Un modèle de cotation pour la veille informationnelle en source ouverte

Thomas Baerecke (*), **Thomas Delavallade** (**),**Marie-Jeanne Lesot** (*), **Frédéric Pichon** (***), **Herman Akdag** (*), **Bernadette Bouchon-Meunier** (*)**, Philippe Capet** (**), **Laurence Cholvy** (****),

Thomas.Baerecke@lip6.fr, Thomas.Delavallade@fr.thalesgroup.com, Marie-Jeanne.Lesot@lip6.fr, Frederic.Pichon@thalesgroup.com, Herman.Akdag@lip6.fr, Bernadette.Bouchon-Meunier@lip6.fr, Philippe.Capet@fr.thalesgroup.com, Laurence.Cholvy@onera.fr

(*) LIP6, Université Pierre et Marie Curie –Paris 6 UMR 7606 (France),
(**) Thales Defence & Security & C4I Systems (France),
(***) Thales Research & Technology (France),
(****) ONERA (France)

Mots clefs :
Confiance, Incertitude, Fiabilité, Extraction d'événements, Renseignement, Intelligence économique
Keywords:
Trust, Uncertainty, Reliability, Event Extraction, Intelligence, Competitive Intelligence
Palabras clave :
Confianza, Incertidumbre, Confiabilidad, Extracción de eventos, Investigación, inteligencia económica

Résumé

La cotation de l'information a principalement été étudiée à l'aune des documents doctrinaux du renseignement militaire dont les limites sur le sujet sont cependant manifestes. Nous étudions ici, à des fins de modélisation, la cotation inscrite dans le cadre général de la veille informationnelle en source ouverte, qui préoccupe des domaines tant militaires que civils. En partant d'informations extraites automatiquement par analyse linguistique de documents textuels, nous proposons une modélisation de l'ensemble du processus de cotation, afin d'obtenir une estimation de la confiance à accorder à ces informations d'origine. Cette modélisation s'appuie d'une part sur l'évaluation de l'incertitude associée à une information prise isolément, en affaiblissant l'incertitude qu'exprime sa source par la fiabilité de cette dernière. Elle requiert d'autre part l'agrégation des incertitudes individuelles associées aux différentes informations portant sur un même événement. Nous abordons en outre les difficultés techniques soulevées par notre modélisation et proposons des éléments de solution. Un exemple semi-fictif illustre enfin la démarche adoptée.

1 Contexte et enjeux

Nombreuses sont les administrations publiques ou les entreprises présentant le besoin d'avoir bonne connaissance des informations qui les concernent, afin non seulement de mener à bien leurs missions, mais aussi de valoriser leur image et de se protéger d'éventuelles attaques dans le champ immatériel de l'information et de la communication. La montée en puissance de l'Internet, qui permet l'expression et la diffusion de la plus grande part de ce que le domaine du renseignement désigne par sources *ouvertes*, par opposition aux sources confidentielles et non publiques, renforce la nécessité d'outiller cette veille informationnelle. Sur le Web, l'accroissement des volumes d'information, l'intensification des flux, et surtout la multiplication de sources originelles mal cataloguées, impliquent chez beaucoup d'acteurs de se doter de moyens techniques en adéquation avec ces mutations technologiques, tout particulièrement dans les cercles d'intelligence économique et de renseignement.

L'activité de veille informationnelle est composée de multiples sous-activités dont certaines sont prépondérantes pour que fonctionne l'ensemble d'un tel système de veille. L'évaluation de l'information, souvent nommée *cotation* de l'information, est l'une d'entre elles. Elle regroupe deux mesures, distinctes quoique non totalement indépendantes l'une de l'autre : l'estimation de la *fiabilité* des sources, et l'évaluation de la *véracité* de l'information colportée. Dans le domaine militaire, des pratiques et des doctrines existent déjà au niveau national comme à celui de l'OTAN. Toutefois, ces méthodes restent peu pratiquées car elles exigent un travail manuel très rébarbatif dès que la masse d'information croît par trop. De plus, les doctrines disponibles présentent de nettes incohérences, rendant de fait inexploitable l'aspect coté de l'information. (Pour plus de détails concernant les limites des documents doctrinaux, le lecteur pourra se reporter à [5], [1] et [8].) D'où la nécessité de mieux définir ce qu'on entend par cotation de l'information, puis de la modéliser afin de l'inscrire dans un système de veille qui soit à terme semi-automatisé. C'est là l'objet de cet article.

Dans ce système de veille, et en amont de la cotation, nous considérons par hypothèse de travail qu'un processus de collecte automatisée permet de récupérer un ensemble de documents aux formats HTML et XML issus de l'Internet et publiés sur des sites considérés comme pertinents par l'utilisateur. Au fil de la collecte, des sites potentiellement pertinents mais non sélectionnés par l'utilisateur lui sont proposés pour étendre le domaine de recherche. Les pages Web collectées sont ensuite filtrées afin de repérer le contenu textuel proprement informationnel (suppression de la publicité, des liens de navigation…). Un moteur sémantique, s'appuyant sur une ontologie du domaine étudié, procède ensuite d'une part à la sélection des documents pertinents, écartant ceux qui ne traitent pas des thématiques suivies par l'utilisateur, et d'autre part à l'extraction des événements potentiellement pertinents *via* une analyse linguistique de ces contenus textuels. Ce sont ces événements que le modèle proposé évalue lors du processus de cotation.

Ainsi notre modélisation se démarque des modèles existants par son insertion dans un système applicatif global de veille informationnelle. Ceci implique en particulier une nécessaire prise en compte des contraintes liées à l'extraction semi-automatique à partir de données textuelles des informations à coter. Notons que notre hypothèse de travail reflète des outils déjà existants et d'un bon niveau de maturité : l'hypothèse n'est donc pas trop ambitieuse ni futuriste.

En aval de la cotation, d'autres sous-activités plus évoluées sont à envisager pour répondre aux multiples besoins du système de veille informationnelle. Ainsi dans une optique de surveillance d'une menace diffuse, des fonctions évoluées de valorisation de l'information en renseignement sont à envisager : signaux faibles de crise, réseaux sociaux à visée criminelle en cours de constitution, diffusion de rumeurs voire tentatives de désinformation sont autant d'éléments qu'il s'agit de déceler précocement au sein de la masse de l'information collectable en source ouverte, et des méthodes et outils doivent permettre d'instrumenter ces investigations. Or pour être menées à bien, ces fonctions reposent fondamentalement sur les résultats du processus de cotation, fonction centrale d'un système global de veille.

2 Modélisation de la cotation : approches usuelles et démarche adoptée

Tout au long de cet article nous nous appuyons sur un exemple réel qui nous permet d'une part d'illustrer notre présentation théorique de la modélisation de la cotation et d'autre part de montrer comment notre modèle peut s'appliquer concrètement à des données textuelles telles que celles qui auraient à être traitées dans le cas d'un système de veille informationnelle.

L'exemple choisi concerne le double attentat à la bombe commis le 29 mars 2010 dans le métro de Moscou. Nous supposons que l'utilisateur souhaite identifier l'auteur de cet attentat à partir d'informations collectées en source ouverte. Nous considérons que les éléments informationnels suivants ont pu être recueillis. Peu après l'attentat, le même jour, un responsable des services de sécurité russes (FSB) évoque la piste du groupe terroriste « L'Émirat du Caucase ». mais le lendemain leur chef « Dokou Oumarov » dément être à l'origine des attentats. Malgré ce démenti, le surlendemain, le ministre de l'Intérieur russe affirme que ce groupe rebelle est effectivement impliqué dans ces attentats. À ces trois éléments informationnels pertinents vis-à-vis de la requête de l'utilisateur nous avons choisi d'en ajouter deux supplémentaires ne faisant pas référence aux auteurs de l'attentat de Moscou afin d'introduire du bruit au sein de notre base de connaissances. Nous avons ainsi les cinq éléments informationnels suivants :

(i1) Alexandre Bortnikov, patron du FSB, le 29/03/2010 : « L'Émirat du Caucase pourrait être impliqué dans l'attentat de Moscou de ce matin ».
(i2) Dokou Oumarov, chef du groupe rebelle islamiste l'Émirat du Caucase, le 30/03/2010 : « L'Émirat du Caucase n'est pas responsable de l'attentat de Moscou du 29/03/2010 »
(i3) Ministre de l'Intérieur russe, le 31/03/2010 : « Les rebelles du Caucase Nord sont impliqués dans l'attentat de Moscou du 29/03/2010 »
(i4) Ministre de l'Intérieur russe, le 01/04/2010 : « Selon un bilan préliminaire, pas moins de 26 personnes ont été tuées dans l'attentat de Moscou du 29/03/2010 »
(i5) Faisal Shahzad le 21/06/2010 : « Je me déclare cent fois coupable de l'attentat raté de Times Square du 1er mai 2010 »

Une telle situation, avec des déclarations successives contradictoires mais pourtant pertinentes, mêlées à d'autres informations dont le contenu sémantique est proche mais pourtant inutile au regard de la question que se pose l'utilisateur reflète assez bien la difficulté qu'il peut y avoir à estimer la véracité d'un énoncé. Cet exemple est loin de constituer un cas particulier, exceptionnel. De tels exemples, bien plus complexes (notre exemple a été bien évidemment volontairement simplifié), apparaissent en effet quotidiennement sur le Web.

