Collecte, analyse et évaluation d'informations en sources ouvertes

Yann Mombrun, Alexandre Pauchet, Bruno Grilhères, Stéphane Canu

Collecte, analyse et évaluation d'informations en sources ouvertes

Yann Mombrun[1,2], Alexandre Pauchet[2], Bruno Grilhères[1] et Stéphane Canu[2]

[1] EADS Defence & Security – System Design Center – Val de Reuil
prenom.nom@eads.com

[2] LITIS – EA 4108 – INSA de Rouen – Saint-Étienne-du-Rouvray
prenom.nom@insa-rouen.fr

Résumé : Avec les chaînes d'informations continues et le nombre toujours croissant de sites Internet, la quantité d'informations disponibles sur les sources ouvertes augmente de jour en jour. Cependant, avant d'exploiter toute information publiée, la question de sa crédibilité doit être posée. Dans cet article, nous présentons une architecture dédiée à l'exploitation des sources ouvertes, afin d'assister les utilisateurs dans l'évaluation des informations collectées.

Mots-clés : Cotation de l'information, Extraction d'information et d'événements

1 Contexte

Durant une session de recherche d'informations, l'évaluation de la qualité des résultats est un besoin récurrent. Que ce soit dans le cadre privé ou pour le renseignement, utiliser une information erronée présente des risques pouvant se révéler importants. Il est donc nécessaire d'évaluer toute information avant de l'exploiter.

On appelle *sources ouvertes* l'ensemble des gisements d'informations accessibles légalement et publiquement sur Internet, sur les radiotélévisions et sur papier. Pour les analystes du ROSO (Renseignement d'Origine Sources Ouvertes), la quantité d'information disponible et exploitable augmente de jour en jour. Cependant, seule une infime partie de ces informations peut être utilisée à des fins de renseignement.

Une information peut être fausse pour diverses raisons. Elle peut avoir été vraie mais être devenue obsolète. Elle peut se trouver erronée involontairement, par manque d'expertise de l'auteur. Enfin, il est possible qu'une information obtenue soit un leurre, une manœuvre de contre-renseignement ou de désinformation.

Plus encore que les médias traditionnels, Internet est un vecteur de *buzz*, d'effet écho. Les informations y sont souvent reprises sans vérifications. Ainsi, les analystes du ROSO doivent minutieusement évaluer un grand nombre d'informations avant de les utiliser. Cependant leurs méthodes, usuellement manuelles, semblent avoir atteint leurs limites. Aider les exploitants à coter ces informations est donc une nécessité.

Dans cet article, nous présentons une architecture dédiée à l'exploitation des sources ouvertes. L'objectif est d'assister les exploitants lors de recherches, d'analyses et surtout d'évaluations d'informations afin de faciliter l'élaboration de rapports de renseignements.

Nous nous intéressons tout d'abord aux travaux connexes autours de l'évaluation de l'information sur Internet et en sources ouvertes. Finalement, nous présentons notre plateforme dédiée à l'évaluation d'informations issues de sources ouvertes.

2 État de l'art

2.1 Évaluation de l'information sur Internet

Dans le domaine de l'information médicale, le risque engendré par l'utilisation d'informations erronées par les patients est important. L'approche principale proposée, pour l'évaluation des sites Internet, consiste à les faire valider par des tiers de confiance, des experts du domaine médical et de la recherche sur Internet (Darmoni *et al.*, 1998). Pour l'évaluation de l'information scientifique ou généraliste, il existe diverses approches. L'utilisateur peut suivre un guide méthodologique (Place *et al.*, 2006). Il peut se fier à un label de qualité apposé par un tiers de confiance (Karkaletsis *et al.*, 2006), par d'autres utilisateurs (Bizer & Cyganiak, 2009) ou par une approche mixte (Archer *et al.*, 2009a). Cette validation externe est d'ailleurs un des cas d'utilisation d'une récente recommandation du W3C, POWDER, un protocole dédié à l'annotation des ressources sur le Web (Archer *et al.*, 2009b). Cependant, la crédibilité des utilisateurs non experts est toujours sujette caution et les sites validés par des tiers de confiance restent peu nombreux, par manque de moyen.

