjective matters.

Once these definitions are set, we propose a theoretical formalisation of the scoring process and of its expression, the score. Using the expressiveness of multivalued logics, we choose to set our solutions in this formalism. To reintroduce the important distinction between impossibility of measure and a neutral, yet expressed, measure, we extend this formalism by adding a new truth-degree. Within this new framework of an extended symbolic logic, we define combination operators to represent the entire collection of proposals we offered and formalise credulity-modeling.

We then consider the implementation of our model in the extraction and scoring of symbolic information. We first examine the transposition of information scoring to the problem of knowledge extraction from text. We describe successively the scoring of information extraction and that of their fusion, examining for both how the scoring dimensions translate. We then develop a prototype for implementing our model. Finally, we apply both model and prototype to a real-world usecase consisting of the extraction and scoring of a social-network from a corpus of published texts.

Keywords: Information scoring, trust, information quality, information fusion, many-valued logic, information extraction, social-network extraction

Si vous me voyiez, Grand'mère !

The mind is not a vessel to be filled but a fire to be kindled.
Plutarch

Granny, you'd have been chuffed.

Il faut inventer en même temps que l'on apprend.
Plutarque

Remerciements

J'ai rêvé de nombreuses années de recherche. Je tiens à remercier particulièrement deux personnes de m'avoir permis de réaliser ce rêve. Bernadette Bouchon-Meunier m'a reçu, alors que je ne faisais encore que rêver, a peut-être vu quelque chose chez moi qui l'a intriguée et, après que le temps lui a permis de confirmer son intuition, m'a prodigué un indéfectible soutien, une aide scientifique, un réconfort intellectuel et émotionnel et m'a fait confiance pour devenir ce que je voulais être, dans des conditions que je n'osais imaginer. Je suis heureux d'avoir pu – et de pouvoir encore – collaborer avec Bernadette et d'avoir mérité sa confiance, presqu'autant que je le suis qu'elle ait accepté de présider le jury de ma thèse.

Mon aventure de chercheur a aussi été façonnée par mon directeur de thèse, Herman Akdag. Herman a accompagné mes premiers pas dans le milieu et m'a enseigné qu'il ne suffit pas de réfléchir ou de trouver, encore faut-il transmettre. Il m'a montré comment passer de mes feuilles pleines d'équations à des articles bien articulés qui, en plus d'exposer des résultats, racontaient ce qu'une démonstration veut dire. Durant tout le reste de notre cohabitation, Herman m'a témoigné une immuable assurance de la valeur qu'il me voyait. Parfois mal assuré dans mon nouvel élément, j'ai toujours pu compter sur son estime et sa conviction de la qualité mon travail.

Je remercie aussi chaleureusement Laurence Cholvy d'avoir accepté de rapporter mon travail. Son incontournable compétence sur le sujet de la cotation et des logiques non-classiques font de ses commentaires une critique experte, expertise qui m'a permis de défendre avec conviction mes propositions et ma façon de voir le problème.

Toute ma reconnaissance va également à Salem Benferhat, second rapporteur de ma thèse. Ses commentaires avisés sur les aspects théoriques de mes propositions me permettent d'envisager de nouveaux pans d'études à mener et m'encouragent dans leur exploitation. Peut-être que ces extensions formelles à venir nous permettront de nouvelles collaborations.

J'ai partagé nombre de discussions avec Olivier Poirel, tant sur le sujet sur lequel il m'encadrait à l'ONERA que sur n'importe quel autre, ceux qui le connaissent savent qu'Olivier parle aisément de n'importe quel sujet. Je le remercie pour sa disponibilité lorsque j'allais le voir et pour l'ensemble des vibrionnants échanges que nous avons eus.

Depuis que j'ai rejoint le Laboratoire d'Informatique de Paris 6, Marie-Jeanne Lesot m'a accordé une aide continue. Je me suis enrichi de nos discussions, tant sur le plan scientifique que littéraire ou culturel, et lui sais gré d'avoir accepté d'être examinatrice dans mon jury de thèse. Je rends également hommage à sa façon de partager sans compter ses capacités intellectuelles et humaines, reconnues de tous ceux qui la connaissent.

Mes remerciements sincères vont aussi à Philippe Capet, pour l'intérêt qu'il a montré pour mon travail, pour sa lecture attentive et l'ensemble des remarques qu'il m'a faites. Je me réjouis à l'idée que notre attachement partagé pour ce sujet nous permettra, peut-être, de collaborer.

Je remercie enfin le colonel Michel Goya d'avoir accepté de s'éloigner de son domaine de recherche habituel, l'histoire, pour donner son avis d'expert ès choses militaires sur mes travaux. Je me doute de l'effort qu'il a dû fournir et lui suis d'autant plus reconnaissant de l'intérêt qu'il y a trouvé.

Je remercie en vrac les personnes que j'ai cotoyées à l'ONERA, des thésitifs et futurs docteurs Antoine, Joseph, Évangeline, Anne-Marie et tous les autres, aux permanents dont Gilles, Fabrice et Fanfan. Je tiens plus particulièrement à remercier Jérome Besombes pour

tous les moments, bons et moins bons, passés ensemble. Sous ses airs d'éternel redresseur de torts Jérôme a une écoute et une compréhension des autres particulières qui, ajoutées à son pragmatisme et ses qualités scientifiques, en font un amer [1] confortable et rassurant et un ami... étonnamment attachant.

Au laboratoire, j'ai eu la chance d'être dans une équipe soudée et riche de personnalités variées. Je repense affectueusement à ceux qui nous ont quittés, Javier, Jason, Thomas (B & D), Thanh Hà, Tri Duc et les autres. Je remercie Maria et Christophe de m'avoir fait une place et Marcin de m'inclure à la conversation. Merci aussi à Frédéric P. pour son amicale surveillance de mes derniers pas. Un mot particulier pour ceux qui ont tant participé à me faciliter la vie Jacqueline, Ghislaine, Catherine, Boubou, Benoît et André. Le LIP6 ne serait pas le même sans eux.

La liste de mes proches qui m'ont soutenu et encouragé tout au long de ce périple est trop longue. Je ne peux, néanmoins, faire autrement que dire à mes parents comme leur aide et leur confiance étaient essentielles dans ce chemin. J'espère être à la hauteur des attentes de ma sœur et de mon frère, les remercie d'être venus à ma soutenance et d'avoir tempéré leur impatience aux derniers dîners de famille.

J'ai débuté ma reconversion suite à une remarque ironique d'une amie très chère. Après la soutenance de Marc-Antoine, j'expliquais à Claire combien je l'enviais et elle me répondit « C'est sûr que toi tu es trop con pour faire une thèse. » Je la remercie de me connaître assez pour savoir où sa remarque m'emmènerait. Je les remercie tous les deux et le reste de *la bande*, Thomas et Maytal, de m'avoir (presque) toujours supporté, dans les hauts et les bas. Je remercie aussi Juliette d'avoir été là au début et de n'avoir cessé de croire en moi. Je termine par les amis rencontrés depuis le début de ma thèse, patients, aimants et toujours là, Claire et Stefane, Lotfi et tous les autres.

Enfin, je dédie ce mémoire à celle qui m'a porté à bout de bras en nous inventant des trésors de force insoupçonnés, m'a poussé quand il fallait, m'a chéri quand j'en avais besoin et m'a aussi montré que je pouvais rêver à l'après. À Marie.

1. (2)**II.** AMER (*r* se prononce) n. m. XVIIe siècle. Dérivé du normand *merc*, « borne de séparation », emprunté du néerlandais.
Point fixe sur la côte ou en mer, visible de loin et servant de repère aux navigateurs. *Prendre ses amers. Ce phare et ce clocher servent d'amers.*
Titre célèbre : *Amers*, recueil de vers de Saint-John Perse (1957).
In : **Le dictionnaire de l'Académie** *(9ème édition)*

Table des matières

Introduction générale **1**

Cadre des travaux 3

Organisation du mémoire 4

I La cotation : contexte, problématique et état de l'art **5**

Introduction à la première partie **7**

1 Contexte et problématique **9**

1.1 Introduction 9

1.2 Contexte applicatif : évaluation pour le renseignement 10

1.2.1 Motivations 10

1.2.2 Constitution de connaissances : le cycle du renseignement 11

1.2.3 Des informations brutes au renseignement : la phase d'exploitation 12

1.2.4 En quête de certitudes insaisissables : l'évaluation de l'information 14

1.2.5 La cotation, une notion essentielle mais obscure 16

1.3 La cotation, un problème général 18

1.3.1 Comprendre la cotation 18

1.3.2 Formalisation du problème 19

1.4 Conclusion 21

2 Confiance, certitude et qualité comme influences pour la cotation **23**

2.1 Introduction 23

2.2 La confiance comme objet d'études 25

2.2.1 Aperçu de la confiance dans le cadre tactique 25

2.2.2 Aperçu de la propagation de confiance 28

2.3 Qualité des données 32

2.3.1 Domaines de recherche liés 32

2.3.2 Problématiques . . . 34
2.4 Représentations de l'incertitude . . . 36
2.4.1 Incertitude probabiliste . . . 37
2.4.2 Incertitude cognitive . . . 37
2.5 Fusion d'informations . . . 39
2.6 Conclusion . . . 40

II Nouveaux critères, méthode et formalisme pour la cotation 43

Introduction à la deuxième partie 45

3 Dimensions de la cotation 47
3.1 Introduction . . . 47
3.2 Évaluation générale de la source : la fiabilité . . . 50
3.3 Évaluation contextuelle de la source : la compétence . . . 52
3.4 Évaluation générale du contenu : la plausibilité . . . 54
3.5 Évaluation contextuelle du contenu : la crédibilité . . . 56
3.6 Expression globale de la confiance . . . 58
3.7 Conclusion . . . 59

4 Combinaison des dimensions : une architecture modulable 61
4.1 Introduction . . . 61
4.2 Architecture de la cotation : caractéristiques . . . 62
4.3 Architecture de la cotation : description . . . 64
4.3.1 Fiabilité de la source . . . 65
4.3.2 Compétence et plausibilité . . . 65
4.3.3 Crédibilité . . . 67
4.4 Personnalisation de la cotation : modélisation de crédulités variables . . . 67
4.4.1 Fiabilité de la source . . . 68
4.4.2 Compétence et plausibilité . . . 69
4.4.3 Crédibilité . . . 71
4.5 Discussion . . . 72
4.6 Conclusion . . . 73

5 Formalisation multivalente de la cotation 75
5.1 Introduction . . . 75
5.2 Formalisme multivalué . . . 76
5.2.1 Degrés de vérité . . . 77

5.2.2 Raisonnement : combinaison de degrés de vérité 79
5.3 Ignorance multivalente : proposition d'extension 80
5.3.1 Problématique 80
5.3.2 Proposition : introduction d'un degré additionnel 81
5.3.3 Discussion 83
5.4 Opérateur de combinaisons à la baisse 84
5.5 Opérateur de combinaisons de recoupement 87
5.5.1 Principes 88
5.5.2 Crédibilité immédiate 89
5.5.3 Crédibilité cumulée 93
5.5.4 Crédibilité, dynamique et flux d'informations : exemple illustratif 97
5.6 Conclusion 102

III Mise en pratique du modèle de cotation 103

Introduction à la troisième partie 105

6 Cotation de connaissances extraites de données textuelles 107
6.1 Introduction 107
6.2 Organisation de l'extraction de connaissances cotées 108
6.3 Intégration de la cotation 110
6.3.1 Cotation de l'extraction d'informations 110
6.3.2 Cotation de la fusion d'informations 112
6.4 Cadre expérimental 113
6.4.1 Données 114
6.4.2 Scénario d'usage 114
6.5 Conclusion 115

7 Mise en œuvre de la cotation pour l'extraction de réseau 117
7.1 Introduction 117
7.2 Modules d'extraction d'informations 118
7.2.1 Extraction d'informations géographiques 120
7.2.2 Extraction de noms propres 121
7.2.3 Extraction et déchiffrage d'acronymes 123
7.2.4 Extraction de relations 126
7.2.5 Module de fusion d'informations 129
7.3 Implémentation de la cotation 130

7.3.1 Cotation de l'extraction d'informations 130
7.3.2 Cotation de la fusion d'informations 132
7.3.3 Éditeur de stratégies 134
7.4 Exemple illustratif 135
7.4.1 Cotation de l'extraction d'informations 136
7.4.2 Cotation de la fusion d'informations 137
7.5 Problématique de l'évaluation 140
7.6 Conclusion 141

Conclusion générale et perspectives **143**

Bibliographie **151**

Table des figures **161**

Liste des tableaux **163**

ANNEXES **165**

A Types de renseignements **167**

B Définitions des étapes du cycle du renseignement, selon l'AJP 2.1 **169**

C Qualité de l'information, quelques définitions **171**

D Stratégie par défaut multivalente **177**

Introduction générale

« Dans **Le Figaro** : portrait du Général Rondot, la vedette du jour au procès Clearstream. Dans le monde de l'espionnage, c'est un maître. Arrestation du terroriste Carlos, libération des otages au Liban : ses faits d'armes sont dans toutes les mémoires. Surnom : "Max", comme Jean Moulin. C'est aussi l'un des meilleurs connaisseurs du Moyen-Orient [...]

Une des questions-clés, c'est comment un homme de sa trempe s'est laissé gruger par Imad Lahoud, présenté comme le faussaire des listings Clearstream... À sa décharge, écrit Stéphane Durand-Souffland, le général n'est pas la seule dupe de M. Lahoud, qui dit avoir personnellement rencontré le chef d'Al Qaïda dans le bureau d'un ministre libanais. A peine s'il ne prétend pas que le terroriste lui a laissé un double des clés de sa grotte. »

***Bruno Duvic**, revue de presse du lundi 5 octobre 2009, France Inter*

Pendant le déroulement du procès de l'affaire Clearstream 2, la question qui taraude le journaliste est comment une personne réputée fiable et aguerrie aux manipulations a-t-elle pu croire à une information peu plausible provenant d'une personne douteuse ? Nous voyons par là combien la question de la confiance est universellement présente dans les choix, décisions et appréciations de la vie courante. Il est également à noter l'influence admise, et ici déçue, de la réputation de la source sur la confiance faite à ce qu'elle dit. La suite montrera en fait, et sans juger d'une affaire pénale, que la confiance portée par le Général Rondot à Imad Lahoud et à ses informations était toute relative.

Comment se forme-t-on une opinion ? Par quel processus en arrivons-nous à être certains de notre avis ? Le procédé est simple et à peu près universel : trouver des sources sûres, croiser leurs allégations, puis en extraire le point de vue qui convient le mieux à nos convictions, nos connaissances et nos intérêts. Dès lors, des questions font surface. Comment estimer la sûreté d'une source ? Comment régler les éventuels conflits entre sources sûres ? À partir de combien de recoupements sommes-nous convaincus ? Ces certitudes sont-elles révisables ? Si dans le cas général ces notions sont variables d'une personne à l'autre, dans un cadre normalisé, elles doivent être étudiées. Si, de plus, on s'intéresse à l'automatisation

du calcul de ce degré de confiance pour des sujets sensibles, la formalisation du problème devient essentielle.

En ce qui concerne l'information générale, nombre de personnes avisées font confiance à ce qui s'appelle en général les médias et, dans certains milieux, les 'sources ouvertes'. De manière quotidienne des compendia de sources ouvertes sont effectués par des opérateurs spécialisés sur des sujets spécifiques. Les journalistes, représentants du quatrième pouvoir selon Tocqueville (1835), se chargent de synthèses et autres revues de presse à l'intention du reste du monde impatient. Les uns comme les autres appliquent ce même procédé : tout d'abord trouver l'information, puis la vérifier (Baud, 2002; Simon, 2005). S'assurer qu'elle provient d'une source sûre et compétente est la première étape. Trouver d'autres personnes, d'autres points de vue afin de l'étayer vient par la suite, afin de renforcer la confiance initiale (Borden & Harvey, 1998). Les scientifiques eux-mêmes cherchent à déterminer les références de qualité pour fonder leurs travaux (Chen et al., 2007).

La confiance est, donc, un sujet d'intérêt capital. Elle s'applique, dans son acception usuelle, aussi bien aux personnes (« Croyance spontanée ou acquise en la valeur morale, affective, professionnelle... d'une autre personne, qui fait que l'on est incapable d'imaginer de sa part tromperie, trahison ou incompétence ») et à soi-même (« assurance que l'on peut avoir en ses ressources propres ou en sa destinée »), qu'à un objet ou une information : « Crédit, foi [...] accordé à quelqu'un ou à quelque chose »[2]. Qu'il s'agisse de la confiance entre individus ou la conviction qu'une information est juste, toute décision l'intègre d'une manière ou d'une autre. Le traitement de sujets sensibles et les décisions qui en découlent augmentent encore la nécessité d'un indicateur de confiance sûr, compréhensible et révisable.

L'évaluation de la performance des algorithmes, techniques et méthodes informatiques est peuplée de scores de confiance. Ceux-ci ont pourtant en général pour objet la méthode plutôt que le résultat. S'il est admis que les traitements automatiques produisent incertitudes ou imprécisions, les mesures usuelles représentent la façon dont les données sont construites plutôt que de qualifier la confiance en l'information produite. S'il est généralement suffisant de savoir qu'un algorithme d'extraction est efficace huit fois sur dix ou que sa production est conforme aux attentes dans les mêmes proportions, dès que l'on s'intéresse à des données sensibles, à la base de décisions potentiellement néfastes, on souhaiterait disposer d'une confiance qualifiant l'information plutôt qu'un doute uniforme sur sa construction. Si, qui plus est, la méthode d'élaboration d'un tel indicateur de confiance est justifiée, lisible voire modifiable, l'utilisateur peut alors envisager d'apprendre à appré-

2. Nota : les définitions proposées ici proviennent du Trésor de la Langue Française, source de confiance s'il en est.

hender le système et utiliser ses indications dans le reste de son intervention.

Cadre des travaux

Notre travail de thèse se situe au cœur des multiples questions posées dans le contexte décrit ci-dessus et s'intéresse au processus d'établissement de la confiance, la *cotation*. Celle-ci décrit la méthode d'acquisition de la conviction qui permet de qualifier la foi à apporter à une information. L'objectif de nos travaux est de proposer une formalisation exploitable de ce processus.

Notre travail de thèse a été réalisé en collaboration entre l'Office National d'Études et Recherches Aérospatiales (ONERA) et le Laboratoire d'Informatique de Paris 6 (LIP6), financé initialement par une bourse ONERA puis par un emploi d'ingénieur de recherches à l'Université Pierre et Marie Curie (UPMC, Paris VI). Il se place dans le cadre de l'assistance aux services de renseignement militaire dans lequel la problématique de cotation est une constante : l'évaluation de la confiance en une information dans ce cadre est indispensable et se pose de manière accrue pour le traitement automatique d'informations symboliques provenant de sources ouvertes.

Nous répondons à l'objectif de formalisation de la cotation en nous attachant à satisfaire des critères de lisibilité, de facilité d'interprétation, de possibilité de justifier les résultats fournis et d'adaptabilité. Nos propositions embrassent trois niveaux, conceptuel, formel et pratique.

Notre analyse conceptuelle s'appuie sur l'étude de l'existant et le recueil de besoins qui nous conduisent à proposer de distinguer la manière de représenter la confiance, la *cote*, d'une part, et le processus permettant son établissement, la *cotation*, d'autre part. Pour la cote nous étudions les prérequis à sa définition, en déduisons les dimensions sur lesquelles elle doit reposer et nous attachons à assurer sa lisibilité par le choix de sa représentation. En ce qui concerne la cotation, nous proposons une architecture de la combinaison décrivant la projection sur la cote de la prise en compte des dimensions retenues.

Au niveau formel, nous faisons le choix d'une logique non-classique expressive, la logique multivalente, pour modéliser la cote et la cotation. Afin de préserver la nuance importante entre impossibilité de mesurer et mesure neutre, nous proposons une extension à ce cadre théorique, en lui ajoutant un nouveau degré de vérité. Nous proposons ensuite des représentations formelles dans ce cadre étendu pour chacune des propositions établies au niveau conceptuel.

Enfin, nous mettons en pratique nos propositions en les appliquant à un problème d'extraction et de cotation d'informations symboliques à partir de données textuelles. Nous étudions successivement la cotation de l'extraction d'informations, puis celle de leur fusion, en examinant pour chacune la transposition des dimensions qui la constituent. Nous avons développé un démonstrateur pour appliquer nos propositions à un problème réel d'extraction et de cotation de réseau social. Ce cadre expérimental soulève le problème spécifique de la multimodalité de la cotation, acteurs et relations étant qualifiés par une cote.

Organisation du mémoire

Ce mémoire est composé de trois parties. La première formalise les objectifs visés par nos travaux en les positionnant à la fois par rapport au cadre applicatif d'origine (chapitre 1) et par rapport aux problématiques scientifiques liées (chapitre 2).

La deuxième partie présente nos contributions conceptuelles puis formelles en trois chapitres. Le chapitre 3 est consacré à l'analyse conceptuelle de la cote, le chapitre 4 à celle de la cotation. Le chapitre 5 décrit leur formalisation théorique dans le cadre choisi de la logique multivalente étendue.

Enfin, la troisième partie présente la mise en œuvre de nos propositions. Le chapitre 6 présente la transposition du modèle à la cotation de connaissances extraites à partir de textes de façon générale. Le chapitre 7 considère le problème de l'extraction et cotation des acteurs et liens d'un réseau social.

Première partie

La cotation : contexte, problématique et état de l'art

Introduction à la première partie

La première partie de ce mémoire s'attache à formaliser l'objectif des travaux présentés ici, la modélisation de la cotation comme processus d'établissement de la confiance. Elle propose une mise en contexte de la problématique de la cotation partant de son cadre applicatif principal, le renseignement militaire. Après avoir décrit les usages de ce cadre, elle examine les problématiques scientifiques liées à leur formalisation puis présente les disciplines voisines de l'établissement de la confiance.

Le chapitre 1 détaille le processus de construction de connaissances dans le cadre du renseignement militaire, au sein duquel la cotation de l'information joue un rôle central. Il propose une analyse des usages s'appuyant sur des entretiens menés avec des professionnels du domaine. À partir de cette étude de l'existant, il établit les difficultés et aboutit à la formalisation du problème de la cotation tel que les propositions présentées dans la suite du mémoire s'attachent à le résoudre.

Le chapitre 2 étudie les problématiques scientifiques liées à l'établissement de la confiance. Il commence par rappeler diverses perceptions existantes de la cotation ainsi que différentes manières d'aborder la confiance. Au-delà de ces travaux traitant explicitement de confiance, ce chapitre présente la cotation comme se situant à l'intersection de trois domaines de recherche : la qualité des données, la représentation de l'incertitude et la fusion de données incertaines. Il étudie les outils que ces disciplines offrent à la modélisation de la cotation et finit de dresser le cadre dans lequel nous situons nos travaux.

Une armée sans agents secrets est un homme sans yeux ni oreilles.

Sun Tzu

L'Art de la Guerre

VI^{ème} siècle avant Jésus-Christ

Chapitre 1

Contexte et problématique

Résumé

Ce chapitre est la présentation et la mise en contexte de la cotation. La première partie offre un panorama des usages dans le renseignement militaire et elle décrit les difficultés de la modélisation de la confiance. La seconde partie aborde la cotation comme un problème plus général, puis présente les principes de la formalisation que nous proposons en réponse aux difficultés de sa modélisation.

1.1 Introduction

Nous présentons dans ce chapitre le cadre de nos travaux qui les a à la fois motivés et orientés. L'objectif de cette étude est de proposer une formalisation de l'établissement de la confiance. Nous débutons ce chapitre par une présentation du cadre applicatif de nos recherches, la cotation dans le renseignement militaire. À partir de cette problématique nous formalisons le problème général de la construction de la confiance à partir d'informations, applicable dans de multiples contextes.

Afin de replacer nos recherches dans leur contexte d'origine, nous présentons les principes généraux de la construction de renseignements. La problématique de l'évaluation de la confiance est une constante du domaine du renseignement où les informations sont sensibles et les décisions qu'elles induisent lourdes de conséquences. Ainsi, la *cotation d'informations* préexiste-t-elle sous une forme particulière dans le cadre du renseignement. Dans cette pro-

cédure d'acquisition de connaissances, nous détaillons en particulier la phase d'analyse où les données brutes sont enrichies puis fusionnées. Nous nous attardons plus spécifiquement sur l'évaluation des informations et approfondissons les principes de la cotation telle que pratiquée dans le domaine.

Après avoir décrit la procédure d'évaluation actuelle, nous exposons les différentes problématiques que nous identifions. De cette synthèse nous dégageons les points centraux de la formalisation que nous nous proposons d'établir. Dans la deuxième partie de ce chapitre nous posons la cotation comme un problème plus général et offrons une première esquisse des réponses que nous proposons dans la suite du mémoire aux questions essentielles identifiées précédemment.

1.2 Contexte applicatif : évaluation pour le renseignement

« Le monde de l'après-guerre froide cède très vite la place à un monde plus mobile, plus incertain et imprévisible, exposé à des vulnérabilités nouvelles. »

***Nicolas Sarkozy**, préface au Livre blanc sur la défense et la sécurité nationale (Commission du Livre Blanc, 2008)*

1.2.1 Motivations

Depuis la fin de la guerre froide et la disparition de l'affrontement des super-puissances, l'évolution des forces et des intérêts a entraîné une modification des conflits et des crises nationaux et internationaux. Les situations à surveiller sont devenues terrorisme et cybercriminalité, les missions le maintien de la paix et la sécurité civile. Ceci a donné lieu à un accroissement des besoins d'analyse amont et de prévision des situations à risques.

Ces évolutions et leur appréhension ont motivé nombre de recherches spécifiques : notamment les travaux de Mouillet (2005) et de Delavallade (2007) ou de Delavallade et al. (2007) sur la détection de conflits intra-étatiques, ceux de Couchariere (2010) traitant de la propagation d'incertitude dans la tenue de situations terrestres, ainsi qu'un projet retenu dans le cadre de l'appel annuel de l'Agence Nationale de la Recherche (2008) sur les « concepts, systèmes et outils pour la sécurité globale ».

L'évolution de ces besoins se traduit par un accroissement des investissements vers les services gouvernementaux de renseignement. Dans son communiqué de juin 2008, exposant le « Livre blanc de la défense et de la sécurité nationale » (Commission du Livre Blanc, 2008), le gouvernement français a annoncé axer de nouveaux efforts vers le développement

d'une nouvelle fonction stratégique « connaissance et anticipation [1] » et motivé le financement accru des services de renseignement.
Ces besoins ont fourni l'objet de cette thèse, réalisée à l'ONERA [2], en collaboration avec le Laboratoire d'Informatique de Paris 6, sur l'étude de l'évaluation de la confiance dans la manipulation de données sensibles. Dans la suite de cette section nous présentons le cycle du renseignement, en nous concentrant par étapes sur la phase de cotation.

1.2.2 Constitution de connaissances : le cycle du renseignement

L'activité de services de renseignement, qu'ils soient militaires, économiques, publics ou privés, est décrite comme la suite d'opérations par lesquelles des informations sont récoltées, assemblées puis enrichies et, enfin, mises à disposition du commanditaire (Ministère de la défense – Commandement de la Formation de l'Armée de Terre (CoFAT), 2001). À la livraison de ces réponses, il est possible que de nouvelles questions émergent. C'est pour cette raison que l'on parle de *cycle* du renseignement. Ce cycle comporte, selon les cultures et habitudes, un nombre variable d'étapes. La version présentée ici, à la figure 1.1, comporte les quatre étapes que l'on retrouve dans toutes ses formulations et que nous décrivons succinctement ci-dessous.

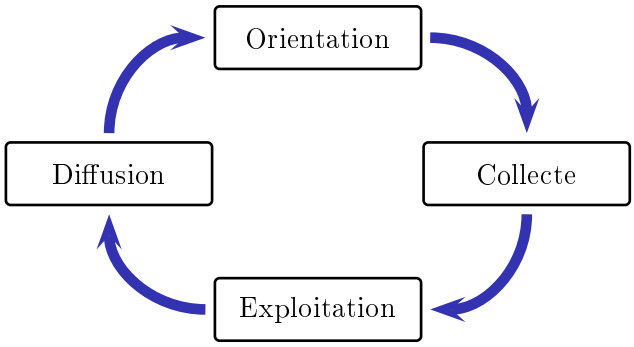


FIGURE 1.1 – Le cycle du renseignement

La phase initiale d'orientation est celle où l'utilisateur exprime le besoin. Dans le cadre du renseignement militaire, le commanditaire établit les zones d'intérêt, qu'elles soient géographiques ou thématiques (économiques ou politiques, par exemple) et le plan de collecte, c'est-à-dire la liste des services à solliciter. Le plan de travail ainsi établi fournit

1. « La fonction "connaissance et anticipation" sera prioritaire. J'ai décidé d'un effort massif, massif, d'investissement sur le renseignement [...] et qui bénéficiera aussi bien aux chefs militaires qu'aux décideurs politiques. » – Le président de la République, discours sur la Défense et la Sécurité Nationale, le 17 juin 2008.
2. EPIC sous tutelle du Ministère de la Défense

l'orientation de la recherche d'informations.

Dans un deuxième temps, la phase de collecte du renseignement, opérée par des agents humains des services du plan de collecte de l'étape précédente, commence par la recherche de sources pertinentes. On distingue généralement des sources documentaires, le renseignement d'origine source ouverte (ROSO ou OSINT, Open Source Intelligence), humaines (ROHUM ou HUMINT, Human Intelligence) et techniques telles que le ROEM, renseignement d'origine électro-magnétique répertoriant les émissions d'ondes, comme celles émises par les instruments de communication radio ou les radars. Une liste détaillée des différents types de renseignements est disponible à l'annexe A, page 167.
Le renseignement de sources ouvertes, qui nous intéresse plus particulièrement dans ce travail, regroupe toute information publiquement disponible : journaux, statistiques, rapports diplomatiques, recensements. Des besoins actuels de connaissances générales permettant l'analyse, éventuellement en dehors de conflits déclarés, de situations géopolitiques découle une nécessité de traiter des volumes toujours croissants de données. Cette constatation engendre à son tour la nécessité de systèmes automatiques ou semi-automatiques. C'est à l'utilisation de ce type de systèmes que nous nous intéressons.

Après l'extraction de données brutes, la phase d'exploitation intervient pour leur enrichissement. Il s'agit, d'une part, de fusionner les informations découvertes, généralement dans le but de les compléter ou de les confirmer, et, d'autre part, de les présenter dans un modèle interprétable par l'utilisateur. C'est ici que l'opérateur évalue les informations recueillies, les regroupe afin d'en extraire un maximum de connaissances et les recoupe avec des éléments déjà connus pour en faire des renseignements exploitables. Cette étape centrale au cycle du renseignement est celle où s'inscrit la problématique de l'évaluation de la confiance, la cotation, qui nous intéresse ici. La section 1.2.3 la décrit donc plus en détails afin d'en dégager les points que nous proposons de traiter par la suite.

Lors de l'étape finale de diffusion, les renseignements construits en réponse à l'interrogation initiale sont fournis au commanditaire ainsi qu'aux éventuels services concernés. Ces destinataires peuvent alors être amenés à réorienter la recherche, aboutissant à une nouvelle itération du cycle du renseignement.

1.2.3 Des informations brutes au renseignement : la phase d'exploitation

Une fois les informations initiales glanées, les services d'analyse interviennent afin de produire des réponses enrichies par leur travail de recoupement et de fusion. Cette phase d'exploitation est décrite comme l'étape de raffinage de l'information et de création de

renseignements exploitables et définie par l'AJP 2.1[3] comme « la production de renseignements à travers l'agrégation, l'évaluation, l'analyse, l'intégration et l'interprétation d'informations et/ou d'autres renseignements[4] » (North Atlantic Treaty Organization (NATO), 2002). Nous détaillons les étapes de ce raffinage dans la suite de cette section. Notons que nous distinguons, à dessein, corrélation d'informations – leur mise en relation – et fusion d'informations – la création d'un nouveau produit à partir d'informations corrélées.

L'agrégation décrit l'assemblage et la consignation de faits corrélés dans le but de faciliter la suite de l'exploitation. À cette étape intervient une nuance importante : la granularité de l'information. En effet, dans le cadre du renseignement, on distingue les informations brutes, ou données, des renseignements, interprétés et enrichis. Il convient donc, afin de mieux envisager la suite, d'introduire les deux définitions suivantes, issues de la littérature liée au renseignement mais que nous adaptons au cadre de l'extraction d'informations à partir de textes :

Définition 1 : Information	**Définition 2 : Renseignement**
Une information est un fait atomique, directement disponible dans le texte d'origine, dont on peut supposer qu'elle représente un fait du monde réel. Une information peut être, dans le monde réel, vraie, fausse ou affectée de tout autre degré de vérité intermédiaire.	Un renseignement est un agglomérat d'*informations* extraites du texte. C'est le produit du recoupement, de l'enrichissement et de la fusion d'informations, apporté en réponse à une requête. Un renseignement est une connaissance construite.

L'étape suivante, l'évaluation, est celle qui nous intéresse ici : lors de celle-ci, les informations sont assorties d'un degré qualifiant, selon les informations collectées, la confiance que leur créateur conseille de leur faire, ce degré est appelé leur *cotation*. Nous revenons dans la section suivante, la section 1.2.4, sur l'expression de cette cotation et sa définition actuelle.

L'analyse, qui intervient ensuite, consiste pour l'opérateur à sélectionner les informations et les éventuels renseignements participant à la réponse à la question d'origine. Il s'agit d'un travail de synthèse et de regroupement sémantique qu'un système automatique peinerait à effectuer dans un cadre général. Il n'est, en revanche, pas déraisonnable de

3. Allied Joint Publication 2.1, document OTAN décrivant les procédures de renseignement. Le document est 'classifié' mais toutes les citations présentes ici sont publiables.

4. La version originelle de cette définition et quelques autres sont disponibles à l'annexe B, page 169.

le comparer aux règles régissant le rapprochement et la sélection de données, préalable à toute fusion.

En effet, la phase suivante d'intégration revient à la combinaison des données préalablement sélectionnées, dans le but d'établir des *renseignements* enrichis et pertinents. Il est question de fusion d'informations, potentiellement de granularités différentes, tenant compte de leur cotation. Il s'agit, bien entendu, de fusion de données hétérogènes, puisqu'il existe un grand nombre de types de renseignements (cf. annexe A, p. 167).

Enfin, lors de la phase d'interprétation, le renseignement construit est mis en contexte et sa portée évaluée. Cette étape vérifie, avant sa remise au commanditaire, la réponse à la question initiale en fonction des connaissances disponibles.

1.2.4 En quête de certitudes insaisissables : l'évaluation de l'information

L'étape du cycle du renseignement qui nous intéresse est donc la phase d'évaluation qui, dans les usages du renseignement, s'applique aux informations plus qu'aux renseignements proprement dits, selon la distinction proposée aux définitions 1 et 2 (voir, notamment, le support de formation aux métiers du renseignement britannique DISS & DISC (2001)[5]). Cette section introduit la grille de cotation existante puis présente la prise en compte de la confiance en un renseignement.

Cotation d'une information

L'annexe du STANAG 2022 définit pour la première fois la grille de cotation du renseignement pour les forces de l'OTAN (North Atlantic Treaty Organization (NATO), 1997). Elle propose une échelle à deux dimensions et six degrés de vérité, décrits à l'aide d'étiquettes linguistiques (cf. tableau 1.1), dans laquelle la confiance en une information est représentée par la combinaison de la fiabilité de sa source et de sa propre crédibilité.

La première dimension d'évaluation de la cotation, la fiabilité, caractérise la source indépendamment de l'information considérée : toute information fournie par une même source bénéficie de la même fiabilité. Celle-ci est estimée par l'opérateur humain en fonction de son historique d'interactions avec la source en question.

Le second axe de cotation évalue la crédibilité de l'information. Le degré maximal de cet axe est affecté aux informations confirmées par d'autres sources. Ceci tend à suggérer que la certitude de l'information se mesure comme un degré de confirmation : plus une information est confirmée, plus elle est certaine et, à l'inverse, plus elle est infirmée, moins

5. Defence Intelligence and Security School (DISS) & Defence Intelligence and Security Centre (DISC).

Fiabilité	Définition
A	Complètement fiable
B	Habituellement fiable
C	Plutôt fiable
D	Habituellement pas fiable
E	Pas fiable
F	Fiabilité ne peut être estimée

Crédibilité	Définition
1	Confirmée par d'autres sources
2	Probablement vraie
3	Possiblement vraie
4	Douteuse
5	Peu probable
6	Crédibilité ne peut être estimée

TABLE 1.1 – Grille de cotation du renseignement

elle est crédible. Nous reviendrons dans la section suivante sur l'interprétation de ce critère et de l'écart entre les valeurs proposées et ce qu'il représente.

La cote d'une information s'exprime donc comme une lettre suivie d'un chiffre, évaluations de ces deux dimensions. Il est à noter que les opérationnels que nous avons rencontrés ont pris l'habitude de n'utiliser que quelques degrés. Il est, par exemple, habituel de coter une information issue d'un capteur $B2$, où il est entendu qu'un capteur *peut* avoir des faiblesses et qu'une information qu'il produit est sûre mais pas encore confirmée. La section suivante revient sur certaines des raisons de cette utilisation limitée.

Évaluation d'un renseignement

Comme nous l'avons vu plus haut, un renseignement est le produit d'une fusion d'informations et, éventuellement, d'autres renseignements. Bien qu'il ne soit pas coté selon la grille de cotation présentée ici, l'évaluation d'un renseignement est effectuée et proposée sous la forme de remarques d'accompagnement. Elle s'exprime comme une annotation textuelle indiquant la mesure dans laquelle l'opérateur chargé de sa production pense que le renseignement est fiable. Elle résulte de son savoir-faire métier et est estimée grâce aux cotes des différentes informations, ainsi que les évaluations des renseignements, le composant.

Les caractéristiques de la cotation dans le cadre de ce contexte applicatif peuvent être résumées par la définition 3 ci-dessous :

Définition 3 : Cotation

Selon les usages du STANAG 2022, la cotation est la mesure de la confiance pouvant être faite à :

- une information : elle est exprimée par un couple combinant la mesure de la fiabilité de la source à la crédibilité de l'information, conformément aux valeurs du tableau 1.1
- un renseignement : elle est alors fournie sous forme d'une annotation textuelle, à l'appréciation de l'opérateur chargé de sa production.

La cotation est un des résultats de l'étape d'interprétation, lors de la phase d'exploitation.

1.2.5 La cotation, une notion essentielle mais obscure

Nous proposons dans cette section une synthèse d'entretiens que nous avons menés avec des professionnels du renseignement et de notre étude de la documentation disponible sur leurs activités. L'objectif de ces entretiens était de nous aider à comprendre l'importance et le fonctionnement de la cotation dans leurs activités mais leur apport s'est également manifesté par une meilleure appréhension de leurs attentes.
Les remarques clefs consécutives qui ont orienté nos recherches par la suite s'articulent autour de la grille de cotation : de façon générale ces observations réclament moins la comparaison de la cotation retournée avec des résultats attendus que l'augmentation de sa lisibilité. Elles peuvent être organisées autour des quatre points suivants : la fiabilité de la source, la crédibilité de l'information, la combinaison de ces deux dimensions qui fournit l'évaluation et, enfin, la granularité des objets manipulés.

La dimension « fiabilité »
L'utilisation de la grille de cotation repose sur l'hypothèse que la première dimension, puisqu'elle décrit la source, est indépendante de l'information. Envisagée ainsi, comme nous l'avons dit plus haut, la valeur de la fiabilité est supposée stable pour toute information que la source fournit. Cependant, d'après les recommandations d'utilisation (DISS & DISC, 2001), elle doit refléter la confiance que l'opérateur fait à la source mais peut également indiquer la conviction que la source a en l'information qu'elle propose, ainsi que sa capacité à juger de ce qu'elle avance. Toutes ces dimensions semblent, en effet, judicieuses pour l'évaluation de la confiance mais sont contradictoires avec les étiquettes linguistiques du tableau 1.1 qui, elles, décrivent seulement les différents niveaux de fiabilité. Cette confusion a été soulignée par les opérationnels que nous avons rencontrés qui préfèrent, donc, se rabattre sur des valeurs par défaut. Ainsi une fiabilité de B est attribuée à un capteur technique, où la non-attribution de l'évaluation maximale, A, permet de prévoir les éventuels problèmes techniques de l'outil.

La dimension « crédibilité »

De la même manière que celle de la fiabilité, l'interprétation de la crédibilité pose problème. Cette dimension est censée représenter la mesure avec laquelle l'information est *crédible*, comme l'indiquent son nom et les étiquettes des quatre derniers niveaux déterminés de l'échelle (notés de 2 à 5). Cependant, puisque le niveau maximal est réservé aux informations 'confirmées par d'autres sources' – et sans aborder le problème de la granularité de l'information sur lequel nous revenons ci-dessous – on constate que la crédibilité peut être considérée comme un indicateur de confirmation. Ces délicatesses d'appréhension de la dimension, potentiellement à l'origine de manques d'objectivité, sont, sans doute, résolues par le savoir-faire des spécialistes, néanmoins les opérateurs que nous avons interrogés sur la question peinent à s'en affranchir. Dans les cas où la crédibilité peut être estimée, ils n'utilisent que les quatre niveaux qualifiant le doute (notés de 2 à 5) et laissent l'intégration de la confirmation (niveau 1) aux services responsables de l'analyse qui effectuent le recoupement d'informations diverses. Ainsi les spécialistes de capteurs techniques attribuent, par défaut, une crédibilité de 2, indiquant par là que l'information est sûre mais pas confirmée.

Combinaison des dimensions pour exprimer la confiance

La cotation sur deux dimensions, en plus de garantir qu'elle représente bien les dimensions souhaitées, est censée augmenter la lisibilité immédiate de son évaluation. Il semble, pourtant, que les opérateurs ne soient pas tous de cet avis : en plus de la difficulté d'attribution d'un niveau de confiance, le problème de sa lecture se présente en conséquence d'un manque de comparabilité. En effet, il est difficile de dire quelle information est la plus crédible si l'une est cotée $B3$ et l'autre $C2$, deux niveaux pourtant censément comparables. Nos interlocuteurs assurent que seules les valeurs connues, les valeurs par défaut, sont expressives et donc interprétables. Une information provenant d'un drone, donc sûre, est évaluée $B2$, établissant une forme d'étalon, malheureusement incomparable.

Un autre handicap de la cotation exprimée comme une telle combinaison provient de l'interprétation de la crédibilité. En effet, considérée comme indicateur de recoupement, l'échelle qu'elle parcourt va de l'infirmation à la confirmation. Il s'agit donc d'une échelle signée, construite autour d'une valeur neutre. La fiabilité, en revanche, décrit une progression d'une activation faible ou inexistante, jusqu'à un maximum absolu. La combinaison de valeurs sur ces deux échelles constituées différemment augmente la marge d'interprétation par les utilisateurs et, avec elle, le risque d'incohérences entre leurs perceptions.

Granularité

Comme nous l'avons dit, la cotation s'applique aux informations et non aux renseignements, bien que le producteur de ces derniers les accompagne d'une évaluation. Le choix

de ne pas mesurer la confiance de manière uniforme sur ces objets de granularités différentes provient sans doute, en partie au moins, de conséquences du système de notation. En effet, la notion de fiabilité de la source pose problème pour un renseignement, puisqu'il est le produit de la fusion d'informations provenant de sources distinctes. Pourtant, cette dimension reste pertinente pour l'établissement de la confiance. De plus, les évaluations des informations n'étant pas, comme nous l'avons vu, comparables, construire la cotation d'un renseignement à partir de leur agrégation n'est pas une opération formalisée. En pratique, l'analyste responsable de cette agrégation démontre son savoir-faire en rédigeant ses remarques d'accompagnement. Ces opérations sont encore compliquées lors de la constitution de renseignements combinant à la fois des informations et d'autres renseignements. L'analyste se forge alors un avis sur la confiance à partir d'indicateurs de natures différentes et, par là-même, incomparables.

Aborder le problème de la granularité soulève, par ailleurs, l'incohérence du niveau 1 de la crédibilité. En effet, faire dépendre la crédibilité maximale d'une information, avant comparaison et fusion, de confirmations par d'autres sources suppose la prise en compte d'autres informations, prise en compte contradictoire pour un tel élément atomique. Notons, qui plus est, que cette divergence introduit également la combinaison de sources dont nous soulignions la difficulté.

1.3 La cotation, un problème général

Les observations sur la cotation faite dans la section précédente nous offrent un premier aperçu des problématiques liées à l'évaluation de la confiance en une information. Dans cette section nous examinons le problème de la cotation de façon générale, au-delà du cadre du renseignement militaire. Après avoir étudié les causes de son manque de lisibilité et d'interprétabilité, nous présentons la formalisation de la cotation que nous proposons et indiquons les principes des réponses que nous apportons aux problèmes soulevés.

1.3.1 Comprendre la cotation

Bien que nous abordions le problème de la cotation comme émanant du contexte applicatif présenté ci-dessus, la question de la confiance en des données construites se pose de manière plus large. Qu'il s'agisse de traitements automatiques, de relations humaines, des sciences sociales les étudiant ou de procédés de vente, savoir comment une information est perçue est une question capitale. Certains professionnels consacrent même un temps

conséquent à essayer d'en orienter la réception. Dans le domaine qui nous intéresse en particulier, la recherche d'informations, pouvoir qualifier de manière prévisible la mesure avec laquelle une information est de confiance fait partie de la réponse à la quête de méthodes d'évaluation idéales. Ce problème d'évaluation s'énonce le plus souvent différemment, soit à l'aide d'une mesure des chances de l'algorithme de fournir une bonne réponse, soit en combinant un facteur de confiance ou de qualité à une mesure de l'adéquation de la réponse. Les outils permettant de représenter la confiance se focalisent, eux, sur la mesure de la *réalité* de faits. Celle-ci est supposée indiscutable alors qu'évaluer la confiance déplace ces mesures vers le domaine du subjectif et notre étude d'un cadre s'y attelant nous permet d'en exprimer les défis particuliers.

L'étude de notre cadre applicatif nous a, en effet, permis d'identifier, à la section 1.2.5, les écueils que nous résumons en l'expression du problème fondamental de la cotation telle qu'elle existe : la difficulté de compréhension de l'évaluation de la confiance. Nous proposons de considérer que cette opacité se manifeste sur deux axes distincts et complémentaires : d'une part, la méthode selon laquelle la confiance est évaluée et, d'autre part, la façon dont cette évaluation est exprimée. Comme la définition d'un vocabulaire commun permet de s'entendre, au moins sur les termes, une méthode mettant en avant, sans équivoque, comment et à partir de quoi la confiance est établie est nécessaire afin d'harmoniser son utilisation, de s'assurer que tous ses utilisateurs l'envisagent, la mesurent et l'interprètent de la même manière, évitant que certains évaluent combien une information est confirmée là où leurs lecteurs lisent les chances qu'elle se réalise. Lever les doutes de cette manière nécessite, cependant, que la méthode soit accessible et acceptée par ceux à qui elle ôte les degrés de liberté et l'interprétation conséquente. Afin d'assurer l'utilité de l'indicateur de confiance, celui-ci doit également être expressif. Si son processus de constitution en garantit la compréhension, son utilisation dans la comparaison de faits n'est possible que si son expression le permet.

Nous nous sommes attaché, dans ce travail, à ces deux problèmes en proposant une formalisation de la cotation, où nous distinguons la valeur, la *cote*, du processus, la *cotation*, selon les directions exposées à la section suivante. Nous avons également veillé, dans l'ensemble de nos travaux, à répondre aux besoins identifiés précédemment, en particulier la facilité d'interprétation et la possibilité de justifier les résultats fournis.