2.1 Variables du modèle

La très grande majorité des travaux de modélisation de la cotation ont recours à un certain nombre de variables intermédiaires qui sont jugées indispensables pour pouvoir estimer de manière suffisamment fine la confiance que l'on peut accorder à une information donnée. Étant donné notre souhait de construire un modèle de cotation global aussi riche que possible, nous avons fait en sorte d'utiliser la plupart de ces variables dans notre modèle. Nous en donnons ci-après une description synthétique.

Fiabilité de la source : cette variable se retrouve dans la plupart des modèles, [1], [2], [5], parfois sous le vocable de validité de la source, [9], et occupe en outre une place prépondérante dans les documents doctrinaux relatifs à la cotation. Elle porte ainsi que le suggère sa dénomination non pas sur l'information elle-même, mais sur la source qui la rapporte. Elle doit refléter la faculté d'une source à délivrer des informations véraces. Nous nous démarquons cependant de la doctrine en considérant une fiabilité contextuelle, qui dépendra du sujet abordé par la source. Une source est en effet plus ou moins compétente à rapporter des informations de telle ou telle thématique et selon son implication dans l'information délivrée elle peut être plus ou moins encline à déformer la réalité. On retrouve ici deux variables importantes associées à la fiabilité et qui sont évoquées dans [9], [13], à savoir la compétence de la source et sa sincérité.

Degré de corroboration de l'information : cette variable occupe une place centrale pour l'évaluation de la véracité d'une information selon les documents doctrinaux et a été de ce fait reprise dans la plupart des modèles de la littérature. Les documents doctrinaux assimilent même presque complètement la véracité (parfois également nommée crédibilité) d'une information avec ce degré de corroboration. Elle doit permettre de juger de la crédibilité de l'information en fonction du nombre de sources indépendantes rapportant des informations confirmant ou infirmant celle que l'on évalue. Peu d'efforts ont été entrepris cependant pour estimer l'indépendance des sources, ce qui est ici crucial. Nous y reviendrons à la section 3.3.
Vraisemblance de l'information : cette variable est mise en évidence dans [1], dans [17] également sous le terme de plausibilité et dans [5] via la notion de contraintes d'intégrité. Elle vise à rendre compte de l'adéquation entre un élément informationnel donné et les connaissances sur le monde, *a priori*, dont on dispose. Ainsi un élément informationnel entrant en contradiction avec nos connaissances sera jugé peu crédible *a priori*. Si le concept de vraisemblance nous paraît également pertinent, sa mise en pratique est plus délicate. Précisons que dans les travaux mentionnant la vraisemblance de l'information qu'aucune méthode permettant d'estimer cette variable n'est mentionnée. Aussi avons-nous choisi de ne pas l'utiliser dans notre modèle.
Incertitude exprimée par la source : cette variable permet de tenir compte des précautions langagières et autres imprécisions qu'une source peut employer pour rapporter une information, révélant par là-même la confiance qu'elle-même accorde à l'information en question. Ainsi dans notre exemple le responsable du FSB emploie le conditionnel pour exprimer l'incertitude qui subsiste quant à l'attribution de la responsabilité de l'attentat de Moscou. L'estimation de cette variable doit s'appuyer, on le voit sur cet exemple, sur des outils de traitement automatisé du langage naturel relativement puissants. Mais de tels outils sont aujourd'hui largement répandus ainsi que nous l'avons précisé en introduction. Cette variable nous paraît extrêmement importante dans le cadre de la modélisation de la cotation. Aussi l'avons-nous introduite dans notre modèle. Notons cependant qu'elle n'a jamais été citée dans la littérature à notre connaissance.

2.2 Sortie du modèle

Le modèle de cotation que nous proposons s'inscrit dans une chaîne globale de traitement dédiée à la veille informationnelle en source ouverte. Dans ce contexte, il est appliqué sur des événements extraits automatiquement de corpus textuels. Chaque événement est une donnée structurée constituée d'un type, d'acteurs et potentiellement d'autres attributs comme le lieu, la date ou encore un marqueur de négation indiquant que l'occurrence de l'événement est niée.
Pour chaque élément informationnel apportant un éclairage sur un événement donné e, nous introduisons une variable $\boldsymbol{e}$ répondant à la question suivante : « e a-t-il eu lieu ? ». Cette variable est binaire et nous notons $E=\{e, \neg e\}$ son domaine de valeurs, avec $\boldsymbol{e}=e$ signifiant que e a bien eu lieu et $\boldsymbol{e}=\neg e$ que e n'a pas eu lieu. Dans notre exemple les trois énoncés pourraient être associés à la même variable $\boldsymbol{e}$ relative à la question : « L'Émirat du Caucase est-il impliqué dans le double attentat de Moscou du 29/03/2010 ? », avec comme valeurs possibles e = « L'Émirat du Caucase est impliqué dans le double attentat de Moscou du 29/03/2010 » et $\neg e$ = « L'Émirat du Caucase n'est pas impliqué dans le double attentat de Moscou du 29/03/2010 ».
Ainsi que le suggèrent les trois premiers énoncés fournis en exemple, statuer sur la véracité d'une information rapportée par une source et exprimée en langage naturel est peu aisé. Les facteurs d'incertitude sont en effet multiples : les sources ne sont pas toujours fiables, les informations délivrées sont imprécises. C'est bien là l'objet du processus de cotation. Il doit permettre d'évaluer la confiance que l'on peut accorder à un élément informationnel donné, ce qui passe par la quantification de l'incertitude associée à la variable correspondante.
Parmi les différentes théories de l'incertain que l'on pourrait utiliser pour y parvenir, nous avons fait le choix de la théorie des possibilités. En effet la théorie des probabilités ne permet pas de modéliser l'ignorance et nécessite dans le cadre de la cotation des probabilités *a priori* subjectives, délicates à obtenir. La théorie de l'évidence qui généralise aussi bien la théorie des probabilités que celle des possibilités ne souffre pas de ces limitations et pourrait être envisagée et est par exemple utilisée à cet effet dans [6], [7] et [13]. La théorie des possibilités offre cependant un éventail plus large d'opérateurs d'agrégation et donc plus de souplesse lorsqu'il s'agit de fusionner des informations portant sur un même événement et issues de sources différentes comme cela est le cas dans l'exemple que nous avons fourni. La théorie des possibilités est en outre compatible avec la théorie des sous-ensembles flous qui peut utilement être mise à profit dans notre contexte pour modéliser les imprécisions du langage naturel.

Dans le cadre de cette théorie, quantifier l'incertitude associée à une variable ***e*** dont le domaine de valeurs est $E=\{e, \neg e\}$ revient à estimer la distribution de possibilité U^E associée à ***e*** et définie sur E. Ainsi notre modèle de cotation appliqué à la variable ***e*** a pour objectif de fournir une estimation des valeurs $U^E(e)$ et $U^E(\neg e)$. À partir de cette distribution il est en effet immédiat d'obtenir les mesures de possibilité et de nécessité associées à ***e*** à partir desquelles on obtient directement la confiance globale accordée à ***e***. Ainsi, contrairement à ce que préconisent les documents doctrinaux nous préférons mettre l'accent en sortie du modèle sur la confiance globale que l'on peut accorder à un événement donné, plutôt que sur le bigramme (fiabilité de la source, véracité de l'information), qui n'est selon nous qu'un couple de variables intermédiaires permettant d'estimer la sortie. Précisons que cette approche mettant l'accent sur un score global reflétant la confiance que l'on peut accorder à une information se retrouve également dans [1] et [17].

3 Proposition d'un modèle global de cotation

Le modèle de cotation que nous présentons plus en détail dans cette section doit pouvoir être intégré dans une chaîne de traitement globale d'un système de veille informationnelle en source ouverte. Il doit permettre d'aider un utilisateur à répondre à certaines de ses interrogations en utilisant une base de connaissances constituée d'événements rapportés sur le Web par un ensemble de sources. Ces événements sont issus d'un processus de structuration des éléments informationnels textuels contenus dans les documents textuels recueillis sur le Web. Cette structuration doit être réalisée par des outils de traitement du langage naturel. Cette phase, essentielle pour la suite des traitements, ne fait pas partie intégrante du processus de cotation et ne sera donc pas évoquée dans la suite.

