12ème atelier sur la Fouille de Données Complexes (FDC)
Extraction et Gestion des Connaissances (EGC 2015)

Organisateurs :

- Cyril de Runz (CReSTIC, Université de Reims Champagne Ardenne)
- Cécile Favre (ERIC, Université Lyon 2)
- Germain Forestier (MIPS, Université de Haute-Alsace)
- Camille Kurtz (LIPADE, Université Paris Descartes)

Le groupe de travail "Fouille de Données Complexes"

La douxième édition de l'atelier sur la fouille de données complexes est organisée par le groupe de travail EGC "Fouille de Données Complexes". Ce groupe de travail rassemble une communauté de chercheurs et d'industriels désireux de partager leurs expériences et problématiques dans le domaine de la fouille de données complexes telles que le sont les données non-structurées (ou faiblement), les données obtenues à partir de plusieurs sources d'information ou plus généralement les données spécifiques à certains domaines d'application et nécessitant un processus d'extraction de connaissance sortant des itinéraires usuels de traitement.

Les activités du groupe de travail s'articulent autour de trois champs d'action progressifs :

- l'organisation de journées scientifiques une fois par an (vers le mois de juin) oú sont présentés des travaux en cours ou plus simplement des problématiques ouvertes et pendant lesquelles une large place est faite aux doctorants ;
- l'organisation de l'atelier "Fouille de Données Complexes" associé à la conférence EGC qui offre une tribune d'expression pour des travaux plus avancés et sélectionnés sur la base d'articles scientifiques par un comité de relecture constitué pour l'occasion ;
- la préparation de numéros spéciaux de revue nationale, dans lesquels pourront être publiées les études abouties présentées dans un format long et évaluées plus en profondeur par un comité scientifique.

Contenu scientifique de l'atelier

Nous avons reçu cette année 10 propositions originales, chacune d'elle a été relue par au moins deux évaluateurs. Pour la majorité des propositions nous avons été en mesure de proposer trois rapports d'experts afin d'offrir un processus scientifique constructif aux auteurs. Vu la qualité soumissions et l'intérêt des discussions qu'elles pouvaient susciter au sein de l'atelier, nous avons retenu cette année l'ensemble des propositions.

Les articles qui vous sont proposés cette année dans les actes qui suivent explorent une grande variété de complexités, aussi bien dans les données que dans les processus de fouille envisagés.

Remerciements

Les responsables de l'atelier souhaitent remercier vivement toutes les personnes ayant contribué à la tenue de cet atelier. En particulier :

- les auteurs pour la qualité de leurs contributions constituant la base essentielle de discussions fructueuses ;
- les membres du comité de programme et plus généralement tous les relecteurs de cet atelier dont le travail d'évaluation était crucial pour assurer la qualité de l'atelier ;
- les organisateurs d'EGC 2015 qui ont mis en place l'environnement et les moyens pour la réussite des ateliers.

Nous remercions enfin vivement les présidents : Jérôme Darmont le président du comité de programme, Benoît Otjacques et Thomas Tamisier les co-présidents du comité d'organisation d'EGC 2015.

Cyril de Runz
CReSTIC
Université de Reims Champagne Ardenne

Cécile Favre
ERIC
Université Lyon 2

Germain Forestier
MIPS
Université de Haute-Alsace

Camille Kurtz
LIPADE
Université Paris Descartes

Membres du comité de lecture

Le Comité de lecture est constitué de :

- Hanane Azzag (LIPN, Univ. Paris 13)
- Nicolas Béchet (IRISA, Univ. Bretagne Sud)
- Aurélie Bertaux (Le2i, Univ. Bourgogne)
- Alexandre Blansché (Univ. Lorraine)
- Omar Boussaid (ERIC, Univ. Lyon 2)
- Guillaume Cleuziou (LIFO, Univ. Orléans)
- Cyril De Runz (CReSTIC, Univ. Reims)
- Sami Faiz (INSAT, Université de 7 Novembre de Carthage)
- Cécile Favre (ERIC, Univ. Lyon 2)
- Germain Forestier (MIPS, Univ. Haute Alsace)
- Pierre Gançarski (iCube, Univ. Strasbourg)
- Mehdi Kaytoue (INSA Lyon)
- Camille Kurtz (LIPADE, Univ. Paris 5)
- Mustapha Lebbah (LIPN, Univ. Paris 13)
- Arnaud Martin (IRISA, Univ. Rennes 1)
- Florent Masséglia (AxIS-Inria Sophia Antipolis)
- Nedra Mellouli (LIASD, Univ. Paris 8)
- Florence Mendes (Le2i, Univ. Bourgogne)
- Christophe Nicolle (Le2i, Univ. Bourgogne)
- Yoann Pitarch (IRIT, Univ. Toulouse 3)
- Chedy Raïssi (LORIA, INRIA)
- Mathieu Roche (UMR TETIS, CIRAD)
- Anna Stavrianou (Xerox Research Centre Europe)
- Abdelmalek Toumi (LabSTICC, ENSTA Bretagne)
- Cédric Wemmert (iCube, Univ. Strasbourg)
- Djamel Zighed (ERIC, Univ. Lyon 2)

TABLE DES MATIÈRES

Session Images

Session Flux de Données

Session Texte

Session Représentation des Connaissances / Ontologie

Corrélation optique pour l'identification de cibles radar

Abdelmalek Toumi*, Ali Khenchaf*

* Lab-STICC UMR CNRS 6285, ENSTA Bretagne,
2 rue François Verny, 29806 Brest Cedex 9, France
toumiab,khenchal@ensta-bretagne.fr,
http://www.ensta-bretagne.fr

Résumé. Ce présent travail s'intègre dans la problématique générale du traitement et de l'exploitation de l'information dans le domaine de la reconnaissance et l'identification de cibles aériennes à partir des images radar. Nous nous intéressons dans ce travail aux méthodes de corrélation optique, particulièrement à la corrélation monovoie et la corrélation multivoie utilisées pour l'identification d'objets. Ces différentes techniques de corrélation optique se distinguent généralement par la nature des filtres qu'elles proposent. Ces derniers sont utilisés pour réaliser la corrélation optique et ainsi la phase de reconnaissance. L'objectif de ce travail est de proposer une méthode de reconstruction de filtre optique caractérisant l'information contenue dans une ou plusieurs images radar en assurant une taille réduite de la base d'images utilisée comme base de références (base d'apprentissage). Pour le critère de corrélation et ainsi la prise de décision (classification), nous proposons une mesure de similarité basée sur un critère énergétique défini par la somme d'énergie maximale normalisée dans le plan de corrélation utilisé.

1 Introduction

La reconnaissance d'objets dans des images radar (satellites, optiques, ...) reste une tâche essentielle pour ce qui est des systèmes de surveillance de zones sensibles civiles ou militaires (Aériennes ou sous-marines). Les exigences imposées à ces systèmes n'ont pas cessé de croître au fil du temps. Elles concernent en particulier leur robustesse, leur fiabilité et leur sûreté de fonctionnement même en milieu hostile. Dans le cadre des radars imageurs, l'un des problèmes fondamentaux est principalement lié d'une part au dispositif de mesure opérant généralement dans un environnement évolutif difficile à maîtriser et d'autre part à la nature des données qui sont incertaines, incomplètes et volumineuses.

Actuellement, les systèmes radars à synthèse d'ouverture (SAR/ISAR) ont offert la possibilité de générer des images radars avec une résolution suffisante en mode fin et/ou ultra fin (en quelques mètres), et ainsi ils apportent une aide précieuse pour des applications de surveillance et de contrôle. L'exploitation des différentes images radars mono et multicapteurs ont contribué ainsi à détecter et à suivre des cibles noyées dans un environnement naturel dynamique, évolutif et même aléatoire.

Dans le cadre de ce présent travail, nous nous intéressons aux images radars, dites ISAR (Inverse Synthetic Aperture Radar). L'architecture proposée repose en première phase sur l'analyse des images ISAR caractérisant des cibles aériennes non coopératives en s'appuyant sur une des approches de l'analyse d'images, et en deuxième phase sur la reconnaissance (classification). Pour la phase d'analyse d'images radars, l'idée la plus répandue pour exploiter et extraire de l'information contenue dans une image s'intéresse généralement au niveau des transitions brusques de luminances ou d'homogénéité caractérisant ainsi des contours ou des régions représentant l'objet. En effet, dans cette phase d'analyse, et afin d'assurer une invariance des descripteurs aux changements géométriques des différents contours et régions, une étape de codage (modélisation) est souvent menée pour constituer une signature finale (vecteur de descripteurs) décrivant l'information utile de l'objet à reconnaître. Dans ce contexte, plusieurs méthodes de modélisation, qualifiées de *bas niveau* peuvent être citées telles que les descripteurs de Fourier et ses variantes, les Moments Invariants (Belkhaoui et al., 2012), et autres. C'est ainsi que plusieurs travaux se sont intéressés à l'extraction des différents types de descripteurs nécessaires pour une classification robuste et satisfaisante en s'appuyant principalement sur les techniques de segmentation. Par ailleurs, les approches de segmentation, les plus répandues, appliquées aux images ISAR souffrent d'une grande sensibilité liée au bruit qui entache les données radars et ainsi les images ISAR. A l'inverse, les techniques, dites globale, qui considèrent l'image dans sa globalité présentent une meilleure alternative. Plusieurs travaux se sont focalisés sur une représentation globale de l'image qui peut être représentée sur son espace d'origine ou sur un nouveau espace transformé tel que l'espace polaire ou log-polaire (Toumi et al., 2012; Kim et al., 2005; Toumi et Khenchaf, 2010). Les techniques de la corrélation optique s'insèrent ainsi dans cette famille de méthodes.