1.3.2 Formalisation du problème

Nous voulons modéliser et formaliser un processus d'évaluation automatique de la cote des données, l'indicateur de la confiance à leur témoigner. Nous séparons, dans le but

de clarifier nos objectifs et façons de considérer le problème, les notions couvertes par la définition 3 (cf. p. 15) : dans la continuité de nos observations, nous articulons nos travaux autour de la définition d'un procédé d'évaluation de la cotation, d'une part, et de la formalisation du système de son expression, d'autre part. Nous présentons ici, de façon synthétique, les principes sur lesquels s'appuient les réponses que nous proposons aux problèmes exposés dans ce chapitre. Pour ce faire, nous proposons les définitions 4 et 5 suivantes :

Définition 4 : Cote
La cote d'une information est l'indicateur de la confiance que l'on peut lui faire. Pour que la cote soit appréhendable sans ambiguïté, il est nécessaire de :
– définir les dimensions selon lesquelles elle est évaluée
– choisir un formalisme pour sa représentation favorisant sa lisibilité

Définition 5 : Cotation
La cotation de l'information désigne le procédé par lequel la cote est évaluée. Elle décrit la manière dont les dimensions choisies pour le calcul de la cote sont combinées. La cotation doit être compréhensible afin d'assurer l'uniformité de son utilisation. Pour parvenir à ce but, sa formulation doit garantir son interprétabilité, c'est-à-dire ne pas laisser de doutes sur la raison de l'aboutissement à une cote. Si la cotation répond à ces exigences, elle offre également la possibilité de justifier une confiance exprimée.

Nous proposons de répondre à ces exigences en nous penchant, dans un premier temps, sur la cote puis sur la cotation de manière distincte et, par la suite, en introduisant un formalisme permettant l'expression et la réalisation de l'ensemble. Revenons, donc, sur la première exigence de la définition 4, la définition des dimensions selon lesquelles l'évaluation doit être pratiquée. Le choix de ces dimensions doit tenir compte de l'existant ainsi que des études attenantes. En nous inspirant d'une interprétation du processus d'établissement de la confiance, nous déterminons les facteurs participant à sa constitution. Nous proposons un découpage de ces dimensions d'intérêt permettant de regrouper les différents critères d'évaluation selon leur objet et leur influence et assurant leur indépendance et leur non-redondance. Nous proposons également d'évaluer chacune de ces dimensions sur une échelle discrète en accord avec l'esprit du tableau 1.1 et veillons à en favoriser l'interprétabilité en portant une attention particulière aux étiquettes linguistiques les illustrant.

Une fois cette élection des dimensions opérée, lorsque ce qui entre en compte dans l'établissement de la confiance est établi, se pose la question de la combinaison des mesures.

Afin de répondre à ce problème, nous proposons une philosophie de l'intégration des dimensions à la confiance, l'architecture de la cotation. Nous proposons et justifions un ordre de prise en compte des dimensions qui nous mène à la définition d'une chaîne de cotation, synthèse de nos propositions pour l'évaluation des critères et de leur impact sur la cotation. Nous examinons en détails et discutons cette influence à la hausse ou à la baisse de l'intégration des dimensions. Nous décrivons, de plus, la souplesse conférée au modèle par ces propositions de combinaison, souplesse permettant l'adaptabilité à différentes perceptions sans pour autant introduire la nécessité d'interprétation et les risques de non-concordance qui en découlent.

Enfin, nous proposons une formalisation théorique du procédé de cotation ainsi que de la cote, son mode d'expression. Celle-ci gagne en lisibilité grâce à une représentation par un indicateur unique, les dimensions la composant étant intégrées selon la méthode proposée précédemment. L'indicateur ainsi produit permet de comparer des niveaux de confiance entre faits et d'ordonner ces faits selon la conviction du lecteur. En outre, l'ordre fourni entre les niveaux d'expression de la confiance lui offre une lisibilité totale qui s'ajoute à la lisibilité du procédé de cotation. Nous proposons également d'uniformiser les évaluations des dimensions afin d'augmenter encore l'interprétabilité du résultat en garantissant la commensurabilité des échelles, évitant par là la combinaison de mesures de natures différentes évoquée plus haut.
Nous choisissons d'exprimer cette formalisation dans le cadre de la logique multivaluée que nous étendons afin de préserver l'expressivité de la distinction entre l'impossibilité de la mesure et une mesure neutre mais exprimée. Dans ce cadre formel étendu nous proposons les opérateurs de combinaison permettant la représentation de la cotation telle que nous la définissions auparavant et montrons également la formalisation de la modélisation de perceptions.

1.4 Conclusion

Dans ce chapitre, nous avons introduit les notions sous-jacentes à la problématique de la cotation. Nous avons commencé ce tour d'horizon par une présentation du contexte applicatif de nos recherches qu'est l'évaluation de la confiance dans le cadre du renseignement militaire. Afin de situer les problèmes, nous avons tout d'abord décrit le processus de construction de connaissances associé. Dans ce processus nous nous sommes attardé sur la phase d'analyse puis, en particulier, sur l'étape d'évaluation des informations. Nous avons terminé cette mise en contexte par une description des points d'achoppement identifiés.

Dans la seconde partie du chapitre, nous avons présenté la facette générale de la cotation et des problématiques associées. Nous sommes revenu sur les obstacles précédemment décrits et en avons offert une décomposition en deux problèmes complémentaires. Enfin, nous avons présenté les principes des solutions que nous proposons pour résoudre chacune de ces questions et que nous détaillons dans la suite de ce mémoire.

Ne te soucie pas de ta trace. Tu es le seul à ne pouvoir l'effacer.
Edmond Jabès

Chapitre 2

Confiance, certitude et qualité comme influences pour la cotation

Résumé

Nous proposons ici un panorama des disciplines voisines de l'établissement de la confiance. Nous commençons par une étude de domaines traitant de confiance, parmi lesquels se situent les travaux existants sur la cotation. Concevant cette problématique comme un glissement de l'évaluation de l'information vers celle de la foi à lui associer, nous nous intéressons ensuite à deux domaines étudiant la nature de cette information : la qualité des données puis la théorie de l'incertain. Enfin, en raison des différences de granularité des éléments à coter, nous abordons le sujet de la fusion.

2.1 Introduction

Nous présentons dans ce chapitre les disciplines qui ont influencé nos recherches sur l'établissement de la confiance en une information. Telle que nous l'avons définie, cette problématique rassemble deux objectifs : fournir un modèle lisible d'expression de la confiance et établir le processus guidant son évaluation. Ces problèmes complémentaires ont des réponses disséminées dans diverses communautés. Les premières que nous étudions sont celles qui s'intéressent expressément à la confiance. Nous commençons par considérer les

travaux sur le sujet dans le cadre de la recherche militaire, où se trouvent quelques perceptions différentes de la cotation ainsi que d'autres formulations de la confiance. Dans un deuxième temps, nous regardons du côté de la réputation dans les réseaux de confiance, où nous identifions des méthodes d'évaluation de la confiance et l'incontournable nécessité de l'adaptation à la subjectivité de l'évaluateur.

Dans la poursuite de la recherche de formalisation de l'établissement de la confiance, nous présentons ensuite le domaine de la qualité des données. Celui-ci s'intéresse à la fois à l'évaluation des modèles de données et à la définition des dimensions de cette évaluation pour garantir la manipulation de connaissances aux normes dans les traitements automatiques. Certaines expressions utilisées dans ces travaux semblent orientées vers la sélection de dimensions de mesure de la confiance en l'information. Les outils et modèles proposés portent sur des facteurs dont nous reconnaissons l'importance dans cette mesure. On trouve également un intérêt pour la personnalisation, l'adaptation des solutions proposées aux attentes de leur destinataire. Cependant, la qualité des données qualifie l'objet sur lequel elle porte et non sa perception. Nous concevons donc la cotation comme une formulation particulière du problème, résultant d'un glissement sémantique de ces techniques vers une abstraction subjective.

Ce même besoin d'abstraction se présente lorsque nous envisageons différentes méthodes de représentation de l'incertitude. En effet, parce que la foi en une information peut être mère de doutes, nous l'envisageons comme une forme de degré d'incertitude. Néanmoins, l'incertitude qualifie elle aussi généralement le contenu informationnel d'un élément. Afin de pouvoir bénéficier de la palette des outils de représentation et expression interprétables de l'incertain, nous devons prendre soin de souligner le glissement opéré. Dans le tableau que nous proposons de ces théories, nous distinguons l'incertitude mesurable de l'incertitude cognitive. Nous expliquons pourquoi la première ne semble pas convenir à la formalisation de la cotation telle que nous l'entendons, alors que la seconde, riche d'outils représentant des concepts de raisonnement humain, nous arme pour la représenter.

Enfin, nous revenons aux problèmes présentés au chapitre précédent de mélanges de constitution de connaissances. Celle-ci nécessite l'assemblage d'informations de granularités potentiellement différentes. Ces questions font partie de la définition de la problématique de la fusion, par laquelle nous terminons ce panorama, en nous concentrant sur la fusion d'informations incertaines qui offre, elle aussi, des outils interprétables de combinaison d'indicateurs de confiance.
Concevant l'expression de la cotation comme nécessitant l'adaptation d'outils de ces différents univers, nous proposons d'envisager le domaine comme chevauchant ces trois thématiques, tel qu'indiqué à la figure 2.1.

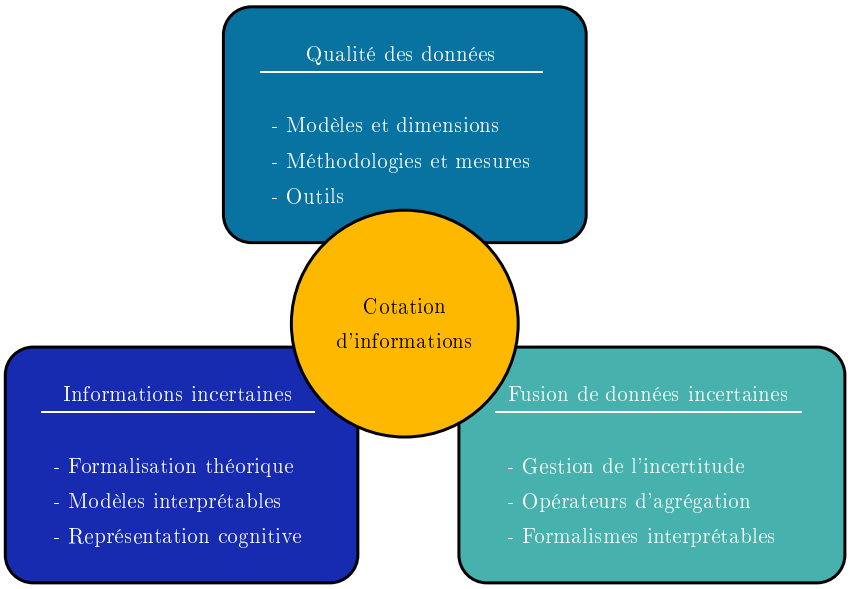


FIGURE 2.1 – La cotation aux confluents des domaines de la qualité des données, des informations incertaines et de la fusion

2.2 La confiance comme objet d'études

Telle que nous la décrivons, la cotation est une représentation de la confiance en une information. Si le sujet n'est pas toujours évoqué en ces termes, la confiance est présente dans de nombreux domaines de recherche. Cette section présente quelques-unes des problématiques abordées par ces travaux. Parmi ceux-ci se situent, donc, la littérature traitant de cotation appartenant aux problématiques de la recherche militaire qui propose également d'autres approches de la question. Nous présentons aussi une formulation alternative du problème de la confiance provenant des communautés qui s'intéressent au 'Trust Management', la gestion de la confiance.

2.2.1 Aperçu de la confiance dans le cadre tactique

Comme nous l'avons vu précédemment, la recherche militaire en général et la recherche sur les problématiques liées au renseignement en particulier sont en plein essor. Nous commençons cette présentation par les articles abordant spécifiquement la cotation telle que définie par le STANAG 2022. Cependant, la cotation ne représente pas le domaine le plus étudié dans ce cadre. Nous nous penchons donc, ensuite, sur des travaux considérant la sécurité comme participant à la confiance. Nous terminons avec une approche formulée autour du problème de la représentation et la communication de l'information.

Extensions du STANAG 2022

Nous avons déjà décrit comment l'accord de standardisation décrivant la cotation du renseignement est ouvert à la critique et à l'interprétation (cf. section 1.2.4, p. 14). Une partie des remarques que nous faisons est présente en introduction de chacun des articles traitant de cotation. Cependant, si cette partie du problème est formulée de manière analogue, la perception générale de la tâche et, donc, les solutions proposées diffèrent. La plupart des auteurs considère la cotation comme une indication sur la crédibilité, c'est-à-dire mesurant la réalité de l'information. Envisagée ainsi, les méthodes qu'ils proposent tiennent le plus souvent du maintien de cohérence dans un système ou de la combinaison de degrés d'incertitude, la fusion d'informations incertaines.

Ainsi, Cholvy et Nimier (2003) s'intéressent essentiellement au maintien de cohérence dans une base de connaissances, en pondérant les distances entre mondes possibles par les fiabilités des sources. Cholvy (2004) reprend la grille de cotation initiale pour qualifier les éléments insérés dans une base de données stockant l'historique des informations et de leurs sources. Les réponses aux requêtes sur la base de données sont choisies par un vote entre les alternatives en conflit. Cet historique contribue à la construction d'une information enrichie par la fusion des éléments stockés. Dans cet esprit, la majorité des articles présentée ici considère le problème de la cotation comme un problème de fusion. Nimier (2004), par exemple, en propose une description par des opérateurs de combinaison dans trois formalismes, le premier probabiliste, le second possibiliste et le dernier basé sur la théorie de Dempster-Shafer. De manière plus détaillée, Besombes et Cholvy (2009) proposent une architecture complète d'un système de fusion d'informations. Afin de procéder à la fusion, les informations sont corrélées entre elles, la mesure de corrélation couvrant toute l'échelle de l'infirmation à la confirmation. Les auteurs introduisent ensuite un calcul de la mise à jour de la cotation en fonction de cette corrélation. S'ils ne détaillent pas le processus de fusion, ils examinent différentes méthodes de calcul de la corrélation.
Ces articles envisagent donc la cotation comme une estimation, dans un monde plus ou moins clos, de la contradiction entre faits connus. Nous avons vu que la crédibilité peut être envisagée comme un indicateur de confirmation essentiel à l'évaluation de la confiance. Néanmoins, nous avons également noté que ne considérer que ce facteur n'est guère satisfaisant et compte au nombre des problèmes de la cotation existante. En outre, ces modélisations, pour riches qu'elles soient, assimilent la cotation d'une information à sa cohérence avec la connaissance du monde, faisant l'impasse sur son mode de production ou le résumant à l'importance relative de la fiabilité de sa source. En cohérence avec certaines des disciplines que nous présentons dans la suite, elles représentent mieux le doute perceptible sur le contenu de l'information que celui concernant la foi en cette même information.

Un autre point de vue est proposé par Bärecke et al. (2010) qui considèrent, en plus des dimensions classiquement prises en compte, la confiance que la source porte à son information, la fraîcheur des informations et les relations d'amitié ou d'hostilité entre sources. Les incertitudes sont représentées par des distributions de possibilité modérées par les fiabilités des sources. La fusion des éléments se fait en fonction de la partition amicale, où l'on considère que deux sources en conflit proposant la même information se corroborent mieux que deux sources amies en faisant autant. Cette formulation reprend, elle aussi, certains aspects de la qualité de l'information pour la détermination des facteurs entrant en compte dans son évaluation. Les auteurs y ajoutent des notions de subjectivité et proposent une modélisation intéressante des répercussions de leur diffusion. Encore une fois, pourtant, l'évaluation de la connaissance concerne plus son contenu informationnel que la confiance déduite de sa constitution, malgré la prise en compte de l'avis de la source, facteur idéalement capital s'il était aisément mesurable.

La sécurité comme facteur de confiance

Les travaux précédents traitent directement de la cotation. Cependant, l'étude de la confiance dans le cadre de données sensibles s'attache également à d'autres points de vue. Parmi ceux-ci, la fiabilité des systèmes informatiques représente la perception de la confiance comme la projection d'une mesure globale, celle du système, sur l'information. Avec cet objectif, les travaux de Hamill (2000) et Beauregard (2001) présentent tous deux des modèles d'assurance de la sécurité de données sensibles. Hamill (2000) offre une modélisation inspirée du 'Value Focused Thinking' de Keeney (1994), permettant de déplacer le choix entre solutions de sécurité des différentes alternatives à la sélection des propriétés désirées pour ces différentes options. Ce déplacement de l'analyse de la sécurité de systèmes vers la théorie de la décision se rapproche du glissement que nous décrivons de l'évaluation de la certitude d'une information, qui concerne son contenu, vers l'appréciation de la foi qu'on lui porte, qui n'en dépend pas nécessairement. Beauregard (2001), quant à lui, propose un modèle permettant de mesurer le compromis entre l'amélioration de la sécurité des systèmes, le gain engrangé sur l'utilisation opérationnelle de l'information et le coût de telles assurances, mettant en avant l'importance des choix dans la construction de la confiance.

Représentation des informations

Une dernière approche que nous souhaitons mentionner se consacre à la représentation des informations (Rein & Schade, 2009). Ces travaux s'intéressent aux mêmes problèmes de confiance en l'information fusionnée pour la surveillance de situations à risques mais optent pour un langage de description formel. Le 'Battle Management Language' ou BML est un standard d'interopérabilité utilisé, notamment, par l'OTAN (Rein et al., 2009). Selon les auteurs, ce langage facilite la représentation et la fusion d'informations en normalisant leur

expression. Ils intègrent à leurs rapports une évaluation de la crédibilité de l'information et de la fiabilité du processus de fusion, dont les dimensions sont discutées dans Kruger et al. (2008). Malheureusement, leur objectif étant la normalisation et la favorisation des échanges entre services et types d'informations, les détails de la combinaison de ces degrés ne sont pas abordés.

De la même manière que nous le faisons, l'ensemble des travaux que nous présentons considère la cotation dans le cadre de la fusion d'informations. Cependant, les objectifs de sa modélisation varient. Pour la plupart, il y est plus souvent question de la mesure de certitude que de la construction, représentation et mesure de la confiance. Les modèles proposés offrent tous des cadres rigoureux à une partie de la cotation et leur complémentarité permettra, à ceux qui les associeront, de proposer une solution générale. Pourtant, aucun d'entre eux n'aborde la cotation telle que nous l'envisageons, comme un indicateur interprétable de la confiance.

2.2.2 Aperçu de la propagation de confiance

Au-delà du cadre tactique, la notion de confiance est largement considérée, en particulier lorsque l'on traite de 'Trust Management'. Derrière cette expression plusieurs communautés s'intéressent à la modélisation de la confiance (Falcone et al., 2001; Weeks, 2001; Krukow, 2004). La confiance étant une affaire relationnelle, elle est surtout étudiée dans le cadre de communications entre agents, que ceux-ci soient des agents logiciels, les nœuds d'un réseau ou des humains dans des systèmes d'échanges. Là où certains travaux présentent l'étendue du 'Trust Management' de façon globale (Khare & Rifkin, 1998; Staab et al., 2004), d'autres se focalisent sur les outils permettant d'exploiter ces 'webs of trust', ces réseaux de confiance. Parmi ceux-ci se distinguent les théoriciens de la mesure et ceux de la réputation : les premiers indiquent comment calculer la réputation d'un acteur, les seconds décrivent différentes attitudes d'échanges entre agents et leur impact sur le système.

Mesure de la confiance

Les échanges sont toujours influencés par la confiance que les participants se prêtent. Avec la naissance, puis l'abondance de la diffusion de données sur internet, l'automatisation du calcul de cette confiance est devenue nécessité. Afin de procéder à ces calculs, des métriques et méthodes d'estimation de la confiance ont vu le jour. Ruderman (2004), par exemple, propose une comparaison de métriques de calcul de réputation d'agents dans un réseau. Il rappelle la similarité des objectifs et contraintes du problème et se prononce sur la supériorité de l'incontournable étalon de la confiance : Google.

On peut difficilement aujourd'hui, en effet, parler de recherche d'information, ni de confiance en une information, sans mentionner Google et son algorithme PageRank (Brin & Page, 1998). Par son omniprésence, le moteur de recherche s'est hissé au-delà de sa fonction principale : la classification de réponses qui, comme peuvent le faire les systèmes de filtrage collaboratif, utilise la fiabilité de la source pour ordonner des résultats construits par ailleurs. Il s'agit donc moins de calculer une confiance générale en l'information que d'ordonner des résultats de manière cohérente.

Hormis l'article fondateur déjà cité et le rapport technique le détaillant (Page et al., 1999), l'algorithme PageRank fait l'objet d'études variées sur sa capacité à qualifier la fiabilité des sources dans un contexte où les volumes et l'intention de tromper influencent l'estimation. Diligenti et al. (2004) proposent une formalisation et une comparaison statistiques de l'algorithme et de ses alternatives. Bien qu'ils replacent ces algorithmes dans leur contexte d'utilisation, la recherche d'informations, ils les envisagent comme estimant la confiance à faire aux sites recensés. Avant eux, Bianchini et al. (2002) offraient une étude topographique de la méthode. À nouveau, les auteurs envisagent PageRank comme une technique d'évaluation de la réputation de sites et donc de la confiance à leur porter. Dans l'esprit de l'exploitation d'un modèle connu de mesure de l'établissement de la confiance que la section suivante présente, les auteurs proposent également des techniques de manipulation de réputation jouant avec la topologie du réseau. La description la plus abordable du fonctionnement de l'algorithme est fournie par Austin (2006) qui décrit comment PageRank établit la confiance dans l'internet envisagé comme un treillis dont la relation d'ordre est donnée par les hyperliens, en propageant de manière récursive la réputation des sources, fonction de la qualité de ses référents, sur l'information.

Parce qu'ils offrent un score de confiance issu d'un calcul de propagation de réputation de sources aux informations, PageRank et les algorithmes équivalents nous ont initialement intéressé comme solution éventuelle à l'évaluation de la confiance. Cependant, la réalité qu'ils représentent est une simplification de celle que nous envisageons. En effet, ce type d'algorithmes ne considère que des relations positives, puisque le calcul de la confiance émane de la réputation et non des informations. Les contradictions sont donc, de fait, exclues a priori.

Exploitation de la réputation

À la frontière avec les méthodes de la partie précédente se situe la problématique de l'évaluation de réputation. Plutôt que de procéder à la diffusion de la confiance dans le réseau pour évaluer la qualité des acteurs, cette problématique particulière s'intéresse à leur qualification en fonction de leurs comportements. Bien qu'il s'agisse de mesurer un facteur équivalent, le glissement opéré de la réputation relative vers la fiabilité en pratique est

une notion d'intérêt en ce qu'elle distingue les axes généraux et contextuels. Toivonen et al. (2006) proposent, par exemple, une formalisation de l'évaluation contextuelle de la confiance. Ils intègrent à une évaluation équivalente à celles de la section précédente deux classes d'attributs, les premiers portant sur la qualité des connaissances sur l'acteur à évaluer et les seconds sur les facteurs contextuels permettant d'évaluer l'acteur en fonction de l'information échangée. Garg et Battiti (2004) et Garg et al. (2004) proposent ROCQ (Reputation, Opinion, Credibility and Quality), un formalisme d'évaluation de la réputation dans un réseau pair à pair par le pair désireux d'échanger. Une fois encore, cette évaluation regroupe des aspects généraux et d'autres contextuels. La nuance avec le formalisme précédent est que la contextualisation tient plus ici de la personnalisation, où le pair évaluateur juge à partir de son expérience avec le pair évalué et son avis sur les pairs recommandeurs. Nous revenons sur cette distinction entre facteurs généraux et contextuels lors de la présentation des dimensions participant à la cotation, dans le chapitre 3.

Quercia et al. (2007) implémentent un système de propagation de confiance distribué à partir d'informations parcellaires pour l'estimation embarquée de l'opportunité de procéder à un échange. Dans la même thématique de projection de la confiance faite à la source sur la décision de partage, Ekström et Björnsson (2002) et Ekström et al. (2005) proposent un système d'évaluation des enchérisseurs dans le commerce électronique. Ces deux formulations du problème soulèvent le point de l'égalité de l'influence des évaluateurs dans la propagation de la confiance dans son calcul automatique. Cette constatation pousse les auteurs à inclure l'utilisateur dans l'évaluation de la confiance faite à la source, dans une mesure semblable à ce que nous proposons dans nos travaux.

Après avoir rappelé différentes manières de voir la réputation, nous en venons maintenant à son utilisation dans l'évaluation d'informations. Nous avons vu que l'algorithme PageRank de Google est utilisé pour établir la réputation de nœuds dans un réseau de confiance. Chen et al. (2007) l'utilisent pour décrire comment trouver l'article initial dans un réseau de citations grâce à son calcul de recommandations. De même, Heath et al. (2007) proposent d'utiliser PageRank pour établir les domaines de compétences et spécialités dans un réseau social. Qu'ils l'utilisent pour qualifier la topologie d'un réseau et la répartition des connaissances sous-jacente ou simplement comme distance, les travaux sur l'exploitation de la confiance s'en servent pour qualifier l'objet de l'échange, une information ou une transaction d'une source sûre profitant donc d'un a priori positif. L'intégration de la confiance faite à la source à l'évaluation d'une information est certainement un facteur incontournable. Cependant, la définition de la portée de ce qu'elle décrit est souvent vague ou très dépendante de l'utilisation. Nous proposons, au chapitre 4, d'étudier comment tenir compte des différents paramètres que nous pensons importants en en précisant la définition

afin de mieux décrire comment la confiance s'établit.

Le dernier pan de la gestion de la confiance sur lequel nous revenons se penche sur la modélisation des politiques de confiance. Celles-ci représentent comment un agent particulier réagit lors d'échanges avec un autre, en fonction de la réputation de ce dernier, de l'attitude de l'agent évaluateur et du contexte social. Le comportement de l'évaluateur est ensuite étudié afin de prédire sa décision d'échanger ainsi que l'impact de son attitude sur le système. Chaque agent dans le système applique sa stratégie de partage et les modèles représentent les gains et pertes des différents comportements. Chu et al. (1997) décrivent un système de gestion de la confiance permettant à chaque utilisateur de spécifier sa politique de sécurité, ses préférences particulières sur l'attribution de confiance. Les auteurs permettent, dans l'écriture des règles de description de ces préférences, d'exprimer la confiance, la défiance et l'impossibilité de se prononcer grâce à une évaluation trivaluée. Les délégations de confiance décrites peuvent l'être, dans le cadre des échanges envisagés, à d'autres systèmes de gestion de la confiance, tels que le système PICS (Platform for Internet Content Selection) de Blaze et al. (1997). Celui-ci intègre la contextualisation de la confiance, notion analogue à celle que nous décrivions plus haut, et lui ajoute la personnalisation par la définition et l'application de politiques de confiance.
La modélisation des politiques de confiance intéresse également la communauté des systèmes multi-agents. Il est généralement supposé qu'un agent n'ayant pas de vue globale sur le système, ses relations sont établies de façon lacunaire. Barber et Kim (2000) proposent de résoudre ce problème à l'aide d'un calcul distribué de mise à jour de la confiance intégrant la subjectivité de l'agent, sa politique de confiance. Un point de vue plus général de la problématique est offert par Sen (2002) qui étudie l'impact sur le système de comportements d'agents égoïstes et le compare à ceux d'agents collaboratifs. Ustymenko et Schwartz (2006) apportent une représentation logique de ces politiques et généralisent le problème au web sémantique envisagé comme un réseau de confiance. L'expression de la confiance se fait sur une échelle discrète parcourant l'espace de la confiance à la défiance, dont la seule valeur non-signée est la valeur neutre médiane. Enfin, Clark et McCabe (2007) proposent la modélisation de la confiance comme un système d'inférences dans une ontologie dont les axiomes formalisent la confiance dite statique de l'agent et les descriptions des instances la confiance dynamique. De la sorte, l'évaluation de la confiance à porter à une nouvelle information découle de la politique locale et la compatibilité des connaissances est préservée par un système de maintien de la cohérence.

Nous regrettions que les systèmes classiques d'évaluation de la confiance ne considèrent que les relations positives entre acteurs. Les modèles de politiques de confiance, quant à eux, envisagent généralement la méfiance dans l'évaluation de leurs dimensions. Cependant,

s'ils prévoient une évaluation négative, la majorité de ces travaux n'aborde pas la mise en œuvre de l'évaluation de cette défiance. Cette prise en compte valide, néanmoins, l'intérêt de la contradiction dans la constitution de confiance. Nous revenons au chapitre 3, sur l'importance de la gradualité de l'échelle d'évaluation et au chapitre 4, sur la personnalisation de l'évaluation de la confiance, problèmes évoqués ici.

2.3 Qualité des données

La qualité des données – ou de l'information [1] – est un sujet de recherche relativement récent et par là même très actif. Il s'intéresse à l'assurance que les informations manipulées dans les traitements automatiques soient conformes à des standards. Cette quête de normalisation de la production des données et informations nous intéresse parce qu'elle propose des outils, théories et études de la description de ce que représente une information de qualité. Nous considérons en partie l'évaluation de la confiance comme un cas particulier de cette démarche, cas pour lequel les facteurs mesurés ne le sont pas de manière absolue mais intègrent une part de subjectivité.
Cette section présente la qualité des données, décrite comme aux confluents de plusieurs domaines d'étude tels que la recherche de modèles ou de techniques. Elle revient, ensuite, sur quelques travaux abordant les diverses problématiques du domaine, problématiques offrant des outils permettant d'assurer ou, du moins, d'augmenter la qualité de l'information.

2.3.1 Domaines de recherche liés

La somme offerte par Batini et Scannapieco (2006) met en évidence l'étendue des recherches en qualité des données qui, de façon générale, s'intéressent à divers aspects de la manipulation de connaissances. Dans leur ouvrage, les auteurs présentent la pluridisciplinarité du domaine éponyme selon la figure 2.2, reproduite ici, que nous décrivons succinctement dans la suite de cette section.

La représentation des connaissances cherche à modéliser les informations d'un point de vue orienté utilisateur : elle assure par là que les données sont lisibles et exploitables. Ces problématiques et les méthodes qu'elles produisent servent à assurer une meilleure qualité aux données, notamment en en garantissant 'l'utilisabilité'.

Les statistiques offrent des outils permettant de mesurer les performances des systèmes

1. Bien que les spécialistes fassent la nuance, nous utilisons les deux termes indifféremment. L'annexe C offre quelques définitions du domaine qui explicitent la différence de granularité.

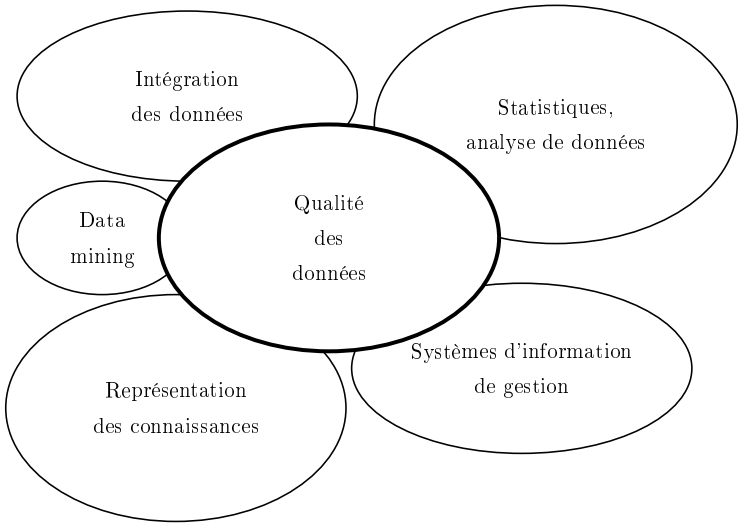


FIGURE 2.2 – La qualité des données dans la recherche (Batini & Scannapieco, 2006)

et l'incertitude des données autant qu'elles cherchent à les expliquer, les résumer et les analyser. Elles sont conçues pour représenter la réalité objective et mesurable et offrent de nombreuses méthodes d'évaluation de l'information (National Institute of Standards and Technology, 2003). On constate, notamment, l'omniprésence des statistiques dans l'évaluation des performances d'algorithmes. Parmi ces mesures, les plus usitées pour l'intégration de ces facteurs mesurables au modèle proposé sont, dans le cadre de la recherche d'informations : la précision (nombre de documents pertinents retrouvés, rapporté au nombre total de documents retrouvés), le rappel (nombre de documents pertinents retrouvés, au regard du nombre de documents pertinents existants) et leur combinaison par moyenne harmonique pondérée, la F-mesure. Nous revenons à la section 3.3, page 52 et dans le chapitre 7, sur la façon d'intégrer ces évaluations de sources de données construites à l'établissement de la confiance.

Comme les statistiques, la fouille intelligente de données, le 'data mining', est à la fois acteur et partie de la qualité de l'information. Elle est d'autant plus efficace qu'elle fouille des données de qualité. Par ailleurs, des outils permettant d'extraire de la connaissance de masses de données permettent, à leur tour, d'en augmenter la qualité espérée.

Enfin, les systèmes d'information de gestion et l'intégration des données constituent

les domaines applicatifs de la qualité des données. Les premiers fournissent les processus d'optimisation de la gestion d'organisations, les outils d'aide à la décision et autres instruments de gestion de l'activité (e.g. Customer Relationship Management, Enterprise Resource Planning). Leur importance dans l'activité de l'organisation réclame qu'ils modélisent, manipulent et présentent des données de qualité. L'intégration de données intervient ici comme représentant des activités principales de la qualité des données. Façade interopérable entre systèmes de gestion de données, elle s'efforce de vérifier la qualité des données stockées afin de garantir celle des résultats produits.

La part plus théorique de la qualité des données offre, donc, moyens et cadres pour l'évaluation de la valeur d'une donnée en tant qu'objet de traitements automatiques. Dans ce contexte, la réalité du fait que décrit la donnée n'est pas considérée, pas plus que la confiance à leur porter. La cotation se situe entre ces deux notions étroitement liées. Elle porte sur l'information produite par un système et non sur le fait que celle-ci décrit, mais ne s'arrête pas à considérer cette information comme une donnée.

2.3.2 Problématiques

Batini et Scannapieco (2006) proposent également une représentation imbriquée des problématiques étudiées par la communauté de la qualité des données, représentation reprise à la figure 2.3. On y découvre non seulement les classes de problèmes mais également les outils transverses que la qualité des données leur applique. Nous examinons plus particulièrement les domaines voisins de la cotation.

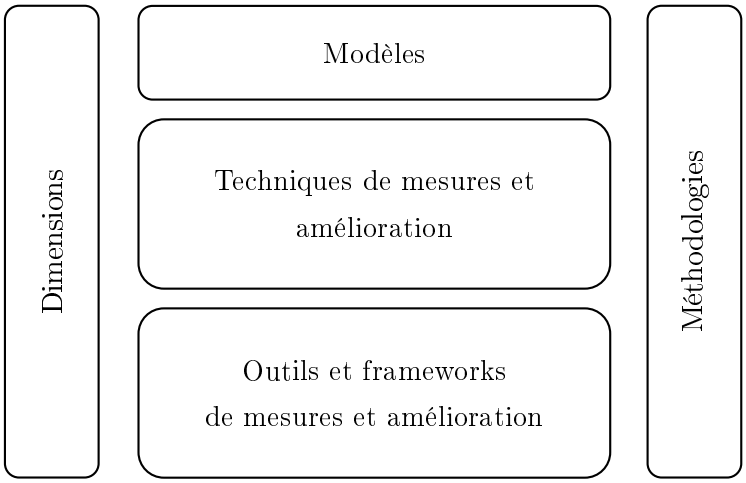


FIGURE 2.3 – Problématiques de recherche en qualité de l'information (Batini & Scannapieco, 2006)

Dimensions de la qualité

De nombreux travaux se penchent sur la définition des critères participant à la qualité et sur leur organisation (Wang et al., 1995; Harrathi & Calabretto, 2006). L'annexe C reproduit, à titre d'exemples, la liste de dimensions proposée par Wang et Strong (1996), le tableau C.1, page 173, ainsi que leur partition fournie par Berti (1999b), au tableau C.2, page 175. Celle-ci propose une décomposition des critères de la qualité des données, critères qu'elle distingue selon leur nature absolue ou relative à la donnée. L'auteur cherche à sélectionner avec pertinence les dimensions selon lesquelles la qualité des données dans une base de données multi-sources peut être évaluée. Elle propose une conception relative de la qualité des données, intégrant la qualité de la source et l'adéquation aux attentes de l'utilisateur à un système de recommandation (Berti, 1998; Berti, 1999a). Sa modélisation de données issues de sources différentes et potentiellement discordantes a également suscité notre intérêt, comme la question de la sélection des dimensions que nous nous posons pour la confiance et à laquelle nous proposons une réponse au chapitre 3.

Mesure de la qualité

S'il est important de définir les axes selon lesquels la qualité doit être estimée, il est essentiel de pouvoir évaluer selon ces dimensions. Ceci fait l'objet des outils et frameworks de mesures de la figure 2.3, autre branche offrant nombre de publications. En effet, un reproche souvent rencontré est la mesure des critères de qualité (Naumann & Rolker, 2000; Kahn et al., 2002). Knight et Burn (2005) répondent à cette doléance en proposant un cadre pour la conception d'un robot d'indexation qualitatif dans la problématique de la qualité sur internet. Nous nous attachons, dans nos propositions, à définir les dimensions d'évaluation de la confiance et nous montrons comment nous concevons leur évaluation.

Amélioration de la qualité : personnalisation et contextualisation

Un autre axe d'études du domaine de la qualité des données concerne les techniques et outils favorisant son émergence. Parmi les solutions proposées, certaines s'attachent à l'adéquation de la donnée aux besoins de son commanditaire. Ainsi, Simmhan et al. (2006) abordent le problème de la sélection de jeux de données pour les problèmes scientifiques disposant d'ensembles de tests concurrents. Diemers (1999) propose un point de vue sociologique constructiviste de ce que représente la conformité de la connaissance aux besoins et attentes de l'utilisateur conduisant à personnaliser les réponses apportées.

Les méthodes visant à augmenter la probabilité de transmission de données de qualité sur internet sont également largement représentées. On retrouve, par exemple, une combinaison de la sélection des critères d'intérêt associés à une recherche dans un contexte médical particulier (Darmoni et al., 1998). La personnalisation par domaine est également présente chez Tang et al. (2003) qui considèrent la prédiction automatique de la qualité d'articles

de presse pour des opérateurs de renseignement. Enfin, les modèles adaptés s'intéressent à la vérification contextuelle de la qualité d'information, comme l'évaluation de la qualité de blogs iraniens, déplacement de l'évaluation de la qualité des données, les posts, vers la source (Kargar et al., 2008). Tous ces travaux ont en commun le besoin de formalisation de la qualité de l'information dans un vaste volume de données librement disponibles et lui apportent la prise en compte du contexte et de la demande pour solution.

La qualité en recherche d'informations

Enfin, certains problèmes plus généraux bénéficient également de réponses qualitatives. Ainsi, Jain et al. (2009) proposent une optimisation orientée qualité de l'extraction d'informations et l'utilisent pour la combinaison de systèmes d'extraction (Jain & Ipeirotis, 2009). Nous utilisons leur conception de l'importance de la pertinence du résultat plutôt que de la méthode lors de notre recherche d'informations sémantiquement riches dans le chapitre 7, page 117. Zhu et Wang (2008), quant à eux, offrent un point de vue sur l'utilisation de la qualité des données dans le renseignement militaire illustrant également la nécessité de contenu pertinent plutôt que volumineux.

Bilan

La qualité des données est une source de méthodes et outils pour l'évaluation d'informations. Cependant, la conception générale du sujet touche plus à une forme de surveillance de la cohérence et complétude de données structurées. Certaines approches du problème ressemblent à notre façon d'entendre la cotation mais nécessitent d'être reformulées pour y répondre pleinement. Quand la qualité des données qualifie l'information, plutôt que le modèle dont elle est issue, elle représente à quel point celle-ci s'intègre dans ce modèle qui, lui, décrit efficacement la réalité. La cotation déplace l'objet de l'évaluation de l'information vers sa réception par son destinataire. Cet éloignement du contenu pour aborder le subjectif se rapproche de celui qu'il est nécessaire d'appliquer aux outils de la modélisation de l'incertain pour les amener à représenter la confiance.

2.4 Représentations de l'incertitude

La confiance en une information est une notion subjective difficilement exprimable. Plutôt que de dire si une information est crédible, elle représente si la façon dont cette information arrive à son auditeur le convainc. Bien que son objet diffère de ceux qu'étudient les travaux sur l'incertitude de l'information, ces derniers disposent d'outils riches permettant la formulation de la confiance, sous réserve d'opérer ce glissement sémantique de leur interprétation. Wang (2001) propose, par exemple, d'envisager la confiance comme

une forme de méta-incertitude.
Nous nous intéressons donc au domaine de la représentation de l'incertain, dans lequel il nous semble important de séparer l'incertitude liée à des phénomènes mesurables de l'incertitude perçue. Le doute que la première décrit émane de l'impossibilité de prédire quelle solution envisagée est la bonne, là où celui de la seconde porte sur la difficulté de définition des solutions possibles. Nous avons choisi, pour procéder à quelques brefs rappels sur les théories de l'incertain, de scinder les approches probabilistes des approches plus qualitatives, que nous qualifions de cognitives dans la mesure où les défauts qu'elles représentent et l'incertitude qu'elles manipulent reflètent, comme nous cherchons à le faire, des principes du raisonnement humain (Bouchon-Meunier & Nguyen, 1996; Wang, 1995).

2.4.1 Incertitude probabiliste

Depuis leur origine, les probabilités traitent d'incertitude. L'étude des jeux de chance par Fermat et Pascal, au milieu du dix-septième siècle, se préoccupe de découvrir un modèle fréquentiste décrivant la répartition d'événements aléatoires. À leur suite, John Graunt pose les bases des études statistiques. Il faut attendre Bayes et, surtout, Laplace pour que les probabilités commencent à être vues comme une 'croyance personnelle' et que l'incertitude probabiliste qualifie la confiance en un événement, plutôt que son acceptation ou son rejet. De là découlent les réseaux bayésiens, exploitant les probabilités conditionnelles pour représenter et calculer les relations entre informations et l'état courant des connaissances que l'on peut rapprocher de l'étape de crédibilité définie précédemment (cf. section 1.2.4, p. 14). Cependant, nous avons montré que la crédibilité n'était pas suffisante pour représenter la confiance en une information. En outre, l'utilisation de ces modèles suppose une connaissance a priori des informations à évaluer, hypothèse forte peu appropriée au cadre de la cotation tel que nous le décrivons.

2.4.2 Incertitude cognitive

Partant de l'historique représentation de l'incertitude par les outils probabilistes, deux voies ont vu le jour afin d'enrichir les représentations de l'incertitude : d'une part, des assouplissements des contraintes probabilistes (Shafer, 1990) et, d'autre part, des extensions des modèles logiques (Zadeh, 1965) [2]. Cette alternative cherche à étendre les formalismes de raisonnement afin d'adhérer aux riches faiblesses de sa version humaine.

2. Où l'on considère que la théorie des ensembles flous donne naissance à la logique floue.

Extensions probabilistes

Parmi la première variante de la modélisation de l'incertain, la théorie de l'évidence de Dempster-Shafer tient un rôle prépondérant. Shafer (1976), utilisant les travaux de Dempster, propose un raisonnement sur les 'croyances' dans un monde fermé. De la connaissance des masses de croyance sur les parties de l'univers est déduite la croyance en et la plausibilité d'un état, la répartition des masses de croyance permettant, en outre, de représenter jusqu'à l'ignorance. Les combinaisons de masses de croyances lors de l'agrégation d'événements généralisent les théories bayésiennes. Les subtilités des règles de combinaisons sont illustrées de manière limpide et joviale par Jøsang et Pope (2011). Par la suite Smets et al. (1991) et Smets (1998) distinguent les niveaux de croyance et de décision en réintroduisant une probabilité ultérieure à la 'combinaison de l'évidence' et l'interprétation pignistique de ladite croyance. Le cadre de la théorie de Dempster-Shafer a servi à la modélisation de la cotation, notamment dans l'article de Nimier et Appriou (1995) évoqué plus haut.

Faisant le lien entre la théorie de l'évidence et les logiques non-classiques, Jøsang (2008) propose la logique subjective, sorte d'adaptation de logiques probabilistes. La logique subjective tient compte explicitement de l'incertitude et de la croyance. La fonction d'affectation est définie comme équivalente à une densité de probabilité de Dirichlet. En introduisant les probabilités dans l'univers de la théorie de l'évidence, l'auteur bénéficie de la modélisation de l'ignorance. Après avoir présenté l'expression 'd'opinions subjectives', il introduit les opérateurs d'inférence permettant de raisonner et propose une représentation de la confiance ainsi qu'une formalisation de l'impossibilité de choix.

Extensions logiques

Plusieurs approches logiques s'attachent à représenter le raisonnement sur des informations dont les imperfections les rendent incertaines ou imprécises. Parsons et Hunter (1998) s'intéressent particulièrement à la correspondance entre variantes des défauts de l'information et du raisonnement et les théories les abordant. Un de ces formalismes est proposé par Zadeh (2005) lorsqu'il rappelle que la logique booléenne est mal équipée pour représenter des unités de mesure pourtant naturelles à un humain. Exploitant ses travaux sur les ensembles flous, Zadeh définit une logique dans laquelle les propositions sont 'plus ou moins' vraies. La théorie émanant de travaux ensemblistes, les opérateurs sur les degrés de vérité sont l'intersection ou l'union. Préservant, aux limites, les résultats de la logique booléenne, la logique floue offre un modèle puissant de raisonnement sur des propriétés mal définies, incomplètes ou imprécises. Cependant, l'utilisation d'un modèle si riche ne se justifie que si l'on sait décrire les partitions floues des variables étudiées, le raisonnement sur les degrés de vérité intermédiaire n'intervenant qu'après leurs définitions.

Une autre version des logiques multivaluées existe pourtant où, malgré la formalisation

en multi-ensembles de Akdag et al. (1992), la description des appartenances est accessoire. Les logiques multivalentes, dites encore symboliques, offrent elles aussi un modèle assoupli de la vérité d'une proposition (Akdag, 1992). Les degrés de vérité s'y expriment selon une échelle totalement ordonnée et les opérateurs sont redéfinis pour maintenir la compatibilité avec la logique classique (Seridi & Akdag, 2000). Ces degrés intermédiaires entre le 'totalement vrai' et le 'parfaitement faux' permettent de manipuler le doute et l'incertitude dans un cadre formel. Bien que les outils du raisonnement logique soient disponibles dans le cadre symbolique, nous nous intéressons principalement ici à l'expressivité des degrés de vérité que nous exploitons au chapitre 5 pour décrire la gradualité de la confiance.