Parmi l'ensemble des événements présents dans la base de connaissances, seul un petit nombre est véritablement utile pour répondre à la question que se pose l'utilisateur. Cette identification des événements pertinents et donc des éléments informationnels associés constitue la première étape de la cotation[1]. Cette étape étant particulièrement délicate et sensible pour les performances du modèle, une étape de validation humaine est nécessaire. Pour chacun des différents éléments informationnels sélectionnés à l'issue de cette étape la confiance que l'on peut leur accorder est évaluée en les considérant un à un, séparément des autres. Ces scores de confiance individuels sont ensuite combinés lors d'une phase de fusion pour obtenir un score de confiance globale. Les éléments informationnels enrichis de leur cote peuvent ensuite être utilisés en aval de la cotation, par tout composant applicatif de veille informationnelle. L'ensemble de ce processus est décrit à la Figure 1. Nous détaillons dans la suite de cette section uniquement les trois principales étapes du processus de cotation: la sélection des éléments informationnels pertinents (section 3.1), l'évaluation d'un élément informationnel isolé (section 3.2) et enfin la fusion d'éléments informationnels (section 3.3).

[1] Elle ne doit pas être confondue avec la phase de filtrage des documents pertinents, ceux qui traitent des thématiques jugées intéressantes par l'utilisateur, qui est réalisée en amont lors de la collecte des documents à partir du Web et qui a été évoquée à la section 1.

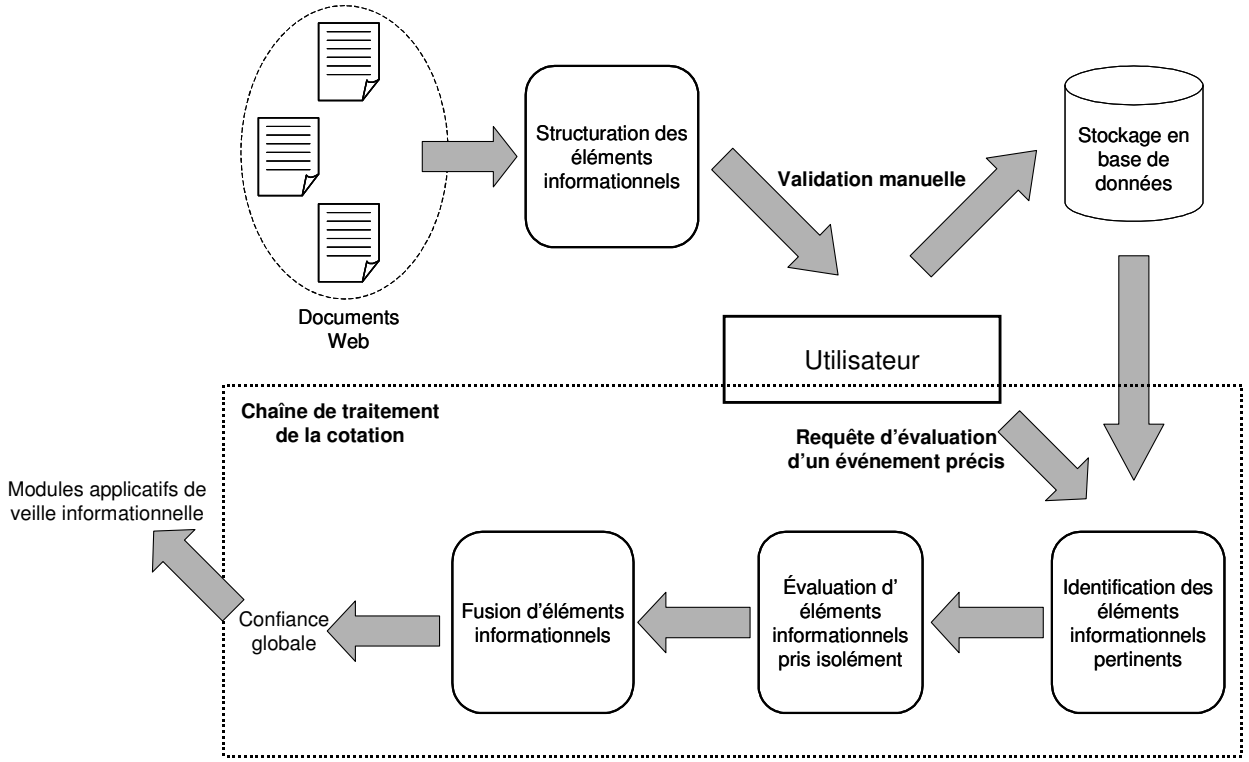


Figure 1 : Architecture globale du processus de cotation

3.1 Identification des éléments informationnels pertinents

Le processus de cotation nécessite de pouvoir identifier, pour une requête relative à la variable e à valeurs dans $E=\{e, \neg e\}$, l'ensemble des éléments informationnels de la base de connaissances portant sur cette variable.

Ainsi que le suggère la Figure 1, ce sont des éléments informationnels structurés extraits de documents textuels qui doivent être analysés par la chaîne de cotation et non pas directement les énoncés exprimés en langage naturel. Chacun de ces éléments informationnels structurés porte sur un type d'événement particulier (dans notre exemple il s'agit des événements de type *attentat*). À chaque type d'événement sont associées certaines propriétés caractéristiques de l'événement. Dans le cas d'un *attentat* on peut ainsi avoir les propriétés suivantes : *auteur*, *date*, *lieu*, *nombre de victimes*… Pour chaque élément informationnel structuré, tout ou partie de ces propriétés peuvent être renseignées.

Nous utilisons ce même formalisme pour représenter une requête utilisateur, c'est-à-dire que nous la considérons comme un élément informationnel structuré particulier. On lui associe ainsi un type d'événement dont certaines des propriétés ont des valeurs précises, les autres n'important pas à l'utilisateur sont laissées indéfinies. L'une des propriétés, que nous nommerons p_{req} joue en outre un rôle particulier : il s'agit de celle dont l'utilisateur cherche à évaluer la valeur. Dans notre exemple il s'agit de la propriété *auteur*.

La tâche d'identification des éléments informationnels pertinents consiste alors à identifier parmi l'ensemble des éléments informationnels structurés de la base de connaissance, ceux qui portent sur des événements de même type que celui de la requête, dont les propriétés ont même valeur que celles définies dans la requête, et ce quelle que soit la valeur de la propriété p_{req}, à condition toutefois que cette propriété soit renseignée. Ce dernier point est important dans la mesure où l'on cherche à identifier tous les éléments informationnels pertinents, qu'ils corroborent ou contredisent la requête. Par contre ceux qui ne permettent pas de statuer sur la valeur de p_{req} sont systématiquement considérés comme non pertinents.
Étant donné la diversité des formes linguistiques permettant d'exprimer une même information, il est cependant peu probable que tous les éléments informationnels réellement pertinents aient exactement les mêmes valeurs pour les différentes propriétés de l'événement considéré. Aussi une identification exacte paraît-elle peu judicieuse. Une solution pour surmonter cette difficulté consiste à effectuer une identification semi-automatique des éléments pertinents ainsi que cela a été proposé dans [2]. Nous proposons pour ce faire de soumettre à l'utilisateur la liste des éléments informationnels les plus proches de sa requête afin qu'il puisse valider manuellement la liste des éléments informationnels qui doit être effectivement retenue.
Pour construire une telle liste nous nous appuyons sur une mesure de similarité entre éléments informationnels structurés. Nous la fixons à 0 pour tous les éléments ne portant pas sur le même type d'événement. Dans le cas contraire, la mesure de similarité retenue consiste en l'agrégation de mesures de similarité élémentaires évaluées propriété par propriété.
Si l'on note q l'élément informationnel structuré correspondant à la requête, T_q le type d'événement associé à q, P_q l'ensemble des propriétés de T_q définies dans la requête, p(q) la valeur de la propriété p pour la requête q, Sim_p, la similarité élémentaire associée à la propriété p et Agg un opérateur d'agrégation des similarités élémentaires, la forme générique de la similarité entre q et un élément informationnel structuré quelconque i peut s'écrire de la façon suivante :

$$Sim(i,q) = \begin{cases} 0 & \text{si } T_q \neq T_e \text{ ou } p_{req}(i) = ? \\ \underset{p \in P_q}{Agg}\left(Sim_p\left(p(i), p(q)\right)\right) & \text{sinon} \end{cases}$$

Les différentes propriétés à considérer pour les événements usuels sont principalement de trois types : il peut s'agir de nombres (nombre de victimes d'attentats par exemple), de dates ou de concepts symboliques (nom d'auteur d'attentat, nom de lieu…). À chacun de ces types doit être associée une mesure de similarité élémentaire.