La corrélation optique est un domaine de recherche actif depuis les premiers travaux de Robertson (Robertson, 1941) qui a montré le principe de l'amnioscopie (Shadow Casting en Anglais) permettant de réaliser en éclairage spatialement incohérent, le produit de corrélation de deux images. En 1964 et depuis l'apparition du laser, d'autres travaux élargissent cette perspective, nous citons les travaux de Vander Lugt (Lugt, 1966) où il proposa le principe de corrélateur holographique en éclairage spatialement cohérent. Par la suite, en 1966, Weaver et Henderson (Henderson et Weaver, 1966) donnèrent le principe de corrélateur à base de la Transformée de Fourier conjointe fonctionnant aussi en éclairage spatialement cohérent.

Le grand avantage inhérent de la corrélation l'optique est qu'elle permet d'obtenir le produit de corrélation de deux images (utilisé comme critère de reconnaissance) de manière quasiment instantanée. Cette propriété a donné beaucoup d'avantages à la corrélation optique pour des solutions embarquées et temps réels. Il est à noter, que cette technique a rencontré une seconde objection à l'utilisation de la corrélation dans la reconnaissance de formes que nous abordons dans ce présent travail. Nous nous intéressons à la qualité de l'opération de corrélation optique dans le cadre de la reconnaissance de cibles radar à partir des images ISAR. Avant de présenter la démarche et la méthode proposée, nous commençons par présenter dans la section suivante les images ISAR utilisées.

2 Imagerie à ouverture de synthèse

Un objet présent dans une image à ouverture de synthèse est caractérisé par sa réflectivité définie comme la distribution spatiale de ses points brillants (Lazarov, 2012; Chen et al., 2008).

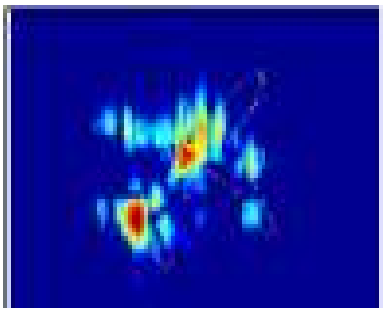


FIG. 1 – *Exemple d'une image ISAR d'un Airbus 310*

Cette réflectivité est exploitée par les techniques de reconstruction des images ISAR telles que la technique à base de la Transformée de Fourier et les méthodes à haute résolution (MUSIC, ESPRIT, ...) (Lazarov, 2012). En effet, l'image ISAR est obtenue par l'analyse temporelle des positions des points brillants suivant l'axe de visée du radar, et l'analyse fréquentielle des points brillants suivant l'axe azimutal (Lazarov, 2012). Nous utilisons dans ce travail une base d'images ISAR réelles issues d'une compagne de mesures. Par ailleurs, dans ce travail, nous ne traitons pas les techniques de reconstruction et les problèmes inhérents à l'imagerie radar. Un exemple d'une image ISAR est donné par la figure 1.

Dans l'exemple illustré par la figure 1 nous signalons que les techniques de segmentation d'image selon les différentes approches d'indexation classiques nécessitent une forte adaptation pour obtenir des résultats satisfaisants. Plusieurs travaux se sont intéressés à la forme en s'appuyant sur la détection de contours par des techniques de traitement d'images peuvent être consultés dans (Belkhaoui et al., 2012; Toumi et al., 2007).

Les différentes techniques proposées dans la littérature pour le traitement des images ISAR pour la reconnaissance de cibles aériennes, sont organisées autour de deux familles. La première famille intègre les approches de segmentation généralement proposées pour extraire des zones d'intérêt (segmentation en régions) ou des contours caractérisant les différentes transitions dans une image. La deuxième famille d'approches dites globales permet de traiter l'image dans sa globalité. Cette dernière famille vise à élaborer une méthode de recherche basée sur une comparaison de l'objet à reconnaître à un ensemble de références introduisant le maximum d'information avec une granularité locale. Ces dernières approches ont donné des résultats très satisfaisants. Dans ce sens, nous pouvons citer les travaux de Kim et al (Kim et al., 2005), et les récents travaux de Toumi et al (Toumi et al., 2012; Toumi et Khenchaf, 2010) sur l'imagerie polaire et log-polaire. C'est dans cette famille d'approches que s'intègre la corrélation optique qui fait l'objet du présent papier et qui est présentée dans la section suivante.

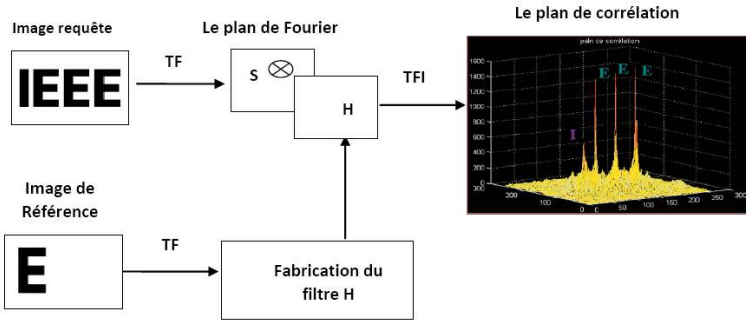


FIG. 2 – *Principe de la corrélation optique*

3 Corrélation optique

3.1 Principe

Le principe de la corrélation optique (Savvides et Kumar, 2003; Savvides et Vijaya Kumar, 2003) consiste à comparer (corréler) une image ISAR à reconnaître (requête) avec une base de références. Cette comparaison est assurée par un passage du domaine temporel au domaine fréquentiel d'une image donnée. Dans ce passage, chaque image est représentée par son spectre. Ainsi, ce spectre est multiplié par un filtre construit lui-même à partir du spectre de l'image de référence. Cette reconstruction de filtre est une étape clé dans la phase de reconnaissance que nous développerons dans la section suivante. Pour la phase de décision, généralement, une transformée de Fourier inverse mène à une décision binaire caractérisée par la présence ou l'absence d'un pic dans le plan de corrélation (le plan de la décision) (Kumar et al., 2006; Alfalou, 1999; Miller et Woodruff, 1995).

Nous illustrons dans l'exemple de la figure 2 le principe de la corrélation optique monovoie où nous cherchons à reconnaître les occurrences de la lettre **E** sur une image contenant le mot **IEEE** à l'aide d'une mono-corrélation et ceci à l'aide d'un filtre construit à partir d'une image de référence de la lettre E.
La corrélation optique présentée dans la figure 2 peut être facilement réalisée par un montage optique connu sous le nom du filtrage $4F$ proposé initialement par Vander Lugt (Lugt, 1966). Cette corrélation se base principalement sur la transformée de Fourier que le lecteur pourrait consulter (Alfalou et Brosseau, 2010; Alfalou, 1999) pour plus de détails.

En effet, l'optique réalise très bien la tâche de reconnaissance grâce à un *détecteur* présenté par le filtre caractéristique des objets (Alfalou et Brosseau, 2010; Savvides et Vijaya Kumar, 2003). Donc, pour deux images similaires, la corrélation de leurs deux filtres respectifs permettra de détecter ainsi la présence et la position de la cible en créant un *pic* qui est l'expression optique du maximum de la fonction de corrélation.

Le critère de corrélation et ainsi la prise de décision dans un système de reconnaissance de formes est basé sur un facteur énergétique PCE (Peak to Correlation Energy) lié à la localisation de l'énergie maximale (pic) (Alfalou et Brosseau, 2010; Kumar et al., 2006) dans le plan

de corrélation. Ce facteur correspond à l'énergie maximale normalisé sur l'énergie totale du plan de corrélation et qui est défini par l'équation (1) :

$$PCE = \frac{Energie\ maximale}{Energie\ totale} = \frac{|\max(f \otimes g)|^2}{|\int\int(f \otimes g)|^2} \tag{1}$$

Où f désigne le filtre de l'image à reconnaitre, g désigne le filtre de l'image de référence et le symbole $\otimes$ représente le produit de convolution.

Il est à noter que la corrélation optique englobe différentes phases de traitement permettant de réaliser la tâche de reconnaissance. L'intérêt dans ce travail porte sur la construction d'un filtre adapté pour la reconnaissance de cibles radar. Plusieurs filtres sont proposés dans la littérature pour répondre aux exigences applicatives dans des contextes bien définis et bien différents de l'application traitée dans le présent travail.

3.2 Filtres optiques

La première phase primordiale pour la corrélation optique est basée sur la multiplication du spectre d'une image de cible (inconnue) par un ensemble de filtres issus d'une base de références. Dans cette dernière, chaque filtre représente une image de la base d'apprentissage. L'objectif est de construire un filtre capable d'assurer une corrélation optimale entre deux images similaires. Nous présentons dans ce qui suit les différents filtres utilisés pour la reconnaissance de cibles.

3.2.1 Filtre classique adapté

Le filtre adapté, noté H est le plus connu des filtres utilisés. Il est équivalent au conjugué de la transformée de Fourier calculé sur une image de référence I dont sa formule est donnée par $H = TF^*(I)$. Une fois le filtre obtenu, une binarisation (équation (2)) est appliquée pour l'adapter à une corrélation binaire.

$$\begin{cases} R(H) > 0 \rightarrow H = 1 \\ R(H) < 0 \rightarrow H = -1 \end{cases} \tag{2}$$

Où $R(H)$ représente la partie réelle du filtre.

Ce filtre est très utilisé pour la détection d'un signal dans un bruit blanc additif car il est capable d'optimiser le rapport signal sur bruit SNR (Savvides et Vijaya Kumar, 2003). L'inconvénient principal de ce filtre est sa sensibilité aux déformations géométriques (rotation, changement d'échelle). En effet, son plan de corrélation peut contenir des lobes secondaires qui influencent son pouvoir discriminant.

3.2.2 Filtre de phase unique

Le filtre de phase unique, noté POF (Phase Only Filter) contient seulement l'information de phase du spectre de la référence (équation (3)).

$$H = \frac{TF^*}{|TF|} \tag{3}$$

Si on applique une binarisation de ce filtre donnée par l'équation (2) on obtient ainsi un filtre binaire de phase unique, ce qui a permis d'introduire le nom BPOF (Binary Phase Only filter). Ce filtre a pour certaines applications un pouvoir discriminant important. Sa faiblisse réside dans sa sensibilité aux déformations géométriques et particulièrement la rotation.