2.5 Fusion d'informations

Nous avons vu que la cotation s'applique à des objets de granularités différentes et qu'elle comprend une étape de création de connaissances enrichies à partir d'informations atomiques, une étape de fusion. Il faut souligner qu'il s'agit de fusion et non d'agrégation, notions distinguées par l'hétérogénéité des données manipulées. En effet, là où l'agrégation combine des termes de même nature, la fusion, elle, assemble des objets potentiellement de natures différentes (Bouchon-Meunier, 1998). Si l'objectif de l'agrégation est généralement de résumer les données en profitant de connaissances distinctes, la fusion a pour but l'un ou plusieurs des points suivants, proposés par Dubois et Prade (2004) :

- améliorer la connaissance sur l'état du monde courant
- mettre à jour les informations sur un point d'intérêt
- dévoiler un consensus, le point de vue d'un groupe
- déceler ou améliorer la connaissance générale à partir de données

La problématique que nous étudions répondant au moins aux premier et dernier de ces points, elle s'inscrit bien dans le cadre de la fusion. Nous nous intéressons à l'évolution de la confiance lors de la constitution de connaissances et nous ne nous attardons donc pas sur la littérature traitant spécifiquement de l'opération de combinaison des éléments atomiques. Un cadre de discussion autour de ce processus de fusion est proposé par le modèle JDL (Llinas et al., 2004; Blasch et al., 2006).

Avant de parler de l'agrégation de degrés de confiance, la problématique qui nous importe, nous revenons sur un problème préalable à la fusion, la comparaison. En effet, avant de pouvoir amalgamer des informations, il faut déterminer lesquelles participent à la même connaissance. Cette notion de corrélation, notamment étudiée par Besombes et Cholvy (2009), se rattache au problème plus général de la similarité, très étudié en dehors

du cadre de la fusion mais qui lui est néanmoins essentiel.
Bouchon-Meunier et al. (2008) présentent l'impact de la recherche sur les similarités dans l'intelligence artificielle en général et des applications réelles utilisant la similarité et la logique floue en particulier. Une variante particulière de la similarité consiste à comparer des objets décrits dans une ontologie. Gligorov et al. (2007) utilisent Google comme mesure pour les appariements approximatifs entre concepts ontologiques. Bernstein et al. (2005) proposent une étude plus détaillée sur les différentes mesures disponibles dans les ontologies. En partant d'un mètre étalon construit en interrogeant des êtres humains, ils regroupent les mesures et exploitent le 'clustering' obtenu pour proposer des mesures de similarité dites personnalisées.

Un autre aspect très débattu de la fusion concerne les opérateurs d'agrégation de degrés de certitude, opérateurs dépendant du cadre formel considéré (Benferhat et al., 2000; Bloch et al., 2001). Nous avons vu plus haut que la théorie de Dempster-Shafer se préoccupe également de la fusion de croyances. Jøsang et al. (2009) donnent, toujours sous l'interprétation probabiliste de la logique subjective, deux opérateurs de fusion, le premier cumulatif pour application à des sources indépendantes mesurant le même phénomène et le second moyennant des avis pour aboutir à un consensus. Ces opérateurs sont opposés à la règle conjonctive de Dempster, applicable aux raisonnements logiques. A contrario, Nimier et Appriou (1995) proposent d'utiliser la règle initiale de Dempster pour fusionner des informations émanant de sources diverses. Faux et Luthon (2007), quant à eux, arguent que la règle prudente de Denœux (2006) donne de meilleurs résultats lorsque les sources ne sont pas indépendantes.
Bloch (1996) offre une classification des opérateurs de fusion dans laquelle elle établit la catégorisation des opérateurs en fonction de leurs comportements et de leur interprétation cognitive. Detyniecki (2000), lui, présente l'ensemble des opérateurs d'agrégation de la logique floue et leurs propriétés. Il introduit également ses propres opérateurs de combinaison de degrés de vérité d'une même proposition, par opposition à la combinaison dans l'inférence, mêlant les degrés de vérité de différents éléments. Les deux auteurs font référence aux comportements des opérateurs allant de la prudence à l'enthousiasme. Nous verrons plus tard que nos propositions sont bien décrites par ces classifications.

2.6 Conclusion

Nous avons évoqué dans ce chapitre différentes disciplines dont les problématiques croisent celles qui nous occupent. Dans un premier temps, nous sommes revenu sur les domaines étudiant la confiance. Nous avons commencé par quelques points de vue éma-

nant de la recherche à connotations militaires. En leur sein, nous avons décrit d'autres formulations du problème de la cotation, des travaux considérant la sécurité des systèmes comme facteur de confiance et un examen de la représentation de l'information cotée. Nous avons pourvuivi l'exploration du domaine de la confiance par les problématiques du Trust Management. Parmi elles, nous avons retenu l'omniprésence de la fiabilité de la source dans l'évaluation de la confiance ainsi que l'utilité de représenter des comportements adaptés au destinataire de l'information.

Dans la suite du chapitre, nous avons décrit la cotation comme se situant à l'intersection de trois domaines de recherche : la qualité des données, la représentation cognitive de l'incertitude et la fusion de données incertaines. Fort de ce constat, nous avons étudié les outils que ces disciplines offrent à la modélisation de la cotation. De la qualité des données, nous avons retenu l'importance de la sélection des attributs permettant l'évaluation, la personnalisation et la contextualisation du modèle. Nous avons également montré pourquoi les outils existants de la qualité de l'information ne suffisent pas à représenter ce que nous souhaitons et décrit le glissement d'interprétation nécessaire à leur utilisation. Après avoir expliqué comment ce glissement se généralise à la représentation de l'incertain, nous avons fourni une brève introduction à ses modélisations. Dans cette introduction nous nous sommes particulièrement intéressé aux théories permettant la représentation de concepts ambigus, comme la gradualité de la confiance. Enfin, revenant sur le besoin de combiner des objets de granularités variables, nous avons effectué un rappel des études en fusion de données portant sur l'expressivité des opérateurs de combinaison de degrés de vérité.

Deuxième partie

Nouveaux critères, méthode et formalisme pour la cotation

Introduction à la deuxième partie

La seconde partie de ce mémoire rassemble nos propositions pour la modélisation de la cotation formalisant le processus d'établissement de la confiance aux niveaux conceptuel et formel. Elle s'articule autour des trois problématiques de la cote, la cotation et leur formalisation théorique.

Le chapitre 3 est consacré à la cote : il étudie les dimensions à prendre en compte dans l'établissement de la confiance en les caractérisant selon leur objet et leur influence. Cette caractérisation guide le choix de quatre dimensions indépendantes et non-redondantes couvrant l'ensemble du spectre décrit. Il propose de les apprécier selon un formalisme de représentation constitué d'une échelle discrète associée à des étiquettes linguistiques.

Le chapitre 4 considère le problème de la combinaison de ces dimensions définissant le processus de cotation. Il propose une philosophie de l'intégration des dimensions à la confiance en établissant l'architecture de la cotation. Celle-ci précise à la fois l'ordre de la prise en compte de ces dimensions et l'influence de leur projection sur l'évolution de la cote. Exploitant la souplesse du modèle, différentes stratégies de cotation sont introduites permettant de modéliser des postures de crédulité variables qui offrent des capacités d'adaptabilité et de personnalisation au modèle.

Le chapitre 5 présente la formalisation théorique du procédé de cotation ainsi que de la cote dans un cadre de logique multivalente étendue : afin d'accroître l'expressivité du modèle en autorisant la représentation de l'impossibilité d'appréciation, un nouveau degré de vérité est ajouté à logique multivalente classique. L'ensemble des propositions des chapitres 3 et 4 est traduit dans ce formalisme, notamment par des opérateurs de combinaison de confiance spécifiques. Grâce à ces opérateurs une nouvelle forme de personnalisation est définie, prenant en compte la dynamique de la conviction en permettant de doter la cotation d'un effet mémoire.

Chapitre 3

Dimensions de la cotation

Résumé

Dans ce chapitre, nous présentons nos propositions de dimensions pour la cotation. Pour chacune d'entre elles, nous offrons une définition et une interprétation de l'échelle discrète d'évaluation. Nous illustrons également la signification de chacune par des exemples d'application sur des données textuelles, puis sur des données issues de capteurs.

3.1 Introduction

Nous allons présenter ici les dimensions que nous retenons pour effectuer la cotation, extension de ce que nous proposions dans Besombes et Revault d'Allonnes (2008). Nos propositions émanent, à l'origine, des critères existants, tels que présentés au tableau 1.1, page 15. Pourtant, elles s'en démarquent sur plusieurs plans que nous détaillons ici. Parmi les critiques usuelles faites à la cotation telle qu'elle existe, les plus systématiques sont la difficulté de lecture et d'appréhension (Cholvy & Nimier, 2003; Cholvy, 2004; Nimier, 2004, voir aussi la section 1.2.5, p. 16). Nous proposons de résoudre le premier de ces points en proposant à l'utilisateur une valeur unique exprimant la confiance qu'il doit avoir en une information. Proposer une valeur sur une échelle graduée permet les comparaisons entre informations et leur tri selon la confiance, une nette amélioration de l'existant.

Notre objectif de lisibilité pour la cote et la cotation ne se limite pas à sa présentation. Nous clarifions les facteurs que nous proposons pour l'évaluer. En général, la qualité d'une

information est mesurée à sa conformité à un modèle dont on assure aussi la qualité (Batini & Scannapieco, 2006). Cette recherche de qualité se fait donc en deux temps. On cherche d'abord à produire un modèle capable de représenter précisément toute l'information attendue. On vérifie ensuite que chaque information est détaillée et non-redondante. Il existe, évidemment, de nombreux autres critères de qualité mais l'expressivité du modèle et la complétude de l'information sont incontournables. Bien qu'essentiels à l'évaluation de la connaissance contenue dans le modèle et l'information, ces critères ne cherchent pas à évaluer si l'utilisateur peut croire en une information en particulier. Un modèle d'excellente qualité assure que la connaissance qu'il décrit est complète, interprétable et exploitable. Un enregistrement de qualité dans un tel modèle fournit un maximum de connaissances. Nous voulons ajouter à ces garanties de qualité une estimation de confiance. Pour évaluer cette confiance, nous proposons de combiner des dimensions de granularités différentes et dont certaines témoignent du processus de création de l'information plus que de son contenu.

Les indicateurs que nous utilisons ont un impact sur la confiance que nous faisons à l'information, mais n'en font pas nécessairement partie : bien que l'information soit censée représenter un fait réel, il n'est pas toujours possible de vérifier son existence. On ne peut avoir foi en sa véracité qu'en se forgeant une opinion avec les moyens disponibles, comme son contexte et mode de production, sa conformité à nos connaissances et attentes ou sa validation par d'autres sources. Ainsi, pour savoir si une personne est grande, il faut non seulement la mesurer avec des outils adaptés, mais également avoir une idée de la répartition des tailles dans la population environnante. De plus, si cette information nous est rapportée, les premiers indices sur lesquels nous fondons notre foi sont liés à sa source.

Parce qu'il n'existe pas de norme absolue d'évaluation de la confiance en une information, nous nous sommes inspiré de principes gouvernant son élaboration. Les questions essentielles que nous proposons de poser pour savoir si nous pouvons avoir confiance en une information sont :

– Qui la fournit ?	– Qu'en sait-il ?
– Est-elle concevable ?	– Est-elle rapportée par ailleurs ?

Les deux premières questions, dans la colonne de gauche, concernent la source de l'information. Elles forment le premier a priori sur la confiance que nous ferons à l'information. Avant même qu'un interlocuteur nous transmette une information, la confiance que nous avons en lui va étalonner la mesure dans laquelle nous allons la croire. Notons que ce premier a priori est lié à la source et ne dépend pas de l'information. Après cette première

étape, la deuxième question porte sur la qualification de la source à s'exprimer sur l'information. Bien que dépendant toujours de la source, cette capacité est variable selon les informations transmises. Les deux questions restantes s'intéressent à l'information. Nous nous proposons de regarder si l'information semble réaliste. Notons qu'il ne s'agit pas ici de vérifier sa réalisation mais de valider sa compatibilité avec la connaissance du monde de l'utilisateur. Ceci ne concerne donc que l'information transmise et le cadre de discernement de l'utilisateur. La dernière question, en revanche, cherche à vérifier si l'information est connue par ailleurs.

On peut noter que ces questions séparent les dimensions liées à la source de celles liées à l'information, dans l'esprit de la séparation entre dimensions de qualité absolue et relative de Berti, 1999b (cf. annexe C.2, p. 175). Les critères que nous retenons se répartissent naturellement dans ces catégories. À cette distinction, nous en ajoutons une autre, selon que la dimension prend en compte le contexte ou non, si son influence est stable ou si elle varie selon le cas considéré. En effet, la transition entre questions sur la source et sur l'information n'est pas suffisante, nos propositions étant également séparables en fonction du métissage de ce qu'elles qualifient. Les quatre questions précédentes croisent ces deux axes de dépendance à la source et dépendance au contexte. De la sorte, les quatre dimensions que nous proposons se répartissent dans le tableau 3.1, où la fiabilité répond à la première question, la compétence à la seconde, la plausibilité à la troisième et la crédibilité à la dernière.

	Général	Contextuel
Source	Fiabilité	Compétence
Contenu	Plausibilité	Crédibilité

TABLE 3.1 – Axes des dimensions de la cotation

Si nous avons choisi cette séparation des dimensions, d'autres explications de leur indépendance et non-redondance sont possibles. Ainsi, on peut citer la subjectivité de l'interlocuteur : la fiabilité de la source et la plausibilité de l'information sont deux dimensions subjectives. De toute évidence, faire confiance à une source dépend de qui la consulte. La plausibilité de l'information, envisagée comme une compatibilité avec les connaissances sur le monde, varie également d'une personne à une autre. Les deux autres dimensions ne sont, en revanche, idéalement pas subjectives. Si sa connaissance n'est pas assurée, la compétence de la source ne dépend pas de qui la regarde. De même, la manière dont deux informations se confirment ne dépend pas du point de vue mais bien de la méthode d'évaluation. Ces

axes de séparation entre les dimensions proposées nous permettent de les considérer comme non-redondantes, ne qualifiant pas les mêmes objets et donc indépendantes.

Nous représentons le niveau d'activation de chacune des dimensions par un degré discret. Pour nous conformer à l'esprit du STANAG 2022, nous considérons une échelle à six niveaux, sans que ce choix soit limitatif. Évidemment, dès que l'on s'éloigne du domaine du renseignement militaire, ou que l'on s'approche de celui du calcul automatique, ces échelles peuvent avoir plus ou moins de degrés selon l'application visée. Nul besoin de se contenter des six degrés définis ici, même s'il a été établi qu'une échelle à plus de sept valeurs perd en explicitation ce qu'elle gagne en précision (Miller, 1955).

Dans ce qui suit, nous détaillons successivement ces quatre dimensions pour la cotation. Pour chacune d'entre elles, nous offrons une définition et une interprétation de l'échelle discrète d'évaluation. Nous illustrons également la signification de chacune par deux exemples d'application, d'abord sur des données textuelles, puis sur des données issues de capteurs.

3.2 Évaluation générale de la source : la fiabilité

Définition

La qualité de la source est une qualité intrinsèque de l'information, selon les dénominations de Wang et Strong (1996) et de Berti (1999b), rappelées dans les tableaux C.1 et C.2 (pp. 173 et 175). C'est un critère incontournable qui sert, habituellement, à pondérer l'interprétation de l'information. L'algorithme PageRank de Brin et Page (1998), par exemple, utilise la confiance faite à la source, le site web, pour trier les réponses à une requête, à pertinence attendue égale.

Nous proposons de distinguer dans cette qualité ce qui est général de ce qui est contextuel et discutons ici sa dimension générale, que nous appelons *fiabilité*. Celle-ci ne dépend ni de la question posée, ni d'informations déjà connues et s'applique de la même manière à toutes les informations que la source fournit. La 'fiabilité de la source' qualifie donc exclusivement la source de manière indépendante de l'information.

Il faut noter que ce terme est fréquemment utilisé pour diverses notions : on y trouve tout à la fois la réputation ou confiance réelle [1] comme nous la comprenons, l'aptitude que la source a à fournir l'information ainsi que ses éventuels doutes sur l'information elle-même (DISS & DISC, 2001; Department of the Army Headquarters, 2005). Ces dimensions confondent donc les composantes générales – ne concernant que la source – et contextuelles

1. 'Trustworthiness'

– liées à d'autres dimensions. Dans l'acception que nous en proposons, la fiabilité ne dépend que de la source et non de l'information qualifiée. Il est, toutefois, notoire que la fiabilité dépend de là où l'on se place et, en conséquence, de qui l'évalue. Nous verrons dans nos exemples que si la fiabilité ne dépend pas de l'information elle est intimement liée à l'utilisateur.

Nous proposons d'exprimer la fiabilité sur une échelle à cinq niveaux, comme indiqué dans le tableau 3.2, représentant les degrés de fiabilité envisagés ainsi que leurs interprétations. Afin de conserver un lien à l'échelle d'origine, nous représentons ces degrés à l'aide de lettres. Nous complétons cette échelle par une valeur permettant de représenter l'impossibilité de mesurer la fiabilité, F. En effet, il arrive que les connaissances disponibles sur la source soient insuffisantes pour estimer sa fiabilité. Ceci se produit, par exemple, à la rencontre d'un interlocuteur jusqu'alors inconnu. La fiabilité estimée de ce dernier peut ainsi être distinguée de celle d'un interlocuteur connu mais dont la fiabilité est moyenne, i.e. ne joue ni en sa faveur ni contre lui. Un tel interlocuteur serait associé à un niveau de fiabilité C. Notons que ce niveau complémentaire figure dans les échelles proposées pour chacun des critères considérés, comme détaillé par la suite.

Fiabilité	Définition
A	Complètement fiable
B	Habituellement fiable
C	Moyennement fiable
D	Habituellement pas fiable
E	Pas fiable
F	La fiabilité ne peut être appréciée

TABLE 3.2 – Fiabilité de la source : qualité intrinsèque appliquée à toute sa production

Il faut, par ailleurs, souligner le caractère dynamique de la fiabilité de la source, amenée à évoluer avec sa fréquentation. Ainsi, on peut en apprendre davantage sur une source et vouloir réévaluer sa fiabilité. Cette mise à jour pose le problème de l'évolution de l'évaluation de la confiance en l'information. Deux possibilités se présentent alors : recalculer cette confiance ou conserver la valeur actuelle comme témoignage du moment où on l'a calculée. Bien entendu, ces alternatives ne se présentent que pour les informations émises depuis le changement de fiabilité. Considérons, par exemple, une source dont on découvre, à une date t_β, qu'elle souffre de troubles de la mémoire depuis une date t_α (où $t_\alpha < t_\beta$).

Toute information fournie entre t_α et t_β est affectée par ce changement de fiabilité. Les informations émises à une date antérieure à t_α ne sont, en revanche, pas réévaluées.

Données textuelles

Dans le cas de l'extraction d'informations à partir de données textuelles, les sources des informations sont les auteurs des textes. Si l'on considère des textes issus d'un agrégateur de nouvelles, comme Google Actualités par exemple, les sources correspondent aux différents journaux traités, comme le Monde ou Voici. L'évaluation de leur fiabilité respective peut être déléguée à l'utilisateur pour prendre en compte sa composante subjective et permettre une adaptation personnalisée à l'utilisateur. Celui-ci peut affecter, a priori et indépendamment de toute information, une fiabilité maximale (niveau A, « complètement fiable ») au Monde et un niveau moins élevé (e.g. D, « habituellement pas fiable ») à Voici. On peut envisager d'intégrer des procédures d'estimation de réputation de source pour proposer une aide à l'utilisateur, telles que celles présentées à la section 2.2.2, page 28.

Données issues d'un capteur

On peut considérer que les capteurs électroniques sont des sources d'informations, ces dernières étant les mesures qu'ils fournissent. La fiabilité de la source est alors interprétée comme correspondant à l'état du capteur : un capteur en parfait état de marche est associé au niveau de fiabilité A, là où son équivalent installé sur un réseau d'alimentation aléatoire sera 'habituellement pas fiable' (soit de fiabilité D). Les capteurs électroniques sont fréquemment munis d'indicateurs de bon fonctionnement qui peuvent permettre d'évaluer automatiquement leur niveau de fiabilité. Cette fiabilité est ensuite exploitée quelle que soit l'information produite et a la même influence sur toutes les mesures effectuées.

3.3 Évaluation contextuelle de la source : la compétence

Définition

Après avoir défini la composante générale de la source, sa fiabilité, nous souhaitons intégrer à notre processus un axe souvent confondu avec elle. Nous avons vu que, telle que nous l'entendons, la fiabilité de la source s'applique de la même manière à toutes les informations qu'elle produit. Pourtant, on est souvent tenté, lorsque l'on considère l'influence que la source a sur la confiance que l'on fait aux informations qu'elle fournit, d'inclure une estimation de ses connaissances sur le sujet. On prendra toujours, par exemple, les échappées philosophiques de spécialistes scientifiques *cum grano salis*.

Nous proposons de dissocier cette dimension que nous qualifions de contextuelle. En effet, la capacité qu'une source a de fournir une information dépend de l'information elle-

même. Il est toujours possible qu'une source donne une information hors de son domaine d'expertise. Une telle information ne doit pas, pour autant, être rejetée pour cette unique raison, mais le destinataire doit en tenir compte lors de la classification de l'information ou lors de l'évaluation de sa confiance.

Nous appelons ce facteur la *compétence* de la source et proposons de la mesurer selon l'échelle donnée dans le tableau 3.3. On retrouve ici un niveau exprimant l'impossibilité d'évaluation. Il se peut, en effet, qu'on ignore l'étendue des domaines de compétence d'une source. Il arrive, par exemple, que certains informaticiens aient une culture philosophique insoupçonnée et soient parfaitement qualifiés à la partager. Nous proposons à nouveau de distinguer ainsi ces experts inattendus de ceux dont nous connaissons le vernis culturel.

Compétence	Définition
1	Expert
2	Compétent
3	Partiellement compétent
4	Insuffisamment compétent
5	Incompétent
6	La compétence ne peut être appréciée

TABLE 3.3 – Compétence de la source : fonction de l'information considérée

Données textuelles

Dans le cas de l'extraction d'information, la définition de la compétence est délicate. Quand l'évaluation est opérée par un être humain il peut estimer, dans la mesure de ses connaissances, la compétence de la source. On peut, par exemple, déterminer que la compétence du Monde en affaires politiques internationales est élevée et le dire compétent. De même, Voici est probablement incompétent en analyses géopolitiques mais expert en rumeurs et vedettariat. Pour un système automatique, évaluer une telle dimension est impossible. C'est sans doute pour cette raison que la compétence n'est pas un critère de qualité incontournable.

Pourtant, la compétence qualifie la faculté d'une source à produire une information sur un sujet. Dans un système d'extraction d'informations, il arrive que l'on distingue les sujets : un algorithme est chargé d'extraire les entités nommées, un autre les noms de lieux et un troisième de déchiffrer des acronymes. Envisagé de la sorte, chaque algorithme possède un domaine de compétence défini. En outre, les outils d'évaluation de performance

classiques nous offrent des mesures de cette compétence. Si l'algorithme d'extraction de noms de lieux a une F-mesure avoisinant 1, sa compétence est maximale et on le dit expert. Selon que l'on souhaite favoriser les vrais positifs ou les faux négatifs on peut choisir, comme à l'accoutumée, la précision ou le rappel pour évaluer la compétence des algorithmes.

Données issues d'un capteur
Un capteur électronique est étalonné pour fonctionner sur une certaine plage mais ceci ne l'empêche pas de fournir des informations au-delà. Un radar de contrôle de la vitesse sur le réseau routier est configuré pour fournir des vitesses accessibles aux véhicules y circulant. Confronté à un déplacement hors norme il fournit tout de même une mesure, mesure de moins en moins crédible qu'elle s'éloigne de sa plage de compétence initiale.

Une autre interprétation de la compétence d'un capteur électronique est son contexte d'utilisation. Prenons, par exemple, les conditions météorologiques. Ces conditions influencent toutes les observations faites pendant qu'elles s'appliquent. Un radar perd en compétence dans le brouillard. De même, il arrive à tout un chacun d'utiliser un appareil électronique dans des conditions hors des recommandations constructeur. On ne s'attend généralement pas pour autant à un refus de fonctionner de l'appareil, même si on n'accorde pas le même crédit a priori au résultat.

Dans l'un ou l'autre de ces cas de figure, le système doit intégrer cette estimation de la compétence pour pondérer la confiance.

3.4 Évaluation générale du contenu : la plausibilité

Définition
Après avoir tenu compte des facteurs influents de la source sur la cotation, nous nous intéressons à l'information proprement dite. Une fois encore, nous proposons de séparer la dimension générale d'une autre plus contextuelle. Un critère essentiel de la croyance en une information est sa *plausibilité*. Avant même d'en envisager les détails ou confirmations, nous rejetons toute information qui semble incompatible avec notre perception du monde. Cette première évaluation s'opère en ne considérant que l'information en question. La recherche de renseignements complémentaires et de confirmation interviendra plus tard.

Ainsi, dans le processus intellectuel dont nous nous inspirons, on n'étudierait pas la sûreté éventuelle d'une information dénuée de toute plausibilité. Ce réflexe de préservation intellectuelle sert, par exemple, aux détracteurs des mouvements créationnistes qui en proposent une contradiction par l'absurde avec leur *Flying Spaghetti Monster* (Van Horn & Johnston, 2007).

Pour estimer la plausibilité d'un fait, nous nous référons à sa compatibilité avec notre connaissance du monde. Il ne s'agit pas de déterminer si le fait est avéré ou non, mais d'estimer dans quelle mesure il *pourrait* se produire. Un exemple d'évaluations pour ce critère est présenté dans le tableau 3.4. Puisqu'il est possible que nos connaissances ne nous permettent pas d'évaluer la plausibilité d'une information, nous avons encore une fois choisi de proposer le niveau 6, représentant l'impossibilité de mesure. Les grands sauts de paradigmes scientifiques émanent certainement de génies ayant reconnu des situations inexplicables par leur connaissance actuelle du monde.

Plausibilité	Définition
1	Certaine
2	Réaliste
3	Possible
4	Peu possible
5	Impossible
6	La plausibilité ne peut être appréciée

TABLE 3.4 – Plausibilité de l'information : conformité à la connaissance extérieure

Données textuelles

La plausibilité mesure donc l'adéquation entre l'information et notre connaissance de l'état du monde. Pour l'extraction d'informations, nous l'interprétons comme sa correspondance au modèle, tant en terme de cohérence que d'exactitude. Afin d'intégrer la plausibilité à notre calcul de confiance, nous devons donc être en mesure d'évaluer cette conformité à la 'connaissance du monde' du système. Nous supposons, donc, que nous disposons de connaissances extérieures, par exemple stockées dans une ontologie.

Nous désignons par connaissances extérieures la connaissance sur l'état du monde, indépendamment des informations extraites. Si une information produite est en désaccord avec ces connaissances extérieures, nous considérons que cette contradiction est différente d'une contradiction avec des informations acquises. En l'occurrence il est vraisemblable qu'un doute en l'information émerge dans la mesure où notre connaissance du monde est, elle, supposée sûre.

Données issues d'un capteur

Si le capteur est, par exemple, un radar de vitesse braqué sur une route, en connaissant

la route et quelques rudiments de lois de la physique, il est possible de mettre en doute certaines mesures. Lorsque le radar observe un véhicule sortant d'un virage étroit sur un chemin de campagne à 150 km/h, nous supposons que la mesure est peu possible et en évaluons la plausibilité à 4. Ceci est différent de la comparaison avec d'autres mesures puisque, notre connaissance du monde étant sûre, elle ne doit pas être remise en cause par les informations que nous évaluons. En revanche, les informations acquises peuvent – et doivent – être réévaluées.

3.5 Évaluation contextuelle du contenu : la crédibilité

Définition

Après avoir été convaincu des capacités de la source – de manière générale puis au regard de l'information considérée – et avoir vérifié que l'information n'était pas trop incompatible avec notre connaissance du monde, nous procédons généralement à la vérification du fait. Lorsque nous apprenons une nouvelle, la majorité d'entre nous va chercher à la confirmer afin de finir de s'assurer qu'elle est crédible. Ce phénomène de recherche de confirmations explique, sans doute, en partie le succès des suggestions des moteurs de recherches, construites à partir des requêtes les plus fréquentes et peut être aussi quelques 'buzzes'. Cette ultime vérification est une recherche de recoupement et de corroboration de l'information encore inconnue, provenant, de préférence, de sources fiables et compétentes.

En raison du degré maximal de la *crédibilité* de la grille de cotation existante (voir tableau 1.1, p. 15), nous choisissons de l'évaluer comme un indice de confirmation par les informations acquises. Nous devons souligner que les faits auxquels nous comparons l'information étudiée sont eux aussi soumis à la même procédure d'évaluation. Ils proviennent tous de sources plus ou moins fiables, plus ou moins compétentes et leur plausibilité est sans doute comparable à celle de l'information considérée. À chaque confirmation ou infirmation, la combinaison des degrés de confiance a priori (i.e. la cote issue des dimensions précédentes) va pondérer l'impact de la confirmation et nous permettre d'établir enfin si nous croyons l'information ou non. De plus, les informations auxquelles nous confrontons celle que nous estimons bénéficient, elles aussi, de cette mise à jour. Ceci correspond à la qualité relative aux données homologues, telle que décrite par Berti (1999b). Cependant, la crédibilité, en tant qu'indicateur de confirmation, ne dépend pas de la cote a priori. Celle-ci joue sur l'impact de la confirmation plutôt que sur sa mesure. La crédibilité sera donc évaluée en utilisant les mesures de correspondance ou de conflit avec les autres informations, puis intégrée en les combinant avec leurs cotes.

Nous insistons à nouveau sur la distinction faite entre les comparaisons entre informations acquises et leur compatibilité avec le modèle de raisonnement. Les informations que nous comparons en évaluant la crédibilité sont toutes construites et ont donc toutes un indice de confiance révisable, par opposition aux connaissances du monde dont la confiance est fixe et maximale.

Le tableau 3.5 présente les différents niveaux que nous proposons pour évaluer la crédibilité. Parmi ces degrés de confirmation on remarque la confirmation et l'infirmation partielles (niveaux 2 et 4). Nous avons proposé une procédure cumulative de confirmation selon laquelle il n'est pas nécessaire de n'être confirmé que par des informations certaines pour atteindre une crédibilité maximale (Revault d'Allonnes et al., 2007b). L'accumulation de suffisamment de confirmations par des informations moins sûres apporte le même résultat, bien que moins rapidement, comme détaillé à la section 5.5.3, page 93. Cependant, ce que désignent confirmation et infirmation partielle tient plus de la question de l'évaluation de la corrélation entre informations. Quiconque a cherché confirmation d'une nouvelle complexe sait qu'il est rare d'en trouver confirmation complète et parfaite, surtout à insister sur des provenances indépendantes. Il arrive aussi qu'on ne puisse faire le lien entre deux informations tout en pressentant qu'elles traitent du même sujet. Pour ces cas, nous proposons une fois de plus une crédibilité inestimable de niveau 6.

Crédibilité	Définition
1	Totalement confirmée
2	Partiellement confirmée
3	Insuffisamment confirmée
4	Partiellement contredite
5	Totalement contredite
6	La crédibilité ne peut être appréciée

TABLE 3.5 – Crédibilité de l'information : indice de confirmation entre informations produites

Données textuelles

Revenons à notre exemple initial de traitement de nouvelles issues d'un agrégateur, comme Google Actualités. Ces agrégateurs offrent généralement différentes sources pour une même nouvelle. Un système de traitement automatique désireux d'établir un résumé d'une de ces nouvelles doit faire appel à une batterie d'algorithmes d'extraction puis de traitements.

Il va, par exemple, chercher à dater l'événement et à établir une liste de protagonistes. L'extraction de dates fait partie des tests communs d'évaluations de systèmes d'extraction (ACE07 2007; 2007). La difficulté principale de la datation automatique d'un événement réside en la confirmation et la complétion des dates. En effet, la majorité des articles de presse comporte des dates. On suppose, en général, que la date de l'événement décrit revient plus fréquemment et sous une forme plus complète. Le système d'extraction des dates extrait donc toute date potentielle de l'événement. Chacune de ces dates provient d'une source à la fiabilité supposée connue. Il peut être plus compétent pour l'extraction des jours de la semaine que pour ceux du mois, pouvant être confondus avec une autre quantité. La plausibilité de sa production peut être vérifiée à l'aide d'un calendrier. Chaque élément extrait par le système possède donc une cote a priori. Lors de la recherche de corroboration, le système cherche d'une part à construire des dates complètes (e.g. « lundi 21 mai 1973 ») puis à les confirmer. Le procédé proposé ici calcule le degré de confirmation entre deux dates potentielles en fonction de leur recoupement et dira que le '21/05/73' confirme totalement le 'lundi de la 21 $^{\text{ème}}$ semaine de 1973'.

Données issues d'un capteur

De manière analogue, la crédibilité dans le traitement de données issues de capteurs électroniques intervient lors de la fusion d'informations. Si, après la mesure peu plausible réalisée précédemment (cf. section 3.4) un second capteur constate la présence d'un autre bolide, l'information initialement douteuse est confirmée et nous chercherons à vérifier le potentiel déroulement d'un rallye dans la région.

3.6 Expression globale de la confiance

À partir des quatre dimensions définies et décrites dans les paragraphes précédents notre procédé propose, en sortie, un degré de confiance associé à l'information considérée, sa *cote*. Celui-ci exprime, en un score unique et comparable, la combinaison des critères ci-dessus. Nous proposons d'exprimer la cote selon le tableau 3.6.

Il serait certainement possible d'intégrer d'autres dimensions au calcul de la cotation d'une information. Nous avons, par exemple, choisi de ne pas tenir compte de la conviction de la source. Dans la mesure où nous serions à même d'estimer ce doute, il va sans dire qu'il serait capital de l'inclure à la cotation. Ceci poserait, cependant, des questions sur le calcul de l'importance du doute, selon la fiabilité et la compétence d'une source.

Dans le même esprit, on pourrait envisager un facteur reflétant l'influence du bouche à oreille sur la déformation de l'information. Outre la difficulté de détection puis d'évaluation

Confiance	Définition
❶	Extrêmement probable
❷	Probable
❸	Envisageable
❹	Douteuse
❺	Improbable
⑥	La confiance ne peut être appréciée

TABLE 3.6 – Cote finale : degré de confiance associé à une information

automatiques de l'impact d'un tel facteur, nous pensons que les doutes soulevés par ce genre de techniques le seraient également par la nôtre. Si, toutefois, il s'avérait possible de reconstruire l'information d'origine, voire de localiser les responsables des altérations (Shah & Zaman, 2009), l'apport à la cotation serait précieux.

3.7 Conclusion

Nous avons présenté dans ce chapitre les dimensions selon lesquelles nous proposons de mesurer la cotation. Nous concevons ces dimensions comme les éléments incontournables influant sur la confiance en l'information. À la différence des habitudes, notamment chez les spécialistes de la qualité de l'information, où l'on cherche à qualifier la mesure dans laquelle le modèle peut produire une information de qualité ou dans laquelle la donnée remplit les critères de ce modèle, nous proposons de projeter ces dimensions sur l'information.

Nous avons expliqué pourquoi ces critères nous semblaient importants et comment nous proposons de les évaluer. Désirant intégrer l'influence de la source à la conviction d'un utilisateur, nous avons préféré en distinguer l'aspect général, la fiabilité, de la compétence, qui ne peut s'estimer qu'en regard de l'information. La cotation de l'information n'étant pas envisageable sans la considérer, nous proposons ensuite d'intégrer une qualité générale, sa plausibilité, et une autre contextuelle, sa corroboration ou infirmation par les connaissances acquises.

Nous pensons que la combinaison et la projection de ces facteurs sur l'information permettent d'automatiser la cotation. Dans le chapitre suivant, nous montrons comment

nous proposons de les combiner et nous introduisons également un aspect de modélisation utilisateur.

Chapitre 4

Combinaison des dimensions : une architecture modulable

Résumé

Nous proposons une architecture de la cotation : nous commençons par présenter l'ordre de combinaison des dimensions, puis nous présentons la synthèse générale de notre proposition, sous la forme d'une chaîne de cotation. Nous expliquons ensuite l'influence des dimensions sur l'évolution de la cote. Nous introduisons également le principe de stratégie grâce auquel la chaîne de cotation peut être adaptée à différents utilisateurs.

4.1 Introduction

Dans ce chapitre nous revenons sur le modèle cognitif qui sous-tend nos propositions. Nous nous fondons sur ce formalisme pour proposer une architecture lisible de la cotation. Nous commençons par établir l'ordre naturel d'intégration des dimensions. De cet ordre, nous exprimons une structure séquentielle de la cotation. Nous synthétisons ensuite nos propositions d'architecture en introduisant une chaîne de cotation.

Nous présentons, par la suite, l'influence de chaque dimension sur la cotation. Trois classes se distinguent : l'initialisation, les dimensions décroissantes et la crédibilité, seule dimension non-monotone.

Exploitant les degrés de liberté qu'offrent les intégrations des dimensions, nous introduisons la notion de postures de crédulité de l'utilisateur et l'outil servant à les modéliser, les stratégies. Nous présentons et illustrons l'emploi des stratégies et leur utilité à l'aide de trois postures caractéristiques dans un exemple.

Enfin, nous terminons ce chapitre par une discussion revenant sur le formalisme séquentiel et évoquant la possibilité de développement de la mise à jour automatique de la fiabilité de la source.

4.2 Architecture de la cotation : caractéristiques

Nous avons déjà évoqué le procédé qui motive nos propositions d'évaluation de la cotation. Jusqu'alors nous présentions le principe général de construction d'opinion, dans l'introduction générale à ce mémoire, puis en extrayions les axes selon lesquels nous proposons d'évaluer la cote, dans le chapitre précédent. Le processus sous-jacent est central à nos propositions sur la cotation. Il structure non seulement les dimensions selon lesquelles nous cotons l'information, mais également la façon dont nous les combinons, de l'ordre d'intégration à sa conséquente séquentialité. Nous revenons ici en détails sur ce modèle qui sous-tend nos travaux et ses implications, puis discutons, à la section 4.5, de certaines de ses conséquences.

Nous considérons la cotation de l'information comme un indicateur de la confiance qui lui est faite. Afin d'établir un processus de cotation interprétable, nous nous donnons un modèle d'évaluation de cette confiance. C'est ce modèle que nous résumions, au chapitre précédent, par les quatre questions appliquées à une information :

1. Qui la fournit ?
2. Qu'en sait-il ?
3. Est-elle concevable ?
4. Est-elle rapportée par ailleurs ?

Rappelons que nous ne cherchons pas à évaluer la certitude de l'information : nous ne nous intéressons pas à savoir si l'information s'est réalisée mais bien à déterminer dans quelle mesure elle nous convainc, qu'elle soit précise, vague ou incertaine. Pour arriver à ce résultat, nous considérons le processus nous permettant de nous faire un avis sur une information jusqu'alors inconnue. La définition des facteurs que nous intégrons à l'évaluation

de la cotation établie, nous nous intéressons maintenant à l'ordre de leur prise en compte qui dérive de notre modèle. Pour illustrer ce processus, considérons ce qui se passe lorsque nous apprenons une nouvelle.

Ordre d'intégration des dimensions

Quelle que soit l'information à laquelle nous nous intéressons, le premier a priori qui influence notre avis est la confiance que nous faisons à l'organe nous la présentant. De fait, lorsque nous cherchons une information, nous privilégions les sources sûres, comme le font pour nous les moteurs de recherche. De la même manière, notre avis est déjà orienté par la personne qui nous la transmet, en l'occurrence, supposons-le, un ami très sûr.

Ce n'est qu'ensuite que nous envisageons la compétence de la source sur le sujet de l'information, aidé par notre connaissance de ses spécialités, voire de l'intersection de nos goûts. Si l'information traite de recherches émergentes en théorie des nombres, on pourra considérer que notre ami – artiste photographe hermétique à toute notion scientifique – n'est pas la meilleure source, sans pour autant remettre en question sa fiabilité usuelle.

Qu'elle soit rapportée comme certaine ou, plus probablement ici, de manière vague et imprécise, nous considérons ensuite si l'information contredit ou non nos connaissances. L'annonce, par exemple, de la découverte de la démonstration originelle par Fermat de son théorème, bien que surprenante, n'est pas inenvisageable.

Enfin, pour parfaire notre avis, nous cherchons à vérifier si l'information est recensée ailleurs. La recherche de confirmations suit un cheminement similaire et les recoupements qui en découlent nous permettent enfin de décider si nous croyons l'information.

Nous insistons ici sur le processus d'intégration des dimensions à la cotation. La confiance que nous faisons à notre ami est un facteur indépendant, comme déjà souligné au chapitre 3, de ce qu'il nous dit, dont nous projetons l'effet sur tout ce qu'il nous confie. Sa compétence sur la question abordée, ou du moins notre perception de cette dimension, n'intervient qu'après dans la constitution de notre avis. De même, la plausibilité de l'information transmise puis sa corroboration interviennent successivement par la suite. De ces constatations sur la formation d'un avis, validées notamment par la littérature sur le renseignement (DISS & DISC, 2001) et sur le journalisme (Borden & Harvey, 1998, par exemple), nous déduisons l'ordonnancement de la chaîne de cotation représentée à la figure 4.1. De là provient aussi la notion de projection des dimensions *sur* l'information, puisque celles-ci sont constitutives de notre conviction plus qu'elles ne le sont de l'information proprement dite.

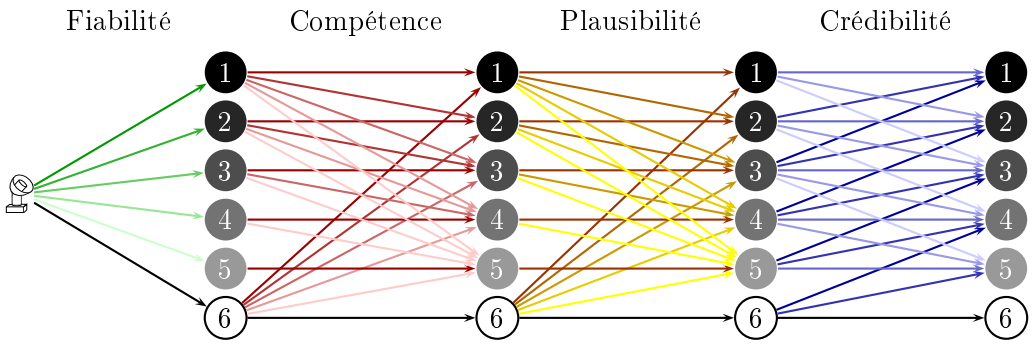


FIGURE 4.1 – La chaîne de cotation proposée

Séquentialité de la chaîne de cotation

La chaîne de cotation présentée à la figure 4.1, que la suite de ce chapitre propose d'éclairer, formalise donc l'élaboration de la croyance. Outre l'ordre d'intégration des dimensions, la chaîne de cotation la présente également comme un processus séquentiel. L'acceptation de notre modèle autorise cette conception et ce choix offre une plus grande interprétabilité à la cotation. Les exemples illustratifs proposés par la suite mettent en avant l'intérêt d'un processus compréhensible de construction d'opinion. En outre, la disponibilité à tout moment d'un degré reflétant l'avancement dans le processus permet de suivre, comprendre et valider le résultat obtenu. Les calculs séquentiels ne sont pas plus coûteux et l'évolution de la cote traduit bien la projection des dimensions.

4.3 Architecture de la cotation : description

L'architecture que nous proposons pour la cotation est donc une formalisation de l'établissement de la confiance que nous lui faisons représenter. La chaîne de cotation de la figure 4.1 synthétise son fonctionnement. Ce schéma représente à la fois l'évaluation des critères ainsi que leur impact sur l'évolution de la cote.

Le processus séquentiel de cotation est décomposé en quatre étapes. La cote courante, après intégration des dimensions déjà évaluées, y est représentée par son niveau actuel, chiffre de 1 à 6 encerclé en nuances de gris pour chaque étape. L'échelle complète est donc présente quatre fois dans la figure 4.1. Le capteur représente la source, dont la fiabilité peut être activée à des niveaux différents. Les flèches qui en émanent symbolisent l'impact de leur niveau d'activation sur la première estimation de la cote, à laquelle elles aboutissent. Par la suite, les flèches figurent les niveaux d'activation de la dimension considérée et en indiquent l'influence sur la cote. L'intensité de l'activation du critère est symbolisée par

une variation de la teinte : plus la flèche est sombre, plus l'évaluation de la dimension est intense. À l'inverse, les flèches claires représentent une intensité minimale.

Rappels sur l'évaluation des dimensions
Comme nous l'avons vu au chapitre précédent, l'évaluation de chaque critère, comme celle de la cote, s'opère sur une échelle différente mais ayant la même construction, i.e. les degrés d'activation des critères vont de 1 ≡ 'haute' à 5 ≡ 'basse', auxquels nous ajoutons un degré d'indécision 6 ≡ 'inquantifiable'. Cependant, chaque évaluation possède sa propre interprétation, comme les tableaux 3.2 à 3.5, pages 51 à 57, le précisent.
Ce rappel veut attirer l'attention du lecteur sur deux points : le comportement particulier à la rencontre d'une cote 'inquantifiable', d'une part, et les multiples flèches représentant l'impact de la prise en compte des dimensions correspondent à ces niveaux d'activation, d'autre part. En conséquence des contraintes imposées à l'évaluation, certaines flèches sont confondues, comme détaillé par la suite.

Nous détaillons, à présent, l'utilisation de la chaîne de cotation et présentons les contraintes attendues sur l'influence de l'intégration de chaque dimension. Nous revenons, à la section 4.4, sur les possibilités offertes par le jeu sur ces influences.

4.3.1 Fiabilité de la source

Dans le modèle d'évaluation de la confiance et dans la chaîne de cotation, nous proposons que le premier critère à entrer en jeu soit la fiabilité de la source nous apprenant une information. Nous utilisons donc la fiabilité de la source comme initialisation de la chaîne de cotation.

Supposons que nous ayons une nouvelle information entrante, donc jusqu'alors non-évaluée. Puisque cette information est inconnue, elle est crue à la hauteur de la confiance faite à sa source. La première étape du processus attribue, ainsi, à une information inconnue le niveau de confiance (tableau 3.6 page 59, niveaux de gris dans la figure 4.1) équivalent à la fiabilité de sa source.

4.3.2 Compétence et plausibilité

L'étape d'intégration de la compétence dans notre processus a un impact négatif sur la confiance. En effet, l'initialisation de la cotation s'opérant au niveau de la confiance faite à la source, on aura sans doute supposé que ladite source 'sait de quoi elle parle' ou, du moins, que nous lui faisons confiance à hauteur de cette supposition. Il est donc naturel que le seul

impact possible de la compétence soit négatif. Une source de toute confiance fournissant une information hors de son domaine d'expertise nous pousse à prendre l'information avec circonspection, alors qu'une source connue pour ses mensonges ou ses manipulations aura du mal à convaincre, même sur sa spécialité.

De manière analogue, la prise en compte de la plausibilité de l'information tend à faire baisser le score global. Si nous croyons une information à hauteur de sa source et si sa source est compétente en la matière, le fait que l'information soit *possible* ne doit pas renforcer notre confiance, puisque nous avions commencé par la supposer au moins possible. En revanche, si, toutes choses égales par ailleurs, elle paraît improbable, nous commencerons à en douter.

Les flèches représentant l'impact de la compétence et de la plausibilité orientent donc toutes la cotation à la baisse. Plus l'activation du critère est importante – i.e. plus la source est compétente ou plus l'information est plausible – moins la baisse est significative. Tous les niveaux sont envisagés, bien qu'il n'y ait pas de flèche pour chacun. Ce phénomène s'explique d'une part par l'implication de l'absence totale de connaissances et, d'autre part, par le fait que les contraintes que nous posons sont sur les tendances, pas sur l'ampleur de l'impact. Nous verrons, à la section 4.4, que cette souplesse permet une adaptativité à l'utilisateur.