Similarité entre nombres. Elle pose assez peu de problème, elle peut être basée simplement sur la différence entre ces nombres, avec une similarité maximale pour une différence nulle, qui décroît ensuite dès que cette distance augmente, avec un seuil au-delà duquel les deux valeurs sont considérées comme complètement différentes.
Similarité entre dates. On pourrait *a priori* imaginer une solution semblable à celle décrite pour deux nombres quelconques. Cependant dans bien des cas la date est fournie de manière imprécise. Dans certains textes, seuls l'année, ou le mois seront connus mais pas le jour exact, alors qu'en général une requête utilisateur aura une date définie de manière précise, au moins au niveau de la journée. Il sera très fréquent que nous ayons à comparer deux dates définies avec des niveaux de précision différents. Pour traiter les dates en toute généralité nous proposons d'utiliser une mesure de similarité entre sous-ensembles flous, en associant à chaque date un sous-ensemble, triangulaire pour une date précise[2] et trapézoïdal pour une date imprécise. Nous noterons D_q celui associé à la date de la requête q et D_i celui associé à la date de l'élément informationnel i. Nous considérons, de plus, que la mesure de similarité ne doit pas être symétrique ; la requête q tient en effet lieu de référence et il s'agit d'évaluer à quel point un élément e est compatible avec cette référence. Ces considérations nous ont poussés à nous inspirer des M-mesures de

[2] Une date est dite précise si on connaît non seulement l'année, le mois et le jour, mais également l'heure de la journée.

satisfiabilité [4]. Cependant afin d'éviter de pénaliser des dates trop imprécises, ce qui risque, après agrégation, de rendre l'élément informationnel correspondant trop peu similaire à la requête, même si toutes les autres propriétés correspondent exactement à ce qui est attendu dans la requête, nous avons choisi de considérer une mesure légèrement modifiée dès lors que la date de la requête est plus précise que la date de l'élément informationnel considéré. Nous avons défini pour cela l'échelle de niveaux de précision suivante, noté $np(D)$ pour une date D : 1 pour une date dont on ne connaît que l'année, 2 si l'on connaît en plus le mois, 3 si l'on connaît en plus le jour et 4 lorsque l'heure de la journée est précisée.
Si l'on note M une mesure de sous-ensemble flou (voir [4]), la mesure de similarité entre dates choisie est la suivante (en prenant l'aire comme mesure de sous-ensemble flou) :

$$Sim_{date}(i,q) = \begin{cases} \dfrac{M(D_i \cap D_q)}{M(D_i)} & \text{si } np(D_i) \geq np(D_q) \\ \dfrac{M(D_i \cap D_q)}{M(D_q)} \times \dfrac{np(D_i)}{np(D_q)} & \text{sinon} \end{cases}$$

Similarité entre concepts. Nous supposons que tous les concepts que nous aurons à traiter sont définis dans une ontologie du domaine, ontologie qui est utilisée pour réaliser l'extraction à partir des textes des éléments informationnels structurés. De nombreux travaux se sont intéressés à la question de la similarité entre concepts d'une taxonomie et ont ensuite été étendus aux cas des ontologies. Nous avons retenu la similarité de Wu et Palmer, introduite dans [19], dont le comportement est tout à fait satisfaisant et qui est très simple à mettre en œuvre. Soient C_q et C_i les concepts correspondant aux valeurs d'une propriété donnée pour q et i. Soit l la longueur du plus court chemin dans l'ontologie entre C_q et C_i et soit p la profondeur dans l'ontologie du plus petit généralisant de C_q et C_i (concept le plus spécifique qui subsume C_q et C_i). La similarité de Wu et Palmer s'écrit alors :

$$Sim_{\mathrm{p}}(i,q) = \frac{2p}{l+2p}$$

Agrégation des similarités élémentaires. Deux éléments informationnels q et i, se référant à un même type d'événement, ne seront considérés comme proches que si les valeurs de toutes leurs propriétés sont proches. Dès lors qu'ils diffèrent grandement sur l'une de ces propriétés leur similarité globale ne doit pas pouvoir être élevée. Il nous faut donc recourir à une agrégation sans compromis, de nature conjonctive. Nous avons opté pour le minimum. Une similarité élémentaire entre q et e pour une propriété donnée $p \in P_q$ est non définie lorsque cette propriété n'est pas renseignée pour i. Dans ce cas de figure, cette similarité élémentaire n'est pas prise en compte lors de l'agrégation. Il faut alors atténuer la mesure de similarité globale pour tenir compte du fait que certaines propriétés ont été écartées. Nous proposons dans ce cas de simplement pondérer le minimum par le facteur $1 - \dfrac{nb\,\mathrm{Pr}\,opriétésManquantes}{nb\,\mathrm{Pr}\,opriétésTotal}$.

3.2 Évaluation d'un élément informationnel isolé

L'objectif ultime et ambitieux de l'activité de cotation consiste en l'estimation de l'incertitude quant à la valeur réellement prise par la variable ***e*** dans l'ensemble *E*. Cet objectif peut être atteint en fusionnant les incertitudes sur *E* associées à chaque élément informationnel (i_j). Soit $U^E[i_j]$ la distribution de possibilité représentant l'incertitude sur *E* étant donné l'élément informationnel (i_j). Dans cette section, nous traitons du problème particulier de la construction de $U^E[i_j]$, tandis que la fusion de ces distributions de possibilité individuelles sera traitée à la section 3.3.

Niveau de confiance exprimée par la source. En premier lieu, nous pouvons remarquer qu'un élément informationnel, comme ceux considérés dans l'exemple, correspond à une déclaration du type « la source *S* a une confiance *C* dans le fait que ***e*** prenne pour valeur A », ou plus synthétiquement « S :(A,C) », avec *A* un élément de *E*. Par exemple, pour l'élément informationnel (i_1), la source S est Alexandre Bortnikov, son niveau de confiance *C* est exprimé par « pourrait être » et nous avons $A=\{e\}$.
Une source S manifeste son niveau de confiance ou de certitude C par l'emploi de certains marqueurs linguistiques. À titre d'exemple, la source peut exprimer différents niveaux de certitude par des adjectifs tels que « prouvé », « certain », « vraisemblable », ou encore « improbable » (par ordre décroissant de certitude). Plus généralement, divers types de marqueurs peuvent être utilisés : verbes, adjectifs, noms, adverbes, structures idiomatiques ou complexes. Il existe par ailleurs des modificateurs, par exemple dans la phrase « il est très probable que » le mot *très* renforce la confiance modérée donnée par le marqueur *probable*.
Afin de déterminer la certitude C d'une source S dans un élément informationnel, nous utilisons l'approche décrite dans [16]. Cette approche analyse l'ensemble des marqueurs linguistiques présents dans un élément informationnel à l'aide de patrons linguistiques et les synthétise en une valeur de certitude C, qui peut être « faible », « modérée », « forte » ou « absolue ». Dans le modèle quantitatif de cotation que nous proposons, ces quatre valeurs sont traduites sous forme numérique (respectivement 0.3, 0.5, 0.7 et 1).
Une distribution de possibilité $U_S^E[i_j]$ représentant l'incertitude de la source *S* sur *E* dans (i_j), peut alors être construite à partir de la déclaration « S :(A,C) » en appliquant le principe du minimum de spécificité (principe similaire à celui du maximum d'entropie en théorie des probabilités bayésiennes). Nous avons : $U_S^E[i_j](A)=1$ et $U_S^E[i_j](\neg A)=1-C$. À titre illustratif, pour le cas de l'élément informationnel (i_1), le marqueur linguistique « pourrait être » est transformé en un faible degré (0.3) de certitude que la valeur de ***e*** soit *e*, c'est-à-dire un faible degré de certitude que l'Émirat du Caucase soit impliqué dans l'attentat. La distribution de possibilité $U_S^E[i_1]$ associée est alors définie par $U_S^E[i_1](e) = 1$ et $U_S^E[i_1](\neg e) = 0.7$.