3.2.3 Filtre harmonique circulaire

Parmi les exigences principales pour la reconstruction des filtres, c'est d'assurer l'invariance à la déformation plane. Pour répondre à cette exigence, les auteurs dans (Tribillon, 1998) ont proposé le filtre harmonique circulaire (CHF). L'idée de ce filtre est de projeter la transformée de Fourier en cordonnées polaires et d'utiliser ses harmoniques comme un filtre pour assurer une invariance aux changements géométriques et ainsi remédier aux inconvénients des deux précédents filtres.

Considérons une image $I(x, y)$ et $F(u, v)$ sa Transformée de Fourier en coordonnées cartésiennes. Notons $F(r, \theta)$ la transformation polaire de $F(u, v)$. Par conséquent, la décomposition en série de Fourier de $F(r, \theta)$ est définie par l'équation (4) :

$$F(r, \theta) = \sum_{k=-\infty}^{\infty} F_k(r) e^{jk\theta} \tag{4}$$

Les harmoniques circulaires sont définies via l'équation (5).

$$F_k(r) = \frac{1}{2\pi} \int_0^{2\pi} F(r, \theta) e^{jk\theta} d\theta \tag{5}$$

L'avantage de ce filtre réside principalement dans son invariance aux transformations géométriques mais il présente une faiblesse en discrimination (Alfalou et Brosseau, 2010; Tribillon, 1998). Pour surpasser cette faiblesse nous proposons dans un premier lieu une variante du filtre CHF qui est le filtre adapté classique circulaire. L'idée de ce filtre est de transformer l'image dans un plan polaire (Toumi et Khenchaf, 2010) avant d'appliquer la construction de filtre. L'invariance aux changements géométriques est assurée par l'image polaire/log-polaire. La construction du filtre est ainsi réalisée sur l'image polaire. Le filtre reconstruit n'est que le conjugué de la TF.

4 Corrélation par monovoie optique

Le principe de la monovoie optique est la construction d'un seul filtre à partir d'une seule image et la reconnaissance est effectuée en réalisant une multiple corrélations entre le filtre de la cible à reconnaître et les filtres créés à partir de la base d'images références. Le filtre optimal permettant d'obtenir le meilleur PCE possible correspond au filtre de l'image la plus similaire. Le filtre optimal serait donc celui qui permet de placer la totalité de l'énergie du résultat de la corrélation dans le pic. Le principe de création du filtre monovoie est illustré par la figure 3.

Une fois la phase de construction de filtre est réalisée, nous avons mené quelques simulations sur une base d'images réelles afin d'évaluer et analyser les performances de la méthode décrite. Nous présentons dans ce qui suit quelques résultats obtenus.

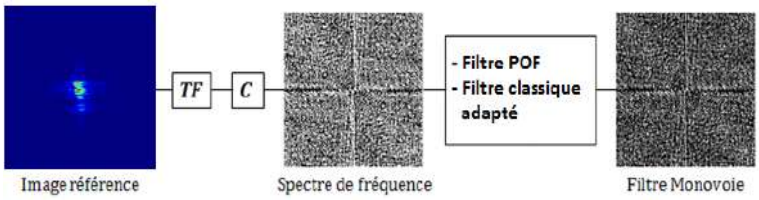


FIG. 3 – *Corrélation par monovoie*

4.1 Simulation et résultats

Les simulations réalisées dans le cadre de ce travail sont effectuées sur une base de données réelles de 11 cibles radar aériennes. Les cibles considérées ont une échelle $1/48^{me}$ et sont : A10, F104, F4 , F117, F14, Harrier, F15, Mig29, F16, Tornado et F18. Chaque cible est représentée par 160 images dont chacune correspond à un angle donné dans $[-5^o \ \ 95^o]$. La base d'images est constituée ainsi de 1760 images. La base d'images est divisée en base de test (BT) et en base d'apprentissage (BA) selon un nombre d'images représentatif de chaque classe (cf. tableau 1). Chaque classe est représentée par un nombre d'images identique choisi selon un pas de sélection.

N^{bre} d'images par classe	4	8	16	32	40	80
Base de test (en images)	1716	1672	1540	1408	1320	880

TAB. 1 – *Taille de la base de test/apprentissage*

La phase de reconnaissance est réalisée en s'appuyant sur le principe du k-ppv (k plus proche voisin). Des filtres optiques sont reconstruits à partir de chaque image de la base de références. Pour chaque image de la abse de test, un filtre est créé (spectre de fréquence) et corrélé avec tous les filtres de la base de références. Une transformée de Fourier inverse est appliquée à chaque résultat de corrélation pour avoir une réponse dans l'espace direct afin d'obtenir le PCE permettant d'estimer le taux de corrélation. Pour chaque classe de la base de références, le plus grand PCE est maintenu référençant l'image la plus similaire à l'image représentant la cible à reconnaître. Au final, pour chaque image de la base test, nous obtiendrons 11 PCE. L'identité de l'image requête est obtenue suivant le PCE le plus élevé. Les résultats obtenus en termes de taux moyens de bonne classification en fonction des différentes tailles de la base d'apprentissage/test sont présentés dans la figure 4.

En utilisant les différents types de filtres présentés précédemment, les résultats obtenus sont satisfaisants bien que le filtre POF fournit un meilleur taux de bonne classification en le comparant au filtre classique adapté qui permet l'obtention d'un taux de bonne classification d'environ 99.95% pour 16 images par classe.

Cependant, l'inconvénient majeur réside dans la sensibilité de ces filtres aux changements géométriques et le temps de calcul nécessaire pour l'indexation *en-ligne* qui présente un handicap majeure comme nous pouvons le constater sur la figure 5 qui présente le temps de recherche d'une seule image en utilisant le filtre POF.

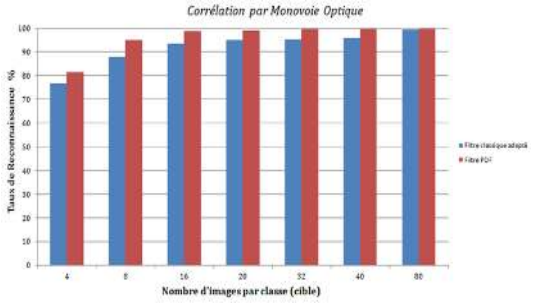


FIG. 4 – *Taux moyens de bonne reconnaissance en fonction de la taille de la base de références*

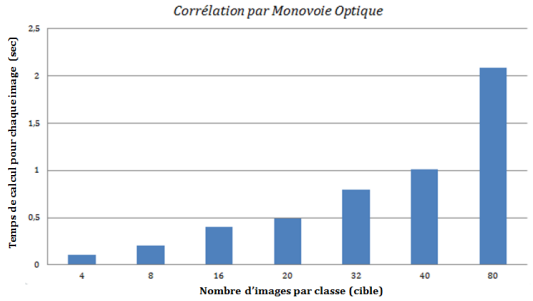


FIG. 5 – *Taux moyens de bonne reconnaissance en fonction de la taille de la base de référence*

Afin d'assurer un temps de recherche dans une base de références importante et une invariance aux changements géométriques, plusieurs variantes ont été proposées dans le domaine de la corrélation optique. Dans ce contexte, nous pouvons citer les techniques de la corrélation multivoie. Nous pouvons aussi recenser les techniques multivoies séparées, les techniques multivoies accumulées et les multivoies segmentées. Nous nous limitons dans ce travail à présenter notre approche de corrélation multivoie accumulée.

5 Corrélation par une corrélation multivoies accumulées

L'idée de la corrélation multivoie est de construire le filtre de corrélation à partir de plusieurs images candidates issues de la base d'apprentissage. Au lieu d'avoir un seul filtre par image, quatre images de la même classe (cible) composeront ainsi un seul filtre appelé *filtre composite*. Cette démarche est sensée augmenter le pouvoir discriminant des filtres et assurer un temps de calcul plus faible. Dans ce cas de figure, le nombre de filtres dans la base de références (apprenstissage) est 4 fois moins important que dans une base de références constituée de filtres monovoies. Par conséquent, un gain de 3/4 dans le temps de calcul est réalisé.