Pour expliquer ce phénomène, notons que nous imposons un traitement spécial à une dimension inquantifiable. Parce qu'une absence d'information n'enrichit pas la connaissance, la cotation ne souffrira pas de l'impossibilité d'évaluer l'une ou l'autre des dimensions. Si, par exemple, la plausibilité de l'information 'ne peut être appréciée', la cotation, après son intégration, n'aura pas évolué, ceci quelle que soit la cote courante.

En outre, nous affirmons que si la cote a pu être évaluée à une étape, elle peut l'être à toutes les suivantes. Ainsi, aucune flèche ne pointe vers une cote de 6, inquantifiable. De ce fait, le niveau accessible le plus bas est le niveau 5, 'Improbable'. En conséquence, puisque compétence et plausibilité ont un impact négatif, une information cotée 5 avant l'une ou l'autre de ces étapes n'évolue pas, en dépit de l'activation de la dimension. Les flèches pour tous ces niveaux sont donc confondues à l'horizontale. De même, hormis pour une information 'extrêmement probable' (cotée 1), certains niveaux d'activation sont confondus.

Il nous reste à expliquer le comportement particulier pour une information dont 'la confiance ne peut être appréciée'. Pour atteindre une telle cote, il faut qu'aucune dimension antérieure n'ait pu être évaluée. En l'absence totale de connaissance, le premier critère pouvant être évalué prend le rôle d'initialisation. Ceci explique que tous les niveaux de confiance sont accessibles à une information n'ayant pu être cotée, en fonction de l'activa-

tion de l'initialisation.

4.3.3 Crédibilité

À la différence des étapes précédentes, l'ultime phase de corroboration de l'information offre une possibilité d'augmentation de la cote. En effet, l'évolution de la confiance en fonction des confirmations et infirmations est un procédé naturel. Cependant, puisque nous envisageons la contradiction de l'information, ce critère peut faire baisser la confiance aussi bien qu'il peut la faire augmenter. Cette dernière étape de la figure 4.1 fait donc évoluer la cotation de manière singulière.

Naturellement, une information cotée 1, même abondamment confirmée, ne peut être crue davantage. En revanche, notons que la baisse de sa confiance est plus limitée qu'aux étapes précédentes. En effet, les cinq degrés témoins d'une corrélation entre les informations considérées se répartissent entre une influence positive (la confirmation) et une influence négative (l'infirmation). À supposer que la crédibilité activée à 'confirmation insuffisante' soit neutre, l'impact négatif se réduit aux deux niveaux minimaux. Par ailleurs, l'infirmation ne pouvant augmenter la cote, seules sont retenues les confirmations d'une information jusqu'alors inqualifiable ou de cote actuelle minimale.

Enfin, nous revenons un instant sur l'intégration pratique de la corroboration à la cotation. Rappelons qu'à cette étape nous comparons des information cotées. Si la confiance faite aux informations corrélées n'entre pas dans l'évaluation de l'activation de la crédibilité, il est sensé d'en tenir compte dans l'évolution de la cotation. De la même manière que nous prêtons plus de foi à une tribune d'Andrew Wiles qu'aux affirmations de notre ami sur le théorème de Fermat, l'intégration de l'infirmation ou de la confirmation sera pondérée par la cote a priori des informations corrélées, comme le proposent Besombes et Cholvy (2007).

4.4 Personnalisation de la cotation : modélisation de crédulités variables

L'architecture que nous proposons est pourvue d'une discrète souplesse, enrichissante pour l'utilisateur. En effet, les flèches de la figure 4.1 représentent les tendances que nous préconisons pour l'influence de chaque estimation de critère. Cependant, nous n'imposons pas la vigueur de ces influences. Grâce à ces degrés de liberté, l'impact des facteurs peut être différent selon l'importance que l'utilisateur leur porte. De plus, deux utilisateurs peuvent avoir des configurations différentes. Ces degrés de liberté permettent, donc, d'adapter la

méthode à l'utilisateur et de représenter différentes attitudes face à la confiance comme nous le décrivions dans Revault d'Allonnes et Besombes (2009).

Nous introduisons la définition suivante :

Définition 6 : Stratégie
Une stratégie d'utilisation de la chaîne de cotation est un jeu d'intensités d'influence de critères. Pour chaque dimension, la stratégie donne le résultat de la projection de son niveau d'activation sur la cote courante.

Cette définition nous mène à ce que nous permettent de représenter les stratégies :

Définition 7 : Posture de crédulité
Une posture de crédulité est un choix, a priori, de l'utilisateur sur ses attentes quant au comportement du système. Elle représente la sensibilité de l'utilisateur aux dimensions et sa facilité de conviction. Le choix d'une stratégie détermine la posture préférée de l'utilisateur.

Il va sans dire que, si la cotation est en cours, on ne peut changer de stratégie.

Dans la suite de cette section, nous illustrons par des exemples la notion de posture de crédulité pour chacune des étapes en considérant deux stratégies extrêmes et une neutre.

4.4.1 Fiabilité de la source

Jusqu'ici, nous avons présenté l'initialisation de la chaîne de cotation comme une égalité entre la fiabilité de la source et le premier a priori sur la confiance. Cependant, le point qui nous importait était moins l'équivalence entre les niveaux que l'importance du premier avis. En effet, on peut imaginer qu'une personne d'une méfiance extrême refuse de croire sur parole, même une personne à qui il accorde toute sa confiance.

Le tableau 4.1 propose, en colonnes, trois stratégies différentes pour la fiabilité de la source. Il fournit la cote initiale pour chaque niveau d'activation de la fiabilité de la source (de A à F, indexant les lignes). La stratégie S_1 est la stratégie par défaut de la figure 4.1,

à laquelle nous associons une posture de crédulité neutre. Celle-ci associe A⇔1, ..., F⇔6. Notons qu'à part la nécessité d'avoir une équivalence entre les degrés 'inquantifiables' – i.e. F⇔6 – toute autre stratégie acceptable – i.e. où A⩾B⩾C⩾D⩾E – peut être utilisée.

On constate que la stratégie S_2 n'accorde pas plus de crédit à une source complètement fiable, qu'à une source de fiabilité légèrement inférieure. De même, au moindre doute sur la fiabilité (niveaux *D* et *E*), elle ôte toute sa confiance (cote de 5). Cette stratégie extrême représente une posture de méfiance de l'utilisateur. À l'inverse, la stratégie S_3 reste ouverte à une source à la fiabilité branlante. Elle commence à légèrement douter si la source n'est absolument pas crédible, mais est prête à croire toute personne à peu près digne de confiance. Cette stratégie, encore plus rare que la précédente, correspond à un utilisateur crédule.

	S_1	S_2	S_3
A	1	2	1
B	2	2	1
C	3	3	2
D	4	5	3
E	5	5	4
F	6	6	6

TABLE 4.1 – Trois stratégies différentes pour l'impact de l'intégration de la fiabilité de la source.

Considérons trois utilisateurs (U_1, U_2 et U_3) manifestant chacun l'une des trois postures de crédulité. U_1 est neutre, U_2 méfiant et U_3 crédule. Supposons que, malgré leurs divergences de crédulité, les trois utilisateurs aient le même avis sur notre ami photographe, qu'ils considèrent parfaitement fiable. Après l'initialisation, chacun a un a priori de $c_1 = 1$ (pour U_1), $c_2 = 2$ (pour U_2) et $c_3 = 1$ (pour U_3).

4.4.2 Compétence et plausibilité

La compétence et la plausibilité ont des impacts similaires sur l'évolution de la cotation. Pour cette raison et pour alléger la lecture de ce document, le tableau 4.2 présente les mêmes postures de crédulité pour les deux dimensions. Cependant, ceci n'est pas une contrainte de notre modèle. Un utilisateur peut, par exemple, préférer se fier à la compétence de la source, plutôt qu'à sa connaissance personnelle du sujet.

Ce tableau propose une formulation différente des stratégies, puisqu'on suppose que la

cotation a été initialisée. De ce fait, l'influence de l'évaluation de la dimension (de 1 à 6, indexant les lignes sur fond noir) se calcule à partir du niveau courant de la cote, c, et la combinaison est bornée par l'étendue de l'échelle d'évaluation, c'est-à-dire que la cote évaluée doit rester inférieure à 5.

Par souci de clarté, le cas où $c = 6$ n'est pas représenté dans le tableau. Dans ce cas, on applique les stratégies d'initialisation du tableau 4.1, où les niveaux d'activation de la fiabilité sont remplacés par leur équivalent pour la dimension considérée.

	S_1	S_2	S_3
1	c	c	c
2	$\min(c+1,5)$	$\min(c+2,5)$	c
3	$\min(c+2,5)$	$\min(c+3,5)$	$\min(c+1,5)$
4	$\min(c+3,5)$	$\min(c+4,5)$	$\min(c+2,5)$
5	$\min(c+4,5)$	$\min(c+4,5)$	$\min(c+3,5)$
6	c	c	c

TABLE 4.2 – Trois stratégies différentes pour l'impact de la compétence de la source et la plausibilité de l'information, où c est l'indice de confiance courant, supposé différent de 6.

Supposons que nos trois utilisateurs ignorent que le photographe n'est pas versé dans les choses de la mathématique et qu'ils le supposent donc compétent, soit de niveau 2. L'intégration de la compétence à leur cotation de ses dires va donc donner : $c_1 = \min(c+1,5) = 2$, $c_2 = \min(c+2,5) = 4$ et $c_3 = c = 1$. On constate déjà des divergences de point vue selon la posture de crédulité.
Les trois utilisateurs, en parfaite harmonie, estiment également que bien que surprenante, la découverte de la preuve de Fermat est 'réaliste'. La nouvelle mise à jour de la cote est : $c_1 = \min(c+1,5) = 3$, $c_2 = \min(c+2,5) = 5$ et $c_3 = c = 1$. Les avis sont partagés entre le crédule qui croit et le méfiant qui nie. U_1, neutre lui, endosse le rôle de l'agnostique.

Notons que l'on peut également représenter une stratégie par sa table d'association pour chaque dimension. La stratégie par défaut S_1 aurait pour table d'association pour la compétence et la plausibilité celle représentée dans le tableau 4.3. On remarque que dès lors qu'on ne représente plus qu'une stratégie, il est aisé d'inclure toutes les combinaisons, notamment lorsque $c = 6$. On retrouve aussi le développement de la première colonne du tableau 4.2.

Cote actuelle \ Activation de la dimension	1	2	3	4	5	6
1	1	2	3	4	5	1
2	2	2	3	4	5	2
3	3	3	3	4	5	3
4	4	4	4	4	5	4
5	5	5	5	5	5	5
6	1	2	3	4	5	6

TABLE 4.3 – Stratégie par défaut de la figure 4.1 : calcul de la mise à jour de la confiance, en fonction de l'évaluation de la compétence ou de la plausibilité

4.4.3 Crédibilité

Comme nous le disions plus haut, l'intégration de la crédibilité à la cotation est une affaire particulière. D'une part, il s'agit de la seule dimension permettant l'amélioration de la confiance. D'autre part, lors de sa projection sur l'information, la cote a priori (i.e. la cote hors crédibilité) de l'information corrélée est prise en compte. Le tableau 4.4 propose, lui aussi, trois stratégies différentes pour la crédibilité. Une fois de plus, le niveau d'activation de la dimension indexe, sur fond noir, les lignes du tableau. On remarque une variation dans la disparate des utilisateurs, où la posture méfiante favorise l'infirmation sur la correction. Ici encore, nous retrouvons l'expression des influences en fonction de la confiance actuelle. Dans un effort de lisibilité, nous supposons que ces expressions incluent la cote a priori des informations corrélées. Les modifications sont à nouveau bornées pour rester dans l'éventail des niveaux de confiance autorisés.

	S_1	S_2	S_3
1	$\max(c-2,1)$	$\max(c-1,1)$	$\max(c-2,1)$
2	$\max(c-1,1)$	c	$\max(c-2,1)$
3	c	$\min(c+1,5)$	c
4	$\min(c+1,5)$	$\min(c+2,5)$	c
5	$\min(c+2,5)$	$\min(c+3,5)$	$\min(c+1,5)$
6	c	c	c

TABLE 4.4 – Trois stratégies différentes pour l'impact de la crédibilité de l'information, où c est l'indice de confiance courant.

Après avoir appris que la preuve élégante de Fermat avait, enfin, été découverte, les trois utilisateurs se précipitent – sans doute dans des buts différents – à la recherche de confirmation. Dès leur première recherche, ils trouvent la tribune d'Andrew Wiles, concédant l'élégance de la nouvelle preuve. Une source fiable ($\delta_1 = 1$, $\delta_2 = 2$, $\delta_3 = 1$) et diablement compétente ($\delta_1 = 1$, $\delta_2 = 2$, $\delta_3 = 1$) confirme un fait réaliste ($\delta_1 = 2$, $\delta_2 = 4$, $\delta_3 = 1$), où δ_i représente la cote de la nouvelle information pour l'utilisateur U_i.
U_1 prend donc parti ($c_1 = \max(c-2, 1) = 1$), U_2 reste dubitatif ($c_2 = \max(c-1, 1) = 4$) et U_3, n'ayant guère évolué, reste en accord avec lui-même ($c_3 = max(c-2, 1) = 1$).

4.5 Discussion

Nous revenons brièvement, dans cette section, sur les points de notre modèle dont la contrainte peut paraître importante. Nous évoquons également une première évolution que nous envisageons à nos travaux.

Ordre et séquentialité
Le modèle proposé ici présente la cotation comme un processus ordonné et séquentiel. Il se peut que cette réduction de la constitution de confiance soit restrictive aux yeux de certains. Pourtant, nous rappelons qu'il n'existe pas de modèle universellement reconnu de ce phénomène. Cependant, la présentation que nous faisons ici de la chaîne de cotation est plus rigide que telle que nous la concevons. Pour nous, la cotation n'est pas un processus fini. Les dimensions pouvant être revues à l'usage, la cote peut elle aussi être réévaluée (cf. section 3.2, p. 51). De plus, bien qu'il nous semble naturel et justifié, l'ordre d'intégration des dimensions que nous recommandons n'est pas incontournable. La présence pour toutes les évaluations d'un niveau inquantifiable permet, notamment, de repousser l'inclusion d'une dimension.

Mise à jour
Après l'étape de crédibilité de la chaîne de cotation, l'information considérée a été cotée en fonction de la fiabilité de sa source ainsi que de sa compétence sur le domaine de définition de l'information, la plausibilité de cette information en regard de notre perception du monde et, enfin, un indice de confirmation entre informations incertaines. À ce moment, nous pouvons envisager de mettre à jour la fiabilité de la source de l'information, dans la veine de ce que propose Delmotte (2007) lorsqu'il met à jour la fiabilité des sources, lors de la fusion d'informations, en fonction de sa contradiction cumulée avec les autres sources. Cette rétroaction de la confiance en l'information sur la fiabilité de sa source ne fait pas partie de la chaîne mais a un effet central. Si une source de confiance fournit

systématiquement des informations cotées peu probables, il peut être judicieux de revoir son jugement, comme le propose la figure 4.2. Nous ne proposons pas encore de règles de mise à jour, notamment parce que celle-ci dépend beaucoup de la source, ainsi que de l'application. Nous envisageons certaines procédures d'étude de réputation de sources, que nous décrivions dans l'état de l'art (cf. section 2.2.2, p. 28), pour apprendre à automatiser leur qualification. Ceci constitue un enjeu important pour le traitement de gros volumes de données, comme les sources ouvertes.

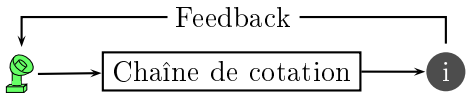


FIGURE 4.2 – Mise à jour de la fiabilité de la source

4.6 Conclusion

Dans ce chapitre, nous avons présenté l'architecture de la cotation, la méthode d'intégration des dimensions à l'évolution de la cote. Nous avons commencé par rappeler le modèle sur lequel nous bâtissons nos propositions. De ce procédé, dont nous avions précédemment extrait les dimensions de la cotation, nous avons extrait l'ordre naturel de leur prise en compte ainsi que la structure séquentielle de la cotation. Nous avons ensuite proposé la synthèse de ces propositions d'architecture, sous la forme d'une chaîne de cotation.

Dans la suite du chapitre nous avons dévoilé l'influence de chaque dimension sur la cotation. Nous avons expliqué l'influence décroissante des premiers critères et les particularités du dernier. Enfin, nous avons défini la notion de postures de crédulité de l'utilisateur et l'outil servant à les incarner, les stratégies. Nous avons explicité l'emploi des stratégies et montré leur utilité en les illustrant de trois postures caractéristiques dans un exemple.

Ce chapitre se clôt sur une discussion revenant sur la séquentiatié du formalisme proposé et évoquant l'éventualité de la mise à jour automatique de la fiabilité de la source.

Chapitre 5

Formalisation multivalente de la cotation

Résumé

Dans ce chapitre nous proposons une formalisation du processus de cotation dans le cadre de la logique multivalente. Après avoir rappelé les principes de ce cadre formel, nous en introduisons une extension permettant de représenter l'impossibilité d'appréciation. Nous présentons ensuite les opérateurs de combinaison de vérité qui modélisent la projection de l'évaluation des dimensions sur l'information considérée. Enfin, nous concluons ce chapitre par un exemple explicitant les stratégies définies précédemment dans le cadre formel.

5.1 Introduction

Nous proposons ici un cadre formel à la mise en œuvre de la chaîne de cotation décrite au chapitre précédent. Ce cadre doit permettre de représenter et de manipuler un score de confiance interprétable pour toute information construite. Nous considérons qu'un tel score est un parent proche des degrés de certitude classiquement étudiés (Bouchon-Meunier, 1993). En effet, nous proposons de construire automatiquement des objets représentant des faits supposés du monde réel et d'assortir ces faits d'une confiance, semblable à la certitude.

Parmi les outils classiques de traitement de l'incertitude, certains offrent des possibi-

lités accrues de lisibilité (cf. 2.4.2, p. 37). Les outils de raisonnement avec des variables linguistiques (Zadeh, 1975), voire avec des mots (Zadeh, 2002) offrent des possibilités de raisonnement sur des attributs immédiatement interprétables. Cependant, bien que la majorité des critères que nous proposons d'utiliser soit mesurable, il nous semble inapproprié d'utiliser les outils de la logique floue. En effet, les partitionnements de l'espace des mesures que nous considérons représentent moins des ensembles flous que des degrés de confiance. Notons également que la majorité des théories de l'incertain offre des outils de raisonnement complexes. Notre ambition n'allant pas jusqu'au raisonnement, nous avons concentré notre étude sur les modèles offrant une grande lisibilité. La logique multivaluée, avec ses degrés de vérité discrets, semblait donc toute indiquée pour représenter ces degrés, ainsi que pour les manipuler.

Nous rappelons donc ici les principes et utilité de la logique multivaluée, puis décrivons le formalisme que nous avons choisi pour représenter nos travaux. Nous présentons, ensuite, une extension de ce cadre formel permettant de représenter l'impossibilité de choix. Après la mise en place du cadre d'évaluation de notre modèle, nous proposons des opérateurs de combinaison vérifiant les propriétés souhaitées pour modéliser la chaîne de cotation, telle que présentée à la figure 4.1, page 64.

5.2 Formalisme multivalué

Que l'on manipule la véracité d'une proposition ou la confiance associée à une information, raisonner dans un univers manichéen est réducteur et insuffisant. Comment, alors, ordonner ses certitudes face à ses doutes ? Pourtant, si l'on se souvient que le discours d'Aristote (4ème siècle av. JC) soutient qu'une proposition est soit *vraie* soit *fausse* et ne peut être *ni l'un, ni l'autre* et, a fortiori, *les deux*, on constate que la logique classique nous y contraint et l'on perçoit la nécessité de la notion intermédiaire d'incertitude. L'histoire attribue, d'ailleurs, à Aristote la première remise en cause du principe de bivalence ou du tiers-exclu, lorsqu'il aborde la vérité anticipée d'un événement à venir. Le seul cas envisageable pour lequel on ne peut trancher entre le vrai et le faux est le cas où cette véracité n'est pas encore déterminée. Le doute provient de l'évolution à venir et non de l'événement considéré, néanmoins, on peut alors considérer que la proposition n'est *ni vraie, ni fausse*. Aristote ne propose pas d'évaluation pour ces cas et les travaux fondateurs des logiques multivaluées découlent plus explicitement et formellement des articles de Łukasiewicz (1920) et de Post (1921).

Ces derniers proposent d'adjoindre un degré de vérité 'indéterminé' à ceux d'Aristote.

Cette apparition de la logique trivalente annonce l'arrivée des logiques multivaluées, permettant de modéliser des états de connaissance imparfaite. Les premières intégrations de degrés de vérité intermédiaires furent discrètes, proposant un nombre fini de représentations de l'état de la connaissance. Cependant, 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 efficaces, ces degrés doivent néanmoins répondre à certaines obligations, tant d'un point de vue pratique, que théorique et sémantique. Leur existence comme extensions de la logique classique les contraint à maintenir une cohérence de raisonnement avec celle-ci. À l'inverse, leur nécessaire intervention pour l'augmentation de l'interprétabilité du raisonnement a ses exigences. Parmi ces prérequis, nous nous intéressons particulièrement à :

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

Le premier point est fondateur des logiques non-classiques, bien qu'il étende la comparaison évidente du choix binaire antérieur. Il est également essentiel à notre étude puisque l'intérêt principal d'un indicateur de confiance provient de l'ordre qu'il impose aux informations qu'il qualifie. Disposer d'outils d'agrégation est également capital en ce que l'évolution de la cote représente bien une combinaison d'évaluations. Les logiques non-classiques offrent nombre d'opérateurs de combinaison et leur prêtent, souvent, une sémantique riche. Enfin, le sens des degrés discrets de la logique symbolique est formulé grâce à des étiquettes sémantiques.

Dans ce qui suit, nous rappelons successivement les principes des degrés de vérité de la logique multivaluée, puis la modélisation des opérations logiques sur ces degrés.

5.2.1 Degrés de vérité

Par convention, nous posons une différence entre les logiques multivaluées sous toutes leurs formes et celle qui nous intéresse, discrète, dite aussi symbolique, que nous appellerons désormais logique multivalente. Celle-ci permet de raisonner dans une algèbre de De Morgan, c'est-à-dire un treillis distributif et fini, comme, par exemple, celui présenté à la figure 5.1, muni d'un opérateur unaire involutif satisfaisant les lois de De Morgan. À l'instar de l'algèbre booléenne, la logique multivalente possède un opérateur de négation remplissant cette fonction. Plus encore qu'un treillis, la logique multivalente propose d'utiliser les M degrés de vérité d'un ensemble $\mathcal{L}_M = \{\tau_0, \ldots, \tau_{M-1}\}$ totalement ordonné. $\mathcal{L}_M$

est dit totalement ordonné puisque $\tau_\alpha \leqslant \tau_\beta \Leftrightarrow \alpha \leqslant \beta$ et assure donc la comparabilité des degrés de vérité.

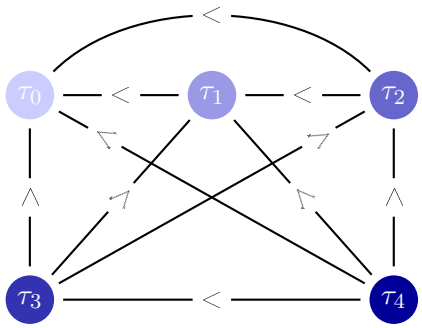


FIGURE 5.1 – Un exemple de treillis de De Morgan : $(\mathcal{L}_5, \wedge, \vee, >)$

Les degrés de vérité de $\mathcal{L}_M$ parcourent, à une granularité variable avec M, les variations de véracité de ce qu'ils qualifient, de τ_0, représentant 'faux', à τ_{M-1}, désignant 'vrai'. Les autres τ_α représentent les valeurs intermédiaires telles que 'peut-être vrai'. Chaque degré de vérité est associé à une étiquette sémantique (cf. tableau 5.1) favorisant ainsi la lisibilité et l'interprétabilité. La vérité devenant graduelle, l'échelle se divise entre les degrés désignant une évaluation favorable et les degrés indiquant un doute. La bascule s'opère au milieu de l'échelle autour de la valeur $\tau_{\frac{M-1}{2}}$, potentiellement inexistante. On choisit, généralement, de disposer d'un élément neutre ; on impose donc que $\tau_{\frac{M-1}{2}}$ appartienne à l'échelle en considérant un nombre M, impair, de degrés à $\mathcal{L}_M$.

Degré	Étiquette sémantique
τ_4	Vrai
τ_3	Plutôt vrai
τ_2	Moyennement vrai
τ_1	Plutôt faux
τ_0	Faux

TABLE 5.1 – Un exemple de degrés de vérité dans $\mathcal{L}_5$

Les échelles multivalentes – et donc les degrés de vérité – permettent de représenter efficacement les connaissances incertaines, mal définies ou difficilement évaluables, comme le permettent les variables linguistiques de Zadeh (1975). En effet, en considérant l'équivalence entre l'expression « Thomas est plutôt grand » et « 'Thomas est grand' est 'plutôt

vrai' », l'incertitude liée à la proposition originelle est représentée par le degré de vérité τ_3 qualifiant la véracité du fait sous-jacent (Truck & Akdag, 2005).

5.2.2 Raisonnement : combinaison de degrés de vérité

Outre l'expressivité et la lisibilité des degrés de vérité de la logique multivalente, des outils permettent de raisonner sur ces valeurs incertaines. Les généralisations usuelles de la conjonction et disjonction binaires, telles que proposées par Post (1921), permettent de construire ces opérateurs. Quels que soient les degrés de vérité de $\mathcal{L}_M$, on définit formellement la conjonction ($\wedge$) et la disjonction ($\vee$) par :

$$\tau_\alpha \wedge \tau_\beta \equiv \min(\tau_\alpha, \tau_\beta)$$

$$\tau_\alpha \vee \tau_\beta \equiv \max(\tau_\alpha, \tau_\beta)$$

On se munit également d'une négation que l'on écrit

$$\neg\tau_\alpha \equiv \tau_{M-1-\alpha}$$

Notons que, puisque $\tau_\alpha \wedge \neg\tau_\alpha = \min(\tau_\alpha, \tau_{M-1-\alpha})$, le principe de non-contradiction ne s'applique qu'aux cas limites, pour τ_{M-1} et τ_0.

Des travaux proposent de nouveaux opérateurs afin de combiner les degrés de vérité intermédiaire et donc de représenter des processus cognitifs évolutifs. Tous ces opérateurs doivent fournir un résultat dans $\mathcal{L}_M$ et leurs définitions assurer leur clôture. Akdag (1992), par exemple, introduit un opérateur d'implication, qu'il dit matériel, afin de raisonner sur une base de connaissances incertaines. Il propose une version multivalente de l'implication de Łukasiewicz et l'écrit

$$\tau_\alpha \rightarrow_L \tau_\beta = \min(\tau_{M-1}, \tau_{M-1-(\alpha-\beta)})$$

S'inspirant des travaux de Darwiche et Ginsberg (1992), qui posent les bases d'une arithmétique symbolique, Seridi et Akdag (2001) développent les opérations élémentaires sur les degrés de vérité. Les auteurs proposent, notamment, des versions multivalentes de l'addition et de la soustraction, basées sur l'implication de Łukasiewicz, qu'ils écrivent :

$$ADD(\tau_\alpha, \tau_\beta) = \neg\tau_\alpha \rightarrow_L \tau_\beta$$

$$SUB(\tau_\alpha, \tau_\beta) = \neg(\tau_\alpha \rightarrow_L \tau_\beta)$$

Pour augmenter encore l'expressivité de $\mathcal{L}_M$, Akdag et al. (2001) proposent d'introduire des modificateurs jouant sur l'intensité des degrés. Ces notions, émanant des travaux de

Zadeh (1975), sont généralisées par Truck et Akdag (2005) et une perspective générale en est donnée par Truck et Akdag (2006). Bien que nous les concevions différemment, ces manipulations de la granularité de $\mathcal{L}_M$ sont proches du formalisme des stratégies que nous proposons, détaillé à la section 4.4, page 67.

5.3 Ignorance multivalente : proposition d'extension

5.3.1 Problématique

Parce que notre objectif est d'utiliser les degrés de vérité de la logique multivalente pour leur caractère lisible et interprétable, nous devons distinguer l'impossibilité d'évaluation de la possibilité de réalisation. Cette distinction est nécessaire afin de lever cette confusion entre valeur de vérité d'une proposition et interprétation épistémique qui a parasité la théorie dès l'origine, selon Dubois (2010).

Les premières versions de la logique multivalente, comme par exemple la logique trivaluée de Łukasiewicz (1920), intégraient un degré de vérité de plus que la logique classique. Ce degré désignait tout à la fois l'incertitude sur la vérité de la proposition qu'il qualifiait (i.e. le fait est possible mais ni certain ni impossible) et l'impossibilité de se prononcer (i.e. il est impossible de dire si le fait est vrai, faux ni même possible). Ce degré, étant donné sa légitimité contestée, servait donc de fourre-tout incertain.

Par la suite, lors de l'apparition de logiques à plus de degrés de vérité, l'imposition d'un ordre total entre ces degrés, liée à la préservation aux limites de la logique classique, contraignait les utilisateurs à attribuer ce rôle au milieu de l'échelle de vérité. La notion de 'possibilité' d'un prédicat recouvrait nécessairement le fait qu'il était impossible d'évaluer sa certitude.

Au contraire de ce que prétend De Glas (1987), nous pensons que cette ambiguïté est levée dans le cadre de degrés de vérité continus, tant possibilistes que flous. En effet, en logique possibiliste, comme le dit Dubois (2010), l'incertitude est équivalente à une possibilité totale et une nécessité nulle. En revanche, l'auteur prétend que toutes les logiques multivaluées, continues ou discrètes, pâtissent de cette confusion résolue par la distinction entre nécessité et possibilité. Pourtant, une interprétation de la théorie des sous-ensembles flous est telle que, lorsque l'incertitude est totale, la seule appartenance certaine est à l'ensemble de discernement (cf. fig. 5.2). Si rien n'est connu sur un événement, tout est également envisageable.

Bien qu'à notre avis la modélisation de l'incertitude en théorie des sous-ensembles

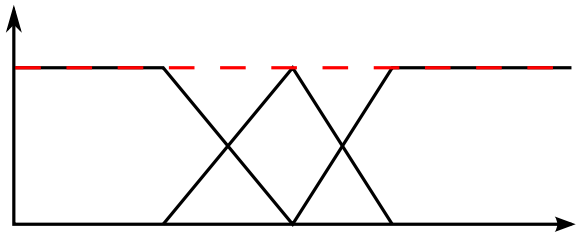


FIGURE 5.2 – Incertitude totale sur une partition floue (*en pointillés*)

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. Dans ce qui suit, nous proposons donc d'introduire un degré de vérité supplémentaire, pour lever cette ambiguïté et nous en examinons les contraintes et conséquences, en extension de ce que nous montrions dans Revault d'Allonnes et al. (2010).

5.3.2 Proposition : introduction d'un degré additionnel

Nous proposons donc d'ajouter un degré de vérité exception à la règle de l'ordre total. Ce degré représentant une connaissance inexistante, il est de nature différente des autres éléments de $\mathcal{L}_M$.

Définition 8 : Degré de vérité inquantifiable
Tau inconnu, noté $\tau_?$, est un degré de vérité spécifique exprimant l'impossibilité d'appréciation.
Tau inconnu a les propriétés suivantes :

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

où $\odot$ représente un opérateur quelconque sur les degrés de vérité, tel que $\wedge$, $\vee$, ADD, SUB, ...

Par définition, $\tau_?$ ne préserve pas l'ordre sur $\mathcal{L}_M$. Il est également important de noter que, en conséquence, $\tau_?$ est défini comme élément neutre de tous les opérateurs. Parce que

$\tau_?$ représente l'ignorance totale, quoi qu'on apprenne par la suite, nous quittons cet état. Afin de préserver les propriétés mathématiques de la logique symbolique, nous proposons donc de représenter la cotation sur $\mathcal{L}_M$ étendu de $\tau_?$, que nous notons $\mathcal{L}_M^e = \mathcal{L}_M \cup \{\tau_?\}$.

Propriétés

On démontre 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 $\neg(\tau_? \vee \tau_\alpha) = \neg\tau_? \wedge \neg\tau_\alpha$

Ces deux lois maintenues ne permettent, néanmoins, pas de définir une algèbre de De Morgan grâce à $\mathcal{L}_M^e$. En effet, celui-ci ne permet de définir que les deux demi-treillis présentés aux figures 5.3 et 5.4.

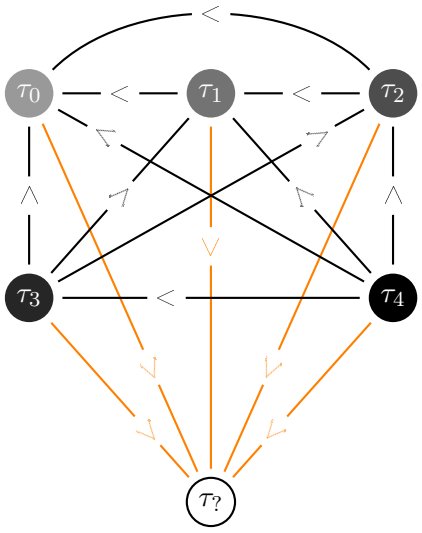


FIGURE 5.3 – Demi-treillis supérieur $(\mathcal{L}_5 \cup \{\tau_?\}, \vee, >)$

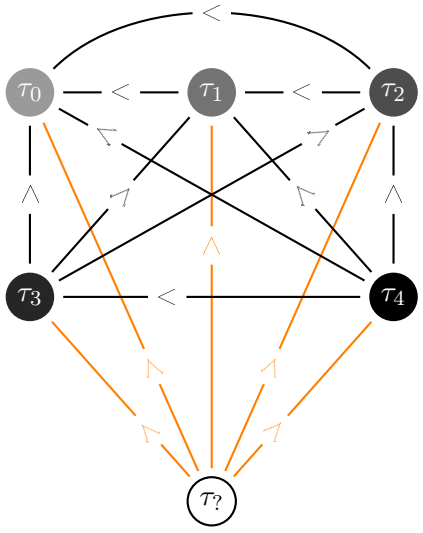


FIGURE 5.4 – Demi-treillis inférieur $(\mathcal{L}_5 \cup \{\tau_?\}, \wedge, <)$

Ceci provient de la non-préservation de la dualité des opérateurs $\vee$ et $\wedge$, qui induit le non-respect de l'ordre sur $\mathcal{L}_M$ par $\tau_?$. De ce fait, les axiomes de l'algèbre de Boole que $\tau_?$ ne préserve pas sont :

1. Lois d'identité : $\forall\tau_\alpha \in \mathcal{L}_M, \tau_{M-1} \wedge \tau_\alpha = \tau_\alpha$ et $\tau_0 \vee \tau_\alpha = \tau_\alpha$
   contredites par
   - $\tau_{M-1} \wedge \tau_? = \tau_{M-1}$
   - $\tau_0 \vee \tau_? = \tau_0$

2. Lois d'inversion : $\forall \tau_\alpha \in \mathcal{L}_M, \neg\tau_\alpha \vee \tau_\alpha = \tau_{M-1}$ et $\neg\tau_\alpha \wedge \tau_\alpha = \tau_0$
   contredites par
   - $\neg\tau_? \vee \tau_? = \tau_?$
   - $\neg\tau_? \wedge \tau_? = \tau_?$
3. Lois d'absorption : $\forall \tau_\alpha, \tau_\beta \in \mathcal{L}_M, \tau_\alpha \vee (\tau_\alpha \wedge \tau_\beta) = \tau_\alpha$ et $\tau_\alpha \wedge (\tau_\alpha \vee \tau_\beta) = \tau_\alpha$
   contredites par
   - $\tau_? \vee (\tau_? \wedge \tau_\alpha) = \tau_\alpha, \forall \tau_\alpha \in \mathcal{L}_M$
   - $\tau_? \wedge (\tau_? \vee \tau_\alpha) = \tau_\alpha, \forall \tau_\alpha \in \mathcal{L}_M$

Nous expliquons ce comportement par l'incomparabilité désirée de $\tau_?$, sur laquelle nous revenons dans la section suivante.

5.3.3 Discussion

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 multivalente. Bien que la comparabilité des degrés de vérité soit la première des contraintes que nous voulons imposer à notre méthode de représentation de la cotation, nous pensons que le gain d'expressivité de l'ajout de tau inconnu en compense la perte ponctuelle.

Perte de la comparabilité

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, l'évaluation de la véracité n'a pu être établie, il semble raisonnable que le degré final ne soit pas une confiance moyennement indéterminée. En revanche, dès qu'une hypothèse a été formulée, les évolutions successives resteront déterminées. Puisqu'on a su, à un moment, se prononcer sur la véracité du fait, on ne peut revenir à l'impossibilité de l'évaluer.

Gain d'expressivité

En conséquence, la comparaison d'un fait 'moyennement vrai' avec un autre le replace bien dans la hiérarchie sémantique de $\mathcal{L}_M$. De même, un fait de vérité $\tau_?$ n'est ni plus, ni moins vrai qu'un autre. Il est, en revanche, à étudier de plus près. Nous renforçons donc la comparabilité *dans* $\mathcal{L}_M$, en en sortant l'inconnu.

Cohérence

Lors de l'établissement ou de la modification d'un cadre de raisonnement logique, il est nécessaire de vérifier la cohérence du système résultant. On dit d'un système qu'il est incohérent si l'on peut inférer une proposition et son contraire. Dans le cadre des logiques

multivaluées, cette assertion est déjà remise en cause par l'abandon des principes de bivalence et de non-contradiction. Pourtant, si l'ajout de $\tau_?$ n'implique pas l'incohérence du système, la raison en est ailleurs. Puisque nous définissons $\tau_?$ comme élément neutre de tous les opérateurs de combinaison, toute incohérence constatée après son introduction persisterait à sa suppression.

5.4 Opérateur de combinaisons à la baisse

Dans la suite de ce chapitre, nous nous intéressons à l'explicitation de la chaîne de cotation présentée au chapitre 3 (figure 4.1, page 64). Dans ce but, nous proposons une formalisation multivalente du processus tel que nous l'avons décrit. Les combinaisons de ce procédé n'étant pas spontanément disponibles dans ce cadre formel, nous proposons nos propres solutions et en discutons les apports.

Nous proposons donc d'évaluer la cote et les critères de son évaluation dans $\mathcal{L}_M^e$. Nous nous intéressons plus précisément maintenant aux opérateurs qui, à partir d'une cote courante τ_α et de l'évaluation d'une dimension τ_β, toutes deux évaluées dans $\mathcal{L}_M^e$, fournissent la cote d'entrée τ_γ à l'étape suivante. Nous devons définir des opérateurs de mise à jour $comb(\tau_\alpha, \tau_\beta)$, concordant avec notre modèle de cotation.

Parce que la chaîne de cotation n'intègre pas toutes les dimensions de la même manière, nous devons séparer les étapes, afin de définir les opérateurs adéquats. Nous débutons cette étude par le cas combiné de la compétence et de la plausibilité, puisqu'elles exhibent des comportements semblables, dans la chaîne de cotation.

Définition

Telles que nous les avons définies, la compétence de la source et la plausibilité de l'information sont des dimensions aux comportements similaires. En effet, leur influence sur la cotation est monotone, c'est-à-dire que, si l'une ou l'autre de ces dimensions influe sur la cote, elle le fait uniquement vers sa baisse. Notons toutefois que l'étendue de la baisse de la cote dépend du niveau d'activation de la dimension. Plus la source est incompétente, par exemple, plus le doute qui s'installe est important.

En outre, nous précisons les règles de traitement de la méconnaissance. Convaincu qu'on ne peut décider à partir du néant, si la cote courante ne peut être évaluée, l'initialisation de la cotation dépend de la première dimension à pouvoir l'être. De manière analogue, si l'on ne peut estimer la compétence de la source ou la plausibilité de l'information, l'évaluation de la confiance ne change pas.

Nous proposons donc de formaliser ces opérateurs de combinaison similaires grâce au même

opérateur, déléguant à la stratégie le soin de spécifier les éventuelles nuances de comportement. Nous définissons donc un opérateur de combinaison F, pour formaliser ces comportements, de la manière suivante :

Définition 9 : Opérateur de combinaison baissier
Soit F, une fonction telle que $F : \mathcal{L}_M^e \times \mathcal{L}_M^e \longrightarrow \mathcal{L}_M^e$
$(\tau_\alpha, \tau_\beta) \longmapsto \tau_\gamma$
où
- τ_α est la cote actuelle
- τ_β est l'évaluation de la compétence ou de la plausibilité
- τ_γ est la cote mise à jour

ayant les propriétés suivantes :

1. $F(\tau_\alpha, \tau_\beta) \leq \tau_\alpha, \forall \tau_\alpha, \tau_\beta \in \mathcal{L}_M$
2. Si $\tau_\alpha \leq \tau_\beta$, alors $F(\tau_\alpha, \tau_\gamma) \leq F(\tau_\beta, \tau_\gamma), \forall \tau_\gamma \in \mathcal{L}_M$
3. Si $\tau_\alpha \leq \tau_\beta$, alors $F(\tau_\gamma, \tau_\alpha) \leq F(\tau_\gamma, \tau_\beta), \forall \tau_\gamma \in \mathcal{L}_M$
4. $F(\tau_?, \tau_\alpha) = F(\tau_\alpha, \tau_?) = \tau_\alpha, \forall \tau_\alpha \in \mathcal{L}_M^e$

L'opérateur F étant borné par la valeur d'entrée de la cotation, la première propriété garantit bien que la compétence et la plausibilité ont une tendance à la baisse. Quelle que soit l'activation de l'une ou l'autre des dimensions, la cote d'entrée de l'étape suivante est, au mieux, égale à la cote précédente, inférieure sinon.
Les deux propriétés suivantes introduisent la monotonie de l'opérateur. En effet, bien qu'à tendance baissière, F est croissante en ses deux arguments. Il semble raisonnable de considérer que, provenant, par exemple, de sources également compétentes, nous croyons plus en l'information dont la source était la plus fiable. À l'inverse, entre deux informations crues de la même manière, on doute naturellement plus de celle dont la plausibilité est la moins élevée.
Enfin, la dernière propriété assure les comportements décrits plus haut, face à l'inquantifiable. Pour un fait qu'on ne sait pas encore coter, la première source d'information nous sert à initialiser le processus. De même, parce que nous supposons un a priori neutre pendant la cotation, on ne pénalise pas une information sous prétexte qu'on ne sait estimer la compétence de sa source ou sa plausibilité.

Fonction proposée

Les propriétés attendues de l'opérateur de combinaison à la baisse ainsi décrites nous permettent de proposer un opérateur compatible avec notre modèle de cotation. À partir de ces exigences et en tenant compte de la nécessité d'intégrer à F les stratégies, nous nous munissons d'un jeu de paramètres κ_α^γ, dont la sémantique est commentée ci-dessous, et définissons l'opérateur suivant :

$$F(\tau_\alpha, \tau_\beta) = \begin{cases} \min\{\tau_\gamma \in \mathcal{L}_M | \tau_\beta \leqslant \kappa_\alpha^\gamma\} & \text{si } \tau_\alpha, \tau_\beta \in \mathcal{L}_M \\ \tau_\alpha & \text{si } \tau_\beta = \tau_? \\ \tau_\beta & \text{si } \tau_\alpha = \tau_? \end{cases}$$

Les paramètres κ_α^γ définissent les seuils de transition de la cotation actuelle τ_α vers sa mise à jour τ_γ. Cet opérateur F autorise, ainsi, le passage de τ_α à τ_γ si $\tau_\beta \leq \kappa_\alpha^\gamma$. La cohérence de notre proposition a des conséquences sur les κ_α^γ. Puisqu'ils décrivent les passages successifs vers les degrés inférieurs, il est nécessaire que, si $\tau_\gamma < \tau_\delta$, alors $\kappa_\alpha^\gamma \leq \kappa_\alpha^\delta$. En effet, pour assurer la monotonie croissante de l'opérateur, il est nécessaire que les seuils soient ordonnés. Afin de lever toute ambiguïté sur la transition, il est possible d'imposer que la seconde inégalité soit stricte. Nous devons également souligner que, selon la stratégie, certaines transitions peuvent être exclues, comme nous le verrons dans les exemples de la section 5.5.4.

Notons, toutefois, que ces contraintes d'ordre sur les κ_α^γ s'appliquent pour une valeur τ_α donnée. En effet, nous n'imposons pas que $\kappa_\alpha^\delta = ADD(\kappa_\alpha^\gamma, \kappa_\gamma^\delta)$. De ce fait, il se peut que $F(\tau_\alpha, \tau_\delta) \neq F(F(\tau_\alpha, \tau_\gamma), \tau_\delta)$, c'est-à-dire que le passage direct de τ_α à τ_δ n'est pas, nécessairement, équivalent à la succession des transitions de τ_α à τ_γ, puis de τ_γ à τ_δ. En ce sens, l'opérateur F n'est donc pas associatif. Puisqu'on n'intègre qu'une fois la compétence de la source et la plausibilité de l'information, cette propriété n'engendre pas de conséquences. Cependant, nous verrons plus tard que nous dispenser de quelques propriétés de combinaison enrichit notre modèle.

La figure 5.5 illustre l'intégration des stratégies à la fonction F, décrite ci-dessus. Les degrés de vérité de $\mathcal{L}_5$ y représentent les valeurs potentielles de la cote et les flèches, assorties des κ_α^γ, les transitions de τ_α vers τ_γ, telles qu'indiquées par la stratégie pour la dimension étudiée.

Comparaison avec les opérateurs existants

Parce que l'opérateur de combinaison F est monotone croissant, on pourrait le confondre avec une norme triangulaire. En réalité, il n'en est rien puisque nous n'imposons ni sa

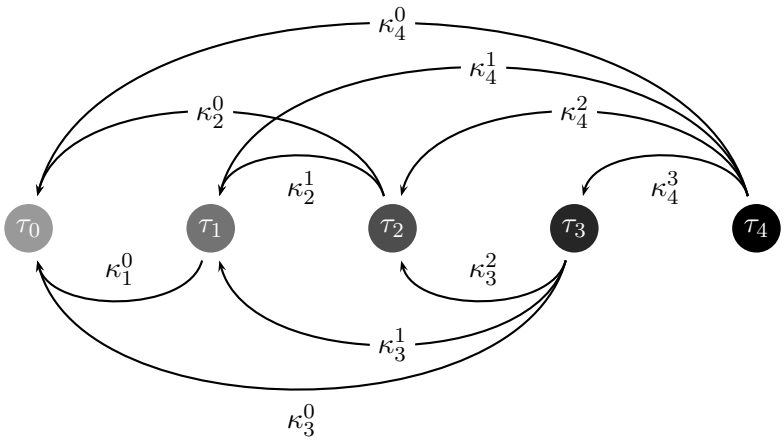


FIGURE 5.5 – Représentation graphique des conditions stratégiques de $F(\tau_\alpha, \tau_\beta)$

commutativité [1], ni son associativité, comme nous venons de le signaler. Nous revenons, à la section 5.5.4, sur les raisons de ce choix et en explicitons l'utilisation.