Les travaux liés aux moteurs de recherche sont axés sur l'évaluation automatique. Divers algorithmes exploitent la structure en graphe du Web pour évaluer l'importance d'une page (Brin & Page, 1998). La qualité des liens entrant peut être également évaluée par des tiers de confiance (Wu *et al.*, 2006a). Le contexte thématique doit également être pris en compte car un site traitant d'un sujet n'intéressant que peu de personnes, sera mal référencé (Wu *et al.*, 2006b).

WISDOM est un système d'analyse de l'information disponible en japonais sur Internet (Akamine *et al.*, 2009). Il permet d'obtenir des statistiques sur les opinions exprimées dans un ensemble de documents, résultats d'une recherche thématique. Les traitements indépendants du contexte, comme l'extraction des auteurs, sont effectués à l'indexation. L'analyse des opinions autour du thème défini a lieu lors de la recherche.

Enfin, divers travaux visent à définir la qualité de l'information sous la forme d'une taxonomie de concepts, mais aucune ne fait consensus dans son intégralité. Ainsi, par exemple, (Naumann & Rolker, 2000) considèrent trois catégories différentes, suivant que les critères soient basés : (1) sur le sujet, l'utilisateur, ce sont les critères subjectifs ; (2) sur l'objet, sur l'information en elle-même, ce sont les critères objectifs et indépendants du contexte ou (3) sur le processus, c'est sont les critères dépendants du contexte dans lequel on souhaite exploiter les informations. Pour (Stvilia *et al.*, 2007), la qualité d'une information se définit principalement par (1) sa qualité intrinsèque, celle de l'information en elle-même, (2) sa qualité relationnelle, par rapport aux autres

et (3) sa qualité réputationnelle, celle de sa source. Les différentes modélisation ont pour constante la nécessité de ventiler les différents critères d'évaluation utilisés sur l'ensemble des composantes afin d'obtenir une qualification globale.

2.2 Évaluation de l'information issue de sources ouvertes

L'OTAN a publié des manuels et recommandations sur lesquels se fondent les armées pour leurs procédures d'évaluation. (NATO, 1997) décrit la cotation de manière théorique, avec deux critères : fiabilité de la source et crédibilité de l'information. (NATO, 2001) énonce des principes à respecter pour exploiter les informations issues de sources ouvertes. Enfin, (NATO, 2002) présente comment exploiter Internet pour le ROSO. Cependant, cette procédure comporte de trop nombreux critères devant être vérifiés *manuellement*. De plus, ces critères ne tiennent pas compte de nombreux types de sites, d'intérêt opérationnel, comme les blogs, forums, wikis ou réseaux sociaux.

EBR a proposé un système automatisé dédié à l'exploitation d'Internet pour l'élaboration de renseignement (Noble, 2004). Des événements sont extraits des pages. La crédibilité d'un événement est évaluée en se fondant sur trois critères : la tonalité du texte, la fiabilité historique de sa source et sa cohérence vis-à-vis de faits confirmés et des autres événements extraits. Cependant, pour éviter que la base des événements extraits, utilisée pour évaluer leur cohérence, ne s'écarte de la vérité terrain, il est nécessaire d'intégrer l'exploitant par un processus de validation (Noble, 2005).

Différents travaux visent à formaliser la méthodologie de (NATO, 1997) afin de l'automatiser (Cholvy & Nimier, 2003). Elle reste cependant difficilement applicable aux sources ouvertes. (US Army, 2006) a adapté cette méthodologie au ROSO : le critère de crédibilité de l'information a ainsi été redéfini. Plus récemment, la compétence de la source et la plausibilité de l'information ont été introduits et la définition de la fiabilité d'une source a été restreinte (Revault d'Allonnes & Besombes, 2009).