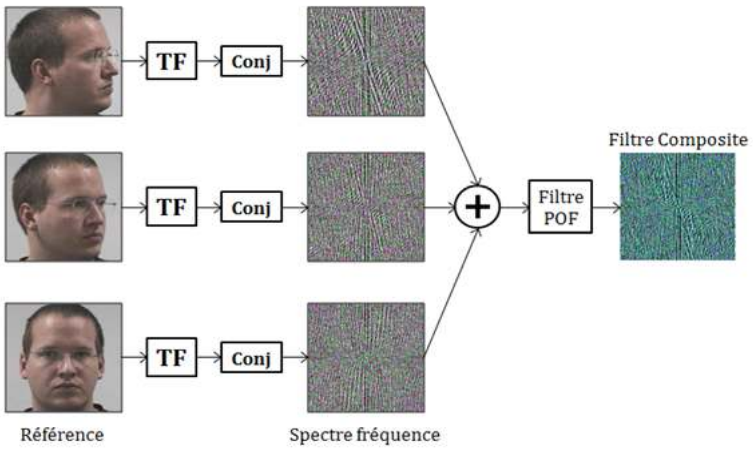


FIG. 6 – *Principe de la reconstruction de filtre composite*

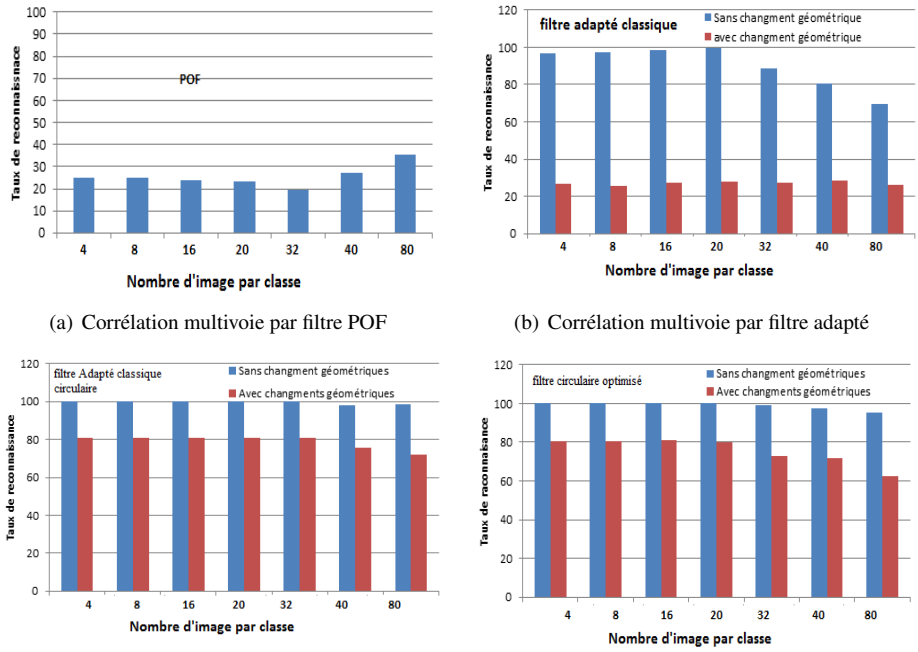
5.1 Filtre composite

La réalisation d'un filtre composite s'obtient en combinant linéairement les spectres R_i de références r_i suivant des poids différents. Ainsi, une combinaison linéaire est obtenue de tous les spectres (multiplexage fréquentielle) avec des poids w_i :

$$R = \sum_{i=1}^{N} w_i R_i \;\; avec \;\; w_i <= 1 \tag{6}$$

Pour l'exemple où le poids choisi est l'unité, toutes les références ont la même importance pour la création du filtre composite. Enfin, le codage du filtre F à partir de R se réalise en utilisant les filtres présentés précédemment. La figure 6 illustre le principe de construction d'un filtre composite à partir des filtres POF pour une reconnaissance faciale (Kumar et al., 2006; Savvides et Vijaya Kumar, 2003). Notons qu'une optimisation sur le filtre peut être réalisée. L'optimisation des filtres composite se fait par des méthodes d'estimation des poids w_i. L'objectif de cette optimisation est de définir la pondération des références pour la construction du filtre composite. Nous avons pour cela recherché la meilleure combinaison qui assure une meilleure uniformité sur les PCE issus des corrélations entre le filtre et les références qui le composent. Nous définissons un critère d'optimisation basé sur une estimation du coût du filtre défini par le rapport entre la moyenne des PCE et leur écart-type. Plus ce rapport est élevé, plus les PCE sont importants d'une grandeur proche les unes des autres.

Le PCE_i représente le PCE associé à la corrélation entre la référence r_i et le filtre (reconstruit de l'image requête). Nous supposons qu'il existe une étroite relation entre le poids w_i et le PCE_i. La première étape réalisée consiste à calculer les $PCE_{i(0)}$ lorsque tous les poids sont mis à 1. Ensuite nous supposons qu'il existe une puissance optimale β dans l'intervalle $[0.5 \;\; 1.5]$ tel que pour toutes les images utilisées (4 images dans composites du filtre) $w_i = PCE_{i(0)}^{-\beta}$ représente le poids optimal. Par un processus d'itération avec un pas de 0.1



(a) Corrélation multivoie par filtre POF

(b) Corrélation multivoie par filtre adapté

(c) Filtre adapté circulaire

(d) filtre circulaire optimisé

FIG. 7 – *Taux moyens de reconnaissance par corrélation multivoie*

entre 0.5 et 1.5 pour β, Nous sélectionnons la valeur de β pour laquelle les poids w_i sont optimaux.

5.2 Simulations et résultats

Les résultats obtenus et présentés dans cette section sont réalisés sur la même base de données présentée dans la section 4.1. Les premiers résultats présentés par la figure 7(a) sont obtenus pour les filtres composites POF. Ainsi, il est facile de constater que la corrélation multivoie avec ce type de filtre donne des taux de reconnaissance très faibles. Contrairement au filtre classique adapté qui permet d'obtenir des résultats meilleurs en absence de changements géométriques (Cf. figure 7(b)). Les filtres circulaires (Cf. figure 7(c) et circulaires optimisés 7(d) présentent un meilleur taux de reconnaissance avec une invariance aux changements géométriques. Généralement, la corrélation optique classique par multivoie accumulée présente un taux de reconnaissance très satisfaisant. Il reste à signaler que cette corrélation peut présenter dans certains cas une saturation. En effet, une fois les spectres références sommés, le filtre résultant de la combinaison entre eux peut avoir une intensité d'énergie très élevée qui engendre une saturation. Cette saturation peut invalider les résultats, si elle est assez forte. Dans le cadre de ce travail, les inconvénients de ces techniques sont traités par le biais de la pondération proposée pour le calcul des filtres composites.

6 Conclusion et perspectives

Nous avons présenté dans ce travail, une approche de corrélation optique pour la reconnaissance de cibles radar à partir des images ISAR. Les résultats obtenus sont très encourageants et satisfaisants pour la base de données utilisée. Même si la corrélation monovoie fournit des performances très satisfaisantes en termes de taux de bonne reconnaissance, elle présente deux inconvénients majeurs dus à la sensibilité des filtres aux transformations géométriques et au temps de calcul important nécessaire pour réaliser la tâche de reconnaissance (corrélation). Par ailleurs, la corrélation multivoie se présente comme une solution adéquate pour pallier les lacunes de la corrélation monovoie. Nous n'avons présenté dans ce travail que la corrélation multivoie accumulée, par conséquent nous présenterons dans un travail futur l'apport des autres corrélations telles que les multivoies séparées et les multivoies segmentées et une étude comparative.

Références

Alfalou, A. (1999). *Implantation optique des corrélateurs multivoies appliques à la reconnaissance des formes*. Thèse de doctorat, Université de Renne 1.

Alfalou, A. et C. Brosseau (2010). Understanding correlation techniques for face recognition: From basics to applications. In M. Oravec (Ed.), *Face Recognition*. InTech.

Belkhaoui, B., A. Toumi, A. Khenchaf, A. Khalfallah, et M. Bouhlel (2012). Segmentation of radar images using a combined watershed and fisher techniques. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pp. 400–403.

Chen, Y., G. Chen, R. Blum, E. Blasch, et R. Lynch (2008). Image quality measures for predicting automatic target recognition performance. In *2008 IEEE Aerospace Conference*, pp. 1–9.

Henderson, G. et C. Weaver (1966). Optical properties of evaporated films of chromium and copper. *Journal of the Optical Society of America 56*(11), 1551–1559.

Kim, D.-H., D.-K. Seo, et H.-T. Kim (2005). Efficient classification of ISAR images. *IEEE Transactions on Antennas and Propagation 53*(5), 1611–1621.

Kumar, B. V., M. Savvides, et C. Xie (2006). Correlation pattern recognition for face recognition. *Proceedings of the IEEE 94*(11), 1963–1976.

Lazarov, A. (2012). ISAR signal formation and image reconstruction as complex spatial transforms. In S. G. Stanciu (Ed.), *Digital Image Processing*. InTech.

Lugt, A. V. (1966). Practical considerations for the use of spatial carrier-frequency filters. *Applied Optics 5*(11), 1760–1765.

Miller, P. et C. Woodruff (1995). Real-time automatic target recognition using an optical correlator. In *Electronic Technology Proceedings.*, pp. 189–193.

Robertson, J. K. (1941). *Introduction to Physical Optics*. Van Nostrand.

Savvides, M. et B. V. Kumar (2003). Quad phase minimum average correlation energy filters for reduced memory illumination tolerant face authentication. In *Audio-and Video-Based*

Biometric Person Authentication, pp. 19–26.

Savvides, M. et B. V. K. Vijaya Kumar (2003). Efficient design of advanced correlation filters for robust distortion-tolerant face recognition. In *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, pp. 45–52.

Toumi, A., B. Hoeltzener, et A. Khenchaf (2007). Using watersheds segmentation on ISAR image for automatic target recognition. In *2nd International Conference on Digital Information Management (ICDIM)*, Volume 1, pp. 285–290.

Toumi, A. et A. Khenchaf (2010). Log-polar and polar image for recognition targets. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1609–1612.

Toumi, A., A. Khenchaf, et B. Hoeltzener (2012). A retrieval system from inverse synthetic aperture radar images: Application to radar target recognition. *Elsevier, Information Science. 196*, 73–96.

Tribillon, J.-L. (1998). *Traitement optique de l'information & reconnaissance des formes par voie optique*. Teknea.

Summary

The present work is part of the general problem of information processing in the field of radar targets recognition. The present work deals with an optical correlation methods, especially in the single-channel and multichannel optical correlation. The principle of these techniques is based on a correlation between the image representing the unknown object with a composite filters reconstructed from a training dataset of images. Indeed, a similarity criterion is proposed which is based on the sum of the highest energies on the correlation plane.

Quantification non supervisée de la sous- et sur-segmentation pour la classification d'images de télédétection

Andrés Troya-Galvis

ICube UMR 7357, 300 bd Sébastien Brant - CS 10413 - F-67412 Illkirch Cedex
troyagalvis@unistra.fr

Résumé. L'analyse d'images basée objet est devenue une technique de prédilection quand il s'agit de travailler avec des images à très haute résolution spatiale. Néanmoins, les résultats dépendent fortement de la segmentation d'images ; par conséquent il est essentiel d'évaluer sa qualité. Les approches d'évaluation supervisées consistent à mesurer la similarité des résultats avec une vérité-terrain. En revanche, les approches non supervisées se servent uniquement des propriétés intrinsèques à l'image et aux régions pour estimer leur qualité. Les métriques non supervisées conçues dans un contexte de télédétection se limitent à une évaluation globale des résultats. Nous proposons une nouvelle métrique non supervisée pour évaluer localement la qualité de chaque région en fonction d'un critère d'homogénéité donné. De plus, nous définissons une variante permettant d'estimer la qualité globale en combinant les différents scores de qualité locale. Finalement, nous analysons le comportement des métriques proposées et validons leur applicabilité pour trouver des segmentations qui minimisent les erreurs de sous- et sur-segmentation.