Prise en compte de la fiabilité des sources. Il est clair que lorsque la source est parfaitement fiable dans la thématique *D* de ***e***, nous avons $U^E[i_j] = U_S^E[i_j]$. À l'autre extrême, lorsque la source n'est pas fiable du tout à propos de cette thématique, l'information incertaine $U_S^E[i_j]$ qu'elle fournit sur *E* n'est pas pertinente et ne doit donc pas être prise en considération. Nous représentons les connaissances disponibles sur la fiabilité de *S* dans la thématique *D* de ***e*** à l'aide d'une distribution de possibilité $U^{F_{S,D}}$ sur l'espace $F_{S,D} = \{\text{fiable},\neg\text{fiable}\}$. Cette méta-connaissance sur la source permet, *via* l'application à $U_S^E[i_j]$ d'une opération dite d'affaiblissement telle que celle proposée dans [20], de construire la distribution de possibilité $U^E[i_j]$ représentant l'incertitude sur *E* étant donné l'élément informationnel (i_j).
Formellement, soit *r* le degré de certitude que la source est fiable dans la thématique *D*, que nous supposons connu pour l'instant. La distribution de possibilité $U^{F_{S,D}}$ sur la fiabilité de la source associée à ce degré de certitude est $U^{F_{S,D}}(\text{fiable})=1$ et $U^{F_{S,D}}(\neg\text{fiable})=1-r$. Selon l'opération d'affaiblissement proposée dans [20], nous avons :

$$U^E[i_j](e) = r \times U_S^E[i_j](e) + 1 - r$$
$$U^E[i_j](\neg e) = r \times U_S^E[i_j](\neg e) + 1 - r$$

Évaluation de la fiabilité des sources. Nous avons jusqu'à présent supposé l'existence d'un degré de certitude *r* que la source est fiable dans une thématique *D* donnée. Deux méthodes peuvent être distinguées afin de déterminer ce paramètre : celle consistant en une évaluation subjective par un expert de la fiabilité (éventuellement à travers l'évaluation puis l'agrégation de différentes dimensions de la fiabilité telles que la compétence et la sincérité de la source évoquées à la section 2.1) et celle procédant par apprentissage automatique de la fiabilité en confrontant l'information fournie par une source aux observations sur le terrain et nécessitant donc un jeu de données étiquetées dont la constitution est assez coûteuse. L'évaluation subjective par un expert de la fiabilité d'une source présente l'avantage de ne justement pas nécessiter un tel jeu de données. Toutefois, notons que la méthode basée sur l'expertise présente, elle aussi, un certain nombre d'inconvénients : son coût peut également être élevé si le nombre de sources à évaluer est grand, elle doit bien évidemment être effectuée de manière rigoureuse et enfin elle est difficile à mettre en place de manière dynamique, car la fiabilité d'une source évolue au fil du temps et doit donc être réévaluée périodiquement, ce qui nécessite le retour de l'expert. La méthode automatique de ce point de vue semble avantagée. Nous détaillons dans le paragraphe suivant la méthode automatique d'apprentissage du degré de certitude *r* retenue dans notre modèle et inspirée de [14].

L'apprentissage de *r* nécessite d'avoir à disposition un corpus dans lequel sont placées les valeurs réellement prises par *N* variables $\boldsymbol{e_k}$, *k=1,…,N*, à valeurs dans $E_k = \{e_k, \neg e_k\}$ et relatives à un même domaine *D*. Un tel corpus sera enrichi au fur et à mesure. Nous prévoyons en effet que l'utilisateur puisse à tout moment valider ou invalider un énoncé donné indiquant ainsi qu'il a acquis la certitude que l'élément informationnel correspondant est avéré ou erroné. Chaque action de cette nature effectuée par l'utilisateur permettra ainsi d'ajouter un exemple de plus au corpus d'apprentissage.

Soit $V(\mathbf{e}_k) \in E_k$ la vraie valeur prise par $\boldsymbol{e_k}$, à laquelle nous associons la distribution de possibilité $U_V^{E_k}$ définie par $U_V^{E_k}\left(V(\mathbf{e}_k)\right) = 1$ et $U_V^{E_k}\left(\neg V(\mathbf{e}_k)\right) = 0$. Soit $U_S^{E_k}$ l'incertitude de la source S quant à $\boldsymbol{e_k}$[3]. Le problème d'apprentissage consiste à trouver la valeur de r qui minimise la fonction suivante :

$$D(r) = \sum_{k=1}^{N} d\left(Aff\left(U_S^{E_k}, {}_1U^{F_{S,D}}\right), U_V^{E_k}\right),$$

où :

- $Aff\left(U_S^{E_k}, {}_1U^{F_{S,D}}\right)$ désigne la distribution de possibilité sur E_k résultant de l'affaiblissement de $U_S^{E_k}$ par $U^{F_{S,D}}$, avec $U^{F_{S,D}}$(fiable)=1 et $U^{F_{S,D}}$(¬fiable)=1-r
- et *d*(,) est une distance entre distributions de possibilité telle que celle utilisée dans [14].

3.3 Fusion de plusieurs éléments informationnels

Pour répondre à la question de l'utilisateur sur la confiance globale associée à la variable $\boldsymbol{e}$, il faut agréger les distributions de possibilité individuelles $U^E[i_j]$, associées aux éléments informationnels i_j se rapportant à la même variable $\boldsymbol{e}$ et calculées selon le processus décrit à la section 3.2. On obtient ainsi une estimation de la distribution de possibilité globale U^E associée à $\boldsymbol{e}$.

L'approche de fusion la plus simple consiste à appliquer un opérateur de compromis aux distributions de possibilité individuelles, définissant par exemple la distribution de possibilité globale comme la moyenne des distributions de chacun des éléments informationnels. Toutefois, l'opérateur doit tenir compte de

[3] Nous faisons l'hypothèse ici qu'il n'existe pas plusieurs éléments informationnels dans lesquels une même source S pourrait s'être exprimée sur la même variable $\boldsymbol{e_k}$. Toutefois, si un tel cas se produit, l'élément informationnel le plus récent est utilisé. De plus, s'il n'existe pas d'élément informationnel dans lequel la source S s'est exprimée sur $\boldsymbol{e_k}$, nous supposons que $U_S^{E_k}(e) = U_S^{E_k}(\neg e) = 1$, c'est-à-dire que la source est ignorante par rapport à $\boldsymbol{e_k}$.

dimensions complémentaires qui permettent de réaliser des fusions plus nuancées. En particulier, il doit tenir compte du degré de corroboration de l'information fournie par les différents éléments informationnels : il doit considérer d'une part la dimension temporelle des éléments à fusionner, qui fait intervenir leur ancienneté, et d'autre part les éventuelles relations de dépendance entre les sources qui les fournissent. En effet, la confiance dans une information est très élevée si elle a été fournie avec des confiances élevées par de multiples sources indépendantes.

Prise en compte de la dimension temporelle : pondération par points d'actualité. La prise en compte de la temporalité consiste à pondérer les éléments informationnels en fonction de leur ancienneté. Celle-ci est en particulier importante dans des domaines où de nombreuses informations se succèdent très rapidement, comme par exemple dans le domaine du terrorisme : les informations initiales deviennent très vite obsolètes.
Pour cette pondération, nous proposons d'associer à chaque élément informationnel un nombre *n* de *points d'actualité* le jour de son émission, qui décroît progressivement ensuite, d'un point par jour. Un élément informationnel est donc invalidé après *n* jours. Le nombre de points initiaux doit être adapté au domaine d'application : de petites valeurs conviennent aux domaines dynamiques, subissant des changements rapides, des valeurs plus élevées aux domaines plus stables. Enfin pour éviter une dévaluation pendant une période sans information, il semble opportun qu'aucun point ne soit perdu après que la dernière information a été émise. Ces poids peuvent ensuite être utilisés dans la procédure d'agrégation, pour ajuster l'influence des éléments informationnels sur le résultat final.

Prise en compte des relations entre sources : fusion par partition. Les relations entre les sources constituent une autre information clef nécessitant d'être représentée d'une manière appropriée. Dans notre modèle, les sources peuvent avoir deux types de relations, hostilité et affinité, constituant un graphe dont les arêtes peuvent être étiquetées par ces deux valeurs. Nous supposons qu'un tel graphe est fourni par des experts du domaine. On peut alors en extraire le degré de dépendance et de redondance entre les informations rapportées : l'idée est que des sources indépendantes qui fournissent la même information donnent plus de poids à cette information qu'un ensemble de sources en relation d'affinité qui produisent naturellement une information plutôt redondante.
Nous proposons de décomposer le graphe des sources ayant rapporté des informations, afin d'identifier des sous-groupes de sources selon les contraintes suivantes : si elles sont liées, les sources d'un même sous-groupe doivent être en relation d'affinité, les sources de sous-groupes différents doivent être en relation d'hostilité. Cette décomposition peut être faite avec des méthodes telles que celles qui sont décrites dans [3]. À partir de cette décomposition du graphe, la fusion comporte alors deux étapes, détaillées ci-dessous : d'abord une fusion partielle à l'intérieur de chaque groupe amical, puis la fusion des résultats fournis par les différents groupes. Nous dénommons ce procédé la fusion par partition.
La fusion partielle s'effectue entre des sources en relation d'affinité, dont on peut attendre qu'elles fournissent généralement des éléments informationnels redondants, sans que cela indique une corroboration de l'information. On peut quantifier le degré de redondance au sein d'un groupe amical en utilisant par exemple des mesures de connectivité dans le groupe. Pour de telles sources, l'unanimité ne doit donc pas renforcer le degré de certitude comme cela serait le cas pour des sources indépendantes, ce qui conduit à utiliser des opérateurs d'agrégation de type compromis, tels que la moyenne.
L'étape suivante doit fusionner les résultats fournis par les différents groupes, qui sont donc indépendants ou en relation d'hostilité. Aussi, dans cette étape, la redondance des informations fournies peut être interprétée comme un renforcement, et des opérateurs d'agrégation de type renforcement peuvent être appliqués [18], [11]. Par ailleurs, il semble approprié de tenir compte de la cardinalité des groupes considérés : en effet, un groupe plus grand doit avoir plus d'importance qu'un groupe réduit, sans que toutefois sa contribution au résultat final domine celle du second groupe. Pour cela, nous proposons d'associer à chaque groupe un poids, défini en fonction de sa cardinalité. Une fonction linéaire amenant un comportement de domination, tel que l'influence des groupes de petite taille est négligeable dans l'agrégation, nous proposons d'utiliser la racine des cardinalités, qui atténue l'influence des groupes les plus importants sans ignorer les groupes plus petits.