Le projet CAHORS propose une plateforme dédiée à l'exploitation des informations issues d'Internet (Delavallade & Capet, 2009). Les traitements sont séparés en deux phases. La première est consacrée à la collecte d'événements : acquisition de documents sur Internet, filtrage et extraction d'événements dans du contenu non structuré. La seconde est dédiée à l'évaluation et l'exploitation des informations extraites.

2.3 Discussion

Notre système doit assister l'analyste chargé d'évaluer des informations. Une des priorités est l'automatisation, lorsque c'est possible, des extractions et des vérifications proposées par les méthodologies précitées, du domaine militaire comme civil.

En général, les travaux relatifs à l'évaluation de l'information dédiée au renseignement ne sont pas adaptés aux particularités du ROSO. Dans le cas des sources ouvertes, il y a de nombreuses ambiguïtés dans les concepts manipulés par le standard OTAN : comment définir la source d'une information issue d'un document sur un site Web, comment s'assurer de l'indépendance de deux sources, comment prend-on en compte l'évolution d'une source (changement de propriétaire du site, par exemple), *etc.* Ces

questions se posent relativement peu pour les autres catégories de renseignement et sont donc rarement prises en compte dans les travaux.

EBR a fait des propositions intéressantes pour répondre à certaines de ces spécificités, par exemple en définissant une typologie des sources (presse, université, *etc.*) et de leurs auto-évaluations (témoin oculaire, déduction d'après un témoignage, *etc.*) ou en associant des fiabilités différentes aux entités et attributs détectés dans un même document en fonction de la tonalité de la phrase. CAHORS définit notamment l'information à évaluer comme étant une entité de type événement, à laquelle on associe non seulement sa source au sens habituel, mais également sa source de publication relayant cette information.

Même si différents systèmes proposent la mise en commun de résultats d'évaluations, en utilisant notamment POWDER, les sources (et ressources) évaluées ne sont pas assez nombreuses pour que les analystes puissent dans une application réelle utilise ces informations. Heureusement, d'autres systèmes à base de recommandation ou de partage — comme les sites de réseau social ou de partage de marque-pages — sont plus fournis en utilisateurs et en informations qui pourront être exploitées aider lors de l'analyse d'une source ou d'un document.

Les méthodologies manuelles comportent nombre de critères qui peuvent être automatisés — et qui parfois le sont dans certains systèmes. Parmi ceux-ci, certain sont intéressants pour les analystes. Ces calculs automatisés pourront être répartis à la manière de ce qui est fait dans WISDOM. Dans ce système, les premiers résultats d'une recherche sont traités à la volée. Ainsi, Lors de l'indexation, seuls les traitements de bas niveau sont appliqués ; alors que les traitements avancés sont effectués à chaque fois qu'un document fait parti des résultats d'une requête utilisateur.

Les modélisations de la qualité de l'information étant nombreuses, nous avons choisi de peu contraindre le système à un formalisme donné. Seules les catégories de plus haut niveau du modèle de (Stvilia *et al.*, 2007) sont utilisées : *qualité intrinsèque*, *qualité relationnelle* et *qualité réputationnelle*. Celles-ci seront définies ci-après. Nous ne proposons pas d'associer à ces qualités une valeur calculée automatiquement, mais plutôt un ensemble d'indicateurs permettant à l'analyste de juger l'information vis-à-vis de chacune de ces catégories.

Finalement, une telle application nécessite de nombreuses interactions avec les utilisateurs. Ils ont besoin de visualiser les documents et les annotations qu'il portent : notamment les événements détectés, les métadonnées extraites ou les critères d'évaluation calculés. De plus, il est nécessaire de laisser la possibilité de modifier, supprimer ou valider des annotations. Tout calcul ou extraction automatique doit pouvoir être corrigé par l'utilisateur. Le système que nous proposons a pour objectif d'assister les opérationnels dans leurs tâches d'évaluation et d'exploitation des informations issues de sources ouvertes, pas de les remplacer. En effet, le risque engendré par l'utilisation d'un système totalement automatisé est trop grand dans le cadre d'une application au renseignement.