Les autres opérateurs de la logique multivalente permettent les opérations élémentaires sur les degrés de vérité et ainsi de modéliser les combinaisons d'informations et le doute associé. Parce que nous cherchons à modérer ces combinaisons selon notre perception de la confiance, nous avons dû définir un opérateur particulier. Les stratégies de combinaison que nous avons présentées plus haut (cf. § 4.4, p. 67) permettent de représenter un a priori de l'utilisateur sur l'évolution de sa conviction. Pour les intégrer au formalisme multivalué, nous avons redéfini les opérations de manière conditionnelle : la vitesse d'évolution de la confiance, en fonction de l'estimation de la compétence de la source ou de la plausibilité de l'information, est conditionnée par la force nécessaire au changement d'avis de l'utilisateur.

5.5 Opérateur de combinaisons de recoupement

La crédibilité requiert quelques efforts supplémentaires. Lors des étapes précédentes, un critère était évalué afin de déterminer l'évolution de la cote. Dans le cas de la crédibilité, il s'agit de comparer l'information à toutes les autres et de faire évoluer sa cote pour chaque information homologue [2]. Avant d'aborder l'opérateur proprement dit, nous revenons sur les différences de cette dimension présente avec les précédentes.

1. En effet, nous n'imposons pas que $\kappa_\alpha^\gamma = \kappa_\gamma^\alpha$.
2. Par homologue on désigne une information pouvant être rapprochée de l'information étudiée, soit en la corroborant, soit en l'infirmant, cf. Berti (1999a).

5.5.1 Principes

Nous avons vu qu'en cas d'information infirmante, la cote de l'information diminue ; à l'inverse, si une information connue la corrobore, ceci *peut* augmenter la cote (cf. section 4.3.3, page 67). Ainsi, la crédibilité a une influence différente des autres dimensions en ce qu'elle peut améliorer la cote. L'opérateur F précédent, dit 'à la baisse', ne convient donc pas et l'introduction d'un nouvel opérateur est nécessaire.
L'autre différence majeure de la crédibilité par rapport aux dimensions précédentes réside en son intégration à la cotation. Comme nous l'évoquions à la section 3.5 (p. 56), la crédibilité représente le niveau de confirmation, positif ou négatif, de l'information considérée par une information homologue. Cependant, nous souhaitons également tenir compte de la cote a priori[3] de ladite information homologue. En effet, s'il est bien évident qu'une confirmation parfaite de l'information considérée renforce sa cote, si cette confirmation est douteuse son impact doit être moins élevé que celui d'une contradiction extrêmement probable.

L'opérateur de combinaison que nous proposons pour l'intégration de la crédibilité à la cotation a donc trois arguments – $\tau_\alpha, \tau_\beta, \tau_\gamma \in \mathcal{L}^e_M$, respectivement la cote actuelle de l'information considérée, le niveau d'activation de la crédibilité et la cote a priori de l'information homologue comparée – et sa sortie est la cote mise à jour après l'intégration de la crédibilité, $\tau_\delta \in \mathcal{L}^e_M$. τ_α et τ_δ sont de même nature, représentant la cote d'entrée de l'information évaluée et sa version actualisée. τ_γ est également une cote a priori mais qualifie l'information homologue en fonction de laquelle la confiance est mise à jour. τ_β, en revanche, représente la crédibilité, c'est-à-dire le niveau de recoupement entre l'information cotée et son homologue envisagée. Rappelons que si la crédibilité est supérieure ou égale à $\tau_{\frac{M-1}{2}}$ l'information homologue est une confirmation et qu'elle est une infirmation dans le cas contraire.

Enfin, pour intégrer l'impact des informations homologues, deux options se présentent. Soit on considère l'évolution immédiate de la crédibilité, soit on choisit de tenir compte de la dynamique des recoupements. En effet, on peut considérer que la cote évolue pour chaque information homologue, dans la mesure de son impact. Dans ce cas, si une information douteuse vient confirmer l'information dont nous étudions la confiance, la majorité des stratégies n'en tiendra pas compte. C'est ce que nous appelons crédibilité immédiate. Cependant, il est possible qu'une succession de telles confirmations survienne. Bien que, prises séparément, ces confirmations ne soient pas convaincantes, nous pouvons vouloir prendre leur nombre en compte. Pour ce faire, nous devons choisir de nous souvenir des

3. i.e. cote hors crédibilité

confirmations et contradictions collectées et changer dès que l'un ou l'autre des niveaux requis pour son évolution est atteint. Ceci nécessite d'intégrer un effet mémoire à la crédibilité et permet la prise en compte de la *vox populi*, propriété qui fait que nous l'appelons crédibilité cumulée.

Dans la suite de cette section, nous proposons d'abord le fonctionnement général de l'opérateur de crédibilité, de manière semblable à ce que nous avons présenté précédemment pour l'opérateur F. Ensuite, nous étendons cet opérateur pour lui ajouter l'effet mémoire que nous recherchons pour intégrer la dynamique de la conviction et définissons la crédibilité cumulée. Enfin, nous discutons les conséquences de cette modélisation et en envisageons une variante simplifiée. Ces différentes propositions sont enfin illustrées par un exemple.

5.5.2 Crédibilité immédiate

Le premier opérateur que nous présentons pour la crédibilité varie, donc, à chaque comparaison avec une information homologue. Comme nous l'avons vu, cet opérateur tient compte à la fois de la crédibilité en tant que dimension – c'est-à-dire en tant qu'indicateur du taux de recoupement des deux informations homologues – et de la cote a priori de l'information comparée. Ceci se conçoit aisément si l'on considère la différence d'évolution de conviction que l'on tire de la confirmation par une discussion de comptoir de la découverte de la démonstration originelle de Fermat comparée, par exemple, à la lecture de la tribune d'Andrew Wiles, telle qu'évoquée dans l'exemple du chapitre 4.

Crédibilité pondérée

Puisque sa cote a priori influence la mesure dans laquelle nous tenons compte d'une confirmation comme, à l'inverse, d'une contradiction, nous proposons d'intégrer le taux de recoupement d'une information homologue en le pondérant par la confiance que nous faisons à cette information. Cette variante pondérée de la crédibilité indique l'ampleur de l'évolution de la cote. Pour faire ceci, nous définissons un opérateur de pondération de la manière suivante :

Définition 10 : Opérateur de pondération de la crédibilité

Soit g, une fonction telle que $g : \mathcal{L}^e_M \times \mathcal{L}^e_M \longrightarrow \mathcal{L}^e_M$

$$(\tau_\alpha, \tau_\beta) \longmapsto \tau_\gamma$$

où
- τ_α est le niveau d'activation de la crédibilité
- τ_β est la cote a priori de l'information homologue
- τ_γ est la crédibilité pondérée de l'information homologue

ayant les propriétés suivantes :

1. $g(\tau_\alpha, \tau_\beta) \leq \tau_\alpha, \forall \tau_\beta \in \mathcal{L}_M$, si $\tau_\alpha \geq \tau_{\frac{M-1}{2}}$
2. $g(\tau_\alpha, \tau_\beta) \leq \neg\tau_\alpha, \forall \tau_\beta \in \mathcal{L}_M$, si $\tau_\alpha < \tau_{\frac{M-1}{2}}$
3. $\forall \tau_\alpha, \tau_\beta, \tau_\gamma \in \mathcal{L}_M$
   - si $\tau_\beta \geq \tau_\alpha \geq \tau_{\frac{M-1}{2}}$, alors $g(\tau_\alpha, \tau_\gamma) \leq g(\tau_\beta, \tau_\gamma)$
   - si $\tau_\beta \leq \tau_\alpha < \tau_{\frac{M-1}{2}}$, alors $g(\tau_\alpha, \tau_\gamma) \leq g(\tau_\beta, \tau_\gamma)$
4. Si $\tau_\alpha \leq \tau_\beta$, alors $g(\tau_\gamma, \tau_\alpha) \leq g(\tau_\gamma, \tau_\beta), \forall \tau_\alpha, \tau_\beta, \tau_\gamma \in \mathcal{L}_M$
5. $g(\tau_?, \tau_\alpha) = g(\tau_\alpha, \tau_?) = \tau_?, \forall \tau_\alpha \in \mathcal{L}_M^e$

La première propriété indique que cet opérateur de pondération est borné par le niveau d'activation de la crédibilité. En effet, supposons qu'une information absolument certaine confirme partiellement l'information que nous cherchons à coter. Il semble naturel de n'en tenir compte dans l'évolution de la confiance qu'au mieux à la hauteur du recoupement.
De manière analogue, si une information extrêmement probable infirme en partie l'information à coter, sa crédibilité pondérée doit avoir une influence moindre que si elle la contredisait absolument. La seconde propriété propose donc une borne supérieure à l'impact de l'information homologue, symétriquement par rapport à $\tau_{\frac{M-1}{2}}$, de façon à respecter la gradualité de la crédibilité. Dans $\mathcal{L}_5$, par exemple, l'influence d'une infirmation de crédibilité τ_1 sera inférieure à τ_3.
En revanche, plus le recoupement est important, plus l'information influe sur la confiance. C'est ce que représente le fait que g soit croissante en son premier argument pour les confirmations et décroissante pour les infirmations (propriété 3). De plus, bien que bornée par le taux de recoupement, parce que nous pondérons l'impact de la crédibilité par la confiance faite à l'information homologue, plus celle-ci est sûre, plus son influence est grande. De ce fait, g est également croissante en son second argument (propriété 4).
Par ailleurs, si l'on ne peut estimer la crédibilité, on suppose que les informations ne sont pas homologues et on ne les compare pas. Enfin, si l'information que l'on cote est, par exemple, parfaitement confirmée par une autre dont on ne sait évaluer la cote, on ne peut faire évoluer la confiance, puisqu'on n'apprend rien de l'inquantifiable. Ces deux considérations sont rassemblées dans la propriété 5.

Définition

Nous utilisons l'opérateur de pondération ainsi défini pour exprimer l'intégration de la crédibilité. Cependant, avant d'envisager comment nous écrivons cette intégration, revenons un instant sur les propriétés que nous voulons qu'elle présente. Nous définissons l'opérateur d'intégration de la crédibilité ainsi :

Définition 11 : Opérateur de crédibilité immédiate

Soit G, une fonction telle que $G : \mathcal{L}_M^e \times \mathcal{L}_M^e \times \mathcal{L}_M^e \longrightarrow \mathcal{L}_M^e$

$$(\tau_\alpha, \tau_\beta, \tau_\gamma) \longmapsto \tau_\delta$$

où

- τ_α est la cote actuelle
- τ_β est le niveau d'activation de la crédibilité
- τ_γ est la cote a priori de l'information homologue
- τ_δ est la cote mise à jour

ayant les propriétés suivantes :

1. $G(\tau_\alpha, \tau_\beta, \tau_\gamma) \geq \tau_\alpha, \forall \tau_\alpha, \tau_\gamma \in \mathcal{L}_M$, si $\tau_\beta \geq \tau_{\frac{M-1}{2}}$
2. $G(\tau_\alpha, \tau_\beta, \tau_\gamma) \leq \tau_\alpha, \forall \tau_\alpha, \tau_\gamma \in \mathcal{L}_M$, si $\tau_\beta < \tau_{\frac{M-1}{2}}$
3. G est croissante en ses trois arguments
4. Conditions sur $\tau_?$:
   (a) $G(\tau_?, \tau_\beta, \tau_\gamma) = g(\tau_\beta, \tau_\gamma), \forall \tau_\beta, \tau_\gamma \in \mathcal{L}_M$
   (b) $G(\tau_\alpha, \tau_?, \tau_\gamma) = \tau_\alpha, \forall \tau_\alpha, \tau_\gamma \in \mathcal{L}_M$
   (c) $G(\tau_\alpha, \tau_\beta, \tau_?) = \tau_\alpha, \forall \tau_\alpha, \tau_\beta \in \mathcal{L}_M$

Rappelons que la crédibilité parcourt une échelle allant de l'infirmation, pour les valeurs basses, à la confirmation, dans ses degrés élevés. Nous choisissons de considérer que le milieu de l'échelle, $\tau_{\frac{M-1}{2}}$ à la frontière entre corroboration et contradiction, marque le début des confirmations. Toute évaluation inférieure de la crédibilité est donc intégrée comme contredisant l'information à coter, alors qu'une crédibilité supérieure ou égale renforce la confiance.

Si la crédibilité ne peut être estimée, les informations considérées ne sont pas homologues et la cote n'évolue pas.

Fonction proposée

Fort de ces propriétés attendues pour l'intégration de la crédibilité, nous décrivons un

opérateur, noté G, dont le comportement conditionnel est décrit par deux fonctions, respectivement $\tilde{F}_i$ pour les infirmations et $\tilde{F}_c$ pour les confirmations, commentées ci-dessous. Nous écrivons donc l'opérateur G ainsi :

$$G(\tau_\alpha, \tau_\beta, \tau_\gamma) = \begin{cases} \tau_\alpha & \text{si } \tau_\beta = \tau_? \\ \tilde{F}_i(\tau_\alpha, g(\tau_\beta, \tau_\gamma)) & \text{si } \tau_\beta < \tau_{\frac{M-1}{2}} \\ \tilde{F}_c(\tau_\alpha, g(\tau_\beta, \tau_\gamma)) & \text{sinon} \end{cases}$$

L'impact à la baisse des infirmations pondérées $\tilde{F}_i$ est semblable à ceux de la compétence et de la plausibilité F. Toutefois, l'ampleur de l'influence suit la gradualité de la crédibilité pondérée, là où elle était inversement corrélée à l'activation de la compétence et de la plausibilité. C'est-à-dire que plus τ_δ est élevé, plus $\tilde{F}_i(\tau_\alpha,\tau_\delta)$ est faible, alors que $F(\tau_\alpha, \tau_\delta)$ est élevée. Ceci signifie que les contraintes sur les κ_α^β, comme leur comparaison avec τ_β, sont inversées dans l'opérateur $\tilde{F}_i$ défini de façon analogue à F (cf. page 86). On écrit donc :

$$\tilde{F}_i(\tau_\alpha, \tau_\beta) = \begin{cases} \min\{\tau_\gamma \in \mathcal{L}_M | \tau_\beta \geq \kappa_\gamma^\alpha\} & \text{si } \tau_\alpha, \tau_\beta \in \mathcal{L}_M \\ \tau_\alpha & \text{si } \tau_\beta = \tau_? \\ \tau_\beta & \text{si } \tau_\alpha = \tau_? \end{cases}$$

La prise en compte des confirmations, quant à elle, se fait de manière analogue avec une tendance à la hausse. On définit donc l'opérateur $\tilde{F}_c$ de la même manière que F (cf. définition 9) mais en réécrivant la première propriété ainsi : $\tilde{F}_c(\tau_\alpha, \tau_\beta) \geq \tau_\alpha, \forall \tau_\alpha, \tau_\beta \in \mathcal{L}_M$. On a donc :

$$\tilde{F}_c(\tau_\alpha, \tau_\beta) = \begin{cases} \max\{\tau_\gamma \in \mathcal{L}_M | \tau_\beta \geq \kappa_\gamma^\alpha\} & \text{si } \tau_\alpha, \tau_\beta \in \mathcal{L}_M \\ \tau_\alpha & \text{si } \tau_\beta = \tau_? \\ \tau_\beta & \text{si } \tau_\alpha = \tau_? \end{cases}$$

De la même manière qu'à l'introduction de l'opérateur F, la figure 5.6 représente l'intégration des stratégies à la fonction $\tilde{F}_c$. Une fois encore, les valeurs de la cote y sont représentées par les éléments de $\mathcal{L}_5$. Notons tout de même que les seuils κ_α^β, bien que toujours indicatifs de l'effort nécessaire pour aller de τ_α à τ_β, vérifient un ordre inversé par rapport à la figure 5.5 reflétant bien l'impact à la hausse de $\tilde{F}_c$.

La figure 5.7 conclut cette introduction de l'opérateur d'intégration de la crédibilité par son illustration intuitive, superposition des figures 5.5 et 5.6. Nous la présentons pour expliciter les flèches de l'étape de crédibilité de la chaîne de cotation de la figure 4.1, page 64. Dans un souci de lisibilité de cette combinaison de deux opérateurs conditionnels, nous en avons ôté les seuils de transition κ. À leur place, les variations de bleu représentent l'influence de l'activation de la crédibilité pondérée.

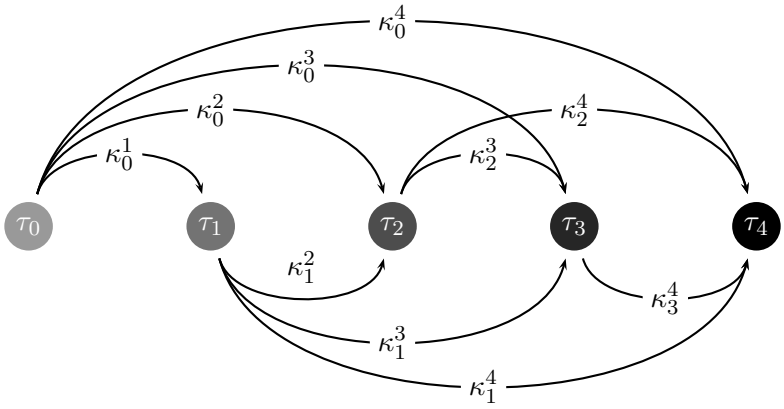


FIGURE 5.6 – Représentation graphique des conditions stratégiques de $\tilde{F}_c(\tau_\alpha, \tau_\beta)$

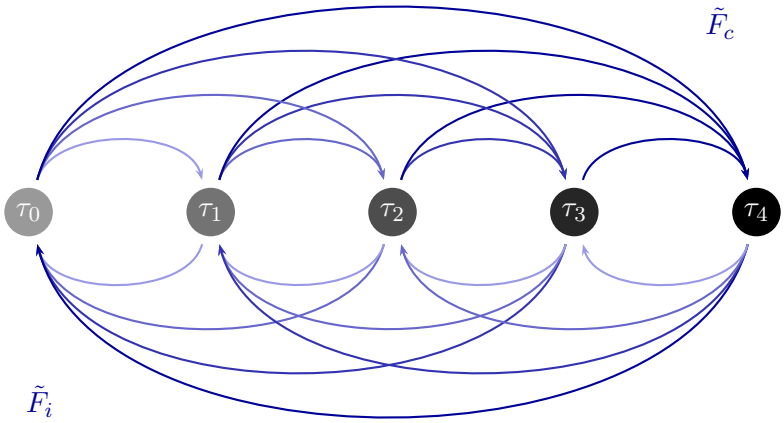


FIGURE 5.7 – Représentation graphique de $G(\tau_\alpha, \tau_\beta, \tau_\gamma)$, allégée des seuils

5.5.3 Crédibilité cumulée

Souhaitant poursuivre notre modélisation du processus de constitution d'avis, nous voulons tenir compte de la dynamique de la conviction. En effet, nous pensons qu'une succession abondante de confirmations, même très douteuses, a toujours un impact sur l'avis. Les rumeurs persistantes sont un exemple de ce cas de figure. Que l'on doute ou pas des informations qu'elles rapportent, l'afflux incessant de confirmations joue nécessairement sur notre point de vue.

Cependant, l'opérateur G précédent ne permet pas de représenter cette attente. Pour illustrer cette limitation, supposons que nous cherchions à évaluer une information cotée τ_3. Supposons maintenant qu'elle soit corroborée par une succession d'informations chacune de crédibilité pondérée τ_1. Si la stratégie est établie pour que seules les confirmations

au-delà de τ_2 influent positivement sur la cotation (i.e. $\kappa_3^4{=}\tau_2$) et que nous appliquons G, ces confirmations ne peuvent augmenter la confiance. S'il est vrai que la fonction des stratégies est de modérer les comportements intempestifs, il paraît anormal de ne pas se souvenir des confirmations – ainsi que des infirmations – pour profiter pleinement de toutes ces informations.

Afin de représenter et intégrer la dynamique de la conviction à notre processus de cotation, nous proposons d'enrichir l'opérateur G en le munissant d'un effet mémoire. En procédant ainsi à un cumul des comparaisons successives, nous sommes à même de laisser la rumeur nous atteindre.

Principe de la solution proposée

Nous proposons donc d'enrichir l'opérateur G en lui permettant de garder la trace des informations homologues à l'information à traiter. Pour tenir compte de la même façon des confirmations et des infirmations, nous le dotons de deux accumulateurs, notés g_- et g_+ et détaillés par la suite. Le premier de ces accumulateurs conserve le cumul des crédibilités pondérées des contradictions et le second celui des approbations. Dès lors, les fonctions $\tilde{F}_i$ et $\tilde{F}_c$ opèrent sur la cote courante et sur g_-, pour les contradictions, et g_+, pour les validations et non plus sur la valeur immédiate résultat de g.
La rémanence des accumulateurs ne permet plus de représenter l'intégration de la crédibilité comme une fonction. Nous l'envisageons donc comme un algorithme de mise à jour de la cote, interrogeable à tout instant pour obtenir la valeur courante.

Algorithme

Nous présentons ici plus formellement l'algorithme G^*(cf. p. 95) de mise à jour de la cote lors de la prise en compte d'une succession d'informations homologues. Puisque nous parlons d'une procédure de mise à jour, l'objet auquel elle s'applique fait partie des variables internes à la procédure et doit donc être initialisé avant l'application de l'algorithme. La cote d'entrée de cette phase est donc la cote a priori de l'information que nous traitons, soit τ_α.
De la même manière, les deux accumulateurs g_- et g_+ sont des paramètres internes de l'algorithme. Nous les initialisons à $\tau_?$, puisqu'avant de recevoir la première confirmation la dynamique de la confiance est indéterminée.

Après l'initialisation des variables internes, l'algorithme est appliqué tant que des informations homologues arrivent. Rappelons qu'une information de crédibilité $\tau_?$ n'est pas homologue à l'information à coter. L'algorithme traite donc une telle information en ne modifiant aucune de ses variables internes.

Ensuite, l'accumulateur de crédibilité pondérée adéquat, selon la prise en compte d'une

Algorithme G^* : Mise à jour de la cote par accumulation de crédibilités pondérées

if $\tau_\beta = \tau_?$ **then**
  $\tau_\alpha \leftarrow \tau_\alpha$
  $g_- \leftarrow g_-$
  $g_+ \leftarrow g_+$
else if $\tau_\beta < \tau_{\frac{M-1}{2}}$ **then**
  $g_- \leftarrow ADD(g_-, g(\tau_\beta, \tau_\gamma))$
  if $\tilde{F}_i(\tau_\beta, g_-) \neq \tau_\alpha$ **then**
    $\tau_\alpha \leftarrow \tilde{F}_i(\tau_\beta, g_-)$
    $g_- \leftarrow \tau_?$
    $g_+ \leftarrow \tau_?$
  end if
else if $\tau_\beta \geq \tau_{\frac{M-1}{2}}$ **then**
  $g_+ \leftarrow ADD(g_+, g(\tau_\beta, \tau_\gamma))$
  if $\tilde{F}_c(\tau_\beta, g_-) \neq \tau_\alpha$ **then**
    $\tau_\alpha \leftarrow \tilde{F}_c(\tau_\beta, g_-)$
    $g_- \leftarrow \tau_?$
    $g_+ \leftarrow \tau_?$
  end if
end if

infirmation ou d'une confirmation additionnelle, est augmenté tant que la dynamique de la conviction ne cède pas dans un sens ou dans l'autre, c'est-à-dire tant que ni $\tilde{F}_i(\tau_\beta,g_-)$ ni $\tilde{F}_c(\tau_\beta,g_-)$ ne permettent de mettre à jour la cote τ_α. Dès que l'un des seuils est atteint par g_- ou g_+, la cote courante est mise à jour et les deux accumulateurs réinitialisés.

Remarques

Nous commentons ici les alternatives envisageables de réinitialisation des variables internes de l'algorithme proposé. Par défaut ces réinitialisations internes sont totales, c'est-à-dire que les accumulateurs g_- et g_+ sont réinitialisés à $\tau_?$, mais nous envisageons d'autres alternatives que nous discutons ci-après. Nous abordons également, dans les remarques qui suivent, l'impact de l'ordre d'intégration des informations homologues sur la cote,

conséquence de l'effet mémoire. Pour nous permettre de garantir la stabilité de la cote de sortie, nous proposons enfin une version appauvrie de la prise en compte de l'effet mémoire, où le déclenchement ne s'opère qu'après que toutes les comparaisons ont été effectuées.

L'ajout des accumulateurs g_- et g_+ permet donc d'ajouter à G l'effet mémoire désiré. La version par défaut de l'algorithme, présentée ci-dessus, tient ainsi compte de la dynamique de la conviction. Cependant, lors de l'étape de mise à jour de la cote, la nécessaire réinitialisation des accumulateurs proposée l'est dans sa version la plus radicale, à $\tau_?$. Une autre possibilité, plutôt que de revenir à l'état initial d'inertie de la conviction, est de choisir d'en garder la trace. Si, par exemple, la cote augmente suite à de nombreuses confirmations, il n'est pas obligatoire de se départir du début de doute que les contradictions accumulées avaient créé. De même, si $g_- > \kappa_\alpha^\gamma$, conserver le surplus de confiance engrangée (i.e. $SUB(g_-, \kappa_\alpha^\gamma)$) témoigne du fait que la dernière évolution de la cote était vers le doute.

Quelle que soit la version choisie, la prise en compte de la dynamique de conviction induit une dépendance à l'ordre d'arrivée des informations. Sauf dans les cas particuliers où les stratégies les rendent associatifs, voire commutatifs, les opérateurs que nous présentons ici ne garantissent pas un résultat pour un même lot d'informations homologues, présentées dans un ordre différent. Ainsi, l'intégration d'une confirmation, suivie d'une contradiction, puis d'une seconde confirmation ne produit pas nécessairement le même résultat que deux confirmations, suivies d'une infirmation. En extension de la manière dont nous proposons d'utiliser les stratégies pour représenter les postures de crédulité, nous pensons que cette particularité enrichit notre modèle en lui intégrant différentes perspectives sur la prise en compte de propos rapportés. Si l'on se place face à un flux d'informations, nous constatons que chacun ne réagit pas de la même manière. L'art de contredire son adversaire au bon moment pour mieux convaincre, lors de débats politiques ou devant les tribunaux, compte sur l'intersection de comportements potentiellement très variés.

Si nous concevons la possibilité d'adapter notre procédé à différentes personnalités comme un atout, il peut, néanmoins, arriver qu'un tel fonctionnement ne soit pas acceptable. Si nous supposons, dans un cadre d'extraction automatique de connaissances, que nous traitons des informations issues d'une même source, nous voyons que, lors du traitement, le calcul ne peut être immédiat, avec effet mémoire ou sans. Dans ce cas, en effet, la garantie de la stabilité du résultat prime sur l'expressivité de la procédure. En effet, s'il est envisageable de recevoir, successivement, différentes informations provenant de sources différentes, les traitements d'une même source doivent être concomitants.
Pour parer cette inquiétude, il est possible d'adapter l'algorithme proposé ici pour que le cumul se poursuive tant que des informations homologues sont disponibles et que la mise à jour de la cote n'intervienne qu'une fois que toutes les comparaisons ont été effectuées.

Plus explicitement, il s'agit d'opérer toutes les mises à jour de g_- et de g_+ sans évaluer $\tilde{F}_i$ ni $\tilde{F}_c$ avant d'avoir pris en compte toutes les informations homologues. Puisque les opérateurs ne sont pas évalués, les accumulateurs ne sont pas réinitialisés. Le seul choix restant, dès lors, porte sur l'ordre d'intégration entre infirmations et confirmations – c'est-à-dire entre $\tilde{F}_c(\tilde{F}_i(\tau_\alpha,g_-),g_+)$ et $\tilde{F}_i(\tilde{F}_c(\tau_\alpha,g_+),g_-)$ – soit selon la posture de crédulité, soit selon l'importance relative de la dynamique. Cet ordre étant fixé pour un utilisateur, la stabilité de la sortie est assurée.

5.5.4 Crédibilité, dynamique et flux d'informations : exemple illustratif

Afin de rendre compte des possibilités offertes par notre modélisation de la crédibilité, nous proposons un exemple de confrontation d'attitudes différentes face à un flux d'informations contradictoires. Nous allons étudier les constitutions d'avis de deux utilisateurs aux stratégies différentes face à une même succession d'informations entrantes.
La cote, la crédibilité, sa version pondérée et les seuils de transition sont tous évalués dans $\mathcal{L}_5$, où les étiquettes sémantiques de chaque évaluation reprennent les interprétations des tableaux 3.2 à 3.6 (pp. 51 à 59).

Mise en place
Le tableau 5.2 présente le détail des stratégies des deux utilisateurs pour la crédibilité. Pour chaque utilisateur, deux colonnes sont présentées, l'une listant les valeurs des seuils de transition à la baisse, l'autre ceux dirigeant la hausse, c'est-à-dire $\tilde{F}_i$ et $\tilde{F}_c$, respectivement. Étant donné les contraintes imposées aux κ_α^β (cf. section 5.4), certaines transitions sont impossibles. En effet, puisque pour l'utilisateur 1, par exemple, $\kappa_4^2{=}\tau_4$, $\kappa_4^2{<}\kappa_4^1$ et $\tau_4{=}\tau_{M-1}$, la cote ne peut pas passer directement de τ_4 à τ_1. Ces cas où la transition est interdite sont représentés dans le tableau par *nd*.

Une présentation plus intuitive des stratégies des utilisateurs est proposée dans les figures 5.8 et 5.9. Pour conserver l'expressivité de la vue d'ensemble de l'opérateur sous-jacent, les cas de transitions indéfinies sont tout de même représentés par des flèches claires dépourvues de seuil. Le soin est laissé au lecteur de superposer les opérateurs pour représenter totalement G.
On constate que pour l'utilisateur 2 les seuils de transition ne sont pas symétriques à la hausse et à la baisse. Par exemple, $\kappa_3^4{=}\tau_3$ alors que $\kappa_4^3{=}\tau_1$. En n'imposant pas de symétrie entre les seuils nous permettons de modéliser des postures cognitives différentes. En effet, une façon de représenter une personnalité méfiante est de favoriser la diminution de la confiance sur son augmentation. Puisque, pour l'utilisateur 2, $\kappa_3^4{=}\tau_3$ et $\kappa_4^3{=}\tau_1$ la dynamique ainsi créée représente de manière adéquate la défiance.

Utilisateur 1				Utilisateur 2			
↓		↑		↓		↑	
κ_1^0	τ_3	κ_0^1	τ_3	κ_1^0	τ_2	κ_0^1	τ_4
κ_2^0	τ_4	κ_0^2	τ_4	κ_2^0	τ_3	κ_0^2	nd
κ_2^1	τ_3	κ_0^3	nd	κ_2^1	τ_2	κ_0^3	nd
κ_3^0	nd	κ_0^4	nd	κ_3^0	τ_3	κ_0^4	nd
κ_3^1	τ_4	κ_1^2	τ_3	κ_3^1	τ_2	κ_1^2	τ_3
κ_3^2	τ_3	κ_1^3	τ_4	κ_3^2	τ_1	κ_1^3	τ_4
κ_4^0	nd	κ_1^4	nd	κ_4^0	τ_4	κ_1^4	nd
κ_4^1	nd	κ_2^3	τ_3	κ_4^1	τ_3	κ_2^3	τ_2
κ_4^2	τ_4	κ_2^4	τ_4	κ_4^2	τ_2	κ_2^4	τ_3
κ_4^3	τ_3	κ_3^4	τ_3	κ_4^3	τ_1	κ_3^4	τ_3

TABLE 5.2 – Paramètres de $\tilde{F}_i$ et $\tilde{F}_c$ pour deux utilisateurs aux postures de crédulité différentes.

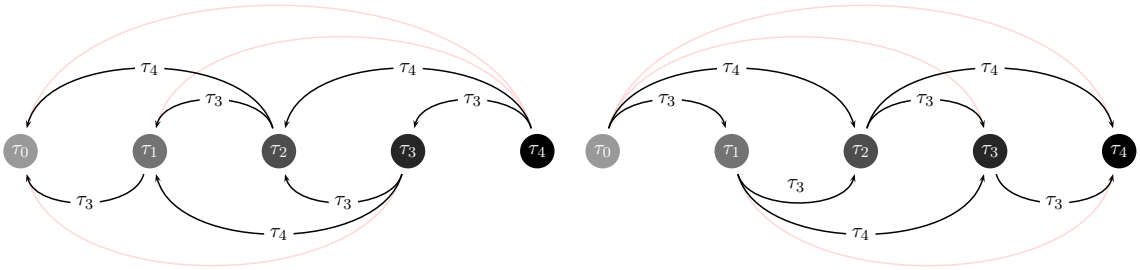


FIGURE 5.8 – Les conditions stratégiques de $\tilde{F}_i(\tau_\alpha, \tau_\beta)$ (à gauche) et de $\tilde{F}_c(\tau_\alpha, \tau_\beta)$ (à droite) pour l'utilisateur 1

L'utilisateur 1, quant à lui, change d'avis plus difficilement que son collègue, mais de façon symétrique à la hausse et à la baisse. Ne montrant pas d'a priori dans un sens ou dans l'autre, il peut être qualifié de convaincu réticent.

Confrontations à un flux d'informations : cas de la crédibilité cumulée

Nous nous donnons, enfin, une succession d'informations homologues à l'information à traiter. Ces informations sont fournies les unes après les autres et sont prises en compte

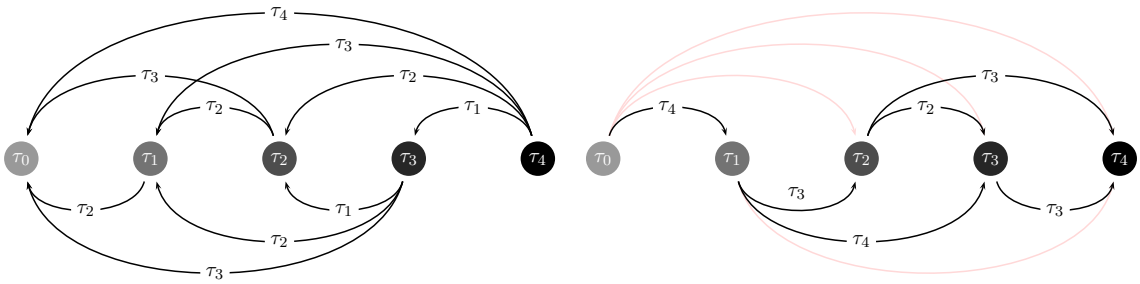


FIGURE 5.9 – Les conditions stratégiques de $\tilde{F}_i(\tau_\alpha, \tau_\beta)$ (à gauche) et de $\tilde{F}_c(\tau_\alpha, \tau_\beta)$ (à droite) pour l'utilisateur 2

dès leur arrivée. Pour alléger la notation, nous n'indiquons pour chaque information que si elle est une confirmation ou une infirmation (i.e. τ_β) ainsi que sa crédibilité pondérée (i.e. $g(\tau_\beta, \tau_\gamma)$). Les tableaux 5.3 et 5.4 présentent l'évolution de la cote à l'intégration de chacune de ces informations, par utilisateur dans la colonne τ_α. Les deux dernières colonnes de ces tableaux détaillent l'évolution des accumulateurs. La cote a priori initiale est de τ_2 (i.e. l'information est *'envisageable'*) et chaque ligne représente la prise en compte d'une nouvelle information.

Flux d'informations		Utilisateur 1		
τ_β	$g(\tau_\beta, \tau_\gamma)$	τ_α	g_+	g_-
τ_4	τ_1	τ_2	τ_1	$\tau_?$
τ_0	τ_2	τ_2	τ_1	τ_2
τ_2	τ_2	τ_3	$\tau_?$	$\tau_?$
τ_1	τ_2	τ_3	$\tau_?$	τ_2
τ_3	τ_3	τ_4	$\tau_?$	$\tau_?$
τ_0	τ_2	τ_4	$\tau_?$	τ_2

TABLE 5.3 – Traçage de la crédibilité cumulée pour l'utilisateur 1 face au flux d'information

Le tableau 5.3 présente l'évolution de la cote de l'information, par l'utilisateur 1, pendant l'intégration du flux d'informations. Notons que, dans ce cas d'intégration d'informations de faible crédibilité pondérée, sa cote évolue entre degrés successifs. Qu'une information permette de sauter un ou plusieurs degré dépend de la stratégie mais, en général, la tendance de la posture de crédulité de l'utilisateur est liée aux $\kappa_\alpha^{\alpha+1}$ et $\kappa_{\alpha-1}^{\alpha}$, en raison des contraintes sur les seuils.

Malgré la régularité de ces transitions immédiates, l'utilisateur 1 ne doit pas être perçu comme excessivement crédule. En effet, le cumul de crédibilités pondérées élevé exigé témoigne de sa réticence à changer d'avis. Il finit par se laisser convaincre (aboutissant à τ_4, soit considérant l'information comme *'extrêmement probable'*), malgré l'alternance entre confirmations et infirmations.

Flux d'informations		Utilisateur 2		
τ_β	$g(\tau_\beta, \tau_\gamma)$	τ_α	g_+	g_-
τ_4	τ_1	τ_2	τ_1	$\tau_?$
τ_0	τ_2	τ_1	$\tau_?$	$\tau_?$
τ_2	τ_2	τ_1	τ_2	$\tau_?$
τ_1	τ_2	τ_0	$\tau_?$	$\tau_?$
τ_3	τ_3	τ_0	τ_3	$\tau_?$
τ_0	τ_2	τ_0	τ_3	τ_2

TABLE 5.4 – Traçage de la crédibilité cumulée pour l'utilisateur 2 face au flux d'information

L'utilisateur 2 est, en revanche, plutôt méfiant. Il est non seulement difficile à convaincre (i.e. les transitions à la hausse sont élevées et celles à la baisse basses) mais, qui plus est, perçoit très mal les changements de cap (i. e. la hausse est d'autant plus compliquée que la cote décroit). De façon peu surprenante, l'utilisateur 2 se laisse convaincre très vite par les infirmations et ne revient pas sur son avis, concluant à une information cotée τ_0, c'est-à-dire *'improbable'*.

Crédibilité cumulée totale

La solution du cumul total, proposée pour le traitement automatique, ne nécessite pas vraiment d'illustration. En effet, puisque la décision n'est prise qu'une fois que toutes les informations ont été traitées, la cote n'évolue pas pendant l'exécution de l'algorithme et les accumulateurs g_- et g_+ ne sont pas réinitialisés. Il suffit donc de prendre une décision lorsque l'agrégation de toutes les crédibilités pondérées est disponibles. Ici, on obtient, au total, $g_-{=}g_+{=}\tau_4$.

Notons, tout d'abord, que puisque le délai avant la décision est plus important, les facteurs en jeu le sont également. De ce fait, les transitions sollicitées sont plus extrêmes qu'auparavant. La stratégie de l'utilisateur 1 étant symétrique, que l'application de $\tilde{F}_i$ ou $\tilde{F}_c$ intervienne en premier ne change pas le résultat, sa cote de sortie est égale à celle

d'entrée, soit τ_2. La seule différence réside en l'extrémité de l'échelle de confiance qu'il visite.

Nous avons constaté que l'utilisateur 2 était plus méfiant que son confrère. Nous notions, en particulier, qu'il était plus facile à dissuader qu'à convaincre. Dans le cas présent, où la somme des arguments en faveur et à l'encontre du fait étudié sont égales, l'ordre de leur intégration va jouer sur l'issue de sa confiance. Si l'on intègre d'abord les contradictions, puis les validations de l'information, l'utilisateur 2 ne croit guère l'information et la cote finale est τ_1. Si, cependant, on commence par prendre en compte les corroborations pour ensuite lui révéler les infirmations, il ne croit absolument plus l'information et lui attribue la cote minimale de τ_0.

Cas de la crédibilité immédiate

Les tableaux 5.5 et 5.6 illustrent l'évolution de la cote pour le même flux d'information dans le cas d'un calcul de crédibilité immédiate.

τ_β	$g(\tau_\beta, \tau_\gamma)$	τ_α
τ_4	τ_1	τ_2
τ_0	τ_2	τ_2
τ_2	τ_2	τ_2
τ_1	τ_2	τ_2
τ_3	τ_3	τ_3
τ_0	τ_2	τ_3

TABLE 5.5 – Traçage de la crédibilité immédiate pour l'utilisateur 1 face au flux d'information

L'utilisateur 2, facilement dissuadé et peu enclin à remettre son avis en jeu lors de la crédibilité cumulée, reproduit le même comportement. L'utilisateur 1, en revanche, privé de la mémoire de dynamique de sa conviction, se révèle plus sceptique qu'auparavant. Il se laisse convaincre plus lentement et, pour un même flux d'informations, finit à une cote moins élevée, de τ_3.

τ_β	$g(\tau_\beta, \tau_\gamma)$	τ_α
τ_4	τ_1	τ_2
τ_0	τ_2	τ_1
τ_2	τ_2	τ_1
τ_1	τ_2	τ_0
τ_3	τ_3	τ_0
τ_0	τ_2	τ_0

TABLE 5.6 – Traçage de la crédibilité immédiate pour l'utilisateur 2 face au flux d'information

5.6 Conclusion

Dans ce chapitre, nous avons proposé une représentation du processus d'évaluation de la cotation dans le formalisme de la logique multivalente répondant à nos exigences de lisibilité formelle. Confronté à une limitation de ce formalisme, nous avons proposé d'assouplir ses contraintes afin de représenter la nuance entre l'impossibilité d'estimer et une mesure neutre exprimée. Nous avons, ensuite, vérifié dans quelle mesure nos assouplissements remettaient en question le formalisme et avons conclu que le gain en expressivité pour la cotation était suffisamment important pour les justifier.

Nous avons proposé des opérateurs de combinaison dans le cadre de la logique multivalente et représentant le procédé de cotation. Après avoir décrit contraintes et comportements attendus pour chaque opérateur, nous en avons donné une expression formelle. Nous avons également discuté leur flexibilité, montrant, notamment, que différentes postures cognitives pouvaient être représentées. Enfin, nous avons proposé des adaptations lorsque l'utilisation l'exigeait.

L'ensemble de ces propositions constitue la traduction formelle et opérationnelle de la chaîne de cotation complète qui permet sa mise en pratique pour une application concrète, présentée dans la partie suivante.

Troisième partie

Mise en pratique du modèle de cotation

Introduction à la troisième partie

La troisième partie de ce mémoire présente la mise en pratique du modèle de cotation proposé. L'application considérée démontre l'intégration de la cotation à l'extraction de connaissances à partir de données textuelles. Un démonstrateur a été développé pour la mise en œuvre de cette problématique réalisant la reconstitution et la cotation d'un réseau social à partir de sources ouvertes.

Le chapitre 6 examine la transposition du modèle général de la cotation à l'extraction de connaissances à partir de données textuelles. Présentant cette problématique comme la succession d'une étape de recherche d'informations symboliques suivie de leur fusion, il décrit l'application d'une chaîne de cotation double et l'adaptation de l'expression des dimensions de la cotation aux spécificités des objets à coter dans ce cadre. Il précise également le cadre expérimental considéré, la reconstitution de réseaux sociaux à partir de sources ouvertes.

Le chapitre 7 décrit le système développé pour la mise en œuvre de nos propositions. Il présente d'abord les modules d'extraction et de fusion d'informations implémentés pour permettre l'identification des acteurs et liens du réseau. Il détaille ensuite l'intégration de la cotation au système dans chacune des deux étapes d'extraction et de fusion. Il explicite les méthodes d'évaluation des dimensions en examinant ce qu'elles mesurent dans ce cas particulier et comment. Enfin, l'outil est mis en application sur des données réelles constituées de la version préliminaire d'un ouvrage collectif portant sur la description de réseaux terroristes.

Chapitre 6

Cotation de connaissances extraites de données textuelles

Résumé

Ce chapitre introduit la mise en pratique du modèle de cotation proposé. Il débute par une présentation générale de l'extraction de connaissances de données textuelles, puis étudie l'intégration de la cotation à ces traitements. Enfin, il se conclut par la présentation du cadre expérimental de la mise en œuvre des propositions.

6.1 Introduction

Nous avons présenté, dans la partie II, un modèle global d'établissement de la confiance en une information. Nous l'avons illustré à l'aide de diverses transpositions à des cadres d'application différents. Nous nous intéressons maintenant plus particulièrement à l'un d'entre eux, l'extraction de connaissances à partir de données non-structurées, à partir de textes.

Notre apport à la problématique est d'appliquer notre chaîne de cotation à l'extraction d'informations. Nous examinons, donc, dans un premier temps, l'architecture usuelle d'un système d'extraction de connaissances, afin de déterminer l'agencement de la cotation dans la chaîne de traitements. Ce premier examen nous amène à délimiter les étapes et acteurs

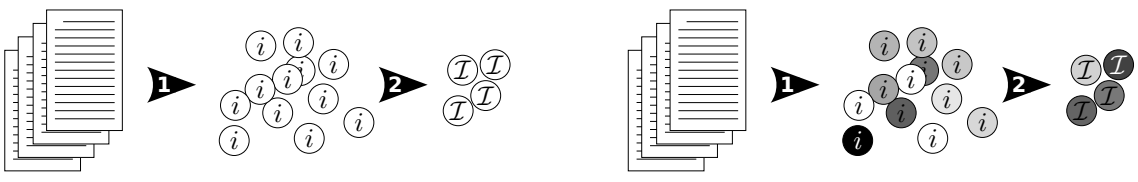
de la recherche de connaissances, afin de mieux y situer la cotation.

Par la suite, nous étudions plus avant la transposition de notre modèle à cette chaîne de traitements. Notons que nous y trouvons un cadre idéal puisque toute information extraite émane d'une source et peut être renforcée ou contredite par d'autres informations extraites ou fournies. À partir de l'étude initiale, nous proposons d'intégrer la cotation à l'issue de chacune des deux phases de la création de connaissances. Nous reprenons également la définition proposée des acteurs de ces traitements, afin de lui faire correspondre les dimensions que nous proposons pour la cotation.

Enfin, nous terminons ce chapitre par la présentation du cadre expérimental que nous nous sommes donné. Cette application est remise en contexte et nous y introduisons les principes nécessaires à la description de nos travaux, détaillée dans le chapitre suivant.

6.2 Organisation de l'extraction de connaissances cotées

Nous désignons par le terme connaissance une information riche extraite automatiquement d'un texte. Si l'on considère, par exemple, l'extraction de connaissances à partir de dépêches, la date, le lieu et la liste des acteurs de l'événement rapporté constituent, ensemble, une telle information riche. La construction de cette connaissance est généralement réalisée en deux temps, comme illustré par la figure 6.1.a : tout d'abord a lieu l'identification des informations simples, dites aussi ici atomiques, la composant, à l'aide de différents algorithmes dédiés, par exemple spécialisés dans la reconnaissance de dates (e.g. '17 décembre 2010'), de lieux (e.g. 'Sidi Bouzid') ou de noms de personnes (e.g. 'Mohamed Bouazizi'). Intervient ensuite la fusion de ces informations atomiques, pour produire l'information enrichie, que nous appelons connaissance (e.g. 'immolation par le feu de Mohamed Bouazizi à Sidi Bouzid, le 17/12/2010', connaissance à rapprocher des débuts de la révolution tunisienne de 2010). Cette distinction se rapproche de celle qui sépare les informations des renseignements, tels que présentés dans les définitions 1 et 2, page 13.



(a) (b)

FIGURE 6.1 – (a) Extraction de connaissances – (b) Extraction de connaissances cotées

Nous nous intéressons à la cotation de ces connaissances, c'est-à-dire, l'estimation de la

confiance qui leur est accordée. Le principe de cet ajout est illustré dans la figure 6.1.b par la coloration des informations, atomiques comme enrichies. Pour parvenir à ce but, nous proposons une double application du processus de cotation, traitant successivement les informations atomiques, puis les connaissances issues de leur fusion. Le procédé représenté à la figure 6.1.b est explicité dans les figures 6.2 pour les informations atomiques, et 6.3 pour les connaissances, commentées ci-dessous : la phase d'extraction, correspondant à la première flèche de la figure 6.1.b, se décompose comme indiqué dans la figure 6.2 ; la seconde flèche, représentant l'étape de fusion, est détaillée dans la figure 6.3.