Combinaison de la temporalité et des relations entre sources. Enfin, la prise en compte de la temporalité et la fusion par partition sont à combiner. Or il n'est pas pertinent d'appliquer la pondération par ancienneté avant de procéder à la fusion par partition, car dans ce cas la temporalité serait perdue. Aussi, nous proposons de

conserver cette information, par exemple sous la forme de l'ancienneté moyenne des éléments informationnels produits dans le groupe, éventuellement en définissant une temporalité floue. Celle-ci peut alors être utilisée comme pondération lors de la fusion des résultats fournis par les différents groupes.

3.4 Évaluation d'un élément informationnel rapporté par des sources successives

Nous n'avons abordé jusqu'à présent que le cas d'éléments informationnels simples dans lesquels l'information est rapportée directement par une source. Cependant les cas d'éléments informationnels plus complexes, dans lesquels l'information est rapportée par une source qui en cite une autre, voire plusieurs autres, sont relativement fréquents. Pour illustrer ceci, considérons une version légèrement différente du premier élément informationnel de notre exemple :

(i6) Selon l'agence de presse Reuters, Alexandre Bortnikov, patron du FSB, a affirmé le 29/03/2010 : « L'Émirat du Caucase pourrait être impliqué dans l'attentat de Moscou de ce matin ».

Cette nouvelle version de (i_1) semble en outre plus réaliste dans la mesure où il serait étonnant que nous ayons eu directement accès aux propos d'Alexandre Bortnikov. Pour simplifier notre exposé, nous noterons S_1 la source *Agence Reuters* et S_2 la source *Alexandre Bortnikov*, nous pouvons alors réécrire de manière schématique notre nouvel énoncé de la façon suivante (i_6)=$S_2 \rightarrow S_1 \rightarrow \boldsymbol{e}$, pour indiquer que S_2 évoque les propos de S_1 relatifs à la variable $\boldsymbol{e}$ à valeurs dans $E=\{e,\neg e\}$. avec e= « L'Émirat du Caucase est impliqué dans l'attentat de Moscou de ce matin ».

Notons que l'on peut réécrire l'énoncé de la manière suivante : (i_6)=$S_2 \rightarrow \boldsymbol{e_1}$, avec $\boldsymbol{e_1}=S_1 \rightarrow \mathbf{e}$, variable dont le domaine de valeurs est $E1=\{e_1, \neg e_1\}$, avec e_1= « Le patron du FSB a rapporté qu'un groupe du Caucase Nord pourrait être impliqué dans l'attentat de Moscou » et $\neg e_1$= « Le patron du FSB n'a pas rapporté qu'un groupe du Caucase Nord pourrait être impliqué dans l'attentat de Moscou ». Différents modèles ont été proposés pour évaluer la confiance globale que l'on peut accorder à $\boldsymbol{e}$ dans un tel cas de figure.

- Dans [6], la confiance à accorder à $\boldsymbol{e}$ est définie en fonction du degré de fiabilité que l'on pense que S_2 a (pour l'information qu'elle produit c'est-à-dire $\boldsymbol{e1}$) et du degré de fiabilité que l'on pense que S1 a (pour l'information qu'elle produit, c'est-à-dire $\boldsymbol{e}$).
- Dans [7], ce modèle est étendu en prenant en compte en plus des degrés qui reflètent l'aptitude des sources à se tromper.
- L'estimation à partir de (i_6) de l'incertitude sur E $U^E[i_6]$ peut également se faire en considérant que l'incertitude sur E, étant donné (i_6), pour S_2, représentée par la distribution de possibilité $U_{S2}{}^E[i_6]$, est égale à l'incertitude sur E étant donné (i_1), représentée par la distribution de possibilité $U^E[i_1]$, soit $U_{S2}{}^E[i_6]=U^E[i_1]$, ceci revient à considérer que la source S_2 fournit l'information $U^E[i_1]$. Pour calculer $U^E[i_6]$, il suffit alors d'affaiblir $U_{S2}{}^E[i_6]$ en utilisant le degré de fiabilité de S_2 dans le domaine de $\boldsymbol{e}$. Notons qu'à la différence des précédents modèles, celui-ci utilise le degré de fiabilité de S_2 dans le domaine de $\boldsymbol{e}$ et non pas vis-à-vis de (i_1). Ceci permet d'homogénéiser les traitements avec le cas d'un événement rapporté par une source unique, simplification qui facilite son implémentation.

4 Exemple illustratif

Dans cette section, nous appliquons le modèle qui a été présenté sur l'exemple concret introduit au début de la section 2, afin d'en illustrer le fonctionnement.

4.1 Identification des éléments informationnels pertinents

Dans l'exemple, la requête de l'utilisateur porte sur le type d'événement *attentat* dont les propriétés attendues sont les suivantes : *date* = 29/03/2010, *lieu* = Moscou, et pour lequel on cherche à évaluer la valeur de la propriété *auteur*. Nous cherchons parmi les cinq éléments informationnels (i_1) à (i_5), ceux qui sont susceptibles de nous renseigner sur cette valeur. Tous portent bien sur un événement de type *attentat*. Cependant pour (i_4), la valeur de la propriété auteur est inconnue, cet élément informationnel est donc exclu a priori de la liste des éléments pertinents. Nous avons donc Sim(i_4,q) = 0.

Nous détaillons ci-dessous, pour les quatre autres éléments informationnels le calcul des valeurs des similarités élémentaires Sim_{date} et Sim_{lieu}, à partir desquelles la similarité globale sera obtenue.

Similarités entre dates : les dates de ces quatre éléments informationnels ainsi que la date renseignée dans la requête sont toutes définies avec la même précision : à la journée. Nous avons donc $Sim_{date}(i,q) = \frac{M(D_q \cap D_i)}{M(D_i)} \quad \forall i \in \{i_1, i_2, i_3, i_5\}$. Étant donné que $date(i_1)=date(i_2)=date(i_3)=date(q)$, on a $M(D_q \cap D_i) = M(D_i) = M(D_q), \forall i \in \{i_1, i_2, i_3\}$ et donc $Sim_{date}(i,q) = 1, \forall i \in \{i_1, i_2, i_3\}$.

Pour (i_5) en revanche la situation est bien différente. Rappelons que $date(i_5)$ = 01/05/2010. Il s'agit d'une date relativement précise, définie à la journée, à laquelle on associe un sous-ensemble flou trapézoïdal dont le noyau correspond au jour en question. Il en va de même pour la date de q, à savoir le 29/03/2010. Selon le choix de modélisation des sous-ensembles que l'on retient, et plus précisément selon l'étendue du support que l'on considère pour les sous-ensembles flous correspondant à un jour particulier il pourrait y avoir intersection entre D_q et D_{i5}. Cependant il paraît raisonnable de supposer que cela ne sera pas le cas en pratique, l'écart entre les deux dates étant trop important (plus d'un mois alors que les dates sont plutôt précises, définies à la journée). Il est en effet assez peu courant de prendre un support de taille supérieure au double ou triple du noyau dont la taille est ici d'un jour. Ainsi $D_q \cap D_{i5} = \varnothing$ et donc $M(D_q \cap D_{i5}) = 0$ et ainsi $Sim_{date}(i_5, q)=0$.

Similarités entre lieux : Supposons que nous disposons de l'ontologie des lieux décrite Figure 2.