3 Exploitation d'informations issues de sources ouvertes

Le système que nous proposons est dédié à l'évaluation et l'exploitation des informations pour le ROSO. Il s'appuie sur les principes de Stvilia *et al.* (qualités intrisèque,

relationnelle et réputationnelle). Les concepts manipulés sont définis ci-après.

Les *sources* émettent des informations. Elles ont une structure hiérarchique : une source *fille* est contenue par une source *mère*. Ainsi Internet, un site Web ou un document tiré de ce site sont des sources. C'est utile pour inférer des évaluations : on aura un *a priori* négatif sur tout article – *document* – publié par un journal considéré non fiable. On extrait de l'*information* de ces documents : entités nommées, événements et relations les unissant. Sans fusion inter-documents, toute information provient d'une unique document, la *source de publication*. Mais celle-ci n'est pas nécessairement la *source de l'information*, dans le cas d'une citation par exemple. Ces sources peuvent être multiples puisque nous fusionnons les informations issues d'un même document.

3.1 Qualité intrinsèque

La qualité intrinsèque d'une information s'évalue indépendamment du contexte. On trouve au sein de cette mesure de qualité des critères liés au document (comme la qualité de l'écriture), des critères mêlant informations et document (par exemple la cohérence des informations extraites) et des critères liés à l'information (notamment la précision des algorithmes utilisés pour l'extraction).

La qualité intrinsèque n'évolue pas avec le temps, car elle est indépendante du contexte. La plupart des critères qui la composent peuvent être calculés automatiquement. Il est donc possible des les évaluer une seule fois, lors de l'acquisition des informations.

3.2 Qualité relationnelle

La qualité relationnelle d'une information se juge en regard des informations préalablement validées. Les critères utilisés sont dépendants des entités, événements et relations extraits. Il est possible, par exemple, d'évaluer la similarité ou la plausibilité d'un événement en le comparant à un ensemble d'événements validés (Saval *et al.*, 2009).

Comme le contexte et le contenu de la base de connaissance évoluent, les critères de qualité relationnelle doivent être réévalués régulièrement, au moment de l'exploitation par exemple. L'historique des valeurs doit être conservé car il est utile à la cotation.

3.3 Qualité réputationnelle

La qualité réputationnelle d'une information dépend uniquement de sa source. Dans notre cas, il s'agit à la fois de la source de publication et des sources de l'information.

Les critères utilisés sont volatiles. Leur valeur doit être connue au moment de la collecte. Par exemple, le propriétaire d'un nom de domaine peut avoir changé. Mais d'autres critères nécessitent d'être réévalués car leur évolution influe sur la cotation (par ex., le nombre de liens pointant vers un document).

3.4 Architecture

Pour évaluer la qualité intrinsèque, relationnelle et réputationnelle des informations, certains traitements doivent être effectués lors de la collecte des informations (voir figure 1) et d'autres lors de toute consultation ultérieure. C'est avec l'aide de toutes les

informations extraites (événements, métadonnées et critères) que l'exploitant pourra juger de la qualité d'une information.

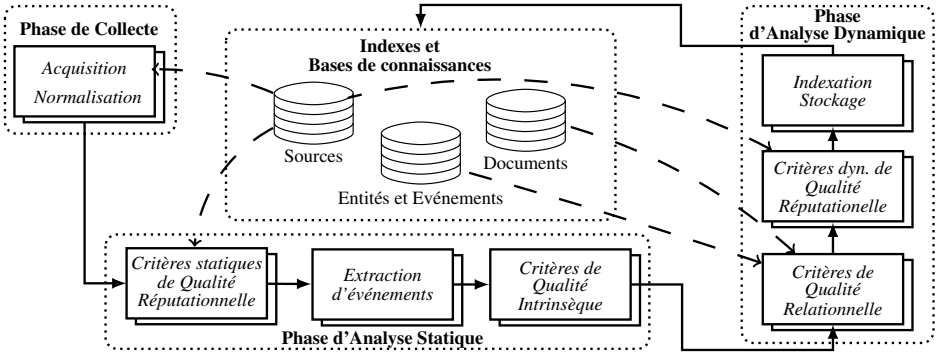

FIGURE 1 – Chaîne de traitements principale constituée de trois phases.