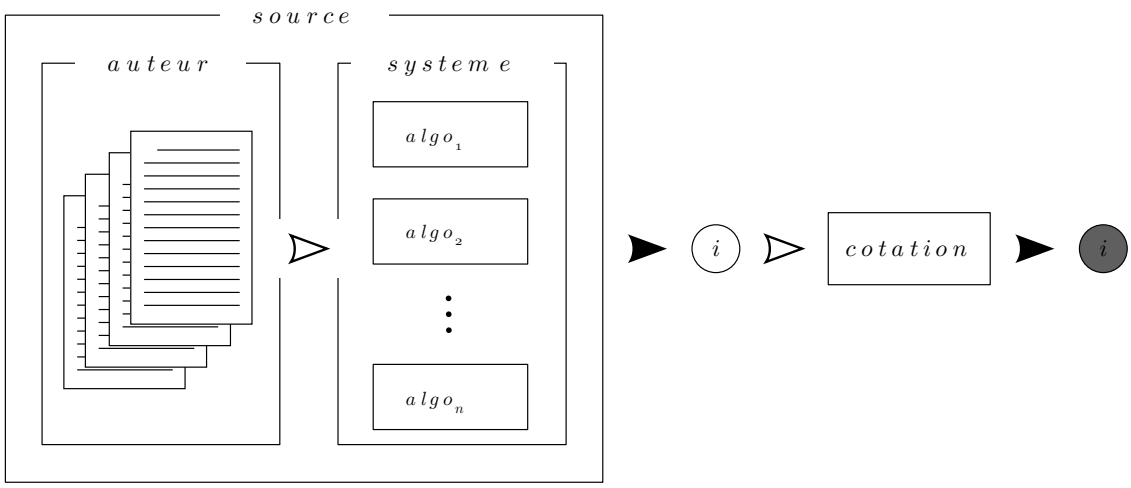


FIGURE 6.2 – Extraction et cotation d'informations atomiques

Les informations atomiques que nous considérons sont extraites du corpus étudié par un jeu d'algorithmes spécialisés. Nous considérons donc que la source d'une de ces informations atomiques est une combinaison de l'auteur du texte dont elle provient et de l'algorithme l'ayant extraite. La première flèche pleine de la figure 6.2 représente la production d'une information atomique par cette source combinée. L'information ainsi produite ne dispose pas encore de l'évaluation de sa cote. Elle est donc fournie en entrée – flèche creuse de la figure – au système de cotation qui la munit de cette évaluation.

Après l'étape d'extraction, nous disposons de toutes les informations atomiques cotées. Celles-ci servent d'entrée au module de fusion – première flèche creuse de la figure 6.3 – qui produit des informations enrichies en combinant les informations atomiques décrivant un même élément. Ces informations riches, bien que construites à partir d'informations atomiques cotées, ne sont pas encore cotées elles-mêmes. Elles sont donc, à leur tour, soumises au processus de cotation. Nous détaillons dans la suite les spécificités de chacune de ces deux étapes de cotation.

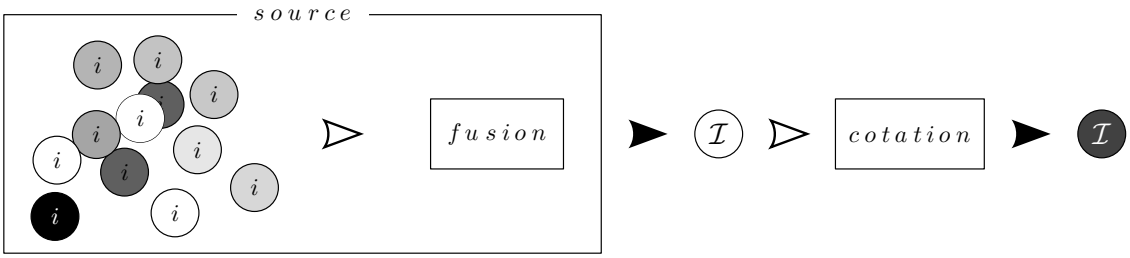


FIGURE 6.3 – Fusion d'informations atomiques et cotation de connaissances

6.3 Intégration de la cotation

Nous présentons ici nos propositions de transpositions du processus de cotation introduit dans la partie II au problème de l'extraction de connaissances à partir de données textuelles. Nous détaillons successivement la cotation de l'extraction d'informations, puis celle de leur fusion, en examinant pour chacune la transposition des dimensions qui la constituent, fiabilité, compétence, plausibilité et crédibilité.

6.3.1 Cotation de l'extraction d'informations

Nous commençons par étudier l'étape de cotation de la figure 6.2, qui se produit suite à l'extraction. Celle-ci s'applique à des informations atomiques telles que, par exemple, le '14 janvier 2011' ou 'Zine el-Abidine Ben Ali, *son épouse* Leïla Ben Ali'. Dans ce dernier exemple, l'information atomique découverte est la relation entre deux acteurs. Elle informe sur l'existence d'un lien entre deux noms identifiés par un autre algorithme spécialisé et permet, lors de la fusion, de confirmer que les deux acteurs de la relation sont bien des personnes.

Fiabilité

Rappelons que les deux premières dimensions de la cotation qualifient la source que nous considérons, pour une information atomique, comme l'association de l'auteur du texte dont elle est issue et de l'algorithme l'ayant extraite. La fiabilité est la dimension générale et non contextuelle de la source : son influence sur la cote reste la même, quelle que soit l'information considérée, tant que l'utilisateur ne revoit pas son jugement sur la source.
Nous l'identifions donc ici à la fiabilité de l'auteur du texte dont elle est extraite. En effet, dans la transposition au texte, la fiabilité de l'auteur s'applique de la même manière à toutes les informations extraites.

Compétence

La compétence de la source est sa dimension contextuelle. Elle est évaluée en fonction du recoupement entre le sujet de l'information et les domaines de compétence de sa source. En ce qui concerne les informations atomiques issues de textes, la compétence de l'auteur n'est pas définie. En effet, tout auteur est compétent à fournir une date ou un nom propre. Son domaine de compétence se définit à un niveau supérieur, où la date qualifie un événement sur lequel l'auteur est à même de se prononcer. Ces notions sont définies pour les connaissances, non pour les informations élémentaires.

De ce fait, la compétence de la source, telle que nous l'avons définie, décrit mieux les algorithmes, responsables de la production des informations atomiques. Nous proposons, donc, d'intégrer la mesure de leurs performances à l'estimation de la cote, en témoignage de la capacité avec laquelle chaque algorithme intégré au système est spécialisé pour une tâche, son domaine de compétence.

Plausibilité

La plausibilité de l'information représente sa concordance aux connaissances extérieures. Dans le cas général, on peut supposer qu'un 'trente février' extrait ait une plausibilité minimale. Cependant, nous cherchons, en particulier, des pépites, des informations sémantiquement prometteuses, dont la découverte nous mène à des connaissances riches. De plus, la certification pour chaque algorithme de la validité de sa production devient vite une charge importante.

Dans le cas particulier qui nous occupe, nous choisissons de considérer que, pour une information atomique, cette concordance n'a pas de sens. Une date découverte par un algorithme spécialisé ne doit pas être en contradiction avec le modèle du monde qu'il parcourt. Cette dimension qualifie, donc, mieux les informations enrichies et, pour les informations atomiques traitées ici, elle est évaluée à $\tau_?$.

Crédibilité

Enfin, la dimension de recoupement, la crédibilité, ne présente pas de spécificité pour le cas de l'extraction d'informations. Elle correspond donc à sa définition initiale, la mesure du recoupement entre informations homologues. Pour la datation d'événement, par exemple, la crédibilité mesure combien le '14/01/2011' et 'un mois après le 17 décembre', deux dates extraites, se confirment ou s'infirment.

Il faut noter que la fusion de ces informations et leur complémentation par d'autres informations atomiques, comme les acteurs de l'événement extraits par d'autres algorithmes spécialisés, intervient dans un deuxième temps. Si l'algorithme d'extraction de dates peut reconstruire une date, il ne produit que des informations de son domaine de compétence, les dates. L'étape de fusion, quant à elle, combine des informations de types différents, afin

d'en extraire une connaissance d'une granularité sémantique supérieure.

6.3.2 Cotation de la fusion d'informations

La seconde étape de la cotation, représentée dans la figure 6.3, évalue la confiance à accorder aux connaissances : elle qualifie des informations enrichies, issues de la fusion d'éléments atomiques. Soulignons qu'à la manière de ce que nous envisagions pour une information atomique, la source d'une telle connaissance est double : elle est constituée, d'une part, des informations atomiques dont elle est issue et, d'autre part, du module de fusion.

Initialisation

Le processus de cotation que nous avons proposé consiste en la mise à jour de la confiance par l'intégration de dimensions pertinentes. C'est pourquoi nous avons présenté la première d'entre elles, la fiabilité, comme la dimension d'initialisation de la cotation pour une information encore inconnue. Cependant, les connaissances que nous considérons maintenant sont construites à partir d'informations cotées. La cotation d'une connaissance a, donc, une initialisation spécifique par l'agrégation des cotes des informations atomiques la constituant.

Fiabilité

Une connaissance étant le résultat de la fusion d'informations élémentaires, la notion de fiabilité de sa source est délicate. En effet, issue de plusieurs informations atomiques, l'identification d'un auteur unique est impossible. De plus, les fiabilités des divers auteurs des informations atomiques sont déjà prises en compte lors de leurs cotations, puis à l'initialisation décrite ci-dessus.

Nous posons, donc, que la fiabilité ne peut être estimée pour une connaissance et l'évaluons à $\tau_?$.

Compétence

Bien qu'il soit impossible d'identifier une source unique d'une information enrichie pour en évaluer la fiabilité, nous savons comment elle est construite. De la même manière que la compétence de la source, pour une information atomique, est identifiée à la performance de l'algorithme l'ayant identifiée, la compétence de la source pour une information enrichie est la mesure de la performance des règles de fusion qui interviennent dans sa constitution.

Plausibilité

La plausibilité de l'information mesure sa conformité avec la représentation de l'état du monde. Une information enrichie décrit la connaissance que le système a construite sur

un fait ou un événement à partir d'informations atomiques. Il est envisageable que cette connaissance soit en contradiction partielle avec la représentation du monde dont dispose le système. Un système de résumé de manifestations à partir de dépêches peut détecter, par exemple, une incohérence lorsque le chiffrage de la participation selon les autorités est supérieur à l'estimation des organisateurs. Une telle contradiction ne peut se produire que lorsque les informations atomiques – les chiffrages – sont comparées et agrégées suite à la fusion et la création de connaissance.
La plausibilité, dans le cas d'informations enrichies, n'est pas définie de manière spécifique et correspond à sa présentation dans le cadre théorique.

Crédibilité
La crédibilité n'a, quant à elle, plus lieu d'être évaluée pour les informations enrichies. En effet, une fois l'étape de fusion achevée, les informations atomiques homologues sont intégrées à une même connaissance. La mesure de confirmation entre informations enrichies n'a, donc, plus de sens et nous évaluons la crédibilité à $\tau_?$.

6.4 Cadre expérimental

Comme nous le disions au chapitre 3, les propositions pour la cotation que nous exposons ici peuvent s'appliquer à différents types de traitements automatiques. Nous nous sommes particulièrement intéressé à la problématique de la recherche et l'extraction de connaissances à partir de textes, parce qu'elle nous rapproche d'opérations et de thématiques présentes dans le renseignement militaire, cadre originel de notre étude. Nous avons vu, au chapitre 1, que les cas d'usage du renseignement sont multiples. Pour une part importante d'entre eux, cependant, le cycle de recherche d'informations concerne une demande précise, sur l'état du terrain ou des forces adverses, par exemple, à laquelle une réponse doit être fournie rapidement. Les réponses à de telles demandes sont, bien entendu, évaluées selon une méthodologie qui peut être décrite dans le cadre de nos propositions. Toutefois, le formalisme que nous présentons ici dévoile son intérêt lors de l'étude de gros volumes de données, lorsque l'évaluation systématique par un être humain devient impossible. Si ceci n'est, généralement, pas le cas pour les demandes précises, le traitement de sources ouvertes pose ces questions.

La problématique du renseignement opérationnel de sources ouvertes vise à établir une connaissance générale d'un domaine, à partir de données publiquement disponibles. Cette activité de veille s'occupe d'établir des dossiers sur la situation politique d'états étrangers, sur les personnalités y participant ou sur des intérêts économiques. Parmi ces sources, on

compte la presse, les études d'instituts statistiques, les communiqués gouvernementaux, ainsi que les informations provenant de bases de données professionnelles. À cette somme colossale d'informations s'ajoutent également toutes celles disponibles sur Internet. De tels volumes de données à traiter ne sauraient l'être manuellement et la sensibilité de cette veille requiert des indicateurs sur les informations manipulées. Nous avons, donc, choisi de montrer l'apport et la faisabilité de l'intégration de la chaîne de cotation dans la recherche de connaissances en développant un démonstrateur dans un cadre d'application contrôlé. Nous présentons ici, dans un premier temps, les données sur lesquelles nous avons travaillé. Avant de détailler, dans le chapitre suivant, les aspects techniques de notre outil, nous en décrivons l'objectif, tel qu'envisagé par un expert.

6.4.1 Données

La collection de textes que nous étudions provient de l'ouvrage collectif 'Guerre secrète contre Al-Qaeda', édité sous la direction d'Éric Denécé [1], directeur du Centre Français de la Recherche sur le Renseignement (CF2R), publié aux éditions Ellipses en 2002. Comme son titre l'indique, ce livre traite, en quatorze chapitres, de la lutte contre le terrorisme née des suites des événements de septembre 2001. Des experts y étudient les attentats, les ripostes et les organisations terroristes impliquées. Chacun des auteurs participant au projet est spécialiste du domaine qu'il aborde.

Les textes sur lesquels nous travaillons sont la version telle qu'envoyée par les auteurs, avant relecture de l'éditeur. À la complexité du traitement automatique de données non-structurées s'ajoutent, donc, les problèmes de fautes de frappe et d'incohérences de notation. On trouve, par exemple, trois orthographes différentes et répétées pour désigner Al-Qaeda et diverses translittérations de noms arabes.

6.4.2 Scénario d'usage

Lors de la récolte des besoins utilisateurs, réalisée sous forme d'entretiens avec Éric Denécé, l'éditeur souhaitait pouvoir accéder à des données qu'il savait présentes dans l'ouvrage, mais trop disséminées pour être facilement localisables. En particulier, il souhaitait pouvoir retrouver les relations décrites entre personnes et organisations.

1. L'auteur nous a confié son ouvrage à titre expérimental mais ne nous a pas autorisé à en reproduire des passages, le livre étant encore en vente. Nous ferons donc référence aux problèmes rencontrés sans en citer d'extraits trop détaillés.

Nous formalisons cette tâche comme l'extraction et la reconstitution d'une forme de réseau social décrit dans un ensemble de textes. Plus précisément, il s'agit de reconnaître les acteurs, personnes ou organisations, et leurs liens. Nous avons proposé d'enrichir ces recherches en leur ajoutant une mesure de la confiance à accorder aux informations extraites, par application du modèle de cotation.

Nous détaillons ci-dessous la transposition dans ce cadre des notions précédentes d'informations atomiques, de connaissances et de fusion. Les informations atomiques que nous cherchons à reconnaître correspondent aux noms des acteurs du réseau, leurs attributs (e.g. le prénom ou le sexe d'une personne) et leurs relations (e.g. deux organisations ayant des buts communs). Les acteurs ne sont pas tous de même nature, qu'ils soient personnes ou organisations. La distinction entre ces deux types est essentielle, puisqu'elle permet l'interprétation du réseau par recoupements et rapprochements. Elle se fait par l'exploitation des relations et attributs extraits.
Les connaissances sont, comme de façon générale, la fusion des informations atomiques et permettent, ici, d'unir tous les détails disponibles sur un acteur. Ces connaissances constituent les nœuds du réseau. Les arcs les liant représentent les relations entre ces entités.
La cotation appliquée à ce type d'exercice porte à la fois sur les arcs et les nœuds. Elle témoigne de la confiance à porter en la solidité des relations et indique, d'une part, si la connaissance présentée est bien conforme aux informations disponibles dans le texte et, d'autre part, si cette connaissance semble centrale au propos.
Le résultat de cette reconstruction peut être représenté comme il l'est à la figure 6.4, où les nœuds et les arcs sont colorés selon la confiance qui leur est faite. Le chapitre suivant, qui présente l'implémentation d'un démonstrateur dédié à cette tâche, revient, en particulier, sur l'obtention de ce résultat.

6.5 Conclusion

Dans ce chapitre, nous nous sommes intéressé à la cotation dans l'extraction de connaissances à partir de données textuelles. Dans un premier temps, nous avons présenté une vision de l'architecture d'un tel système permettant de distinguer les informations simples, matière brute de l'extraction, des connaissances riches, résultat de leur raffinage et production du système. Nous avons proposé de munir chacune de ces phases d'une étape de cotation, profitant de la compatibilité du modèle proposé et de la conduite de cette recherche.

Dans un second temps, nous nous sommes penché sur les spécifications de cette trans-

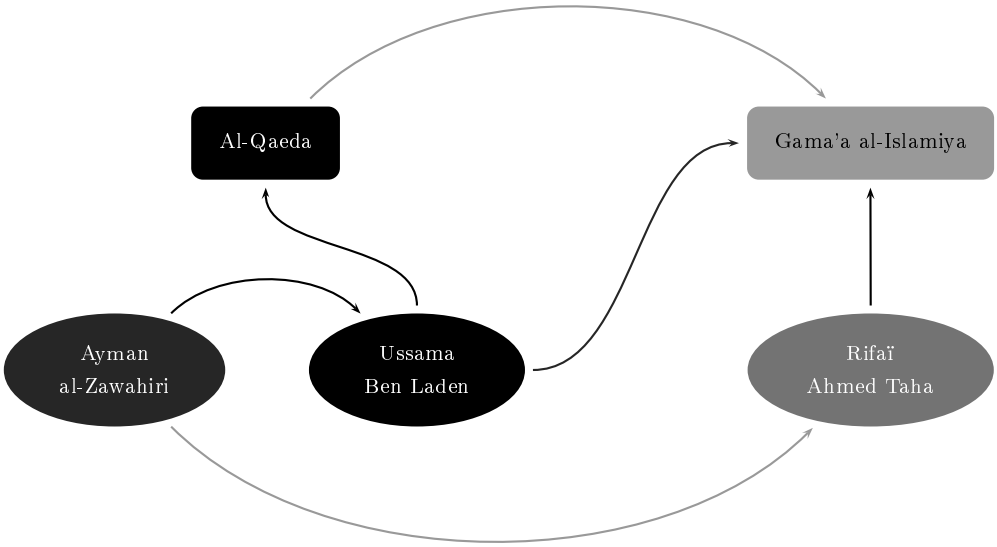


FIGURE 6.4 – Cotation d'un réseau social extrait de textes

position. Pour chacune des étapes de l'extraction de connaissances, nous avons présenté et discuté l'expression des dimensions de la cotation. Les spécificités des objets de la cotation nous ont, parfois, amené à privilégier l'estimation d'une dimension, soit pour les informations atomiques, soit pour les connaissances. Dans ces cas particuliers, la succession des deux phases de cotation et la disponibilité du degré $\tau_?$ nous a permis de jouer sur l'ordre – et l'objet – d'intégration des dimensions.

Enfin, nous avons conclu ce chapitre par une première présentation du cadre expérimental de nos travaux. Avant d'entrer dans les détails techniques de l'implémentation de la cotation dans un système d'extraction de réseau, nous avons rappelé la demande de l'utilisateur, présenté les données et précisé l'objectif. Dans la suite, nous verrons, donc, la mise en œuvre de cet outil et comment la cotation s'y intègre.

Chapitre 7

Mise en œuvre de la cotation pour l'extraction de réseau

Résumé

Ce chapitre décrit le système répondant à l'objectif de reconstitution et cotation d'un réseau social à partir d'un corpus de documents textuels. Il débute par la présentation des modules d'extraction automatique d'informations, puis expose les modules implémentant la chaîne de cotation appliquée au problème considéré. Enfin, la cotation dans l'extraction est illustrée dans un exemple, reprenant toutes les étapes du processus.

7.1 Introduction

Nous avons choisi d'aborder le problème complexe de l'extraction de connaissances à partir de textes comme un problème d'orpaillage. Partant du principe que quelques portions de textes, potentiellement rares, étaient sémantiquement riches, nous nous sommes concentré sur leur extraction. Les pépites que nous cherchons sont chargées de sens parce que leur objet est identifié et que, de cette signifiance même, dérivent les lignes guidant le raffinage. Nous verrons, par exemple, dans ce chapitre comment une seule phrase permet d'unifier plusieurs individus et de les lier à une organisation.

Avant, donc, d'être à même d'illustrer l'extraction de connaissances comme l'orpaillage

que nous décrivons, nous présentons ici les développements réalisés pour le démonstrateur chargé d'extraire et, plus principalement, coter un réseau social à partir de données textuelles. Dans un premier temps, nous présentons les modules d'extraction d'informations, puis de connaissances. Une fois cette mise en place décrite, nous décrivons l'intégration de la cotation en en détaillant l'implémentation. Le procédé décrit au chapitre précédent, séparant la cotation de l'extraction d'informations et la cotation de la fusion, est, donc, repris en détails, chaque dimension étant expliquée à son tour. Nous pouvons, alors, proposer un exemple complet, reprenant l'ensemble du processus et montrant la facture du réseau présenté à la figure 6.4 (p. 116). Enfin, nous concluons ce chapitre par une brève discussion sur la problématique de l'évaluation du modèle de cotation proposé.

7.2 Modules d'extraction d'informations

Cette section présente les modules implémentés pour la manipulation de textes, permettant l'extraction automatique d'informations atomiques et leur fusion. Comme détaillé à la section 6.4.2 (p. 114), les informations atomiques extraites se répartissent par nature : celles qui se rapportent aux acteurs du réseau et renseignent sur leurs attributs, par exemple leur nom ou leur nationalité ; celles qui décrivent les relations des acteurs entre eux, qui permettent de construire le réseau. Les acteurs, les nœuds du réseau, sont soit des personnes, soit des organisations, chacun de ces deux types ayant des attributs le décrivant, dont la figure 7.1 donne un aperçu pour une personne. La distinction entre les types oriente les attributs à chercher et les relations possibles.

Parmi les attributs des acteurs, le nom joue un rôle essentiel en en permettant l'identification. Pour cette recherche, nous proposons l'utilisation de trois modules complémentaires : la reconnaissance d'informations géographiques, l'identification de noms propres et l'extraction puis l'explicitation d'acronymes. Chacune de ces méthodes confère, de plus, à son résultat l'un des types recherchés, sa production étant, donc, de la forme (label, type). Le module de reconnaissance d'informations géographiques peut, parfois, renseigner un second attribut des acteurs, la nationalité.

Les relations, quant à elles, sont extraites par un module dédié et se représentent sous la forme d'un triplet (label, label, relation). Les relations décrivant des liens entre acteurs, comme la relation 'chef de' liant un individu à l'organisation qu'il dirige, ont pour labels les noms de leurs objets. D'autres relations sont également extraites, permettant l'acquisition de valeurs pour les attributs d'acteurs. C'est, par exemple, le cas de la découverte d'alias pour un individu, fournissant une appellation différente du nom courant. Les relations



FIGURE 7.1 – Attributs d'une personne, en gras

portant sur des objets spécifiques, leurs contraintes d'application permettent aussi de typer leurs objets. La figure 7.2 présente la liste des relations décrites dans l'ontologie pour une personne, ainsi que les domaines de définition de la relation de nationalité.



FIGURE 7.2 – Relations d'une personne à gauche, description de la relation 'is_from' à droite

La recherche de distinction entre types spécifiques d'informations est une problématique étudiée, notamment, dans le domaine de l'extraction d'entités nommées. Parmi les tâches classiques de celle-ci, la reconnaissance de dates ou de noms de personnes font l'objet de compétitions (ACE07 2007; 2007). Deux méthodes distinctes sont, généralement, proposées pour l'implémenter, la recherche dans le texte à partir de listes et la reconnaissance à l'aide d'heuristiques. Les modules que nous présentons ici implémentent l'une et l'autre : l'identification des noms de lieux exploite une liste fournie a priori ; la recherche de noms propres est réalisée à l'aide de règles heuristiques.

Dans la suite de cette section, nous décrivons successivement les quatre modules d'identification d'informations atomiques, auxquels nous ajoutons le composant responsable de la fusion, qui combine celles-ci en connaissances. Nous détaillons, pour chaque module, les objectifs recherchés, sa mise en œuvre, puis ses résultats.

7.2.1 Extraction d'informations géographiques

La première recherche d'entités nommées à laquelle nous procédons concerne les lieux. Pour l'établir, nous munissons le système de connaissances sur le monde lui permettant d'identifier les pays, leurs capitales ainsi que les qualificatifs de nationalité. La découverte de noms de villes et de pays les élimine de la liste des noms propres et favorise la reconnaissance de personnes. La nationalité, pour sa part, fait partie des attributs recherchés pour les acteurs du réseau.

Mise en œuvre

À partir de données du Ministère des Affaires Étrangères, nous construisons une structure décrivant les pays du monde et leurs gentilés. Nous avons traduit, sous la forme d'une ontologie OWL, la liste fournie sur le site web du ministère (désormais disponible sur le portail de l'Union Européenne[1]) qui associe à chaque pays les variantes de son nom (e.g. 'Union du Myanmar' ou 'Birmanie'), sa capitale et la dénomination de ses habitants. Pour ces derniers nous disposons des adjectifs féminins et masculins, si une différence existe (e.g. 'burkinabè', forme invariable des nationaux du 'Burkina Faso'), auxquels nous ajoutons une forme plurielle. La figure 7.3 présente un extrait de cette ontologie dans Protégé^{TM}.

L'identification des informations géographiques que nous recherchons consiste donc à trouver les occurrences des termes présents dans l'ontologie. La découverte d'un nom de ville ou de pays fournit donc une information de type 'Place', plus précisément 'City' ou 'Country', alors qu'un gentilé sert à qualifier une information extraite et typée ('Person' ou 'Organisation') par ailleurs.

Résultats

Sur l'ensemble du texte, nous obtenons 1 073 occurrences d'informations géographiques, dont 13% de capitales, 54% de pays et 33% de gentilés. Les erreurs commises par l'extraction à partir de chaînes de caractères connues sont rares, néanmoins, elles se produisent. Le cas le plus fréquent d'erreur est lié à une forme de polysémie des noms propres. Puisqu'ils comportent, en général, une majuscule initiale les noms de pays peuvent être confondus avec un nom de personne. C'est, par exemple, le cas de 'Dominique' qui peut désigner la

1. `http://publications.europa.eu/code/fr/fr-5000500.htm`

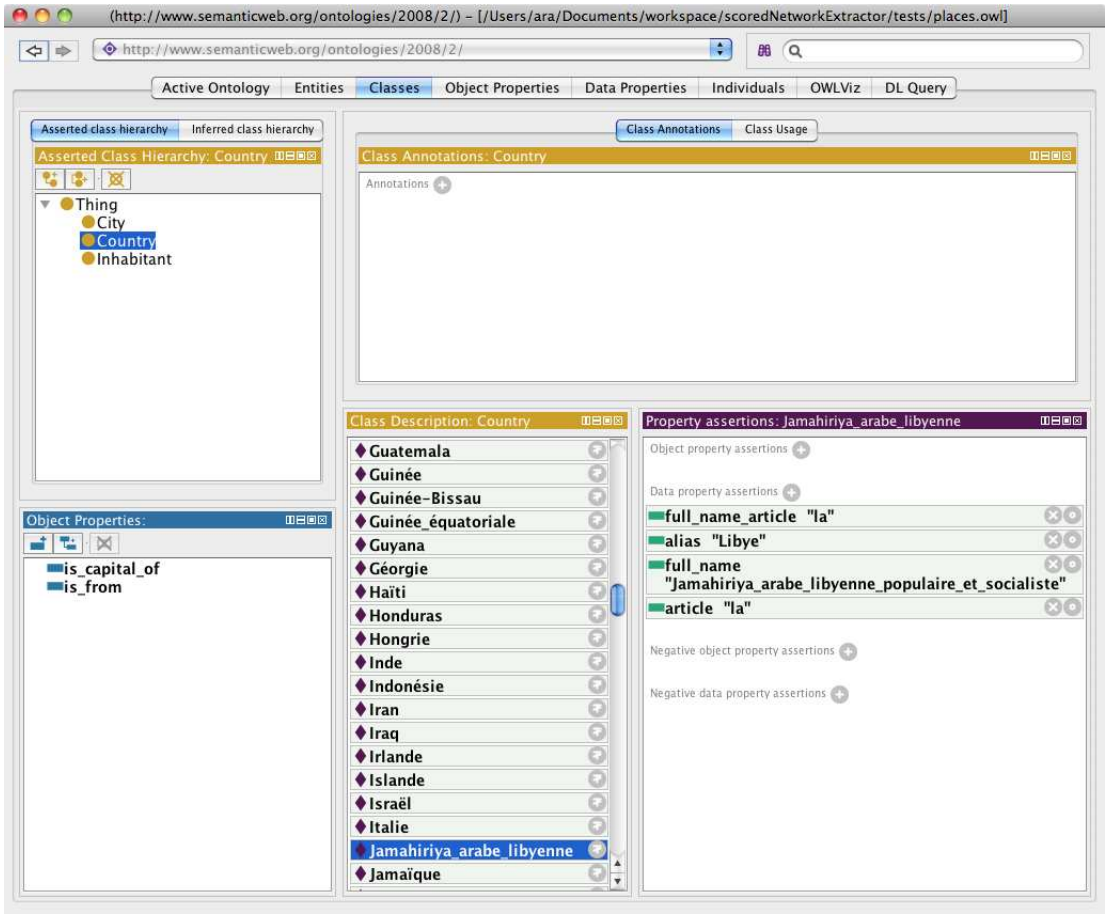


FIGURE 7.3 – Connaissances extérieures : une ontologie de lieux (pays, capitales, dénominations des habitants)

'République dominicaine' comme être un prénom. Comme pour l'ensemble des modules présentés ici, la prise en compte de ces erreurs intervient lors de l'évaluation de l'algorithme, à l'étape de la projection de la compétence sur la cotation. Les conflits éventuels entre chaînes de caractères communes entre informations atomiques sont, eux, résolus à la phase de crédibilité, pendant laquelle les informations homologues sont confrontées.

7.2.2 Extraction de noms propres

La population d'un réseau social requiert la détermination de ses membres. Sa reconstitution à partir de textes nécessite de les y découvrir. Les membres du réseau social que nous étudions sont des personnes et des organisations. Nous procédons, donc, à la recherche automatique de noms propres et l'orientons vers les noms de personnes.

Objectifs

Plutôt que nous donner une liste de noms, nous choisissons de procéder à leur extraction à partir de règles, afin de mieux représenter le travail d'opérateurs de traitement de sources ouvertes. En effet, bien que d'autres travaux dont le contexte est similaire se fondent sur l'exploitation d'une base de connaissances détaillées d'acteurs et de relations (Mannes et Golbeck 2007a; 2007b), l'établissement de nouvelles connaissances ne peut se limiter à l'exploitation de ce que l'on sait. Notre but étant de bâtir automatiquement un tel réseau, partir de la liste de ses membres limiterait la portée de notre démonstrateur.

Mise en œuvre

Pour extraire les noms propres, nous supposons que les règles orthographiques sont respectées. Cette recherche est, donc, implémentée de manière standard, c'est-à-dire, à l'aide d'expressions régulières. Nous définissons les patrons correspondant à des noms de personnes comme des suites de mots débutant par une majuscule. De ces patrons nous excluons les débuts de phrases – sauf si le mot suivant débute également par une majuscule – et les mots entièrement en majuscules. Ces derniers sont considérés comme des acronymes potentiels et sont soumis au traitement détaillé dans la section suivante (cf. § 7.2.3).

L'ouvrage sur lequel nous travaillons comporte 61 196 mots, dont 15% commencent par une majuscule. On y trouve, notamment, trois orthographes différentes et répétées pour désigner Al-Qaeda et diverses autres translittérations de noms arabes. Parce que certains des noms recherchés sont d'origine étrangère, quelques aménagements – notamment sur la capitalisation – sont apportés aux règles initiales. Ainsi, celles-ci considèrent 'Ussama ben Laden' comme contenant deux noms propres, 'Ussama' et 'Laden', puisque 'ben' ne comporte pas de majuscule. Nous les adaptons donc en leur ajoutant une règle d'extension autour de syntagmes intercalaires, tels que 'ben', ici, dont la liste est fixée manuellement.

L'étape cruciale de notre travail est celle du typage des noms propres identifiés, pour différencier les personnes des lieux et des organisations. Conscient que notre méthode ne découvre pas exclusivement les personnes que nous cherchons, nous nous appuyons à nouveau sur un jeu d'heuristiques. De la liste des éléments extraits, nous pouvons déjà exclure les noms de pays identifiés précédemment, à la section 7.2.1. Nous verrons, à la section 7.2.3 suivante, que nous identifions des noms d'organisations à partir du déchiffrement d'acronymes. Ces noms propres sont également exclus du typage par défaut en tant que 'Person' des noms propres restants.

Afin de renforcer la confiance dans ce typage par défaut, nous cherchons également des mots-clés permettant de valider cette supposition. Par exemple, si un nom propre est précédé ou suivi d'un titre (e.g. '[...] du Président Yahya Khan, du Premier ministre Ali Bhutto et du Président Zia.', voir tableau 7.1 pour d'autres exemples) le typage de l'entité

ainsi décrite en est renforcé (cf. section 7.3.1). De même, la découverte de relations, sujet de la section 7.2.4, permet de valider ou préciser le typage des entités reliées.

Patron	Type de l'entité
`monsieur`	*'Person'*
`docteur`	*'Person'*
`cheikh`	*'Person'*
`président`	*'Person'*
`ministre`	*'Person'*
`bank`	*'Organisation'*
`organisation`	*'Organisation'*
`groupuscule`	*'Organisation'*
`faction`	*'Organisation'*
'Organisation' $\oplus$ `terroriste`	*'Terrorist_ Group'*

TABLE 7.1 – Exemples de mots-clés favorisant le typage des noms propres extraits

Résultats

L'extraction de noms propres est relativement efficace avec une précision de 85%. Pour l'établir, nous vérifions la liste de l'ensemble des entités extraites et en ôtons les items mal reconnus, que l'item ne soit pas un nom propre (e.g. 'Préfet Bernard Gérard Première', où le titre 'Préfet' est repéré, grâce à la recherche de mots-clés, mais le début de la 'Première partie' dont il est l'auteur n'est pas séparé) ou que le nom propre reconnu ne soit pas entier (e.g. 'Hassan Zein', plutôt que 'Hassan Zein al Abidine').

Le typage par défaut en tant que personne a un taux de réussite de l'ordre de 60%, établi lui aussi par vérification manuelle des extraits.

7.2.3 Extraction et déchiffrage d'acronymes

La poursuite de l'identification des membres du réseau mène à retrouver les noms d'organisations. L'étude du corpus traité nous révèle que le déchiffrement d'acronymes autorise l'identification d'organisations. En effet, ayant constaté que dans les cas où un sigle était explicité dans le texte, il faisait généralement référence à une organisation, nous

proposons un module de déchiffrage d'acronymes. En outre, si les organisations font partie des nœuds que nous cherchons à identifier et relier, elles permettent également de supposer une relation entre personnes proches de l'une d'entre elles.

Objectifs

Ne pouvant identifier toutes les références aux organisations grâce à l'utilisation de mots-clés – comme, par exemple, 'organisation terroriste' – nous nous penchons sur l'explication d'acronymes. En effet, lorsqu'un acronyme est explicité dans le texte, il désigne, le plus fréquemment, une organisation. Puisque nous souhaitons établir des liens entre personnes, l'extraction d'organisations – et si possible de liens d'appartenance – favorise ces rapprochements.

Des travaux traitant de déchiffrement d'acronymes proposent d'interroger des bases de connaissances extérieurs, tels Jain et al. (2007) qui interrogent Google. Nous proposons plutôt d'effectuer la résolution à partir du texte, considérant, dans l'orpaillage, que l'information est vraisemblablement pertinente si elle est incluse dans le document. Un acronyme ne nécessitant pas d'explicitation n'est probablement pas spécifique au sujet étudié.

Mise en œuvre

Notre traitement des acronymes se découpe en deux étapes : leur détection, puis leur déchiffrement. La première de ces étapes consiste à rechercher les suites d'au moins deux majuscules, éventuellement séparées par des points. De cette manière nous repérons les sigles tels que : 'FIS', 'NAVSPECWARCOM' ou 'C.I.A'. Cette règle ne permettant pas de détecter les acronymes comportant des chiffres, comme 'CF2R', nous lui ajoutons cette possibilité. Certaines portions de texte, entièrement en majuscules, produisent des erreurs lors de la détection des acronymes et sont donc rejetées avant l'étape d'explicitation.

La seconde étape, de déchiffrement, est plus complexe et nous proposons donc de la traiter à l'aide d'heuristiques. De façon générale, ces heuristiques s'appuient sur le fait que l'explicitation de l'acronyme est à proximité. Nous recherchons donc la forme étendue de l'acronyme détecté dans la même phrase, en distinguant les cas particuliers de la présence de parenthèses. Plus précisément, en leur présence nous considérons les cas suivants :

- l'acronyme est seul dans une paire de parenthèses : sa forme étendue se trouve généralement à proximité avant la parenthèse ouvrante
- l'acronyme est accompagné de texte dans la paire de parenthèses : on cherche d'abord son explication dans la parenthèse
- l'acronyme est suivi d'une parenthèse : son expansion est, probablement, entre les parenthèses

La résolution consiste en la sélection de mots consécutifs – avec un fenêtrage glissant en cas de besoin – dont les premières lettres composent le sigle. Notons que, en général, les mots concernés ne comportent pas de majuscules dans la phrase et la recherche d'explicitation ne peut se fonder sur de tels indices. Sur un point plus technique, recherche et explication d'acronymes sont implémentées à l'aide d'expressions régulières.

Résultats

La figure 7.4 présente le traçage de la méthode produit pendant le traitement d'un texte : l'acronyme étudié y est présenté à la ligne *'Resolving'* ; l'expression régulière construite pour son déchiffrement à la ligne *'Trying'* ; la portion de phrase privilégiée pour sa découverte est donnée à la ligne *'On'* ; et la ligne *'Matched'* présente le résultat.

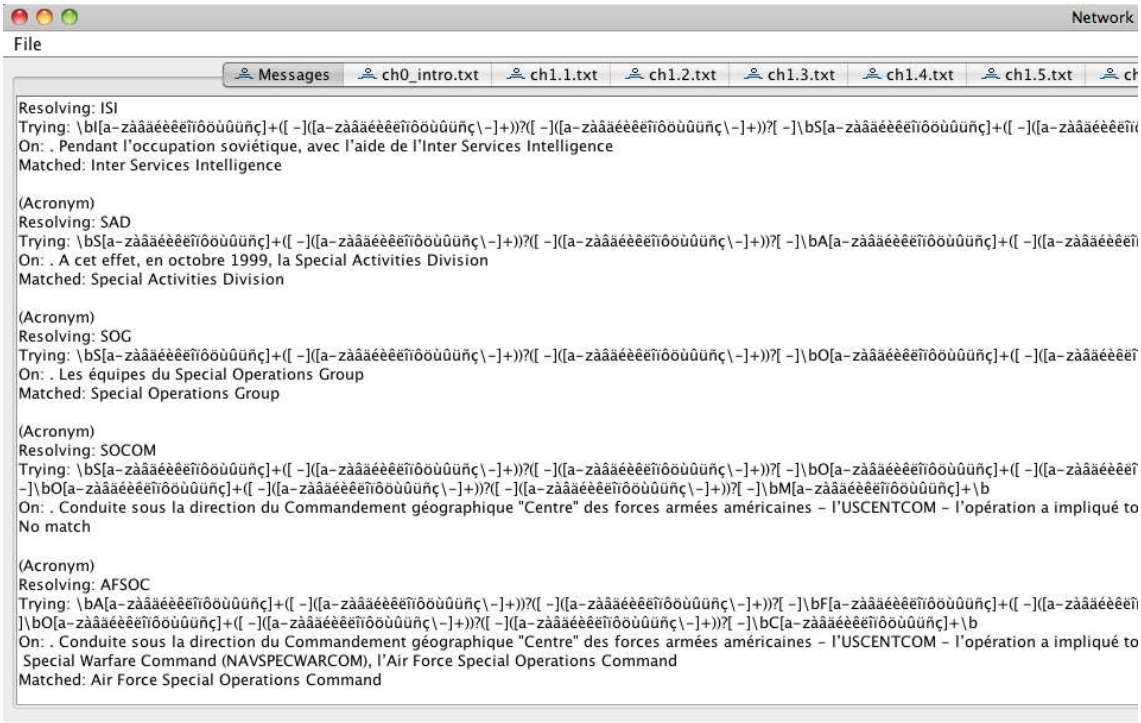


FIGURE 7.4 – Déchiffrement d'acronymes lors de l'extraction d'organisation

La figure 7.5, quant à elle, donne les statistiques de détection et de déchiffrage des acronymes, dans le cas de la présence de parenthèses : parmi les cent six cas où l'acronyme est seul entre parenthèses, le module en résout soixante et un ; deux des quatorze cas où l'élicitation est entre parenthèses sont résolus ; sur les trente-quatre autres cas, où acronyme et nom complet partagent les parenthèses, vingt solutions sont trouvées. À partir de ces résultats nous définissons une précision de 53%. Le rappel, de 50%, est établi par parcours du texte et pointage des acronymes solubles.

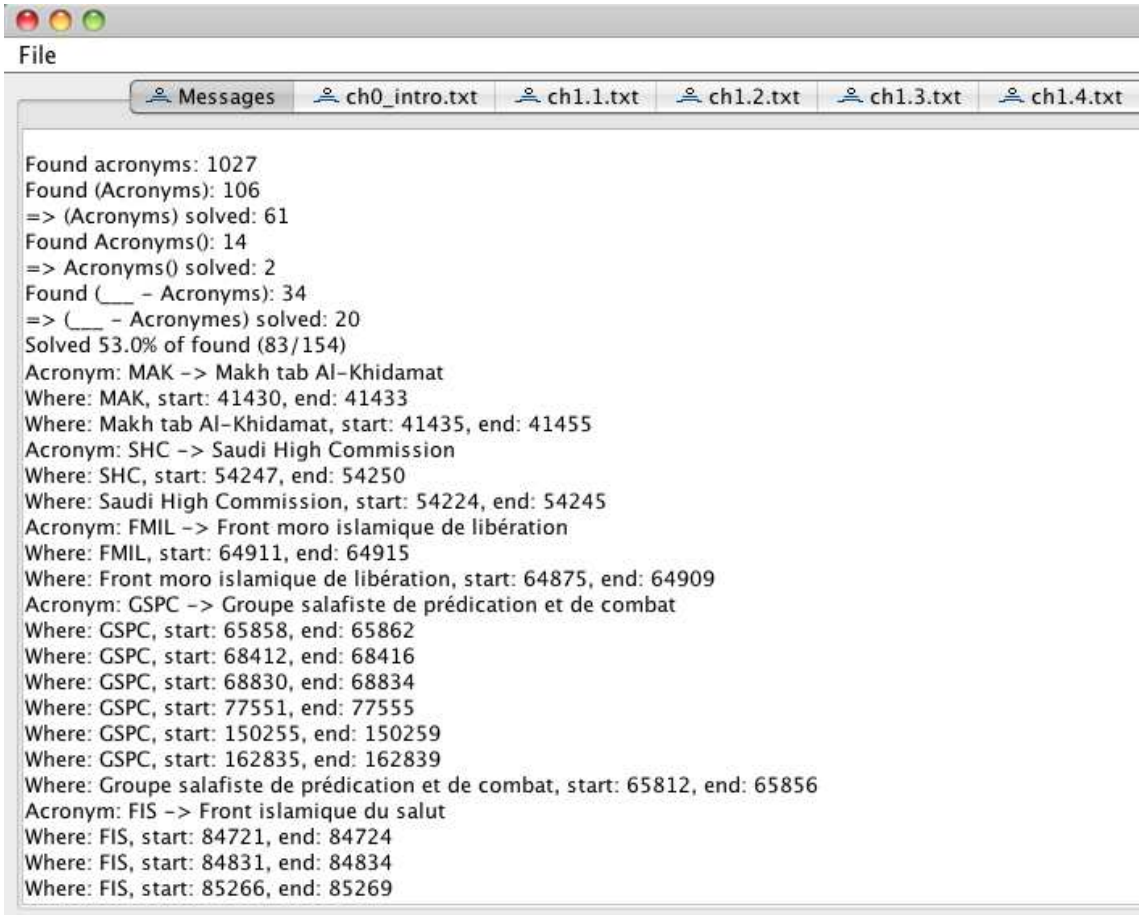


FIGURE 7.5 – Statistiques de la résolution d'acronymes

7.2.4 Extraction de relations

La dernière étape de reconstitution d'un réseau est l'établissement de liens entre ses acteurs identifiés. Parce que les liens recherchés ont un sens, il se peut qu'ils apportent des informations additionnelles au système, soit qu'ils identifient, par exemple, l'unicité entre deux acteurs encore considérés comme distincts, soit que la sémantique associée force le typage de l'un d'entre eux, à l'encontre de ce qui était établi. Nous ajoutons, donc, aux extracteurs décrits ci-dessus des règles d'enrichissement des données, permettant à la fois de lever certaines ambiguïtés et de lier les éléments découverts.

Objectifs

Une fois que nous disposons d'entités extraites, nous cherchons à les associer entre elles, afin d'établir le réseau les reliant. Pour parvenir à ce but, nous procédons à la recherche de relations entre objets reconnus.

De plus, parce que les méthodes présentées ici ne sont pas infaillibles, il arrive que l'iden-

tification d'éléments ne corresponde pas à l'a priori associé à l'algorithme. Nous avons, par exemple, évoqué la confusion entre pays et partie de noms de personnes lors de la découverte de 'Dominique'. Certaines des relations que nous cherchons servent également à valider ou préciser le typage associé aux termes ainsi dégagés.

Mise en œuvre

Nous nous intéressons à trois types de relations : les relations dites professionnelles qui lient les personnes entre elles, si elles collaborent, ou à des organisations auxquelles elle adhèrent ; les relations familiales, dont l'exploitation nécessite une description du réseau qu'elles constituent ; les relations de proximité, qui dérivent de la structure du corpus, dans lequel des acteurs évoqués concomitamment ont une forte probabilité d'être en relation. À ce découpage des relations, nous en ajoutons un autre entre relations explicites et relations implicites. Les premières sont identifiées dans le texte, à l'aide de mots-clefs, alors que les secondes sont déduites, à partir d'extraits. Parmi celles-ci, par exemple, se trouvent les relations familiales. Découvrir que 'Muhammad bin Oud Ben Laden' est le père d'Ussama implique la relation inverse, 'Ussama Ben Laden' est le fils de 'Muhammad bin Oud Ben Laden'.