1 Introduction

L'analyse automatique d'images de télédétection est une tâche difficile mais nécessaire, puisque le traitement manuel de telles images devient impraticable à cause du rapide accroissement de leur taille, et leur acquisition de plus en plus fréquente. L'objectif consiste à déterminer la nature des objets présents dans l'image. Une manière de résoudre le problème consiste à analyser chaque pixel de manière individuelle et appliquer des méthodes de fouille afin d'inférer la classe thématique du pixel en fonction de ses valeurs radiométriques ; ou de proposer des clusters de pixels similaires susceptibles de représenter une classe d'objets réels. L'image classifiée résultante peut alors être utilisée par des géographes dans une vaste gamme d'applications, (Pham et al., 2011; Räsänen et al., 2013; Westen, 2013). Néanmoins, les approches orientées pixel arrivent à leurs limites avec l'apparition des images à haute et très haute résolution spatiale. En effet, à de telles résolutions, un pixel représente une région qui peut aller de 0.5m^2 à 2m^2. Par conséquent, la complexité et la variabilité des objets identifiables dans l'image augmentent considérablement. Ainsi, un pixel ne représente plus un objet réel mais plutôt une partie de celui-ci. L'analyse d'images basée objet ou *Object Based Image Analysis* (OBIA) tente de surmonter cette difficulté en regroupant les pixels en objets de plus haut niveau, appelés régions ou segments. Ces régions permettent de calculer des propriétés portant

plus d'information sur les objets dans l'image et permettant donc de les décrire de manière plus précise.

De manière générale, des algorithmes de segmentation d'images sont employés comme pré-traitement afin de construire les segments utilisés dans la phase de classification ultérieure. Il est évident qu'une mauvaise correspondance entre les segments et les objets réels peut conduire au calcul de propriétés peu pertinentes. La segmentation de l'image est donc une étape critique, puisque les erreurs sont propagées tout au long du processus et peuvent conduire à des résultats inattendus et indésirés.

Les métriques supervisées fournissent un score de qualité précis par rapport à une vérité-terrain. Néanmoins, dans le contexte de la télédétection, il est inconcevable d'obtenir une vérité-terrain complète et de bonne qualité. Effectivement, la segmentation manuelle de telles images est peu envisageable à cause de leur grande taille, e.g., $10000 \times 10000 = 10^8$ pixels pour une image de la communauté urbaine strasbourgeoise.

Nous nous intéressons aux métriques non supervisées. Celles-ci, font appel aux propriétés intrinsèques de l'image et des segments afin de quantifier les erreurs de segmentation. Notre objectif, est d'extraire des objets ayant des tailles et des formes différentes ; nous cherchons à partitionner l'image en régions homogènes et sémantiquement significatives. Les métriques classiques se basent majoritairement sur le nombre de segments ou considèrent le contraste entre peu de régions. Ainsi, elles sont mal adaptées à la télédétection car le nombre d'objets présents sur l'image pour une classe donnée peut aller de quelques dizaines à plusieurs centaines. Il existe peu de métriques conçues dans le cadre de la télédétection, et elles évaluent la qualité uniquement de manière globale. Pour répondre à ces problèmes, nous proposons une nouvelle métrique qui évalue la qualité de chaque segment de manière individuelle en fonction de son voisinage spatial ainsi que d'un critère d'homogénéité donné. Nous définissons également une fonction pour mesurer la qualité de manière globale en combinant les scores de qualité locale et en considérant en même temps les erreurs de sur- et de sous-segmentation. De plus, il est possible de raffiner les procédures de post-traitement en gardant une trace des scores de qualité de chaque segment pour améliorer leur classification.

La suite de cet article est structurée comme suit. Dans la section 2, nous rappelons la définition de la segmentation d'images dans le cadre de la télédétection, et nous présentons brièvement les approches existantes pour évaluer la qualité des segmentations. Dans la section 3, nous introduisons notre métrique de qualité non supervisée. Dans la section 4, nous étudions le comportement de la métrique proposée et nous validons son applicabilité en télédétection. La section 5, résume les avantages et les inconvénients de notre métrique, ainsi que nos futures voies de recherche.

2 Évaluation de la segmentation d'images

Comme nous l'avons mentionné auparavant, la segmentation d'images est la première, et l'une des étapes les plus importantes dans les méthodes OBIA. L'objectif est de partitionner une image en régions homogènes selon un critère donné. Il existe un grand nombre de méthodes de segmentation, et des milliers de variantes destinées à des applications spécifiques. Pour une étude complète sur ce sujet, le lecteur peut se référer à Cheng et al. (2001). Dans le contexte de la télédétection, les approches les plus courantes sont celles basées sur l'homogénéité spectrale telles que : *mean-shift* (Michel et al., 2015), *region-growing* (Baatz et Schäpe,

2000), *split-and-merge* (Wang et al., 2010), *watershed* (Derivaux et al., 2010), ou des stratégies hiérarchiques (Peng et al., 2013). Néanmoins, de moindres efforts on été accomplis en termes d'évaluation de la qualité des segmentations.

En télédétection, une segmentation idéale devrait fournir un partitionnement de l'image tel qu'il y ait une correspondance parfaite entre chaque segment et chaque objet représenté dans l'image. À partir de cette définition, nous pouvons caractériser deux types d'erreurs de segmentation :

— la sur-segmentation, lorsqu'un objet correspond à plusieurs segments (i.e., les segments sont trop petits) ; et
— la sous-segmentation, lorsqu'un seul segment correspond à plusieurs objets (i.e., les segments sont trop grands).

D'après Zhang *et al.* (Zhang et al., 2008), il existe des méthodes d'évaluation dites subjectives ; elles consistent basiquement à examiner visuellement les résultats. Bien que cette tâche soit longue est fastidieuse, c'est le seul moyen de s'assurer que les résultats correspondent réellement aux attentes de l'utilisateur. En télédétection, ces approches s'avèrent quasiment impraticables, étant donnée la grande taille des images. Il existe aussi des méthodes dites objectives, celles-ci essaient de quantifier la précision de la segmentation à partir de critères associés à des métriques données.

Les métriques d'évaluation supervisées, comparent les résultats de segmentation avec une ou plusieurs segmentations de réference (vérité-terrain), obtenues généralement de manière manuelle (Martin et al., 2001). Elles consistent typiquement à mesurer la similarité entre segmentations. Dans Paglieroni (2004) et Corcoran et al. (2010), on retrouve les caractéristiques les plus importantes à prendre en compte afin de concevoir de telles métriques de qualité. Dans cette catégorie, il existe des métriques orientées région (Polak et al., 2009; Vojodi et al., 2013; Pont-Tuset et Marques, 2013), orientées contours (Li et al., 2013) ainsi que des approches probabilistes qui peuvent employer plusieurs segmentations de référence (Peng et Li, 2013). Dans Monteiro et Campilho (2012), le lecteur intéressé peut trouver une étude exhaustive des métriques de qualité supervisées.

Dans nos travaux, nous nous intéressons particulièrement aux approches d'évaluation non-supervisées. Celles-ci se basent sur des propriétés intrinsèques, pouvant être calculées directement à partir de l'image et des segments produits. Bien que les approches non-supervisées ne requièrent pas de vérité-terrain, elles ont été étudiées dans une moindre mesure. En effet, la définition de telles métriques représente un vrai défi, puisqu'elles doivent modéliser en quelque sorte la notion de qualité perçue par l'être humain. Ceci n'est pas évident, puisque la qualité d'une segmentation reste un concept subjectif pouvant varier d'une application à une autre, voire d'un individu à un autre, pour la même application.

Généralement, les métriques non supervisées considèrent que chaque segment doit être homogène selon un critère donné. Elles essaient de maximiser l'homogénéité intra-segment et de minimiser la similarité inter-segment. Ainsi, la plupart des métriques consiste essentiellement à combiner deux termes : le premier quantifiant la sur-segmentation, et le second quantifiant la sous-segmentation.

Il existe des méthodes basées sur le contraste, la texture, la moyenne et la variance des réponses radiométriques des pixels (Srubar, 2012), et l'entropie (Zhang et al., 2003; Khan et Bhuiyan, 2014). D'autres approches réalisent des mesures dans des espaces couleurs particuliers qui tentent de modéliser le système visuel en tenant compte des différences entre couleurs

perçues par l'être humain (Chen et Wang, 2004). Pour une étude approfondie des métriques non-supervisées, le lecteur peut se référer à Srubar (2012).

Peu de métriques non-supervisées ont été proposées dans le contexte de l'imagerie de télédétection. Zhang *et al.* ont proposé une métrique considérant la taille de l'image, le nombre de segments ainsi que les valeurs moyennes par segment et par bande spectrale (Zhang et al., 2012). Corcoran *et al.* ont proposé une métrique prenant en compte le domaine spatial ; ils argumentent que la vision humaine segmente les images dans le domaine spatial en cherchant du contraste aux frontières des objets (Corcoran et al., 2010). Johnson et Xie ont proposé une métrique caractérisant l'homogénéité intra-segment comme une somme pondérée des variances d'un ensemble de caractéristiques, et l'hétérogénéité inter-segment à l'aide de l'indice de Moran (Johnson et Xie, 2011).