Notre application se fonde sur l'architecture d'intégration orientée service WebLab [1] qui définit un modèle d'échange et des interfaces de services pour le traitement des documents multimédia (Giroux *et al.*, 2008). Chaque service est capable d'annoter sémantiquement un document au format WebLab afin de l'enrichir.

L'exploitant peut sélectionner certaines sources dans la base de connaissance avant de lancer les traitements. Celles-ci auront été préalablement identifiées et éventuellement annotées (description, catégorisation, évaluation manuelle, *etc.*).

La phase de *collecte* consiste en l'acquisition de documents. Le service composite acquiert les données d'une source Web puis les convertit en document WebLab tenant compte de la structure (menus, publicités, commentaires, *etc.*).

La phase d'*analyse statique* consiste en l'ensemble des services d'extractions d'information non contextuels. Ils sont présentés en trois services composites sur la figure 1 : le calcul de qualité réputationnelle statique, l'extraction d'événements et le calcul de la qualité intrinsèque. Le premier extrait des informations sur la source de publication. Le suivant est une suite d'AGATE, une plateforme dédiée au suivi des catastrophes naturelles, utilisant des techniques de traitement automatique de la langue dans les flux de presse. Ces travaux ont été enrichis pour extraire des informations comme les auteurs ou les citations et affecter des attributs aux entités extraites (Saval & Mombrun, 2010). Le dernier service permet le calcul de critères pour l'évaluation de la qualité intrinsèque.

La phase d'*analyse dynamique* comprend les traitements devant être effectués lors de la collecte *et* de l'exploitation des documents, car dépendants du contexte. Ces services sont présentés sous la forme de trois agrégats : le calcul de la qualité réputationnelle dynamique, celui de la qualité relationnelle et l'indexation/stockage. En plus des calculs de différents critères, le stockage est effectué pour historiser les évaluations.

1. `http://weblab-project.org`

3.5 Comparaison avec les travaux connexes

Le système que nous proposons pour l'évaluation de l'information est similaire dans son approche à ce que proposent CAHORS et EBR. La principale différence est liée à l'acquisition et au filtrage des informations. Ces deux systèmes semblent connectés au Web en général, sans sélection de sources. Cela peut engendrer du bruit dans le système. Pour s'en protéger, CAHORS a introduit un filtrage après la collecte, *via* une méthode à base de classification. Nous avons choisi de gérer les sources en amont, et de n'acquérir qu'une source sélectionnée afin de réduire ce bruit. Cela correspond au fonctionnement de différentes cellules de renseignement, où il n'est pas rare que la personne en charge de la sélection des sources d'intérêt soit différente de celle qui traitera les données collectées.

La répartition des traitements est similaire de celle de WISDOM. Cependant, dans notre système nous avons choisi de stocker les résultats des analyses dynamiques dans les index. Cela permet de conserver un historique utile pour l'analyste lors de l'évaluation. Mais cela permet également de réduire la charge du système en évitant de réévaluer trop fréquemment un même document.

4 Conclusion et perspectives

Dans cet article, nous avons présenté les fondements d'une application dédiée à l'exploitation des sources ouvertes permettant l'évaluation des informations collectées.

Dans la continuité de ces travaux, nous allons modéliser les bases de connaissance dédiées à la description des événements, entités, sources et critères d'évaluation. Le système dont l'architecture a été présentée ici doit ensuite être implémenté. Les services manquants seront développés et intégrés à partir de travaux issus d'AGATE.

Finalement, nous procéderons à l'évaluation du système. L'évaluation technique de certains services composites, notamment l'extraction d'événements, sera conduite à l'aide de campagnes existantes, probablement ACE 5 (Doddington *et al.*, 2004) et ESTER 2 (Galliano *et al.*, 2005). De plus, une évaluation de l'ergonomie et de l'acceptabilité de notre système sera effectuée par des analystes du ROSO.