Parmi les relations entre personnes que nous recherchons, celles qui alimentent directement la reconstruction du réseau portent sur les collaborations professionnelles. Nous extrayons, donc, des relations rapprochant une personne et une organisation, les organisations entre elles ou liant des individus. Des exemples de chacun de ces types de relations sont proposés dans le tableau 7.2. Une forme générale des mots-clefs signifiants est donnée dans la colonne 'Patron' et les types des entités liées sont fournis dans les deux colonnes suivantes.

Patron	Type entité 1	Type entité 2
`sous-groupe de`	*'Organisation'*	*'Organisation'*
`chef de`	*'Person'*	*'Organisation'*
`membre de`	*'Person'*	*'Organisation'*
`bras droit de`	*'Person'*	*'Person'*
`alias`	*'Person'*	*'Person'*

TABLE 7.2 – Quelques mots-clés de relations favorisant le typage des entités extraites

Parmi celles-ci, la relation 'alias' joue un rôle déterminant. En effet, bien que le système stocke les noms alternatifs de pays en tant qu'alias, dans la plupart des cas un alias

décrit une personne. Au-delà de la confirmation du typage de ses objets, que les relations précédentes fournissent également, cette relation permet d'alimenter la connaissance conséquente, en unifiant les informations atomiques de plusieurs labels.

Les relations familiales permettent de supposer que leurs objets sont des personnes, nous employons des règles équivalentes à celles présentées à la figure 7.6, afin de les repérer. Lors de la découverte, par exemple, de 'Mohammad Jamal Khalifa, beau-frère de Ben Laden', les informations atomiques ('Mohammad Jamal Khalifa','Person') et ('Ben Laden','Person') sont inférées et seront comparées aux informations homologues identifiées par ailleurs. En outre, un lien est créé entre elles, après la fusion et la création des connaissances décrivant les personnes. Si l'on découvre aussi que Ben Laden n'a qu'une sœur, on inférera, de plus, que Khalifa est son époux.

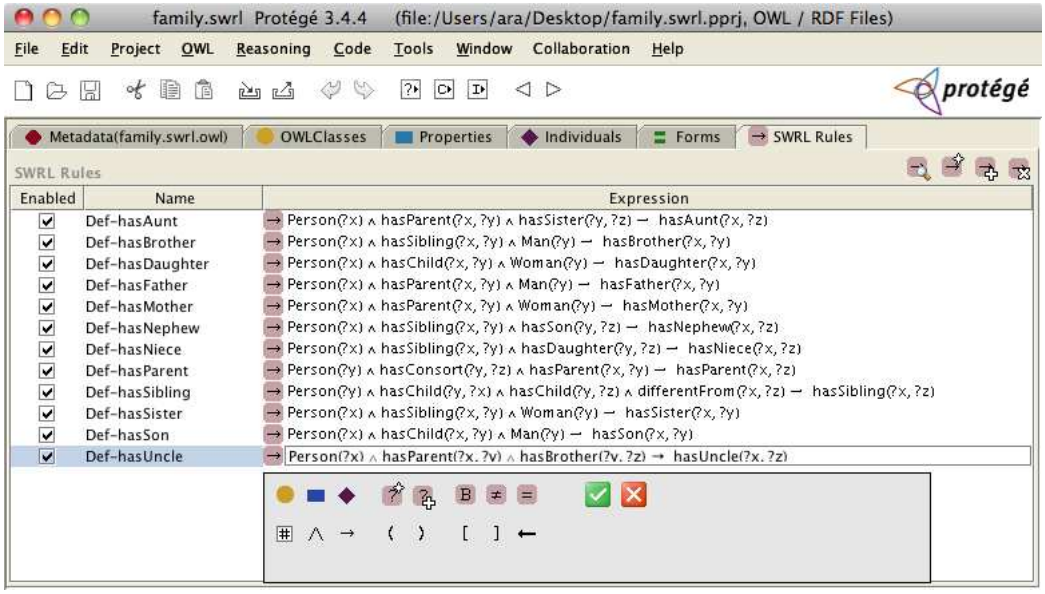


FIGURE 7.6 – Règles de déduction de relations familiales, enrichissant la structure initiale

Enfin, et parce que les liens de parenté sont rares et qu'il n'est pas toujours dit à quel groupe appartient un individu ou qui sont ses collègues, l'établissement de liens entre eux est parfois difficile. C'est pourquoi nous introduisons une règle étudiant les co-occurrences répétées de noms. Celle-ci établit une relation de proximité entre personnes, entre organisations et entre personnes et organisations citées dans les mêmes textes. Nous étendons sa portée au-delà des phrases, en modérant l'importance du lien en fonction de la distance et en limitant la distance maximale au paragraphe.

Résultats

Une fois encore, nous évaluons la performance de nos propositions en vérifiant la cohérence

de leur production. À la reconnaissance d'un des patrons recherchés, la validation de son résultat dépend de l'identification correcte de ses objets et de la présence effective du lien entre eux.

Les différentes classes de règles évoquées ci-dessus ne produisent pas des résultats de qualité égale. L'étude de co-occurences, par exemple, est activée plus souvent que les autres règles, pas toujours à bon escient, bien que son intégration découle de la structure et du propos du corpus. Sa précision est, comme nous le disions, pondérée en fonction de la distance entre les informations, plafonnant à 60% lorsque les noms sont dans la même phrase.

Les autres règles ressemblent plus aux modules que nous utilisons par ailleurs. Déclenchées avec discernement, les entités qu'elles renvoient sont, généralement, conformes aux attentes. Les règles de détection de relations familiales ont, par exemple, une précision de 90%.

7.2.5 Module de fusion d'informations

Une fois les informations atomiques collectées, nous souhaitons en extraire des connaissances. Dans ce but, nous procédons au rapprochement d'informations traitant d'un même sujet afin d'en obtenir une description plus informative.

Objectifs

La fusion d'informations, telle que nous l'entendons, consiste en l'extraction d'un maximum de détails par croisement, recoupement et complémentation. La découverte, par exemple, du nom 'Ussama ben Laden', puis celle de 'ben Laden', identifié en tant que personne, permet de décomposer le nom complet entre prénom et nom de famille. Ces nouvelles informations permettent, à leur tour, d'identifier des relations entre acteurs. La structure attendue des informations enrichies guide, donc, la fusion des informations atomiques, en vue de l'établissement de connaissances riches. Les relations, quant à elles, peuvent être vues comme des attributs communs entre les éléments qu'elle relient.

Mise en œuvre

Afin d'établir que deux informations atomiques traitent du même sujet, nous employons des heuristiques de fusion. L'identification de toutes les dénominations d'une personne permet de rattacher toute information complémentaire à la bonne connaissance. Parmi ces dénominations on s'intéresse principalement au nom, qu'il apparaisse sous sa forme complète ou par parties discriminantes, ainsi qu'aux alias. Les relations établies entre une personne identifiée et une organisation ou entre deux personnes viennent enrichir le contenu

informationnel de chacune.

Résultats
L'évaluation des règles et algorithmes a été effectuée manuellement en vérifiant que les informations liées entre elles décrivaient bien une même connaissance. Afin de favoriser l'émergence d'informations riches à partir de pépites, les quelques règles proposées dont la précision était trop basse ont été supprimées, même lorsque cette perte de précision favorisait le rappel. Nous évaluons la précision globale du module de fusion à 80% en vérifiant, encore une fois, à la main la cohérence des connaissances produites.

7.3 Implémentation de la cotation

Les modules d'extraction ainsi définis nous permettent de disposer d'informations atomiques, les chaînes de caractères extraites du texte, qu'ils enrichissent en les munissant d'un type ou en établissant des relations entre elles. La cotation intervient lors de cet ajout de sémantique et doit mesurer la confiance faite à la connaissance qui provient de cette qualification des extraits et des liens qui les unissent.
Comme exposé au chapitre 6, la cotation de connaissances comporte deux étapes, que nous détaillons successivement ci-après. Pour chacune, nous présentons les modules qui implémentent les quatre étapes du procédé général de cotation appliquées à ce cas particulier. Nous proposons également un aperçu de l'interface de sélection de stratégie, choix central dans l'orientation du processus de cotation.

7.3.1 Cotation de l'extraction d'informations

La cotation de l'extraction s'applique aux informations atomiques. Comme nous l'avons vu, au chapitre précédent et dans le début du présent, la plausibilité ne concerne pas la cotation d'informations atomiques. Cette section détaille, donc, la mise en pratique des trois dimensions restantes.

Fiabilité
Comme détaillé au chapitre précédent, la source d'une information atomique est constituée à la fois de son auteur et de l'algorithme l'ayant extraite. La confiance en une information démarre avec la foi prêtée à son locuteur et ne dépend, ici, que de l'auteur. Parce que cette dimension est subjective, nous en déléguons l'estimation à l'utilisateur, grâce à l'interface présentée à la figure 7.7. Cette fenêtre modale lui est présentée, lors de l'ajout d'un document au corpus traité. La qualification demandée concerne le document entier. Si

celui-ci comporte plusieurs auteurs, l'utilisateur fournit son agrégation subjective de leurs fiabilités. Dans l'interface présentée, on reconnaît la version anglophone des niveaux de fiabilité du tableau 3.2, page 51. La correspondance avec $\mathcal{L}_5^e$ est telle que le maximum 1 correspond à τ_4, le minimum mesurable 5 à τ_0 et l'inestimable 6 à $\tau_?$.

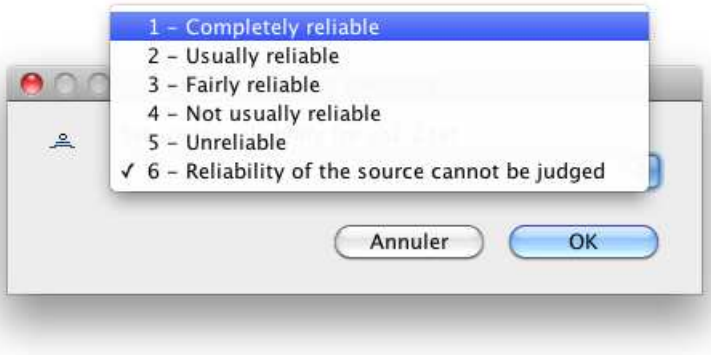


FIGURE 7.7 – Sélection manuelle de la fiabilité de la source à l'ouverture d'un document

Compétence

La compétence de la source sur le domaine de l'information est, en théorie, une dimension objective. Sa qualification à parler d'un sujet ne dépend ni des a priori de l'auditeur, ni de ses affinités avec la source. Cependant, l'estimation automatique de la compétence d'une personne réelle n'est possible qu'en étudiant son influence dans une communauté dédiée au sujet. Nous concentrons nos efforts sur l'évaluation de la confiance pour les informations disponibles dans les textes que nous étudions. La source d'une telle information est, donc, à la fois l'auteur du texte considéré et le processus nous donnant accès à son contenu. Si nous ne sommes pas en mesure d'estimer la compétence de la source, nous savons, en revanche, évaluer les domaines de compétence des algorithmes employés pour extraire l'information, ainsi que leurs performances. La source, envisagée comme la combinaison d'un auteur et d'un système d'extraction d'informations, a donc bien des domaines de compétence dont nous pouvons tenir compte dans l'évaluation de la confiance.

La section précédente présentait les modules d'extraction dont nous disposons et leurs domaines de compétence. Nous y proposions également la méthode d'évaluation de leurs performances. Dès lors, l'évaluation de la compétence du système d'extraction d'information formé de ces différents modules dépend des mesures statistiques des algorithmes proposés. À la recherche de pépites de connaissance, nous choisissons de retenir la mesure qualifiant la qualité des pépites extraites, plutôt que la proportion d'informations découvertes. Dans la majorité des cas, notre estimation de la compétence de la source est donc

fondée sur la précision de l'algorithme, rapportée à son domaine de compétence.
Pour établir la correspondance entre la valeur continue de la précision et le niveau d'activation discret de la compétence, nous partitionnons son intervalle de définition en cinq parties que nous faisons correspondre aux degrés de $\mathcal{L}_5$. La valeur indéterminée, $\tau_?$, n'est pas incluse dans la correspondance, puisqu'elle correspond à une mesure impossible, c'est-à-dire une compétence inconnue. Dans notre implémentation, le choix de la répartition des degrés de compétence, de la division en parts égales de [0,1] à la restriction aux valeurs acceptables, est déterminé pour toutes les évaluations de compétence.
Les modules d'extraction d'informations présentés plus haut offrent donc au système une compétence, répartie par sujet de la manière suivante :

- Extraction d'informations géographiques : τ_4, soit 'expert'
- Extraction de noms propres : τ_3, soit 'compétent'
- Extraction et déchiffrement d'acronymes : τ_2, soit 'partiellement compétent'
- Reconnaissance d'organisations : τ_3, soit 'compétent'
- Extraction de relations : τ_4, soit 'expert'

Notons que ce choix de transposition de la précision des méthodes vers la compétence est le nôtre, représentant une certaine confiance dans le système. Selon le cadre d'application, on pourra choisir une autre partition de l'espace de mesure de la compétence.

Crédibilité
L'estimation de la crédibilité implique une comparaison entre informations homologues. Pour savoir si deux informations sont homologues, nous devons comparer les chaînes de caractères. La mesure du recoupement dépend ensuite de la similarité entre les labels. La nuance entre confirmation et infirmation est, elle, définie grâce à la correspondance de types entre informations.
Ainsi, deux extractions différentes de la personne nommée 'Ussama ben Laden', c'est-à-dire deux couples ('Ussama ben Laden', *'Person'*), se confirment totalement et la crédibilité est évaluée τ_4 (cf. tableau 3.5, p. 57). Chacune d'entre elle, en revanche, contredit partiellement 'Ben Laden Corporation' ('Ben Laden Corporation', *'Organisation'*), identifiée comme une organisation grâce au mot-clé 'Corporation', et la crédibilité est de τ_1 pour toutes les trois. Nous considérons que les spécialisations de types, comme 'Terrorist_Group' sous-concept de 'Organisation' dans l'ontologie de sortie, participent à la confirmation.

7.3.2 Cotation de la fusion d'informations

La seconde étape de cotation du système qui nous occupe s'applique aux connaissances, résultats de la fusion d'informations atomiques. Comme dans la section précédente, nous

présentons maintenant l'implémentation des deux dimensions de la cotation de connaissances, telle que décrite à la section 6.3.2, page 112, la compétence et la plausibilité.

Compétence

De la même manière que pour les informations atomiques, la compétence de la source, pour les informations enrichies, évalue la performance des algorithmes les ayant produites. Ici, chaque règle permettant de reconstruire la connaissance pourrait être estimée indépendamment des autres, selon son efficacité. Cependant, comme nous l'indiquions à la section 7.2.5, nous avons favorisé, dans notre démonstrateur, les règles de fusion de performances comparables. Nous retenons, donc, la précision du module dans son ensemble, dont le résultat est cohérent dans 80% des cas, donnant une compétence de τ_3, soit un module de fusion 'compétent'.

Plausibilité

La plausibilité reflète la compatibilité de l'information enrichie avec la connaissance du monde du système. Cette notion est interprétable de différentes manières, selon que l'on se place dans un monde clos totalement décrit ou dans un cadre plus réaliste où la connaissance du système est limitée. Quoi qu'il en soit, l'évaluation automatique de la plausibilité d'une connaissance ne révèle pas son potentiel de réalisation mais plutôt dans quelle mesure elle satisfait au modèle du monde au travers duquel le système peut l'estimer. Pour qui pense que la terre n'est pas plate, le bord du monde est un lieu peu plausible. Nous avons choisi deux mesures de plausibilité, l'une suivant un a priori sur les connaissances que nous cherchons, la seconde introduisant le principe de réalité.

Nous savons que le corpus que nous étudions décrit, entre autres, les acteurs et les relations d'un réseau. Qui plus est, notre système est construit spécifiquement pour la reconnaissance et l'extraction de ces informations. Nous supposons donc que si une connaissance est centrale au texte, sa description sera détaillée et ses relations importantes et nombreuses. À partir de cette hypothèse, nous déterminons que plus une information enrichie est complète, plus sa plausibilité est élevée. Dans le même esprit, plus une connaissance est liée à d'autres, plus il est plausible que son rôle dans le corpus soit important.

Le deuxième axe d'évaluation de la plausibilité que nous nous donnons concerne la vision du monde fournie au système. En effet, le calcul de plausibilité nécessite la définition de connaissances extérieures sur certains faits inenvisageables. Après que la fusion d'information atomiques a produit une connaissance, leur compatibilité détermine la mesure avec laquelle celle-ci est plausible. Le cas de contradictions entre informations atomiques étant du ressort de l'étape de crédibilité, la compatibilité à laquelle nous nous intéressons ici est d'un niveau sémantique plus élevé. Nous munissons donc le système d'un ensemble

de règles de maintien de la cohérence afin d'estimer la plausibilité.

Nous allons, par exemple, chercher à maintenir la compatibilité entre les attributs des individus reconnus et les inférences réalisées. Ainsi, parmi les attributs des personnes que nous cherchons figure le genre. Si, après la fusion, la connaissance 'Ussama Ben Laden', personne de genre masculin, se trouve être également mère de deux enfants, la plausibilité d'une telle assertion sera minimale et évaluée τ_0.
De même, les organisations, telles que nous les définissons, peuvent être en relation avec d'autres ayant des intérêts communs ou opposés. On se munit donc d'une règle mettant en doute le lien de dépendance ('appartient à' ou 'travaille pour') entre une personne et deux organisations d'intérêts opposés. D'après cette règle, le fait d'avoir un individu travaillant à la fois pour 'Al Qaeda' et la 'CIA' est, à son tour, 'peu possible' et la plausibilité est de τ_1.
Notons que, bien que nous n'ayons pas rencontré le cas, on peut envisager des règles plus complexes. On peut, par exemple, supposer selon la cote des connaissances, que la description d'un tel individu est parfaitement plausible, dans le cas d'un agent infiltré. L'assouplissement de la règle de vérification de la plausibilité implique, donc, parfois des connaissances additionnelles, déductibles pour des informations sûres.

7.3.3 Éditeur de stratégies

L'autre point sur lequel nous avons insisté est que notre modèle s'adapte à l'utilisateur en fonction de son point de vue sur la confiance. Nous avons, pour ce faire, introduit les notions de stratégies de cotation et de postures de crédulité (cf. définition 7, page 68). Nous proposons, donc, un éditeur de stratégies, illustré à la figure 7.8 : celui-ci permet à l'utilisateur de spécifier les transitions d'évolution de la cotation à l'intégration de dimensions.
Les degrés sont représentés par des chiffres, les éléments de $\mathcal{L}_M^e$ restant mal appréhendables pour un utilisateur non expert en logique multivalente, c'est-à-dire que la gradualité est représentée comme à la figure 4.1, page 64. L'interface permet d'ajouter, supprimer ou modifier les flèches. Si l'utilisateur place moins de flèches qu'il n'y a de niveaux d'activation pour la dimension considérée, le système attribue la transition la plus défavorable à τ_0, la suivante, si elle existe, à τ_1 et suppose que les niveaux supérieurs à la dernière flèche fournie sont confondus avec elle. Enfin, par défaut, une activation de $\tau_?$ est envisagée comme une flèche horizontale, c'est-à-dire qui ne modifie pas la cote courante. L'interface vérifie que les contraintes d'influence discutées au chapitre 4 sont bien respectées afin de garantir que le module de cotation fonctionne comme attendu.

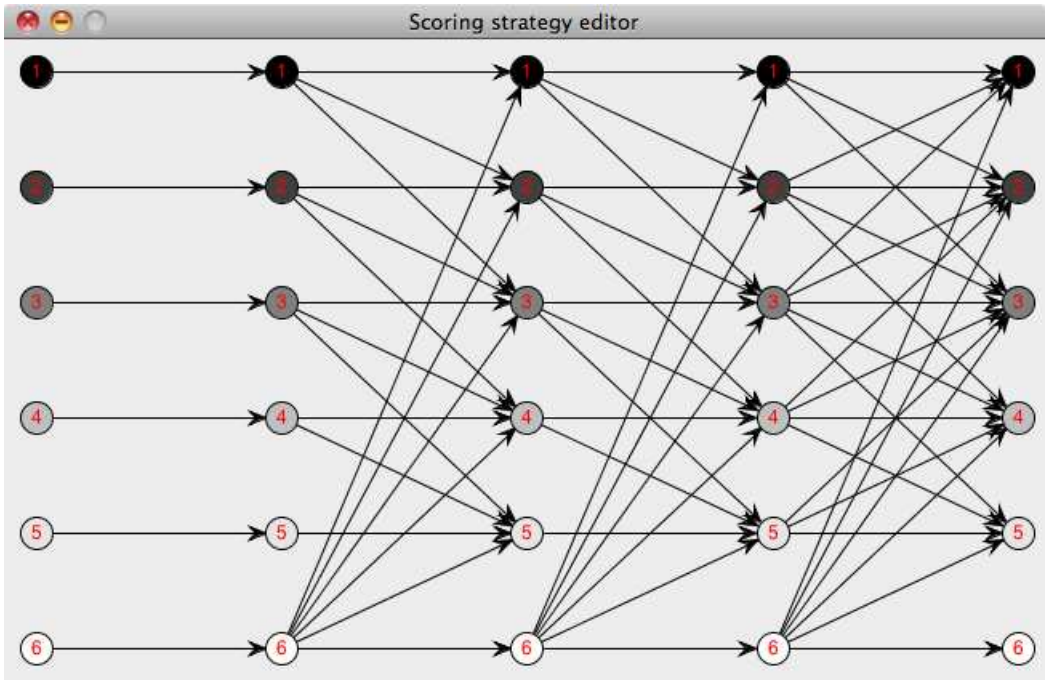


FIGURE 7.8 – Éditeur de stratégies – ici, la stratégie par défaut

Parce que la manipulation de l'interface de modification de stratégies n'est pas aisée, notre outil propose, par défaut, à l'utilisateur les trois stratégies de référence, présentées à la section 4.4, page 67. Ces stratégies sont conçues pour représenter des postures crédule, méfiante et la stratégie par défaut. La figure 7.8 montre l'éditeur de stratégies dans le cas de cette stratégie par défaut.

7.4 Exemple illustratif

Afin de donner un aperçu de l'implémentation de la cotation dans un système d'extraction de connaissances, nous présentons l'ensemble du processus dans un exemple. Ne pouvant représenter l'ensemble des modules ni du réseau, nous choisissons d'illustrer un détail représentatif des capacités du démonstrateur. Nous nous intéressons, plus particulièrement, à la portion de texte suivante, riche en pépites :

'Rifaï Ahmed Taha, alias Abou Yasser, égyptien, chef du Gama'a al-Islamiya'

En exploitant ce morceau de phrase, nous proposons de parcourir l'extraction d'informations atomiques, leur fusion en connaissance et d'aller, enfin, jusqu'à la construction d'une partie du réseau. L'intérêt illustratif de la phrase choisie est non seulement son abondance

d'informations atomiques mais également l'assurance qu'elles traitent du même sujet. Ce deuxième avantage est exploité à la phase de fusion, pour l'extraction de connaissances.

Nous considérons, dans la suite de cet exemple, un utilisateur ayant sélectionné la stratégie S_1, dite par défaut, des tableaux 4.1, 4.2 et 4.4, pages 69 à 71, dont la transposition au formalisme multivalent de la cotation est fournie à l'annexe D, page 177.

7.4.1 Cotation de l'extraction d'informations

La première étape du traitement extrait de cette courte portion de texte les six informations atomiques présentées au tableau 7.3.

Id.	Information atomique	$C(s)$
i_1	('Rifaï Ahmed Taha','Person')	τ_3
i_2	('Abou Yasser','Person')	τ_3
i_3	('Rifaï Ahmed Taha','Abou Yasser','alias')	τ_4
i_4	('Abou Yasser','Égypte','is_from')	τ_4
i_5	('Abou Yasser','Gama'a al-Islamiya','is_head_of')	τ_3
i_6	('Gama'a al-Islamiya','Organisation')	τ_3

TABLE 7.3 – Informations atomiques extraites du texte avec la compétence de la source, $C(s)$ précisée dans la dernière colonne

Les informations i_1 et i_2 sont extraites par l'algorithme d'extraction de noms propres et, donc, reconnues en tant que personnes. L'information i_3, établissant que les deux informations précédentes décrivent la même personne, est identifiée par le module d'extraction de relations. La relation de l'information i_4, pour sa part, provient de l'extraction d'informations géographiques. L'extraction de relations permet également d'établir i_5, information de laquelle on déduit le typage proposé de i_6.

Fiabilité

Puisque toutes ces informations atomiques proviennent de la même phrase et, a fortiori, du même document dont l'auteur est 'expert', la fiabilité de la source est égale pour toutes, évaluée à τ_4. La cote de toutes ces informations est, donc, initialisée à τ_4.

Compétence

La dernière colonne du tableau 7.3 fournit, pour chaque information, la compétence de sa

source, c'est-à-dire l'évaluation de la performance de l'algorithme l'ayant produite.
La projection de la compétence sur la cotation, calculée grâce à la fonction F (cf. définition 9, p. 85), se fait comme indiqué, pour la stratégie par défaut, dans les tableaux 4.3 et D.1, pages 71 et 177 respectivement.
On constate, par exemple, que la cote de i_1 évolue vers $F(\tau_4, \tau_3) = \tau_3$, soit une légère baisse. De même, les informations i_2, i_5 et i_6 provenant de sources de compétences égales, ont également pour cote τ_3. i_3 et i_4, quant à elles, ont une cote de $F(\tau_4, \tau_4) = \tau_4$, après l'intégration de leurs algorithmes experts.

Crédibilité
Dans le cas général, les informations homologues se récoltent potentiellement loin des informations qu'elles confirment ou infirment. Cependant, en ce qui concerne i_1 que nous détaillons ici, la relation décrite par l'information atomique i_3 fournit deux informations supplémentaires, à savoir le typage de ses deux membres en tant que personnes. Cette information cotée τ_4 vient donc confirmer parfaitement l'information i_1. L'application de G donne alors $G(\tau_3, \tau_4, \tau_4) = \tilde{F}_c(\tau_3, g(\tau_4, \tau_4)) = \tau_4$, toujours selon la stratégie S_1. De la même manière, l'information i_2 est confirmée par i_3.
De manière analogue, la relation 'chef de' favorise le typage de ses objets, le premier comme personne, le second en tant qu'organisation.

7.4.2 Cotation de la fusion d'informations

Nous profitons de cet exemple pour insister à nouveau sur la différence entre informations homologues et informations relatives à la même connaissance. Le meilleur exemple de cette différence vient de la relation décrite par l'information i_3. Cette relation porte sur deux informations atomiques distinctes, i_1 et i_2. Ces deux informations décrivent en réalité la même personne mais, à la granularité des informations atomiques, ne sont pas homologues puisqu'elles ne portent pas sur la même chaîne de caractères. Elles sont, en revanche, combinées à l'étape de fusion pour rendre compte de cette unicité d'objet. Cette clarification étant faite, intéressons-nous à la construction de la connaissance qui sourd du tableau 7.3, avant d'en considérer la cotation.

L'extraction de relations permet, donc, au système de relier les informations atomiques. En s'éloignant du niveau des labels typés, la relation 'alias' unifie les deux personnes reconnues en une seule. La relation 'chef de', intervient non seulement sur le typage des informations atomiques comme détaillé précédemment, mais permet, de plus, d'établir un lien entre deux connaissances, une organisation et son dirigeant. Cette relation ne porte plus que sur une personne, la fusion entre les deux alias étant acquise. Enfin, le gentilé

enrichit la connaissance établie en fournissant sa nationalité et son sexe. La figure 7.9 représente cette connaissance telle que décrite dans l'ontologie.



FIGURE 7.9 – Propriétés de la connaissance sur Rifaï Ahmed Taha dans l'ontologie

Les connaissances et relation décrites ci-dessus peuvent être représentées comme un réseau élémentaire, illustré par la figure 7.10. Chacun de ces éléments, connaissances et relation, est soumis au processus de cotation que nous détaillons maintenant.

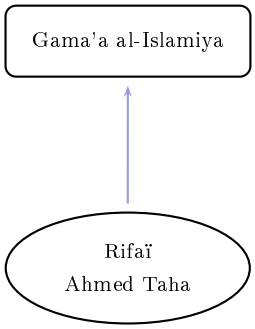


FIGURE 7.10 – Mise en relation de connaissances : Rifaï Ahmed Taha et la Gama'a al-Islamiya

Initialisation

L'initialisation du processus de cotation pour une information enrichie provient de l'agrégation des cotes des informations atomiques la composant. Ceci offre un résumé de sa

production, que nous exprimons par une médiane. La connaissance représentée à la figure 7.9, provenant des informations i_1 à i_5 du tableau 7.3, est donc initialisée à τ_3. La connaissance additionnelle sur l'organisation n'étant issue que d'une information atomique, son initialisation est égale à la cote de i_6, c'est-à-dire τ_3.

Compétence

Comme indiqué à la section 7.3.2, le degré d'activation de la compétence est l'évaluation de la performance du module de fusion, soit τ_3. Sa projection sur les deux connaissances présentées ci-dessus est donc : $F(\tau_3, \tau_3) = \tau_2$.

Plausibilité

Aucune des connaissances ainsi décrites ne contredit la vision du monde dont le système est muni. L'évaluation de la plausibilité repose, donc, sur la complétude des connaissances. La personne, suffisamment détaillée, bénéficie d'une activation maximale de la plausibilité, soit τ_4. Sa cote évolue donc de la manière suivante : $F(\tau_2, \tau_4) = \tau_2$
Si l'organisation est moins bien décrite, sa plausibilité est maintenue à un niveau acceptable par sa liaision à une autre connaissance, elle est considérée comme réaliste, de plausibilité τ_3. En conséquence, sa cote baisse à nouveau, $F(\tau_2, \tau_3) = \tau_1$.

Contextualisation dans le réseau

L'orpaillage permet, donc, à partir d'une phrase riche en pépites d'extraire et de coter des connaissances. Nous avons limité l'illustration précédente à une portion de texte afin de mettre en avant la collecte, la fusion et la cotation d'informations. Cependant, l'extraction du réseau dépend de l'agglomération de connaissances découvertes. Nous proposons, en conclusion de cet exemple, un aperçu de ces rapprochements, grâce à l'inférence de relations.

Un processus similaire permet d'établir un réseau entre Ussama Ben Laden, Ayman al-Zawahiri (son médecin et bras droit) et Al-Qaeda, représenté par les flèches pleines de la partie gauche de la figure 7.11. Nous proposons de relier ces deux réseaux élémentaires grâce à la proximité dans le texte de leurs composants, puis par inférence. Le résultat de ce processus complet d'extraction d'informations atomiques, de leur fusion puis de la cotation des connaissances ainsi construites est représenté à la figure 7.11. Les couleurs des nœuds y représentent la cote des connaissances. Les arcs sont également colorés selon la confiance qui leur est faite et ceux ne dérivant pas de l'extraction par mots-clefs sont pointillés.

Bien qu'à la lecture du texte, les liens entre Ussama Ben Laden et Rifaï Ahmed Taha soient apparents, leur découverte automatique est moins aisée. Toutefois, la proximité dans les textes de mentions simultanées de Ayman al-Zawahiri et de Rifaï Ahmed Taha, permet de soupçonner une relation entre eux. Cette relation n'apparait qu'une seule fois mais est estimée 'envisageable' et cotée τ_2, parce que les deux noms sont découverts côte à côte.

De la même façon, on établit un lien par co-occurrence entre Ben Laden et 'le Gama'a al islamiya (ou Jamaa Islamiya) égyptien', coté τ_3, parce que plus abondamment citée. Enfin, combinées avec le lien existant entre Ussama Ben Laden et Al-Qaeda, ces deux relations conduisent à l'inférence d'un rapprochement entre les organisations, la Gama'a al-Islamiya et Al-Qaeda. Ce lien, provenant de la combinaison de relations de confiance faible, est estimé comme une connaissance et initialisé de la même manière.

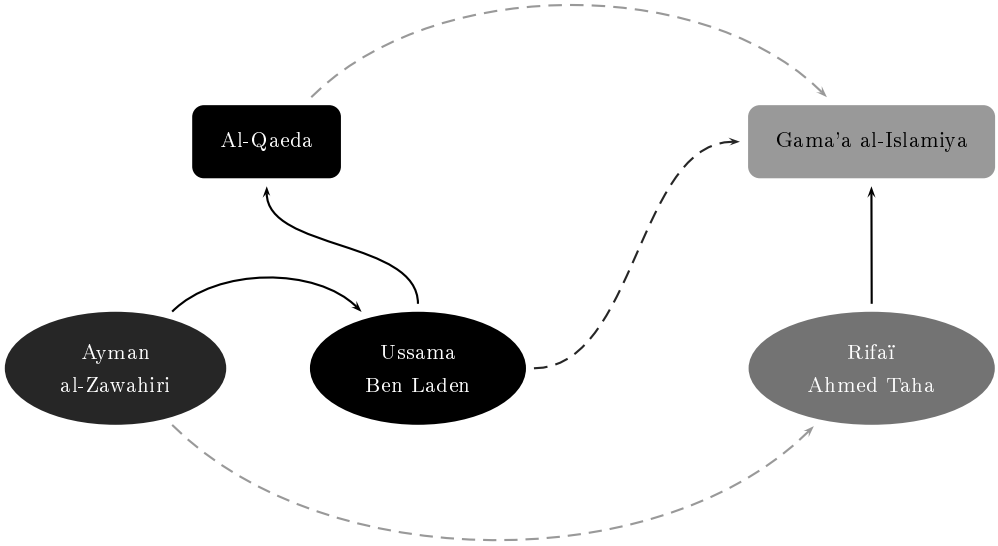


FIGURE 7.11 – Un exemple de visualisation d'une partie du réseau extrait et coté

7.5 Problématique de l'évaluation

Comme pour tout système de traitement automatique, se pose la question de l'évaluation des performances du modèle de cotation proposé ou, à défaut, de son implémentation dans le démonstrateur. Étant donné que notre travail ne traite pas, à proprement parler, d'extraction de connaissances, la comparaison des résultats de nos méthodes, sur des benchmarks établis, à d'autres solutions ne nous permettrait pas d'évaluer nos propositions, n'évaluant le démonstrateur que sur le cadre d'application.

Parce que la cotation représente la confiance faite à une information, elle est nécessairement subjective. Son évaluation exige d'interroger l'utilisateur sur sa satisfaction quant au résultat produit. Ceci soulève le problème de l'interrogation d'expert. Établir un scénario dans lequel un utilisateur aurait en main à la fois toutes les informations et connaissances pour juger le système et, également, un avis sur la cotation est une tâche complexe et d'autant plus coûteuse qu'elle nécessite la participation de plusieurs experts.

En outre, lors de nos entretiens avec des utilisateurs, pendant la récolte de besoins, les demandes exprimées résidaient moins dans la comparaison de la cote retournée avec les résultats attendus que dans l'augmentation de sa lisibilité. Lever l'opacité du modèle existant repose essentiellement sur trois critères, la souplesse, la facilité d'interprétation et la possibilité de justifier les résultats fournis.
La formalisation que nous proposons ici répond au premier de ces objectifs en permettant la personnalisation du modèle à l'utilisateur, par le biais des postures de crédulité, implémentées par les stratégies.
La clarification des dimensions participant à l'évaluation de la confiance, dans ce qu'elles qualifient et ce qu'elles représentent, est notre première contribution à la résolution de la deuxième exigence. La modélisation multivalente offre un cadre formel au processus et son versant sémantique parachève nos propositions pour rendre la cotation interprétable.
Enfin, la traçabilité du processus de combinaison des dimensions retenues, de la chaîne de cotation, offre la possibilité de justification des résultats fournis. Les projections successives des dimensions sur la cote courante permettent, en effet, d'isoler celle par où une éventuelle divergence arriverait.

7.6 Conclusion

Dans ce chapitre, nous avons exposé la mise en pratique du modèle de cotation proposé en nous intéressant particulièrement à la transposition de la chaîne de cotation à l'extraction de connaissances à partir de textes, ici à la reconstitution d'un réseau. La reconstitution d'un réseau impliquant l'identification des acteurs et des liens le composant, nous avons proposé quatre modules permettant la récolte de ces informations et détaillé, pour chacun, objectifs, mise en œuvre et résultats. Nous avons présenté le module de fusion permettant le passage d'informations atomiques aux connaissances qu'elles composent.
Suite à cette présentation technique, nous avons détaillé l'intégration de la cotation au système, dans l'étape d'extraction d'informations, comme dans celle de leur fusion. Nous avons rappelé les dimensions qui composent chacune et en avons donné les méthodes d'évaluation ainsi que leurs valeurs.
Par la suite, à partir d'un morceau de texte signifiant, nous avons illustré l'extraction d'informations et leur fusion en connaissances pour la reconstitution d'un réseau et sa cotation.
Enfin, nous avons conclu ce chapitre par une discussion sur l'évaluation de nos propositions pour la cotation. Nous avons rappelé que la cotation est un indicateur d'un jugement subjectif et évoqué les difficultés de comparaison et d'appréciation de ce type de mesure. Nous avons, néanmoins, montré comment nos propositions répondent aux attentes des uti-

lisateurs, en dotant la cotation des trois propriétés désirées que sont la souplesse, la facilité d'interprétation et la justification des résultats.

Conclusion générale et perspectives

La confiance en une information témoigne de la mesure à laquelle on la croit. Elle ne représente pas une évaluation de la réalité du fait mais la conviction de son évaluateur. La majorité des travaux sur la confiance la résume à une forme d'estimation de la qualité de ce qu'elle caractérise. Pour se conformer à l'esprit de ce qu'elle désigne, l'évaluation automatique de la confiance doit se fonder sur un modèle permettant d'exprimer intelligiblement quels facteurs participent à la foi en son objet et de quelle manière. C'est à cette problématique de modélisation de la cotation que nous nous sommes attaché dans nos travaux.

Contributions

Nous avons proposé de formuler la problématique de la cotation comme l'établissement d'un formalisme compréhensible de la constitution d'opinions. Nous avons traduit cette exigence de lisibilité comme imposant à un tel formalisme d'être souple, facilement interprétable et de permettre d'expliquer comment la confiance a été constituée. Pour répondre à cet objectif nous avons identifié deux axes complémentaires à la représentation de la confiance : la cote, c'est-à-dire la façon dont on l'énonce, et la cotation, c'est-à-dire les principes de sa construction.

Afin de favoriser son interprétabilité, nous avons sélectionné des dimensions indépendantes et non-redondantes, la fiabilité de la source, sa compétence, la plausibilité de l'information et sa crédibilité. Pour cela nous avons croisé deux axes de caractérisation des dimensions, leur objet et leur influence : nous avons étudié la manière dont ce que qualifie chaque dimension dépend du contenu de l'information ou de sa source, d'une part, et si la dimension prend en compte le contexte ou non. Les quatre dimensions proposées couvrent l'ensemble des combinaisons de ces deux axes. Pour assurer la lisibilité des mesures, nous les avons associées à un formalisme de représentation constitué d'une échelle discrète associée à des étiquettes linguistiques explicitant la gradualité mesurée.

À partir de ces dimensions de la cote, nous nous sommes penché sur leur combinaison, la cotation : nous avons proposé une formalisation de l'intégration des dimensions à l'évaluation de la confiance en établissant une architecture à la cotation. En exposant cette architecture, nous avons précisé l'influence de chaque dimension sur l'évolution de la cote en explicitant la monotonie de leur combinaison. Nous sommes également revenu sur l'ordre de prise en compte des dimensions que nous avons proposé afin de faciliter la compréhension du modèle.
De plus, nous avons proposé une formalisation de ces propositions dans le cadre de la logique multivalente que nous avons étendu afin de pouvoir représenter l'impossibilité de mesurer, évaluation jusqu'alors inexistante dans ce formalisme, et, par là, accroître encore l'expressivité du modèle. Nous avons ensuite défini des opérateurs de combinaison permettant de représenter formellement l'ensemble de nos propositions.
Les solutions proposées au problème de la cotation lui confèrent ainsi l'expressivité de la représentation de la cote et à la fois l'interprétabilité et la possibilité de justification propres à la cotation définie, tant dans les principes conceptuels que dans leur traduction formelle.

Pour ce qui concerne l'exigence de souplesse, nos propositions permettent de l'envisager sous deux angles, les postures de crédulité et la dynamique de la conviction. Les premières sont exprimées par les stratégies permettant d'adapter l'établissement de la confiance à différentes conceptions de la conviction. Celles-ci sont formalisées par la présence de paramètres dans les opérateurs de la logique multivalente définis pour la transposition théorique du processus de cotation.
À l'aide de ces opérateurs formels, nous avons défini une autre traduction de la variabilité cognitive de la confiance : les notions de crédibilité immédiate et cumulée qui offrent la possibilité de représenter l'évolution de la confiance dans le temps en tenant compte de l'ordre des variations entre confirmations et infirmations. La crédibilité cumulée munit, en effet, la cotation d'un effet mémoire la rendant sensible à ces dynamiques.

Enfin, nous avons réalisé une mise en œuvre pratique de nos propositions en étudiant la transposition de chacune au cadre de l'extraction de connaissances à partir de textes. Nous avons étudié en détails la cotation d'informations extraites et la cotation des connaissances résultant de leur fusion et proposé l'application d'une chaîne de cotation double. Pour la cotation de l'extraction d'informations et pour la cotation de la fusion, nous avons présenté et discuté l'expression des quatre dimensions de la cotation, la fiabilité de la source, sa compétence, la plausibilité de l'information et sa crédibilité.
Nous avons appliqué les principes exposés à la reconstitution d'un réseau social à partir d'un corpus de documents textuels et réalisé une implémentation d'un démonstrateur. Celui-ci

a nécessité le développement de modules d'extraction et de fusion d'informations, pour des informations géographiques, des noms propres, des acronymes et des relations. Nous concentrant sur la recherche de pépites significatives, nous avons proposé et mis en œuvre des solutions heuristiques efficaces. Partant des nœuds et arcs du graphe ainsi identifiés, la cotation permet alors de mesurer et de représenter la confiance faite aux acteurs et liens du réseau. Cette application permet d'illustrer en pratique l'ensemble des réponses fournies par le modèle théorique de cotation proposé aux attentes de la formalisation d'un processus subjectif de construction de la confiance.

Perspectives

Les perspectives ouvertes par ce travail sur la cotation sont variées et abondantes, tant sur les aspects appliqués que sur le modèle. Nous en proposons une sélection organisée autour de ses enjeux pratiques, des évolutions possibles des dimensions, l'étude interdisciplinaire du modèle et enfin l'exploitation des propositions théoriques.

La cotation en pratique

Nous avons proposé une application particulière du modèle de cotation qui nous a conduit à développer un système permettant la cotation de réseaux sociaux en sources ouvertes. Au-delà de la possibilité d'enrichir l'outil, par exemple en le munissant de connaissances plus fournies à l'aide d'une ontologie décrivant les villes, les régions ou les continents, deux points seraient particulièrement intéressants pour le faire fructifier : la présentation des résultats et le retour utilisateur.

Visualisation Une manière efficace d'accroître encore la facilité de prise en main de la cotation, notamment dans l'outil d'extraction de réseaux sociaux, réside sans doute dans la visualisation. En effet, nombre de travaux mettent en avant le gain en interprétation de données graphiques, comme ceux de Freeman (2000) qui montre de quelle manière le visuel facilite l'appréhension de réseaux sociaux. Il serait bénéfique d'étudier les modes de représentation visuelle de la confiance, problématique riche dans le cadre étudié.
La figuration de l'incertitude dans un graphe peut, par exemple, exploiter l'élasticité des liens que l'on pourrait utiliser pour représenter la confiance faite à une relation extraite, en figurant les relations sûres par la proximité de leurs acteurs. Ceci permettrait, en particulier, d'étudier la mise en évidence de cliques contraintes, groupes topologiquement fortement liés, mais dont certaines relations bénéficient d'une moindre confiance.

Un autre sujet d'étude dans les graphes consiste à identifier les nœuds séparateurs dont la suppression sépare le graphe en composantes disjointes. Dans le cadre d'un réseau social, ces nœuds ont un rôle central sur la diffusion d'informations. L'évaluation de ce rôle pourrait être enrichie par la prise en compte de la confiance faite aux relations, mise en avant par l'élasticité.
En outre, la cote des relations n'est pas la seule information à mettre en avant : les acteurs sont eux-mêmes munis d'un niveau de confiance. Cette multimodalité de la manifestation de la confiance est une autre direction à explorer pour la visualisation de réseaux sociaux cotés.
En dehors de notre cadre de recherche, s'ajoute à ces problématiques celle des algorithmes de représentation de graphes. Ceux-ci utilisent la topologie du graphe et l'élasticité autorisée pour organiser les nœuds de manière optimale. L'établissement d'algorithmes de placement de graphes à l'élasticité localement contrainte induisant d'éventuelles anomalies topologiques présente des défis tant mathématiques que pratiques.

Interprétabilité de l'outil Une telle interface de visualisation développerait la prise en main de l'outil, ouvrant la voie à l'étude de son adoption par les utilisateurs. Dans l'optique de valider l'implémentation du modèle, on pourrait établir des expériences de récolte de ressenti à l'aide de questionnaires afin d'affiner certains des choix que nous avons faits. Parmi ceux-ci, il serait intéressant de connaître la réception de l'évaluation dans $\mathcal{L}^e_5$, c'est-à-dire une échelle à six niveaux, dont le niveau inestimable $\tau_?$. Nous avons parlé des complications liées au choix du nombre de degrés intermédiaires à proposer entre l'activation nulle et totale d'une dimension. La nuance entre expressivité et complexité pourrait être étudiée à l'aide du système pour établir une vérité terrain.
Cette étude pourrait déboucher sur un second point à examiner concernant les échelles, plus précisément leur égalité. Nous avons proposé d'évaluer toutes les dimensions sur des interprétations différentes de la même échelle, $\mathcal{L}^e_5$ en l'occurrence. Le modèle étant suffisamment flexible pour permettre des combinaisons de $\mathcal{L}^e_M \times \mathcal{L}^e_N$, avec $M \neq N$, on pourrait dilater certaines échelles d'évaluation et en contracter d'autres. Il serait, par exemple, possible d'évaluer la crédibilité dans $\mathcal{L}^e_9$ afin d'étendre l'égalité des échelles aux deux signes de celle-ci. De la même manière, on pourrait représenter une partition amicale des sources du point de vue de l'évaluateur et mesurer la fiabilité de la source dans $\mathcal{L}^e_2$, où les sources amies sont considérées comme fiables et les sources ennemies comme peu fiables. L'utilisation de $\mathcal{L}^e_3$ permettrait de conserver la nuance entre sources neutres, ni amies, ni ennemies, et les sources inconnues, de fiabilité $\tau_?$.
Ces éventuelles évolutions des échelles d'évaluation posent également la question de la géné-

ralité de notre implémentation des stratégies et, par là, des postures de crédulité. En effet, quelques uns des choix d'implémentation que nous avons opérés sont sans doute réducteurs par rapport à la souplesse offerte par le modèle, en particulier le traitement des transitions confondues. Le changement de gradualité des échelles impliquerait une redéfinition des paramètres des opérateurs de combinaion κ_α^β. De façon générale, il faudrait s'assurer que chacun trouve à se satisfaire parmi les postures de crédulité ainsi représentées.

Enrichissement des dimensions de la cotation

Nous avons sélectionné les quatre dimensions pertinentes à l'évaluation de la confiance et avons pris soin de définir ce qu'elles décrivent, comme d'exposer comment nous évaluons les critères qu'elles couvrent. Toutefois, quelques évolutions de la fiabilité, la crédibilité et la plausibilité offrent des sujets d'étude intéressants, certains techniques et d'autres plus théoriques.