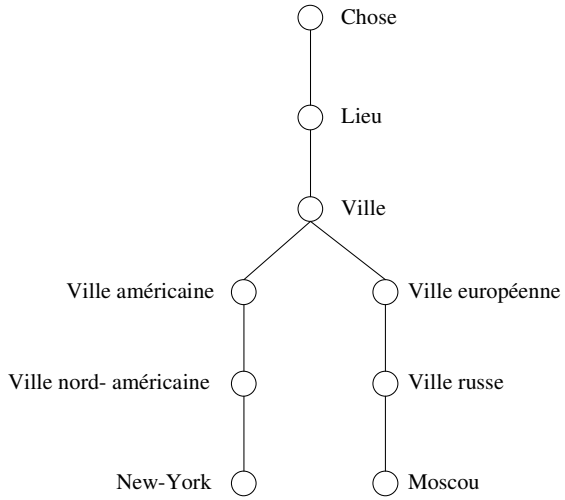


Figure 2 Ontologie des lieux

Les lieux renseignés dans les éléments (i_1), (i_2) et (i_3) sont identiques à celui de la requête, il s'agit de *Moscou*. La similarité de Wu et Palmer donne bien une similarité maximale de 1 puisque la longueur du plus court chemin entre *Moscou* et *Moscou* est nulle. Plus précisément, le plus petit généralisant de Moscou et Moscou est Moscou dont la profondeur est de 6, on obtient donc $Sim_{lieu}(i,q) = \frac{2 \times 6}{0 + 2 \times 6} = 1, \forall i \in \{i_1, i_2, i_3\}$.

Le cas de (i_5) est plus intéressant. On a lieu(i_5)=*New-York*. Le plus petit généralisant entre *New-York* et *Moscou* est le concept *Ville* dont la profondeur est de 2. Le plus court chemin entre les deux villes est de 6, ce qui conduit à $Sim_{lieu}(i_5, q) = \frac{2 \times 2}{6 + 2 \times 2} = 0.4$.

Similarité globale : l'agrégation des similarités élémentaires revient à prendre le minimum de ces similarités. On a donc :

$$Sim(i_1, q) = \min(1,1) = 1$$
$$Sim(i_2, q) = \min(1,1) = 1$$
$$Sim(i_3, q) = \min(1,1) = 1$$
$$Sim(i_5, q) = \min(0,0.4) = 0.4$$

Rappelons que l'on avait $Sim(i_4, q) = 0$. Sont donc sélectionnés comme éléments informationnels pertinents uniquement les trois premiers : (i_1), (i_2) et (i_3).

4.2 Évaluation des éléments informationnels pris isolément

Afin d'améliorer la lisibilité de cet article, les sources « Alexandre Bortnikov », « Dokou Oumarov » et « Ministre de l'Intérieur russe » sont renommées respectivement « S1 », « S2 » et « S3 ». Dans (i_1), la source d'information S1 a une confiance faible dans *e*. Dans i_2, la source d'information S2 a une confiance absolue dans ¬*e*, alors que dans i_3, la source S3 a une confiance absolue dans *e*. Ainsi, nous obtenons les distributions de possibilité sur *E* suivantes pour les sources S1, S2 et S3 dans les éléments informationnels (i_1), (i_2) et (i_3):

$$U_{S_1}^{E}[i_1](e) = 1 \qquad U_{S_1}^{E}[i_1](\neg e) = 0.7$$
$$U_{S_2}^{E}[i_2](e) = 0 \qquad U_{S_2}^{E}[i_2](\neg e) = 1$$
$$U_{S_3}^{E}[i_3](e) = 1 \qquad U_{S_3}^{E}[i_3](\neg e) = 0$$

L'évaluation de chacun des trois éléments informationnels identifiés comme pertinents consiste à affaiblir les distributions de possibilité U_{S1}^{E}, U_{S2}^{E} et U_{S3}^{E} en fonction de la fiabilité de leur source. Afin d'illustrer cette opération, nous supposons que les degrés de certitude de fiabilité des sources S1, S2 et S3 nous sont donnés et valent respectivement 0.5, 0.3 et 0.5. (Ces degrés peuvent avoir été obtenus subjectivement – en considérant par exemple pour S1 que si la source paraît effectivement compétente pour évoquer l'attentat de Moscou, son objectivité est moins évidente – ou automatiquement par apprentissage.) En appliquant la méthode décrite à la section 3.2 nous obtenons :

$$U^{E}[i_1](e) = 0.5 \times 1 + 1 - 0.5 = 1$$
$$U^{E}[i_1](\neg e) = 0.5 \times 0.7 + 1 - 0.5 = 0.85$$

$$U^E[i_2](e) = 0.3 \times 0 + 1 - 0.3 = 0.7$$
$$U^E[i_2](\neg e) = 0.3 \times 1 + 1 - 0.3 = 1$$

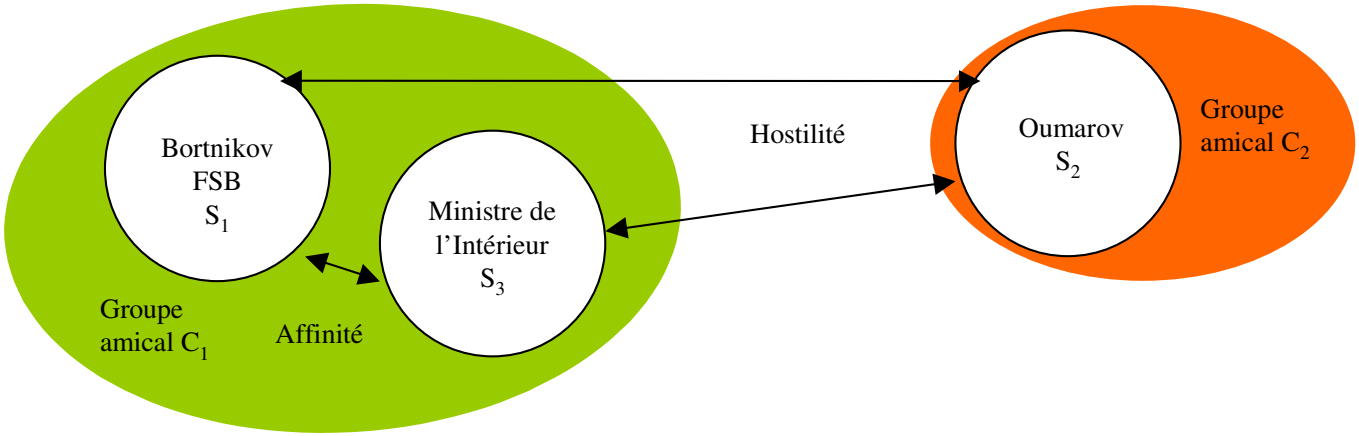
$$U^E[i_3](e) = 0.5 \times 1 + 1 - 0.5 = 1$$
$$U^E[i_3](\neg e) = 0.5 \times 0 + 1 - 0.5 = 0.5$$

4.3 Fusion des éléments informationnels

Étant donné les trois distributions de possibilités précédentes, le module de fusion produit une mesure de confiance agrégée, représentant la confiance globale sur ***e***. Une simple fusion par moyenne conduit, après normalisation, à la distribution de possibilités (1 0.87). Nous détaillons ci-dessous les résultats obtenus en prenant en compte des informations de corroboration, la temporalité des éléments informationnels et les relations entre sources.

Prise en compte de la temporalité. La prise en compte de l'ancienneté des éléments informationnels nécessite de fixer au préalable le nombre de points d'actualité accordé à un élément lors de sa publication. Dans le domaine choisi, les informations se succèdent rapidement, nous utilisons donc une valeur faible, et fixons à 5 le nombre initial de points d'actualité. Les éléments informationnels (i_1), (i_2) et (i_3) obtiennent alors respectivement les pondérations 3, 4 et 5. Le résultat de la moyenne ainsi pondérée est alors, après normalisation : (1 0.84)

Fusion par partition. La fusion par partition repose sur la décomposition de l'ensemble de sources en sous-graphes amicaux illustrée ci-dessous : le sous-groupe C_1 représenté en vert contient les sources S_1 et S_3, tandis que la source S_2 reste isolée.