Discussion

Les métriques décrites ci-dessus, se basent sur des mesures de qualité globales afin de trouver un ensemble de paramètres optimaux pour un algorithme de segmentation donné. Néanmoins, à notre connaissance, il n'existe pas d'algorithme de segmentation capable de segmenter correctement tous les objets de toutes les classes thématiques d'intérêt, quels que soient les paramètres employés. En effet, la taille idéale des segments peut varier considérablement en fonction du niveau sémantique. De plus, à un même niveau sémantique, les objets peuvent varier en taille et en forme. Comme proposé dans Corcoran et al. (2010), les métriques de qualité non-supervisée doivent employer des caractéristiques modélisant correctement le système visuel humain, et doivent tenir compte du domaine spatial. Notre hypothèse de base est que l'évaluation locale (par segment) doit être prise en compte dans les applications OBIA, restant ainsi dans le paradigme basé objets. Effectivement, l'estimation de la qualité par segment peut permettre une meilleure compréhension des résultats et faciliter la prise de décision pendant les phases d'analyses subséquentes. Par ailleurs, les métriques devraient être facilement adaptables pour permettre la sélection des meilleures segmentations à des niveaux de détails différents.

3 Une métrique quantifiant la sous- et la sur-segmentation

Comme nous l'avons évoqué dans les sections précédentes, les métriques de qualité non-supervisées travaillent principalement de manière globale. Néanmoins, les hypothèses des approches OBIA suggèrent de travailler avec des métriques considérant la qualité de chaque segment individuellement. Dans cette section, nous présentons notre principale contribution : une métrique non-supervisée de qualité locale permettant la quantification de la sous- et la sur-segmentation pour chaque segment. Nous proposons également, une fonction permettant d'agréger les résultats de qualité locale en une mesure de qualité globale. Étant donné que notre métrique intègre un indice d'homogénéité comme méta-paramètre, elle est facilement adaptable pour trouver des segmentations à des échelles différentes, ou pour valider la pertinence de tels indices dans un contexte particulier.

3.1 Évaluation locale

Soit, $S = \{R_i \mid 0 \leq i < M\}$ une partition d'un espace image composée de M segments R_i ; soit $\mathcal{N}(R_i)$ l'ensemble des segments dans le voisinage de R_i ; soit $H(R_i)$ une fonction bornée dans l'intervalle $[0, 1]$ caractérisant l'homogénéité du segment R_i, 0 signifiant que le segment est complètement homogène, et 1 signifiant qu'il est complètement hétérogène ; et soit δ un seuil sur les valeurs de $H(R_i)$. La fonction d'évaluation locale ϕ_δ est définie comme suit :

$$\phi_\delta(R_i) = \begin{cases} -1 & \textbf{si } H(R_i) > \delta \\ 1 & \textbf{si } H(R_i) \leq \delta \textbf{ et } \exists R_j \in \mathcal{N}(R_i), H(R_i \bigcup R_j) \leq \delta \\ 0 & \textbf{sinon} \end{cases} \quad (1)$$

Intuitivement, nous considérons qu'un segment R_i est de bonne qualité s'il est bien isolé par rapport à son voisinage à une certaine échelle ; celle-ci est déterminée par le seuil δ. En effet, plus un segment est grand, plus il y a de chances qu'il soit hétérogène (i.e., H aura des valeurs plus élevées). Ainsi, lorsque δ est petit, les segments les plus grands ont plus de chances d'être considérés comme sous-segmentation, et lorsque δ croît, les segments les plus petits ont plus de chances d'être considérés comme sur-segmentation. On retrouve ainsi trois cas de figure possibles :

— R_i est trop hétérogène (i.e., $H(R_i) > \delta$), il est donc considéré comme une région sous-segmentée de l'image. Dans ce cas, le segment est pénalisé avec un poids négatif ;
— R_i est assez homogène (i.e., $H(R_i) \leq \delta$), mais il existe au moins un segment dans son voisinage tel que leur union est elle aussi assez homogène (i.e., $\exists R_j \in \mathcal{N}(R_i)$, $H(R_i \bigcup R_j) \leq \delta$), il est donc considéré comme une région sur-segmentée de l'image. Dans ce cas, le segment est pénalisé avec un poids positif ; et
— R_i est assez homogène et il est bien isolé par rapport à son voisinage (i.e., $\nexists R_j \in \mathcal{N}(R_i), H(R_i \bigcup R_j) \leq \delta$). Dans ce cas, le segment n'est pas pénalisé.

La figure 1 illustre ce principe d'évaluation locale. Considérons un indice d'homogénéité H tel que $H(R_i) = 0$ si et seulement si tous les pixels de R_i ont la même valeur, et $H(R_i) = 1$ sinon ; et définissons $\delta = 0.5$. Le segment R_2 n'est pas homogène car il contient des pixels de deux couleurs distinctes, on a $H(R_2) = 1 > \delta$, R_2 est donc marqué comme sous-segmenté. R_0 et R_5 sont tous les deux considérés comme sur-segmentés, en effet, on a $H(R_0) = 0 < \delta$ mais également $H(R_0 \bigcup R_5) = 0 < \delta$; le même raisonnement peut être appliqué réciproquement pour R_5. En pratique, ces deux segments pourraient être fusionnés pour former un seul segment bien isolé. En suivant le même raisonnement, nous observons que les segments R_1, R_3, et R_4 sont bien isolés par rapport à leurs voisins.

3.2 Évaluation globale

La qualité globale d'une segmentation peut-être vue comme une combinaison de la qualité de chacun des segments qui la composent. Nous proposons d'agréger les différentes mesures de qualité locale afin d'obtenir une estimation de qualité globale.

Nous définissons notre métrique globale UOA (*Under- and Over-segmentation Aware*) comme la somme pondérée des scores de qualité locale. Elle donne un indice direct sur le taux de sur- ou sous-segmentation dans l'image. Cependant, elle peut donner une estimation peu précise de la qualité lorsque les deux types d'erreurs sont présents à égale proportion dans

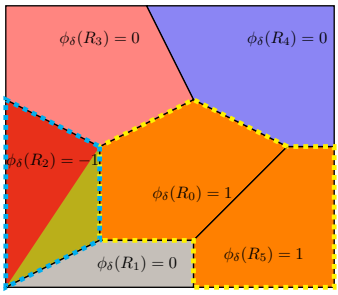


FIG. 1 – *Example fictif illustrant l'approche d'évaluation locale. Les lignes pointillées en jaune montrent un exemple de sur-segmentation. Les lignes pointillées en bleu montrent un exemple de sous-segmentation.*

l'image, puisque les pénalisations auront tendance à s'annuler mutuellement. Elle est définie comme suit :

$$UOA = \sum_i \omega(R_i).\phi_\delta(R_i) \tag{2}$$

où ω est une fonction poids telle que pour tout R_i on a $\omega(R_i) \geq 0$ et $\sum_i \omega(R_i) = 1$. Dans nos différentes expériences, nous considérons $\omega(R_i) = \frac{N_i}{N}$ avec N_i le nombre de pixels dans le segment R_i et N le nombre de pixels total de l'image. Autrement dit, chaque segment est pondéré proportionnellement à sa taille dans l'image ; ce choix permet de prendre en compte la diversité d'objets ayant des tailles différentes. En effet, les objets bien isolés ne sont pas pénalisés, quelque soit leur taille. Les zones sur-segmentées sont formées par des regroupements de segments trop petits ; l'apport des pénalisations de ces segments à la métrique est donc proportionnel à la taille de l'objet sous-jacent. Similairement, les zones sous-segmentées contiennent, par hypothèse, plusieurs objets au sein d'un seul segment ; l'apport de la pénalisation est proportionnel à la taille des objets sous-jacents.

3.3 Propriétés

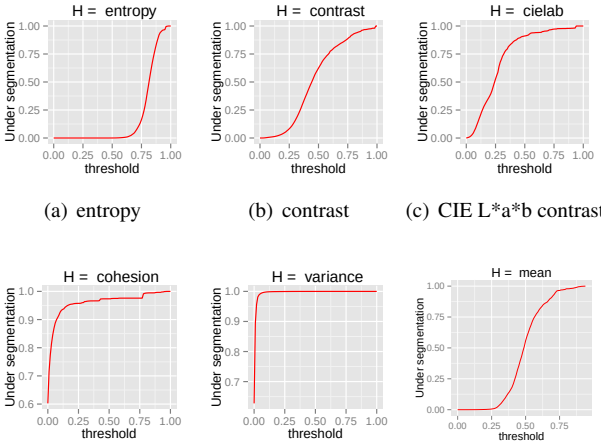
Nous allons maintenant présenter quelques propriétés de notre métrique :

— UOA est bornée, ce qui est une propriété très utile, et souvent manquante dans les métriques non supervisées existantes. En effet, elle varie de -1 (sous-segmentation totale) à 1 (sur-segmentation totale), et trouve sa valeur optimale en 0 ;

— UOA est très expressive lorsque l'on évalue des segmentations très sur-segmentées (une haute valeur positive) ou très sous-segmentées (une haute valeur négative), mais elle peut donner des résultats inattendus si l'image a une proportion similaire de régions sur-segmentées et sous-segmentées ;

— le choix du méta-paramètre H requiert un certain niveau d'expertise. En effet, il est crucial de choisir un indice d'homogénéité pertinent pour l'application en question. De plus, le choix de δ est également important puisqu'il dépend du comportement H et doit être adapté à l'échelle de segmentation voulue ;

— la complexité algorithmique est dans le meilleur des cas $\mathcal{O}(M)$ avec M le nombre de segments dans la segmentation, lorsque la segmentation est totalement sous-segmentée.



(a) entropy (b) contrast (c) CIE L*a*b contrast

(d) cohesion (e) variance (f) mean

FIG. 2 – *Analyse de sensibilité de* Ψ *par rapport à* δ *pour 6 indices d'homogénéité différents.*

En moyenne, la complexité peut être approximée par $\mathcal{O}(MK)$ avec K le nombre de voisins par segment en moyenne. Néanmoins, K est généralement très petit par rapport à M et borné par une faible valeur, ainsi nous pouvons considérer que la complexité est linéaire par rapport à M.