Évaluation de la source et sources successives L'évaluation de la fiabilité pourrait bénéficier d'une prise en compte de l'expérience, selon la boucle de rétroaction discutée au chapitre 4, faisant évoluer la fiabilité en fonction de la réalité observée. Dans ce but, on pourrait adapter certaines des techniques d'évaluation de la réputation présentées au chapitre 2. Utiliser ces méthodes nécessiterait quelques ajustements puisque, par exemple, celles qui fonctionnent à la Google établissent la réputation à partir de celles de référents, requiérant l'existence de relations entre les sources. Une telle automatisation permettrait, d'une part, d'effectuer une mise à jour de la fiabilité de sources connues, étendant la dynamique du processus de cotation. Elle pourrait, d'autre part, être utilisée pour initialiser la fiabilité de nouvelles sources. En effet, dans un cadre réel de traitement de gros volumes de données, l'interrogation systématique d'un opérateur pour qualifier chaque scripteur serait rédhibitoire.
Un autre sujet de recherche actuel d'intérêt est la prise en compte de sources successives (Cholvy 2010a; 2010b). Pour ces problématiques de discours rapporté, dans lequel l'information est fournie par une source en citant une autre, il faut définir quelle source la fiabilité qualifie et tenir compte des sources additionnelles. Afin de ne pas dénaturer la dimension fiabilité, il conviendrait de s'assurer qu'elle ne porte pas sur la dégradation de l'information mais bien sur sa source.

Élargissement de la corrélation L'étape de crédibilité soulève le problème de la mesure de la gradualité de la corrélation entre informations. On peut envisager, au moins,

trois origines à cette gradualité. La première compare le contenu informationnel et pose la question du choix de la mesure de recoupement. Le problème de la similarité des éléments à agréger est un classique du domaine de la fusion qui s'enrichit à l'aide d'opérateurs de comparaison sémantiquement expressifs.
La gradualité de la corrélation peut également provenir d'une évaluation de l'indépendance des sources. On dit que deux sources sont indépendantes si l'on peut démontrer que la connaissance d'une information par l'une ne dépend pas de la connaissance de la même information par l'autre. Ce critère est donc estimé pour plusieurs sources en regard d'une information. Notre partition des dimensions ne représente pas un tel cas mais on pourrait envisager, dans la mesure où l'on saurait mesurer l'indépendance, d'intégrer une telle pondération à la corrélation.
Une troisième manière d'envisager la gradualité de la corrélation émane de la prise en compte de relations d'amitié ou d'hostilité entre les sources. Il s'agit alors d'influencer positivement une corroboration provenant d'un ennemi, où l'on aura sans doute supposé que l'inimitié est une garantie d'indépendance, ou d'une notion équivalente.

Raisonnement ontologique On pourrait, afin de raffiner la plausibilité, exploiter la représentation des informations dans une ontologie et, ainsi, profiter des capacités de raisonnement de ce type de structures. Celles-ci permettraient d'inférer la compatibilité de l'information avec les connaissances, voire sa contradiction avec des individus préalablement acquis.
Notons que la relation de subsomption de ces structures induit des règles de généralisation entre concepts, de la forme « Si X est un terroriste alors X est une personne ». La cotation intégrée à ce type de règles leur conférerait une évaluation de l'incertitude de l'observation, malgré le glissement que nous avons évoqué entre cote et incertitude. Cette notion de règles de description de la connaissance, de raisonnement conséquent et d'observations incertaines est liée à nos travaux sur l'abduction floue (Revault d'Allonnes et al. 2006; 2007a).
Ceux-ci s'intéressent, eux aussi, à augmenter la sémantique associée à un système de raisonnement en prenant en compte les observations réelles. Notre objectif principal étant de proposer un système interprétable, ces propriétés des ontologies pourraient être exploitées pour enrichir la connaissance pendant la fusion, où le rapprochement entre informations et individus est conditionné par la compatibilité de leurs types, comme il l'est lors de la corrélation. De plus, nos travaux permettraient de qualifier la forme de relation existant entre un concept et ses spécialisations (Revault d'Allonnes et al., 2009).

Psychologie du procédé

Il serait intéressant d'effectuer une caractérisation psychologique et cognitive du procédé d'établissement de la confiance proposé. Une telle caractérisation pourrait intervenir à tous les niveaux du modèle. On pourrait, par exemple, généraliser l'interprétation que nous avons faite des trois postures de crédulité, proposées au chapitre 4, que nous avons respectivement qualifiées de crédule, neutre et méfiante. Une caractérisation similaire pourrait s'appliquer à des classes de stratégies, une telle classe pouvant, par exemple, regrouper les comportements symétriques entre la conviction et le doute à la réception de nouvelles informations, dont un exemple est donné à la fin du chapitre 5. À un niveau plus global encore, les tendances proposées à la hausse ou à la baisse lors de l'intégration des différentes dimensions pourraient également faire l'objet de descriptions psychologiques. Le bénéfice tiré de telles caractérisations par des experts cogniticiens serait l'établissement d'une théorisation sémantique des différents éléments du procédé proposé.

Extensions théoriques

Au-delà de la cotation, nous avons fait des propositions formelles dans le cadre de la logique multivalente, dont les pièces maitresses sont les opérateurs de combinaison que nous avons définis et la représentation de l'impossibilité de mesurer. Chacune d'entre elles ouvre la voie à des perspectives théoriques et formelles.

Généralisation des opérateurs Nous avons défini des opérateurs paramétriques de combinaison pour la formalisation de la cotation dans le cadre d'une logique multivalente étendue (opérateur de combinaison baissier, opérateur de pondération de la crédibilité, opérateur de crédibilité pondérée). Il serait intéressant de replacer ces opérateurs dans le cadre de la logique multivalente et de les caractériser par rapport aux opérateurs existants, qu'ils soient opérateurs de combinaison ou de raisonnement, afin de préciser l'enrichissement qu'ils apportent.
Plus généralement, les différentes théories cognitives de l'incertain possèdent un grand nombre d'opérateurs d'agrégation, dont la classe des opérateurs paramétriques est largement étudiée, notamment pour l'expressivité d'agrégateurs variables. L'étude et le positionnement des opérateurs proposés dans les typologies existantes permettrait de les caractériser et par là d'obtenir leur formalisation cognitive. Ceci conduirait à la comparaison de leur expressivité à celle des opérateurs historiques et les ferait bénéficier de la sémantique associée.

Extension de l'incertitude multivalente Nous avons identifié une lacune de la logique multivalente en la confusion entre l'impossibilité d'évaluer la vérité et la valeur neutre de l'échelle. Nous avons, donc, introduit une extension à $\mathcal{L}_M$ en lui adjoignant un degré symbolisant cette ignorance, $\tau_?$, afin d'augmenter son expressivité. Il serait intéressant d'étudier la problématique du raisonnement sur $\mathcal{L}_M^e$ et l'extension nécessaire des opérateurs associés lorsque confrontés à l'inconnu. En levant l'ambiguïté portant sur l'impossibilité de mesurer la vérité, nos travaux participent à la prise en compte explicite de l'incertain dans le raisonnement dans le cadre de la logique multivalente.

Know how to listen and you will profit even from those who talk badly.
Plutarch

Bibliographie

ACE07 (2007). Automatic Content Extraction Evaluation (ACE07).

Agence Nationale de la Recherche (2008). Concepts Systèmes et Outils pour la Sécurité Globale. `http://www.agence-nationale-recherche.fr/AAPProjetsOuverts?lngAAPId=188`.

Akdag, H. (1992). *Une approche logique du raisonnement incertain*. Doctorat d'État, Université Pierre et Marie Curie.

Akdag, H., De Glas, M. & Pacholczyk, D. (1992). A qualitative theory of uncertainty. *Fundamenta Informaticae*, *17*, 333–362.

Akdag, H., Truck, I., Borgi, A. & Mellouli, N. (2001). Linguistic Modifiers in a Symbolic Framework. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS)*, *9*, 49–61.

Aristote (4ème siècle av. JC). *De l'interprétation*. Flammarion.

Austin, D. (2006). How Google finds your needle in the web's haystack. *American Mathematical Society Feature Column*.

Barber, K. S. & Kim, J. (2000). Belief Revision Process Based on Trust : Agents Evaluating Reputation of Information Sources. *Lecture Notes in Computer Science*, 73–82.

Bärecke, T., Delavallade, T., Lesot, M.-J., Pichon, F., Akdag, H., Bouchon-Meunier, B., Capet, P. & Cholvy, L. (2010). Un modèle de cotation pour la veille informationnelle en source ouverte. *Colloque international Veille Stratégique Scientifique & Technologique (VSST'2010)*. Toulouse, France.

Batini, C. & Scannapieco, M. (2006). *Data Quality*. Springer.

Baud, J. (2002). *Encyclopédie du renseignement et des services secrets*. Charles Lavauzelle.

Beauregard, J. E. (2001). *Modeling Information Assurance*. Thèse de doctorat, Air Force Institute of Technology.

Benferhat, S., Dubois, D. & Prade, H. (2000). *Fusion de connaissances ou de préférences en intelligence artificielle*, 243–257. Cepaduès-éditions.

Bernstein, A., Kaufmann, E., Bürki, C. & Klein, M. (2005). How Similar Is It ? Towards Personalized Similarity Measures in Ontologies. *Wirtschaftsinformatik* (pp. 1347–1366). Bamberg, Germany : Physica-Verlag.

Berti, L. (1998). From Data Source Quality to Information Quality : The Relative Dimension. *Proceeding of the 1998 Conference on Information Quality* (pp. 247–264). Cambridge, MA.

Berti, L. (1999a). *La qualité des données et leur recommandation : Modèle conceptuel, formalisation et application à la veille technologique*. Thèse de doctorat, Université de Toulon et du Var.

Berti, L. (1999b). Quality and Recommendation of Multi-source Data for Assisting Technological Intelligence Applications. *Lecture Notes in Computer Science*, 282–291.

Besombes, J. & Cholvy, L. (2007). Information fusion : using an ontology to information evaluation. *Proceedings of the International Colloquium on Information Fusion 2007* (pp. 416–422). Xi'an, China.

Besombes, J. & Cholvy, L. (2009). Information Evaluation in Fusion using Information Correlation. *Proceedings of the 12th International Conference on Information Fusion* (pp. 264–269). Seattle, USA.

Besombes, J. & Revault d'Allonnes, A. (2008). An Extension of STANAG2022 for Information Scoring. *Proceedings of the 11th International Conference on Information Fusion* (pp. 1635–1641). Köln, Germany.

Bianchini, M., Gori, M. & Scarselli, F. (2002). PageRank : A circuital analysis. *In Proceedings of the Eleventh International World Wide Web (WWW) Conference.*

Blasch, E., Kadar, I., Salerno, J., Kokar, M. M., Das, S., Powell, G. M., Corkill, D. D. & Ruspini, E. H. (2006). Issues and Challenges in Situation Assessment (Level 2 Fusion). *Journal of Advances in Information Fusion*, *1*, 122–139.

Blaze, M., Feigenbaum, J., Resnick, P. & Strauss, M. (1997). Managing trust in an information-labeling system. *Special issue of selected papers from the 1996 Amalfi Conference on Secure Communication in Networks.*

Bloch, I. (1996). Information Combination Operators for Data Fusion : A Comparative Review with Classification. *IEEE Transactions on Systems, Man and Cybernetics – Part A*, *26*, 52–67.

Bloch, I., Hunter, A., Appriou, A., Ayoun, A., Benferhat, S., Besnard, P., Cholvy, L., Cooke, R., Cuppens, F., Dubois, D. et al. (2001). Fusion : General concepts and characteristics. *International Journal of Intelligent Systems*, *16*, 1107–1134.

Borden, D. L. & Harvey, K. (1998). *The Electronic Grapevine : Rumor, Reputation, and Reporting in the New On-Line Environment.* Lawrence Erlbaum Associates.

Bouchon-Meunier, B. (1993). *La logique floue.* Presses Universitaires de France.

Bouchon-Meunier, B. (Ed.). (1998). *Aggregation and Fusion of Imperfect Information.* Studies in Fuzziness and Soft Computing. Physica-Verlag.

Bouchon-Meunier, B. & Nguyen, H. T. (1996). *Les incertitudes dans les systèmes intelligents.* Presses Universitaires de France.

Bouchon-Meunier, B., Rifqi, M. & Lesot, M.-J. (2008). *Similarities in Fuzzy Data Mining : From a Cognitive View to Real-World Applications*, vol. 5050/2008 of *Lecture Notes in Computer Science*, 349–367. Springer Berlin / Heidelberg.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, *30*, 107–117.

Chen, P., Xie, H., Maslov, S. & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*, 8–15.

Cholvy, L. (2004). Information Evaluation in fusion : a case study. *Proceedings of the International Conference on Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 2004).* Perugia, Italy.

Cholvy, L. (2010a). Evaluation of information reported : A model in the theory of evidence. *IPMU (1)* (pp. 258–267). Springer.

Cholvy, L. (2010b). Plausibility of information reported by successive sources. *SUM* (pp. 126–136). Springer.

Cholvy, L. & Nimier, V. (2003). Information Evaluation : discussion about STANAG 2022 recommendations. *NATO–IST Symposium on military data and information fusion.* Prague, Czech Republic : Research and Technology Organization Information Systems Technology Panel (RTO–MP–IST–040).

Chu, Y.-H., Feigenbaum, J., LaMacchia, B. A., Resnick, P. & Strauss, M. (1997). REFEREE : Trust Management for Web Applications. *Computer Networks and ISDN Systems*, *29*, 953–964.

Clark, K. L. & McCabe, F. G. (2007). Ontology schema for an agent belief store. *International Journal of Man-Machine Studies*, *65*, 640–658.

Commission du Livre Blanc (2008). Livre blanc sur la défense et la sécurité nationale. `http://www.livreblancdefenseetsecurite.gouv.fr`.

Coucharière, O. (2010). *Fusion de données hétérogènes pour la tenue de situation terrestre.* Thèse de doctorat, Laboratoire d'Informatique de Paris VI (LIP6), Université Pierre et Marie Curie, Paris, France.

Darmoni, S., Leroux, V., Daigne, M., Thirion, B., Santamaria, P. & Duvaux, C. (1998). Critères de qualité de l'information de santé sur l'Internet. *Santé et Réseaux Informatiques, Informatique et Santé*, *10*, 162–174.

Darwiche, A. & Ginsberg, M. L. (1992). A Symbolic Generalization of Probability Theory. *Proceedings of the 10th National Conference on Artificial Intelligence* (pp. 622–627). San Jose, California : The AAAI Press/The MIT Press.

De Glas, M. (1987). Representation of Łukasiewicz' many-valued algebras ; the atomic case. *Fuzzy Sets and Systems*, *14*.

Delavallade, T. (2007). *Évaluation des risques de crise, appliquée à la détection des conflits*. Thèse de doctorat, Laboratoire d'Informatique de Paris VI (LIP6), Université Pierre et Marie Curie, Paris, France.

Delavallade, T., Mouillet, L., Bouchon-Meunier, B. & Collain, E. (2007). Monitoring Event Flows and Modelling Scenarios for Crisis Prediction : Application to Ethnic Conflicts Forecasting. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS)*, *15*, 83–110.

Delmotte, F. (2007). Detection of defective sources in the setting of possibility theory. *Fuzzy Sets and Systems*, *158*, 555–571.

Denœux, T. (2006). The cautious rule of combination for belief functions and some extensions. *Information Fusion, 2006 9th International Conference on* (pp. 1–8).

Department of the Army Headquarters (2005). FM 2–22.3 Human Intelligence Collector Operations.

Detyniecki, M. (2000). *Mathematical Aggregation Operators and their Application to Video Querying*. Thèse de doctorat, Université Pierre et Marie Curie.

Diemers, D. (1999). On the Social Dimension of Information Quality and Knowledge. *Proceedings of the 1999 MIT Conference on Information Quality* (pp. 125–143).

Diligenti, M., Gori, M. & Maggini, M. (2004). A unified probabilistic framework for web page scoring systems. *IEEE Transactions on Knowledge and Data Engineering*, 4–16.

DISS & DISC (2001). Intelligence Wing Student Précis.

Dubois, D. (2010). *Degrees of Truth, Ill-Known Sets and Contradiction*, vol. 249 of *Studies in Fuzziness and Soft Computing*, 65–83. Springer.

Dubois, D. & Prade, H. (2004). On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, *142*, 143–161.

Ekström, M. A., Bicharra Garcia, A. C. & Björnsson, H. C. (2005). Rewarding honest ratings through personalised recommendations in electronic commerce. *International Journal of Electronic Business (IJEB)*, *3*, 392–410.

Ekström, M. A. & Björnsson, H. C. (2002). A rating system for AEC e-bidding that accounts for rater credibility. *Proceedings of the International Council for Research and Innovation in Building and (Conseil Internationnal du Bâtiment) Organization & Management in Construction Symposium (CIB W65)* (pp. 753–766).

Falcone, R., Singh, M. P. & Tan, Y.-H. (Eds.). (2001). *Trust in Cyber-societies, Integrating the Human and Artificial Perspectives*, vol. 2246 of *Lecture Notes in Computer Science*. Springer.

Faux, F. & Luthon, F. (2007). Étude de différentes règles de fusion d'information couleur appliquées à la détection d'un visage en temps réel. *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'2007)*. Cépaduès.

Freeman, L. C. (2000). Visualizing Social Networks. *Journal of Social Structure*, *1*, 4.

Garg, A. & Battiti, R. (2004). *The Reputation, Opinion, Credibility and Quality (ROCQ) Scheme* (Technical Report). University of Trento, Italy.

Garg, A., Battiti, R. & Costanzi, G. (2004). Dynamic Self-management of Autonomic Systems : The Reputation, Quality and Credibility (RQC) Scheme. *Autonomic Communication, First International IFIP Workshop, WAC 2004, Revised Selected Papers* (pp. 165–178). Berlin, Germany : Springer.

Gligorov, R., Aleksovski, Z., ten Kate, W. & van Harmelen, F. (2007). Using Google distance to weight approximate ontology matches. *Proceedings of the 16th International Conference on World Wide Web* (pp. 776–784).

Hamill, J. T. (2000). *Modeling Information Assurance : a Value Focused Thinking Approach*. Thèse de doctorat, Air Force Institute of Technology.

Harrathi, R. & Calabretto, S. (2006). Un modèle de qualité de l'information. *Actes des Journées Extraction et Gestion de Connaissances (EGC)* (pp. 299–304). Lille, France.

Heath, T., Motta, E. & Petre, M. (2007). Computing Word-of-Mouth Trust Relationships in Social Networks from Semantic Web and Web2.0 Data Sources. *Proceedings of the Workshop on Bridging the Gap between Semantic Web and Web 2.0, 4th European Semantic Web Conference (ESWC2007)*. Innsbruck, Austria.

Jain, A., Cucerzan, S. & Azzam, S. (2007). Acronym-expansion recognition and ranking on the web. *Proceedings of the IEEE International Conference on Information Reuse and Integration* (pp. 209–214).

Jain, A. & Ipeirotis, P. G. (2009). A quality-aware optimizer for information extraction. *ACM Transactions on Database Systems (TODS)*, *34*, 1–48.

Jain, A., Ipeirotis, P. G., Doan, A. & Gravano, L. (2009). Join optimization of infor-

mation extraction output : Quality matters. *Proceedings of the 2009 IEEE International Conference on Data Engineering* (pp. 186–197).

Jøsang, A. (2008). Conditional Reasoning with Subjective Logic. *Journal of Multiple-Valued Logic and Soft Computing*, *15*, 5–38.

Jøsang, A., Diaz, J. & Rifqi, M. (2009). Cumulative and averaging fusion of beliefs. *Information Fusion.*

Jøsang, A. & Pope, S. (2011). Dempster's rule as seen by little coloured balls. *Computational Intelligence Journal (à paraître).*

Kahn, B. K., Strong, D. M. & Wang, R. Y. (2002). Information quality benchmarks : product and service performance. *Communications of the ACM*, *45*, 184–192.

Kargar, M. J., Ramli, A. R., Ibrahim, H. & Azimzadeh, F. (2008). Formulating Priory of Information Quality Criteria on the Blog. *World Applied Sciences Journal*, *4*, 586–593.

Keeney, R. L. (1994). Creativity in Decision Making with Value-Focused Thinking. *MIT Sloan Management Review*, *35*, 33–41.

Khare, R. & Rifkin, A. (1998). Trust management on the world wide web. *Computer networks and ISDN Systems*, *30*, 651–653.

Knight, S.-A. & Burn, J. (2005). Developing a framework for assessing information quality on the world wide web. *Informing Science : International Journal of an Emerging Transdiscipline*, *8*, 159–172.

Kruger, K., Schade, U. & Ziegler, J. (2008). Uncertainty in the fusion of information from multiple diverse sources for situation awareness. *Proceedings of the 11th International Conference on Information Fusion.* Koeln, Germany.

Krukow, K. (2004). *On foundations for dynamic trust management* (Technical Report). Université de Åarhus, Danemark.

Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E. & White, F. (2004). Revisiting the JDL Data Fusion Model II. *In P. Svensson and J. Schubert (Eds.), Proceedings of the 7th International Conference on Information Fusion (FUSION 2004)* (pp. 1218–1230).

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

Mannes, A. & Golbeck, J. (2007a). Building a Semantic Web Portal for Counter Terror Analysis. *2007 IEEE Aerospace Conference* (pp. 1–5).

Mannes, A. & Golbeck, J. (2007b). Ontology Building : A Terrorism Specialist's Perspective. *2007 IEEE Aerospace Conference* (pp. 1–5).

Miller, G. A. (1955). The Magical Number Seven, Plus or Minus Two : Some Limits on our Capacity for Processing Information. *Psychological Review*, *101*, 343–352.

Ministère de la défense – Commandement de la Formation de l'Armée de Terre (CoFAT) (2001). Manuel du cadre de contact (TTA150).

Mouillet, L. (2005). *Modélisation, reconnaissance et apprentissage de scénarios de conflits ethno-politiques.* Thèse de doctorat, Laboratoire d'Informatique de Paris VI (LIP6), Université Pierre et Marie Curie, Paris, France.

National Institute of Standards and Technology (2003). *NIST/SEMATECH e-Handbook of Statistical Methods.* NIST/SEMATECH.

National Institute of Standards and Technology (Ed.). (2007). *16th Text REtrieval Conference Proceedings (TREC 2007).* National Institute of Standards and Technology. Gaithersburg, Maryland.

Naumann, F. & Rolker, C. (2000). Assessment methods for information quality criteria. *Proceedings of the International Conference on Information Quality (IQ)* (pp. 148–162). Cambridge, MA.

Nimier, V. (2004). Information Evaluation : a Formalisation of Operational Recommendations. *Proceedings of the Seventh International Conference on Information Fusion* (pp. 1166–1171). Stockholm, Sweden.

Nimier, V. & Appriou, A. (1995). Utilisation de la théorie de Dempster-Shafer pour la fusion d'informations. *Groupe d'Études du Traitement du Signal et des Images (GRETSI)* (pp. 137–140).

North Atlantic Treaty Organization (NATO) (1997). Annex to STANAG 2022 (Edition 8).

North Atlantic Treaty Organization (NATO) (2002). Allied Joint Publication – 2.1, Intelligence Procedures.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The pagerank citation ranking : Bringing order to the web.* (Technical Report 1999-66). Stanford InfoLab.

Parsons, S. & Hunter, A. (1998). *A review of uncertainty handling formalisms*, vol. 1455/1998 of *Lecture Notes in Computer Science*, chapter 2, 8–37. Springer.

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

Quercia, D., Hailes, S. & Capra, L. (2007). Lightweight distributed trust propagation. *Proceedings of the 7th IEEE International Conference on Data Mining.* Omaha, US.

Rein, K. & Schade, U. (2009). How certain is certain ? Evaluation of uncertainty in the fusion of information derived from diverse sources. *Proceedings of the 12th International Conference on Information Fusion* (pp. 1613–1620). Seattle, USA.

Rein, K., Schade, U. & Hieb, M. R. (2009). Battle Management Language (BML) as an Enabler. *International Conference on Communications (ICC2009).* Dresden, Germany.

Revault d'Allonnes, A., Akdag, H. & Bouchon-Meunier, B. (2006). Vers une classification de problèmes abductifs en fonction d'observations possibles. *Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'2006)* (pp. 121–128). Cépaduès.

Revault d'Allonnes, A., Akdag, H. & Bouchon-Meunier, B. (2007a). Selecting Implications in Fuzzy Abductive Problems. *IEEE Symposium on Foundations of Computational Intelligence, FOCI 2007* (pp. 597–602). Hawaii, USA.

Revault d'Allonnes, A., Akdag, H. & Bouchon-Meunier, B. (2009). For a Data-Driven Interpretation of Rules, wrt GMP conclusions, in abductive problems. *Journal of Uncertain Systems (JUS)*, *3*, 280–297.

Revault d'Allonnes, A., Akdag, H. & Bouchon-Meunier, B. (2010). Incertain et inconnu, deux facettes de la cotation. *COTA@IC'2010 : COtation des Informations, Théorie et Applications – atelier des 21èmes Journées francophones d'Ingénierie des Connaissances* (pp. 99–103). Alès, France.

Revault d'Allonnes, A., Akdag, H. & Poirel, O. (2007b). Trust-moderated information-likelihood. A multi-valued logics approach. *Computation and Logic in the Real World, CiE 2007* (pp. 1–6). Sienna, Italy.

Revault d'Allonnes, A. & Besombes, J. (2009). Critères d'évaluation contextuelle pour le traitement automatique. *QDC'09 : Qualité des Données et des Connaissances – atelier des 9èmes Journées Francophones Extraction et Gestion des Connaissances* (pp. A6 13–20). Strasbourg, France.

Ruderman, J. (2004). A comparison of two trust metrics.

Sen, S. (2002). Believing others : Pros and cons. *Artificial Intelligence*, *142*, 179–203.

Seridi, H. & Akdag, H. (2000). A qualitative approach for processing uncertainty. *Uncertainty in intelligent and information systems*, *20*, 46–57.

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

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

Shafer, G. (1990). Perspectives on the theory and practice of belief functions. *International Journal on Approximate Reasoning*, *4*, 323–362.

Shah, D. & Zaman, T. R. (2009). Rumors in a Network : Who's the Culprit ? *Neural Information Processing Systems (NIPS)*.

Simmhan, Y. L., Plale, B. & Gannon, D. (2006). Towards a quality model for effective data selection in collaboratories. *IEEE Workshop on Scientific Workflows and Dataflows*.

Simon, F. (2005). *Journaliste : Dans les pas d'Hubert Beuve-Méry.* Arléa.

Smets, P. (1998). The transferable belief model for quantified belief representation. *Handbook of Defeasible Reasoning and Uncertainty Management Systems : Quantified Representation of Uncertainty and Imprecision*, *1*, 267–301.

Smets, P., Hsia, Y.-T., Saffiotti, A., Kennes, R. & Umkehrer, E. (1991). The transferable belief model. *Proceedings of the European Conference Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU)* (pp. 91–96). Marseille, France : Springer.

Staab, S., Bhargava, B. K., Lilien, L., Rosenthal, A., Winslett, M., Sloman, M., Dillon, T. S., Chang, E., Hussain, F. K., Nejdl, W., Olmedilla, D. & Kashyap, V. (2004). The Pudding of Trust. *IEEE Intelligent Systems*, *19*, 74–88.

Tang, R., Ng, Kwang, B., Strzalkowski, T. & Kantor, P. B. (2003). Automatically predicting information quality in news documents. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 97–99).

de Tocqueville, A. (1835). *De la démocratie en Amérique.* Folio Histoire.

Toivonen, S., Lenzini, G. & Uusitalo, I. (2006). Context-aware Trust Evaluation Functions for Dynamic Reconfigurable Systems. *Proceedings of the WWW'06 Workshop on Models of Trust for the Web (MTW'06).* Edinburgh, Scotland : Sun SITE Central Europe (CEUR-WS.org).

Truck, I. & Akdag, H. (2005). *A Qualitative Approach for symbolic Data Manipulation under Uncertainty*, vol. 181 of *Studies in Fuzziness and Soft Computing*, chapter 2, 23–51. Springer.

Truck, I. & Akdag, H. (2006). Qualitative degrees to handle uncertainty : formal models and applications. *Knowledge and Information Systems*, *9*, 385–411.

Ustymenko, S. & Schwartz, D. G. (2006). An Agent-Oriented Logic for Belief and Trust. *COMPSAC (2)* (pp. 321–326). Chicago, Illinois, USA : IEEE Computer Society.

Van Horn, G. & Johnston, L. (2007). Evolutionary Controversy and a Side of Pasta : The Flying Spaghetti Monster and the Subversive Function of Religious Parody. *GOLEM : Journal of Religion and Monsters*, 1–32.

Wang, P. (1995). *Non-Axiomatic Reasoning System – Exploring the Essence of Intelligence.* Thèse de doctorat, Indiana University.

Wang, P. (2001). Confidence as higher-order uncertainty. *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications* (pp. 352–361).

Wang, R. Y., Storey, V. C. & Firth, C. P. (1995). A Framework for Analysis of Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering*, *7*, 623–640.

Wang, R. Y. & Strong, D. M. (1996). Beyond Accuracy : What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*, 5–33.

Weeks, S. (2001). Understanding Trust Management Systems. *IEEE Symposium on Security and Privacy* (pp. 94–105).

Zadeh, L. A. (1965). Fuzzy Sets. *Information and control*, *8*, 338–353.

Zadeh, L. A. (1975). The Concept of a Linguistic Variable and its Application to Approximate Reasoning – I, II & III. *Information Sciences*, *8/9*, 199–249, 301–357, 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*, 307–324.

Zadeh, L. A. (2005). Toward a generalized theory of uncertainty (GTU) – an outline. *Information Sciences*, *172*, 1–40.

Zhu, H. & Wang, R. Y. (2008). An information quality framework for verifiable intelligence products. *Data Engineering : Mining, Information, and Intelligence*.

Table des figures

1.1 Le cycle du renseignement 11

2.1 La cotation aux confluents des domaines de la qualité des données, des informations incertaines et de la fusion 25

2.2 La qualité des données dans la recherche 33

2.3 Problématiques de recherche en qualité de l'information 34

4.1 La chaîne de cotation proposée 64

4.2 Mise à jour de la fiabilité de la source 73

5.1 Un exemple de treillis de De Morgan : $(\mathcal{L}_5, \wedge, \vee, >)$ 78

5.2 Incertitude totale sur une partition floue 81

5.3 Demi-treillis supérieur $(\mathcal{L}_5 \cup \{\tau_?\}, \vee, >)$ 82

5.4 Demi-treillis inférieur $(\mathcal{L}_5 \cup \{\tau_?\}, \wedge, <)$ 82

5.5 Représentation graphique des conditions stratégiques de $F(\tau_\alpha, \tau_\beta)$ 87

5.6 Représentation graphique des conditions stratégiques de $\tilde{F}_c(\tau_\alpha, \tau_\beta)$ 93

5.7 Représentation graphique de $G(\tau_\alpha, \tau_\beta, \tau_\gamma)$ 93

5.8 Représentation graphique de $\tilde{F}_i(\tau_\alpha, \tau_\beta)$ et $\tilde{F}_c(\tau_\alpha, \tau_\beta)$ pour l'utilisateur 1 . . . 98

5.9 Représentation graphique de $\tilde{F}_i(\tau_\alpha, \tau_\beta)$ et $\tilde{F}_c(\tau_\alpha, \tau_\beta)$ pour l'utilisateur 2 . . . 99

6.1 Processus d'extraction de connaissances 108

6.2 Extraction et cotation d'informations atomiques 109

6.3 Fusion d'informations atomiques et cotation de connaissances 110

6.4 Cotation d'un réseau social extrait de textes 116

7.1 Attributs d'une personne 119

7.2 Relations d'une personne & description de la relation 'is_from' 119

7.3 Ontologie de lieux 121

7.4 Résolution d'acronymes 125

7.5 Statistiques de la résolution d'acronymes 126

7.6 Règles de déduction de relations familiales, enrichissant la structure initiale 128

7.7 Interface de saisie de fiabilité de la source 131

7.8 Éditeur de stratégies 135

7.9 Propriétés de la connaissance sur Rifaï Ahmed Taha dans l'ontologie 138

7.10 Mise en relation de connaissances : Rifaï Ahmed Taha et la Gama'a al-Islamiya 138

7.11 Représentation de réseau coté 140

Liste des tableaux

1.1 Cotation du renseignement (STANAG 2022/AJP2.1) 15

3.1 Axes des dimensions de la cotation 49

3.2 Fiabilité de la source 51

3.3 Compétence de la source 53

3.4 Plausibilité de l'information : conformité à la connaissance extérieure 55

3.5 Crédibilité de l'information : indice de confirmation entre informations produites 57

3.6 Cote finale : degré de confiance associé à une information 59

4.1 Fiabilité : 3 stratégies 69

4.2 Compétence : 3 stratégies 70

4.3 Stratégie par défaut pour l'impact de la compétence ou de la plausibilité . . 71

4.4 Crédibilité : 3 stratégies 71

5.1 Un exemple de degrés de vérité dans $\mathcal{L}_5$ 78

5.2 Deux points de vue différents sur la crédulité 98

5.3 Traçage de la crédibilité cumulée pour l'utilisateur 1 99

5.4 Traçage de la crédibilité cumulée pour l'utilisateur 2 100

5.5 Traçage de la crédibilité immédiate pour l'utilisateur 1 101

5.6 Traçage de la crédibilité immédiate pour l'utilisateur 2 102

7.1 Exemples de mots-clés favorisant le typage des noms propres extraits 123

7.2 Quelques mots-clés de relations favorisant le typage des entités extraites . . 127

7.3 Exemples d'informations atomiques 136

C.1 Critères de qualité des données (Wang & Strong, 1996) 173

C.2 Critères de qualité des données (Berti, 1999b) 175

D.1 Stratégie par défaut multivalente : initialisation 177

D.2 Stratégie par défaut multivalente : intégration de la compétence ou de la plausibilité 177

D.3 Stratégie par défaut multivalente : activation de la crédibilité 178

ANNEXES

Annexe A

Types de renseignements

L'article « List of intelligence gathering disciplines » de Wikipédia, reproduit ici, recense les différents types de collecte de renseignements.

- HUMINT : Human Intelligence–gathered from a person on the ground.
  - Espionage
  - Friendly accredited diplomats
  - Military attaches
  - Non-governmental organizations (NGOs)
  - Patrolling (Military police, patrols, etc.)
  - Prisoners of war (POWs) or detainees
  - Refugees
  - Strategic reconnaissance, as by Special Forces
  - Traveler debriefing (e.g., CIA Domestic Contact Service)
- GEOINT Geospatial Intelligence–gathered from satellite, aerial photography, mapping/terrain data
  - IMINT–Imagery Intelligence : gathered from satellite and aerial photography
- MASINT Measurement and Signature Intelligence
  - Electro-optical MASINT
    - Airborne Electro-Optical Missile Tracking MASINT
    - Tactical Countermortar Sensors
    - Infrared MASINT
    - Airborne Electro-Optical Missile Tracking MASINT
    - Optical Measurement of Nuclear Explosions
    - LASER MASINT
    - Spectroscopic MASINT
    - Hyperspectral Imagery MASINT
    - Space-based Staring Infrared Sensors
  - Nuclear MASINT

        - Radiation survey and dosimetry
        - Space-based Nuclear Energy Detection
        - Effects of Ionizing Radiation on materials
    - Geophysical MASINT
        - Weather and Sea Intelligence MASINT
        - Acoustic MASINT (also known as ACOUSTINT- Acoustic phenomena)
        - Seismic MASINT
        - Magnetic MASINT
        - Gravitimetric MASINT
    - Radar MASINT
        - Line-of-Sight Radar MASINT
        - Synthetic Aperture Radar (SAR) and Inverse Synthetic Aperture Radar (ISAR)
        - Non-Cooperative Target Recognition
        - Multistatic Radar MASINT
        - Passive Covert Radar
    - Materials MASINT
        - Chemical Materials MASINT
        - Biological Materials MASINT
        - Nuclear test analysis
    - Radiofrequency MASINT
        - Frequency Domain MASINT
        - Electromagnetic Pulse MASINT
        - Unintentional Radiation MASINT
- OSINT Open Source Intelligence–gathered from open sources.
  Can be further segmented by source type : Internet/General, Scientific/Technical and various HUMINT specialties (e.g. trade shows, association meetings, interviews, etc.)
- STRATINT – Strategic Intelligence
- SIGINT Signals Intelligence–gathered from interception of signals
    - COMINT–Communications Intelligence
    - ELINT–Electronic Intelligence : gathered from non-communications electronic emissions
        - FISINT–Foreign Instrumentation Signals Intelligence
            - TELINT–Telemetry Intelligence : the collection and analysis of telemetry data from the target's missile or sometimes from aircraft tests.
- TECHINT Technical Intelligence–gathered from analysis of weapons and equipment used by the armed forces of foreign nations, or environmental conditions.
    - MEDINT–Medical Intelligence : gathered from analysis of medical records and/or actual physiological examinations to determine health and/or particular ailments/allergetic conditions for exploitation.
- FININT Financial Intelligence–gathered from analysis of monetary transactions
- VPBINT – Variable Programmable Beacon Intelligence – gathered via telematics from static and dynamic (mobile) variable programmable intelligent beacons.

Annexe B

Définitions des étapes du cycle du renseignement, selon l'AJP 2.1

Les définitions reproduites ici proviennent de l'Allied Joint Publication 2.1 (North Atlantic Treaty Organization (NATO), 2002) et détaillent les différentes étapes du cycle du renseignement présenté à la figure 1.1, page 11.

Définition 12 : Processing
"The production of intelligence through collation, evaluation, analysis, integration and interpretation of information and/or other intelligence."

Définition 13 : Collation
"A step in the processing phase of the intelligence cycle in which the grouping together of related items of information or intelligence provides a record of events and facilitates further processing."

Définition 14 : Evaluation
"A step in the processing stage of the intelligence cycle constituting appraisal of an item of information in respect of the reliability of the source and the credibility of the information."

Définition 15 : Analysis
"A step in the processing phase of the intelligence cycle in which information is subjected to review in order to identify significant facts for subsequent interpretation."

Définition 16 : Integration
"A step in the processing phase of the intelligence cycle whereby analysed information or intelligence is selected and combined into a pattern in the course of the production of further intelligence."

Définition 17 : Interpretation

"The final step in the processing phase of the intelligence cycle in which the significance of information or intelligence is judged in relation to the current body of knowledge."

Annexe C

Qualité de l'information, quelques définitions

Les définitions suivantes proviennent de l'IAIDQ (International Association for Information and Data Quality). Nous présentons ici des définitions librement traduites, les versions originelles étant disponibles à `http://iaidq.org/main/glossary.shtml`.

Définition 18 : Données

1. Symboles, nombres et autres représentations de faits.
2. Les matériaux bruts à partir desquels, en les plaçant dans un contexte, l'information est produite. Voir information.

(Larry English)

Définition 19 : Information

1. Données en contexte : le sens associé aux données ou l'interprétation des données en fonction du contexte.
2. Le produit fini après traitement, présentation et interprétation des données.

(Larry English)

Définition 20 : Qualité de l'information
L'adéquation de l'information à son utilisation. L'information doit répondre aux exigences de ses auteurs, utilisateurs et administrateurs.
(Martin Eppler)

Définition 21 : Évaluation de la qualité de l'information
L'échantillonage aléatoire des données et leur mesure selon les critères prédéterminés, tels

que : précision, complétude, validité, non-redondance ou fraicheur pour évaluer son niveau de qualité ou de fiabilité. Aussi appelée évaluation de la qualité ou audit qualité.
(Larry English)

Définition 22 : Critères de qualité de l'information
Un aspect ou propriété de l'information que l'utilisateur estime influent dans le calcul de la qualité de l'information. Parmi ces critères, on retrouve la complétude, la précision, le la fraicheur, la compréhensibilité, l'objectivité et la présentation. Aussi appelés dimensions de la qualité de l'information.
(Larry English)

Le tableau C.1 suivant propose la liste des critères de qualité, telle que donnée par Wang et Strong (1996), regroupés par classes selon une typologie séparant les critères qualifiant exclusivement la donnée, ceux prenant en compte le contexte et ceux mesurant l'utilisation de la donnée par l'utilisateur. Berti (1999b) propose une autre segmentation des critères de la qualité des données, reproduite au tableau C.2.

Category	Dimension	Definition : 'The extent to which...'
Intrinsic	Believability	data are accepted or regarded as true
	Accuracy	data are correct, reliable and certified error-free
	Objectivity	data are unbiased and impartial
	Reputation	data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	data are beneficial and provide advantages for their use
	Relevancy	data are applicable and useful for the task at hand
	Timeliness	the age of the data is appropriate for the task at hand
	Completeness	data are sufficient in depth, breadth and scope for the task at hand
	Amount of data	the quantity or volume of available data is appropriate
Representational	Interpretability	data are in an appropriate language and unit and definitions are clear
	Ease of understanding	data are clear without ambiguity and easily comprehended
	Consistency	data are presented in the same format and are compatible with previous data
	Concision	data are compactly represented without being overwhelmed
Accessibility	Accessibility	data are available or easily retrieved
	Security	access to data can be restricted and kept secure

TABLE C.1 – Critères de qualité des données proposés par Wang et Strong (1996), tels que cités dans Batini et Scannapieco (2006)

Catégorie	Sous-catégorie	Critère
Qualité absolue	Qualité de la gestion de la donnée par le système	Accessibilité Confidentialité Facilité d'échange Facilité de maintenance Facilité de manipulation Facilité de recherche Sécurité...
	Qualité de la représentation de la donnée par le système	Compréhension Concision Interprétation...
	Qualité intrinsèque de la donnée	Absence d'erreur Cohérence Exactitude Qualité de la source

Qualité relative de la donnée	Qualité contextuelle	Référentiel temporel discret Actualité Fraîcheur... continu Traçabilité Variabilité Volatilité... Référentiel applicatif Criticité Pertinence Utilité...
	Qualité relative à l'utilisateur	Référentiel cognitif Fiabilité Originalité Rareté Vraisemblance... Référentiel affectif Intérêt Préférences...
	Qualité relative aux données homologues	Identité Égalité Redondance Similarité...

TABLE C.2 – Catégories de critères de qualité des données, Berti (1999b)

Annexe D

Stratégie par défaut multivalente

Les tableaux présentés ici indiquent le mode de calcul de la projection de chaque dimension, pour la transposition de la statégie par défaut du chapitre 4 au cadre multivalent.

Fiabilité de la source

Cote initiale	τ_4	τ_3	τ_2	τ_1	τ_0	$\tau_?$
$\tau_?$	τ_4	τ_3	τ_2	τ_1	τ_0	$\tau_?$

TABLE D.1 – Stratégie par défaut de la figure 4.1 : calcul multivalent de l'initialisation du processus à l'intégration de la fiabilité de la source

Activation de la dimension

Cote actuelle	τ_4	τ_3	τ_2	τ_1	τ_0	$\tau_?$
τ_4	τ_4	τ_3	τ_2	τ_1	τ_0	τ_4
τ_3	τ_3	τ_3	τ_2	τ_1	τ_0	τ_3
τ_2	τ_2	τ_2	τ_2	τ_1	τ_0	τ_2
τ_1	τ_1	τ_1	τ_1	τ_1	τ_0	τ_1
τ_0	τ_0	τ_0	τ_0	τ_0	τ_0	τ_0
$\tau_?$	τ_4	τ_3	τ_2	τ_1	τ_0	$\tau_?$

TABLE D.2 – Stratégie par défaut de la figure 4.1 : calcul multivalent de la mise à jour de la cote, en fonction de l'évaluation de la compétence ou de la plausibilité

Cote actuelle \ Activation de la crédibilité	τ_4	τ_3	τ_2	τ_1	τ_0	$\tau_?$
τ_4	τ_4	τ_4	τ_4	τ_3	τ_2	τ_4
τ_3	τ_4	τ_4	τ_3	τ_2	τ_1	τ_3
τ_2	τ_4	τ_3	τ_2	τ_1	τ_0	τ_2
τ_1	τ_3	τ_2	τ_1	τ_0	τ_0	τ_1
τ_0	τ_2	τ_1	τ_0	τ_0	τ_0	τ_0
$\tau_?$	τ_4	τ_3	τ_2	τ_1	τ_0	$\tau_?$

TABLE D.3 – Stratégie par défaut de la figure 4.1 : calcul multivalent de la mise à jour de la cote à l'intégration de la crédibilité

Résumé

La confiance en une information est une mesure de la foi qu'on peut lui apporter, c'est-à-dire qu'elle indique combien on peut la croire. Si la confiance constitue un sujet de recherche très étudié, dans la plupart des cas sa mesure témoigne plus de la qualité de la production de l'information sur laquelle elle porte qu'elle ne révèle si l'on peut s'y fier. Comme on ne se satisfait pas du fait qu'une nouvelle émane d'une source sûre pour la croire, l'évaluation automatique de la confiance nécessite un modèle plus riche, capable de décrire pour quelles raisons son objet est ou n'est pas crédible. C'est à cette problématique que nous nous sommes attaché dans nos travaux.

Après avoir étudié la perception d'un mode de représentation de la confiance, nous proposons d'aborder séparément son expression et le processus gouvernant son établissement, de distinguer la cote de la cotation. Nous étudions les prérequis à la définition de la cote et en déduisons les dimensions incontournables à sa construction. Nous en proposons un découpage permettant de regrouper les différents critères d'évaluation selon leur objet et leur influence et assurant leur indépendance et leur non-redondance. Nous nous attachons également à assurer la lisibilité des mesures participant à l'évaluation de la cote en proposant de les apprécier sur des échelles discrètes explicitées d'étiquettes linguistiques.

Une fois cette sélection des dimensions effectuée, nous nous posons la question de leur combinaison pour modéliser le processus d'établissement de la confiance. Afin de répondre à ce problème, nous proposons une philosophie de l'intégration des dimensions à la confiance : nous façonnons une architecture à la cotation. Nous fournissons à cette architecture une représentation sous la forme d'une chaîne de cotation, celle-ci mettant en avant l'ordre de prise en compte des dimensions d'intérêt et leur influence sur la hausse ou la baisse de la cote. Nous montrons également comment la souplesse de notre modèle permet de représenter différentes postures de crédulité de l'utilisateur, adaptabilité essentielle à la modélisation de principes subjectifs.

Suite à ces définitions, nous proposons une formalisation théorique du procédé de cotation ainsi que de la cote, son mode d'expression. Exploitant l'expressivité de la logique multivaluée, nous choisissons ce formalisme pour exprimer nos propositions. Afin de préserver la nuance importante entre l'impossibilité de mesurer et la mesure neutre, nous l'étendons en lui adjoignant un nouveau degré de vérité. Dans ce cadre de logique symbolique étendue, nous définissons les opérateurs de combinaison permettant de représenter l'ensemble de nos propositions et formalisons la modélisation de la crédulité.

Enfin, nous considérons la mise en pratique de notre modèle dans l'extraction et la cotation d'informations symboliques. Dans un premier temps nous examinons la transposition de la cotation au problème de l'extraction de connaissances à partir de textes. Nous détaillons successivement la cotation de l'extraction d'informations, puis celle de leur fusion, en examinant pour chacune la transposition des dimensions qui la constituent. Nous implémentons, ensuite, un démonstrateur pour la mise en œuvre de nos propositions. Enfin, nous appliquons modèle et démonstrateur à un cas réel d'extraction et de cotation de réseau social.

Mots-clés: Cotation, confiance, qualité de l'information, fusion d'informations, logique multivaluée, extraction d'informations, extraction de réseaux sociaux