S1 et S3 fournissent toutes deux la même information, affirmant que l'attentat a été commis par l'Émirat du Caucase. Comme elles appartiennent à un sous-graphe complet, correspondant à une redondance attendue maximale, ces deux éléments informationnels ne sont pas considérés comme se corroborant et renforçant la confiance en leurs affirmations. Aussi la fusion intra-groupe est-elle réalisée par un opérateur de compromis, la moyenne. Après normalisation, on obtient la distribution suivante : (1 0.68)

Le deuxième sous-graphe contient un seul élément informationnel, celui fourni par la source S_2, qui reste donc inchangé à (0.7 1).
Ces résultats sont ensuite fusionnés. Dans cet exemple, les informations fournies par les deux groupes hostiles sont contradictoires, aussi la fusion réalise un compromis et calcule la valeur finale comme la moyenne pondérée, les poids étant définis par la racine des cardinalités des sous-groupes. Après normalisation, on obtient (1 0.92)

Interprétation des résultats. Les résultats calculés avec les deux méthodes indiquent une très forte incertitude puisque la nécessité est inférieure à 0.2 dans tous les cas, voire 0.08 après la fusion par partition. Nous constatons qu'il est tout à fait possible que l'Émirat de Caucase ait commis l'attentat, mais qu'il est presque aussi possible qu'un autre auteur, inconnu, soit responsable. Quand on compare les résultats de la fusion naïve, qui consiste en une simple moyenne de toutes les valeurs disponibles, avec ceux de la fusion par partition, on observe que la fusion par partition reflète une incertitude plus forte. Celle-ci correspond en effet à un affaiblissement des sources en affinité, ce qui évite en particulier que le résultat puisse être trop influencé par des « cercles d'amis ».

4.4 Évaluation d'un élément informationnel rapporté par des sources successives

Nous ne considérons que le troisième modèle décrit à la section 3.4 et que nous appliquons à l'énoncé (i_6). Ce modèle suppose que $U_{S2}^{E}[i_6]=U^{E}[i_1]$.
Or, d'après la section 4.2, on a : $U^{E}[i_1]\ (e) = 1$,et $U^{E}[i_1]\ (\neg e) = 0.85$.
Pour calculer $U^{E}[i_6]$, il suffit ensuite d'affaiblir $U_{S2}^{E}[i_6]$ en utilisant le degré de fiabilité de S_2 dans le domaine de E. Si on considère par exemple que l'agence Reuters est totalement fiable (compétente et sincère), on peut conclure que l'incertitude $U^{E}[i_6]$ de l'agent sur E est telle que $U^{E}[i_6]=U_{S2}^{E}[i_6]=U^{E}[i_1]$, ce qui semble raisonnable ici.

5 Conclusion

L'exemple précédemment exposé illustre la démarche générale du processus de cotation selon le modèle que nous avons proposé. Toutefois, pour répondre aux besoins en veille informationnelle de nombreux services, et plus particulièrement pour doter sa partie cruciale qu'est la cotation de moyens efficaces et robustes, le modèle présenté ici doit encore être affiné. Les objectifs majeurs consistent en l'algorithmisation du modèle, de façon à bâtir un système semi-automatique de cotation de l'information, suivi d'applications spécifiques aux métiers concernés.
Pour prolonger les travaux brièvement exposés dans cet article, nous serons amenés à détailler le formalisme évoqué afin de le préciser et de le renforcer pour un traitement générique de corpus informationnels. Puis des algorithmes à déterminer, fondés sur ce formalisme, devront être développés pour aller vers le système global visé. Des épreuves d'évaluation devront également être réalisées afin de mesurer la qualité des solutions proposées. Cette phase est cruciale et nécessite la formalisation et la mise en place d'un protocole d'évaluation rigoureux qui, à notre connaissance, n'existe pas dans la littérature. Une piste que nous souhaitons explorer à ce sujet repose sur la constatation suivante : la chaîne de cotation prise dans sa globalité peut être vue comme une source d'information possibiliste dans la mesure où elle fournit pour un élément informationnel donné une mesure d'incertitude sous forme de distribution de possibilité. Ainsi, il est envisageable d'utiliser des méthodes d'évaluation de la qualité de sources ou classifieurs possibilistes telles que celles utilisées dans [10] et [12], afin d'évaluer globalement la

modélisation proposée. La principale difficulté réside dans la détermination de l'exactitude objective des informations traitées. Il faudrait en effet, idéalement, pour chaque information qui sera cotée par la chaîne de traitement proposée, savoir si celle-ci est correcte ou non. Ceci nécessite un important travail d'annotation manuelle d'un corpus de référence. La constitution d'un tel corpus pourrait bénéficier à l'ensemble de la communauté qui est confrontée à cette même difficulté.
L'évaluation individuelle de chacune des briques du composant de cotation pourrait être envisagée. Cependant cela reviendrait à multiplier les annotations et augmenterait donc encore un peu plus le coût de constitution d'un tel corpus. Aussi nous semble-t-il préférable de privilégier une évaluation globale. Notons cependant que l'extraction à partir du texte des événements à coter étant susceptible d'influer fortement sur les performances de la chaîne de cotation, une évaluation spécifique de l'extraction d'événements doit également être menée.

Remerciements
Ces travaux ont été réalisés dans le cadre du projet CAHORS, financé par l'ANR, Agence nationale de la recherche, dans le cadre de l'appel à projets CSOSG (concepts, systèmes et outils pour la sécurité globale) 2008.

6 Bibliographie

[1] BESOMBES J. & REVAULT D'ALLONNES A. : *An extension of STANAG2022 for information scoring*. International Conference on Information Fusion. Cologne, Allemagne, 2008.

[2] BESOMBES J. & CHOLVY L. : *Cotation des informations en fusion de renseignement : utilisation d'une ontologie*. In A. APPRIOU Ed. Gestion de la complexité et de l'information dans les systèmes critiques. Éditions du CNRS. Paris, France, 2009.

[3] BICHOT C.-E. & SIARRY P. *: Partitionnement de graphe : optimisation et application,* Lavoisier, 2010.

[4] BOUCHON-MEUNIER B., RIFQI M. & BOTHOREL S. : *Towards general measure of comparison of objects*. Fuzzy Sets and Systems Vol. 84(2), p 143-153, 1996.

[5] CHOLVY L. : *Information evaluation in fusion: a case study*. International Conference on Information Processing & Management of Uncertainty in Knowledge-Based Systems (IPMU). Perugia, Italie, 2004.

[6] CHOLVY L. : *Evaluation of Information reported : a model in the Theory of Evidence*. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU). Dortmund, Allemagne, 2010.

[7] CHOLVY L. : *Plausibility of Information Reported by successive sources*. International Conference on Scalable Uncertainty Management (SUM). Toulouse, France, 2010.

[8] DELAVALLADE T. & CAPET P. : *Information evaluation as a decision support for counter-terrorism*. NATO symposium on "C3I in Crisis, Emergency and Consequence Management", IST-086. Bucarest, Roumanie, 2009.

[9] DEMOLOMBE R. : *Reasoning about trust: a formal logical framework*. International Conference on iTrust. Oxford, Royaume-Uni, 2004.

[10] DESTERCKE S. & CHOJNACKI E. : *Methods for the evaluation and synthesis of multiple sources of information applied to nuclear computer codes*. Nuclear engineering and design, Vol 238(9), p. 2484-2493, 2008.

[11] DETYNIECKI M. : *Mathematical Aggregation Operators and their Application to Video Querying.* Thèse de doctorat, Université Pierre et Marie Curie. Paris, France, 2000.

[12] DRUMMOND, I., MELÉNDEZ, J. & SANDRI, S. :Assessing the aggregation of parametrized imprecise classification. *Frontiers in Artificial Intelligence and Applications*, Vol. 146, p. 227-235, 2006.

[13] DUBOIS D. & DENOEUX T. : *Pertinence et Sincérité en Fusion d'Informations*. In Rencontres Francophones sur la Logique Floue et ses Applications, p. 23-30, Cépaduès-Editions. Annecy, France, 2009.

[14] ELOUEDI, Z., MELLOULI, K. & SMETS P. : *Assessing sensor reliability for multisensor data fusion within the Transferable Belief Model*. IEEE Trans. on Systems, Man and Cybernetics B, Vol. 34(1):p. 782–787, 2004.

[15] FODOR J.C., YAGER, R. & RYBALOV, A. : *Structure of Uninorms*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 5, p. 411-427, 1997.

[16] GOUJON B. : *Uncertainty Detection for Information Extraction*. International Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria, 2009.

[17] REVAULT D'ALLONNES A. & BESOMBES J. : *Critères d'évaluation contextuelle pour le traitement automatique*. Qualité des Données et des Connaissances (QDC). Strasbourg, France, 2009.

[18] SILVERT W. : *Symmetric Summation: A Class of Operations on Fuzzy Sets*. IEEE Trans. on Systems, Man, and Cybernetics, Vol. 9, IEEE, pp. 659-667, 1979.

[19] WU Z., & PALMER M. : *Verb Semantics and Lexical Selection*. Proceedings of the Annual Meetings of the Associations for Computational Linguistics, p 133-138, 1994.

[20] YAGER R. : *Approximate reasoning as a basis for rule-based expert systems*. IEEE Trans. on Systems, Man and Cybernetics, Vol. 14, p. 636-643, 1984.