4 Expérimentation

Dans cette section, nous analysons le comportement des métriques proposées. En particulier nous étudions leur comportement par rapport au paramètre δ. Pour ce faire, nous avons généré 100 segmentations *Mean Shift* (Christophe et Inglada, 2009) différentes à partir d'une image Pléiades de dimensions 1024×1024 [1]. L'image est un extrait de la ville de Strasbourg avec une résolution spatiale de 60cm^2 et quatre bandes spectrales : rouge, vert, bleu et proche infra-rouge. Cet extrait contient des objets de natures différentes (e.g., différents types de bâtiments, de la végétation, des routes, et des corps d'eau). Cela nous permet de valider la robustesse de notre métrique face à la variabilité des objets dans les images de télédétection.

4.1 Le méta-paramètre H

Comme nous l'avons présenté auparavant, la définition de UOA est générique et permet de choisir un critère d'homogénéité arbitraire ; celui-ci peut varier d'une application à une autre. Nous avons étudié le comportement de UOA avec 6 indices d'homogénéité : l'entropie (Zhang et al., 2003), le contraste, le contraste dans l'espace couleur CIE L*a*b, la cohésion (Corcoran et al., 2010), et la variance (Johnson et Xie, 2011), ainsi que la moyenne de ces 5 indices.

1. Les jeux de données à très haute résolution spatiale ont été fournis par le LIVE UMR CNRS 7263.

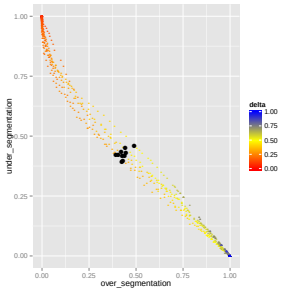


FIG. 3 – *Sous- vs. sur-segmentation avec* δ *allant de* 0 *à* 1*, pour* 10 *segmentations choisies aléatoirement.*

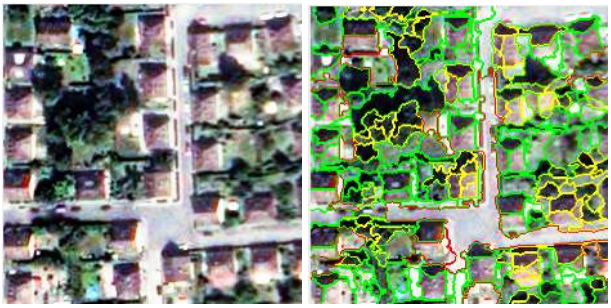
En comptant le nombre de segments sur- et sous-segmentés (pondérés par leur taille dans l'image), nous pouvons quantifier le taux de sur- et sous-segmentation dans l'image. Ainsi, dans cette première expérience, nous avons choisi une segmentation au hasard, composée de 62379 segments. Nous avons calculé chaque indice H pour chaque segment et nous avons observé la variation de la sous-segmentation par rapport à δ. La figure 2 illustre cette analyse de sensibilité pour les 6 indices d'homogénéité. Nous remarquons qu'ils se comportent différemment. En effet, chaque indice est plus sensible à la valeur de δ dans des intervalles différents. L'entropie est très sensible dans $[0.6, 0.9]$ (figure 2(a)). Le contraste est très sensible dans $[0.2, 0.5]$ et varie ensuite de manière moins importante dans $]0.5, 0.9]$ (figure 2(b)). Le contraste dans CIE L*a*b varie presque linéairement dans $[0.1, 0.8]$ (figure 2(c)). La cohésion (figure 2(d)) et la variance (figure 2(e)) sont particulièrement sensibles dans $[0.0, 0.1]$, et varient très peu sur le reste des valeurs. La moyenne des indices varie sur tout l'intervalle $[0.0, 1.0]$ en approchant la fonction cumulative d'une loi normale (Fig. 2(f)). Ces résultats montrent que le comportement de UOA peut varier en fonction de l'indice H choisi, plus particulièrement, δ doit être choisi avec précaution puisque sa valeur optimale dépend de H.

4.2 Le paramètre δ

Dans l'expérience suivante, nous avons employé la moyenne de ces indices, étant donné que cet indice moyen se comporte de manière plus régulière que les autres. Ceci nous permet d'étudier le comportement de UOA par rapport à δ en réduisant les effets de biais pouvant être induits par H. Nous avons calculé UOA avec δ, allant de 0 à 1 par pas de 0.01 ; nous avons ainsi 10000 observations différentes.

La figure 3 illustre le comportement de UOA en fonction de δ, en particulier, nous analysons la variation de la sous-segmentation par rapport à la sur-segmentation pour 10 segmentations choisies de manière aléatoire. Nous observons que les taux de sur- et sous-segmentation ainsi que δ, sont clairement corrélés. Effectivement, lorsque δ augmente, le taux de sous-segmentation décroît et le taux de sur-segmentation augmente et *vice versa.* Les points mis en valeur dans le graphe correspondent aux valeurs optimales de δ trouvées par UOA.

Nous observons que UOA trouve les valeurs optimales de δ entre 0.30 et 0.50, en moyenne les taux de sur- et sous-segmentation sont d'environ 0.40. Nous remarquons qu'il n'y a pas de segmentation ni de δ pour lesquelles les taux de sur- et sous-segmentation soient proches de 0



(a) extrait d'un quartier résidentiel (b) évaluation de la segmentation

FIG. 4 – *Visualisation de l'évaluation locale ; en jaune les frontières des segments sur-segmentés, en rouge les frontières des segments sous-segmentés et en vert les frontières des segments bien isolés par rapport à leurs voisins.*

en même temps. Autrement dit, il n'y a pas de segmentation parfaite, et nous devons gérer cela en essayant tout de même de minimiser les erreurs. Globalement, nous observons que UOA est consistante, et réussit notamment à trouver des valeurs pour δ telles que les résultats sont un bon compromis entre les deux types d'erreurs.

4.3 Validation en télédétection

Afin de valider l'utilité de UOA, nous avons fixé $\delta = 0.37$ puisque c'est une des valeurs optimales trouvées dans l'expérience précédente. Puis, nous avons évalué et trié les 100 segmentations. La figure 5 montre les résultats obtenus ; les lignes rouges représentent les frontières des segments. La segmentation dans la figure 5(a) est clairement sur-segmentée. Les différences entre le résultat le plus sous-segmenté (figure 5(b)), et le meilleur résultat trouvé (figure 5(c)) sont moins évidentes. Dans la zone zoomée (imagette en jaune), nous remarquons que les arbres à coté de la route sont mal segmentés dans la figure 5(b), de même, le parking que l'on voit se retrouve dans le même segment que la route.

Nous avons vu qu'il est peu probable de trouver une segmentation sans erreur. Néanmoins, notre approche de qualité locale peut permettre de mieux gérer et traiter les erreurs. Dans la figure 4, nous voyons à gauche un extrait de l'image Pléiades correspondant à un quartier résidentiel. Nous avons employé UOA pour trouver une segmentation délimitant au mieux les maisons. À droite, nous affichons les frontières des segments en fonction de leur qualité locale ; en jaune les frontières des segments sur-segmentés, en rouge les frontières des segments sous-segmentés et en vert les frontières des segments bien isolés par rapport à leurs voisins. Remarquons que certaines maisons ont été sur-segmentées, et que la métrique de qualité locale le signale correctement. Une opération de fusion entre ces segments pourrait être employée afin d'améliorer la segmentation de ceux-ci, et en espérant une meilleure classification du segment. Similairement, nous observons deux maisons (en haut et en bas à gauche) dont une partie à été sous-segmentée, une opération de rétrécissement pourrait être envisagée pour mieux correspondre aux maisons et améliorer également leur classification. Nous envisageons d'employer ces résultats dans la définition d'une méthode faisant collaborer plusieurs segmenteurs

(a) le plus sur-segmenté (b) le plus sous-segmenté (c) meilleur résultat

FIG. 5 – *Exemples de segmentations trouvées par notre métrique.*

et plusieurs classifieurs afin d'obtenir une segmentation ainsi qu'une classification de l'image, complète et de bonne qualité.

5 Conclusion

Les métriques supervisées fournissent des solutions efficaces pour comparer deux segmentations, mais pour donner une estimation de la qualité, elles requièrent une vérité-terrain qu'il est quasiment impossible d'obtenir dans le contexte de la télédétection. Les métriques non-supervisées se basent sur des propriétés intrinsèques calculées directement à partir de l'image et des segments résultants. La plupart d'entre elles réalisent l'évaluation seulement de manière globale ; nous argumentons que l'évaluation de la segmentation dans les approches OBIA doit rester dans le paradigme basé objets ; la qualité d'une segmentation doit donc être vue comme une combinaison de la qualité des segments qui la composent. Dans cet article, nous avons présenté une nouvelle approche permettant d'estimer la qualité locale de chaque segment, c'est-à-dire de déterminer si le segment est sur-segmenté, sous-segmenté ou bien isolé par rapport à son voisinage. Nous avons également défini UOA, une fonction d'agrégation, permettant d'obtenir une estimation de la qualité globale à partir des scores de qualité locale. Finalement nous avons analysé et validé le comportement des métriques proposées et nous avons suggéré leur utilité pour améliorer les résultats de segmentation et de classification des images de télédétection.

Dans nos travaux futurs, nous envisageons d'apprendre les seuils δ adaptés pour bien segmenter différents types d'objets à partir d'exemples, ainsi que la formalisation et l'implémentation d'une approche collaborative de segmentation-classification pour l'analyse d'images de télédétection.

Remerciements

Ces travaux de recherche ont été possibles grâce au financement de l'*Agence Nationale de la Recherche* dans le cadre du projet COCLICO (ANR-12-MONU-0001).

Références

Baatz, M. et A. Schäpe (2000). Multiresolution segmentation: An optimization approach for high quality multi-scale image segmentation. *Angewandte Geographische Informationsverarbeitung XII*, 12–23.

Chen, H.-C. et S.-J. Wang (2004). The use of visible color difference in the quantitative evaluation of color image segmentation. In *ICASSP. Proc.*, pp. 593–596.