Références

AKAMINE S., KAWAHARA D., KATO Y., NAKAGAWA T., INUI K., KUROHASHI S. & KIDAWARA Y. (2009). WISDOM : a web information credibility analysis system. In *ACL-IJCNLP '09*, p. 1–4.

ARCHER P., FERRARI E., KARKALETSIS V., KONSTANTOPOULOS S., KOUKOURIKOS A. & PEREGO A. (2009a). Quatro Plus : quality you can trust ? In *ESWC'09*.

ARCHER P., SMITH K. & PEREGO A. (2009b). Protocol for Web Description Resources (POWDER). `http://www.w3.org/TR/powder-dr`. W3C Recommendation 2009.

BIZER C. & CYGANIAK R. (2009). Quality-driven information filtering using the WIQA policy framework. *Journal of Web Semantics*, **7**(1), 1 – 10.

BRIN S. M. & PAGE L. E. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, **30**(1-7), 107–117.

CHOLVY L. & NIMIER V. (2003). Information evaluation : discussion about Stanag 2022 recommentations. In *RTO-MP-IST-040'03*.

DARMONI S. J., LEROUX V., DAIGNE M., THIRION B., SANTAMARIA P. & DUVAUX C. (1998). Critères de qualité de l'information de santé sur l'Internet. *Informatique et Santé*, **10**, 162–174.

DELAVALLADE T. & CAPET P. (2009). Information evaluation as a decision support for counter-terrorism. In *RTO-MP-IST-086'09*.

DODDINGTON G., MITCHELL A., PRZYBOCKI M., RAMSHAW L., STRASSEL S. & WEISCHEDEL R. (2004). The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of LREC'04*, volume 4, p. 837–840.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F. & GRAVIER G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Interspeech'05*.

GIROUX P., BRUNESSAUX S., BRUNESSAUX S., DOUCY J., DUPONT G., GRILHÈRES B., MOMBRUN Y. & SAVAL A. (2008). WebLab : an integration infrastructure to ease the development of multimedia processing applications. In *ICSSEA'08*.

KARKALETSIS V., PEREGO A., ARCHER P., STAMATAKIS K., NASIKAS P. & ROSE D. (2006). Quality labeling of web content : The Quatro approach. In *MTW'06*.

NATO (1997). Annex to STANAG 2022 (Edition 8). Information handling services.

NATO (2001). Open source intelligence Handbook.

NATO (2002). Intelligence exploitation of the Internet.

NAUMANN F. & ROLKER C. (2000). Assessment methods for information quality criteria. In *IQ'00*, p. 148–162.

NOBLE D. F. (2004). Assessing the reliability of open source information. In *Fusion 04*, volume II, p. 1172–1178.

NOBLE D. F. (2005). Fusion of open source information. In *Fusion'05*, volume II.

PLACE E., KENDALL M., HIOM D., BOOTH H., AYRES P., MANUEL A. & SMITH P. (2006). Internet Detective : wise up to the Web. Intute Virtual Training Suite.

REVAULT D'ALLONNES A. & BESOMBES J. (2009). Critères d'évaluation contextuelle pour le traitement automatique. In *QDC'09*, p. 9–16.

SAVAL A., BOUZID M., BONDU E. & BRUNESSAUX S. (2009). Risk detection and situation awarness : From anyone to everyone. In *S4'09*.

SAVAL A. & MOMBRUN Y. (2010). Agate : Plate-forme de suivi de catastrophes sur le web. In *RFIA'10*.

STVILIA B., GASSER L., TWIDALE M. & SMITH L. (2007). A framework for information quality assessment. *JASSIST*, **58**(12), 1720–1733.

US ARMY (2006). Field Manual Interim 2-22.9 - Open Source Intelligence.

WU B., GOEL V. & DAVISON B. (2006a). Propagating trust and distrust to demote Web spam. In *MTW'06*.

WU B., GOEL V. & DAVISON B. (2006b). Topical TrustRank : using topicality to combat Web spam. In *WWW'06*, p. 63–72.