Cheng, H., X. Jiang, Y. Sun, et J. Wang (2001). Color image segmentation: Advances and prospects. *Pattern Recognition 34*, 2259–2281.

Christophe, E. et J. Inglada (2009). Open source remote sensing: Increasing the usability of cutting-edge algorithms. *IEEE Geoscience and Remote Sensing Newsletter 35*, 9–15.

Corcoran, P., A. Winstanley, et P. Mooney (2010). Segmentation performance evaluation for object-based remotely sensed image analysis. *International Journal of Remote Sensing 31*, 617–645.

Derivaux, S., G. Forestier, C. Wemmert, et S. Lefèvre (2010). Supervised image segmentation using watershed transform, fuzzy classification and evolutionary computation. *Pattern Recognition Letters 31*, 2364–2374.

Johnson, B. et Z. Xie (2011). Unsupervised image segmentation evaluation and refinement using a multi-scale approach. *ISPRS Journal of Photogrammetry and Remote Sensing 66*, 473–483.

Khan, J. F. et S. M. Bhuiyan (2014). Weighted entropy for segmentation evaluation. *Optics & Laser Technology 57*, 236–242.

Li, H., J. Cai, T. N. A. Nguyen, et J. Zheng (2013). A benchmark for semantic image segmentation. In *ICME, Proc.*, pp. 1–6.

Martin, D., C. Fowlkes, D. Tal, et J. Malik (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV, Proc.*, pp. 416–423.

Michel, J., D. Youssefi, et M. Grizonnet (2015). Stable mean-shift algorithm and its application to the segmentation of arbitrarily large remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing 53*, 952–964.

Monteiro, F. C. et A. C. Campilho (2012). Distance measures for image segmentation evaluation. In *ICNAAM, Proc.*, Volume 1479, pp. 794–797.

Paglieroni, D. W. (2004). Design considerations for image segmentation quality assessment measures. *Pattern Recognition 37*, 1607–1617.

Peng, B. et T. Li (2013). A probabilistic measure for quantitative evaluation of image segmentation. *IEEE Signal Processing Letters 20*, 689–692.

Peng, B., L. Zhang, et D. Zhang (2013). A survey of graph theoretical approaches to image segmentation. *Pattern Recognition 46*, 1020–1038.

Pham, H. M., Y. Yamaguchi, et T. Q. Bui (2011). A case study on the relation between city planning and urban growth using remote sensing and spatial metrics. *Landscape and Urban Planning 100*, 223–230.

Polak, M., H. Zhang, et M. Pi (2009). An evaluation metric for image segmentation of multiple objects. *Image and Vision Computing 27*, 1223–1227.

Pont-Tuset, J. et F. Marques (2013). Measures and meta-measures for the supervised evaluation of image segmentation. In *CVPR, Proc*, pp. 2131–2138.

Räsänen, A., A. Rusanen, M. Kuitunen, et A. Lensu (2013). What makes segmentation good? A case study in boreal forest habitat mapping. *International Journal of Remote Sensing 34*, 8603–8627.

Srubar, S. (2012). Quality measurement of image segmentation evaluation methods. In *SITIS, Proc.*, pp. 254–258.

Vojodi, H., A. Fakhari, et A. M. E. Moghadam (2013). A new evaluation measure for color image segmentation based on genetic programming approach. *Image and Vision Computing 31*, 877–886.

Wang, Z., J. R. Jensen, et J. Im (2010). An automatic region-based image segmentation algorithm for remote sensing applications. *Environmental Modelling & Software 25*, 1149–1165.

Westen, C. V. (2013). Remote sensing and GIS for natural hazards assessment and disaster risk management. In *Treatise on Geomorphology*, pp. 259–298. Academic Press.

Zhang, H., J. E. Fritts, et S. A. Goldman (2003). An entropy-based objective evaluation method for image segmentation. In *Electronic Imaging 2004*, pp. 38–49.

Zhang, H., J. E. Fritts, et S. A. Goldman (2008). Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding 110*, 260–280.

Zhang, X., P. Xiao, et X. Feng (2012). An unsupervised evaluation method for remotely sensed imagery segmentation. *IEEE Geoscience and Remote Sensing Letters 9*, 156–160.

Summary

Object Based Image Analysis (OBIA) has been widely adopted as a common paradigm to deal with very high resolution remote sensing images. Nevertheless, OBIA methods strongly depend on the results of image segmentation. Unsupervised metrics only make use of intrinsic image and segment properties. Furthermore, the few metrics developed in a remote sensing context focus mainly on global evaluation. In this article we propose a novel unsupervised metric which evaluates local (per segment) quality by analysing segment neighbourhood, thus quantifying under- and over-segmentation given a certain homogeneity criterion. Moreover, we propose two variants of this metric, for estimating global quality of remote sensing image segmentation by the aggregation of local quality scores. Finally, we analyse the behaviour of the proposed metrics and validate their applicability for *finding* segmentation results having good trade-off between both kinds of error.

Descripteur de relations spatiales entre régions structurelles pour la reconnaissance d'objets en couleurs

Michaël Clément, Camille Kurtz, Laurent Wendling

Université Paris Descartes, LIPADE – SIP (EA 2517)
45 rue des Saints-Pères, 75006 Paris, France
prenom.nom@parisdescartes.fr

Résumé. Les méthodes classiques de reconnaissance d'objets à partir d'images reposent généralement sur une description statistique ou structurelle du contenu des objets, sous la forme de caractéristiques visuelles telles les contours, la couleur ou encore la texture. De telles caractéristiques sont alors considérées pour des tâches de classification ou de recherche d'images similaires. Dans cet article, nous présentons un nouveau descripteur d'objets complexes représentés dans des images en couleurs. L'originalité de ce descripteur réside dans une représentation homogène d'attributs de formes et de relations spatiales entre régions structurelles extraites des objets à partir d'une méthode robuste de décomposition. Nous proposons également différentes stratégies de comparaison de ces descripteurs, basées sur une mise en correspondance des régions structurelles des objets. Les résultats obtenus sur une base d'images en couleurs montrent que les relations spatiales entre les différentes régions composant des objets complexes constituent des caractéristiques intéressantes pour leur description.

1 Introduction

La reconnaissance d'objets complexes à partir d'images en couleurs est une tâche difficile qui est considérée comme une étape clé dans le domaine de l'analyse et de la fouille d'images. Les méthodes de reconnaissance d'objets reposent généralement sur une description automatique des objets représentés dans les images. Ces tâches de reconnaissance sont généralement organisées en trois étapes (Andreopoulos et Tsotsos, 2013) : (1) détection et extraction de régions d'intérêt dans l'image à l'aide d'un algorithme de segmentation ; (2) description de ces régions à l'aide de caractéristiques quantitatives modélisant différents types d'information (forme, couleurs, texture, etc.) ; (3) reconnaissance des objets d'intérêt par classification des régions décrites. Une limite de cette approche est que ces différents types de caractéristiques peuvent être difficiles à combiner pour discriminer des objets complexes de manière efficace.

La disposition spatiale des différents objets dans une scène, ou des différentes régions d'intérêt d'un même objet, constitue une information particulièrement importante dans la perception humaine de la similarité entre les images. Par conséquent, les relations spatiales entre les régions composant des objets peuvent être considérées comme des caractéristiques discriminantes pour reconnaître la nature de ces objets. Cependant, ces caractéristiques spatiales sont rarement utilisées pour la reconnaissance d'objets à partir d'images en couleurs.

Dans la littérature, les méthodes décrivant le positionnement relatif d'objets dans des images peuvent être réparties en deux grandes familles d'approches : les approches qualitatives et les approches quantitatives. Les approches qualitatives utilisent des relations symboliques de positionnement (à gauche de, au dessus de, etc.) ou topologiques (Freeman, 1975; Egenhofer, 1989; Inglada et Michel, 2009). Ces méthodes offrent des représentations de l'information spatiale qui sont difficilement intégrables dans des processus de reconnaissance de formes. De plus, elles ont souvent recours à des simplifications grossières de la forme des objets, ou se ramènent à un ensemble limité de relations spatiales. Les approches quantitatives regroupent les méthodes qui visent à capturer précisément le positionnement spatial de régions les unes par rapport aux autres. Les méthodes basées sur la logique floue sont fréquemment utilisées dans différents domaines d'application tels le raisonnement spatial dans les images médicales (Bloch et Ralescu, 2003) ou la reconnaissance d'écriture manuscrite (Delaye et Anquetil, 2011). Ces méthodes produisent un « paysage » flou pour chaque direction considérée, mais la combinaison de tels paysages pour produire une représentation globale de la structure spatiale entre régions n'est pas évidente. Une autre approche est basée sur le modèle des histogrammes de forces (Matsakis et Wendling, 1999). Ce modèle permet de prendre en compte des objets composés de plusieurs régions déconnectées et permet de résumer leur position relative le long de toutes les directions.

En se basant sur le modèle des histogrammes de forces, nous proposons dans cet article un descripteur homogène de formes et de relations spatiales adapté à la reconnaissances d'objets représentés dans des images en couleurs. Il s'agit notamment d'une extension des travaux de (Garnier et al., 2012) où ce descripteur avait été proposé pour des images en niveaux de gris. Dans notre extension, les objets sont décomposés en couches structurelles par le biais d'une stratégie couplant l'algorithme de segmentation d'images Mean Shift à l'algorithme de classification non-supervisée . Cette première contribution permet d'extraire de manière pertinente les régions d'intérêt des objets dans des images en couleurs. Notre seconde contribution est de proposer différentes stratégies pour la comparaison de ces descripteurs en se basant sur une mise en correspondance des régions composant les objets lors du calcul d'une mesure de distance. Ces deux contributions permettent d'améliorer la reconnaissance d'objets structurés à partir d'images complexes.

La suite de cet article est organisée comme suit. En section 2, nous présentons la méthodologie proposée pour la reconnaissance d'objets dans des images en couleurs. Les validations expérimentales de cette méthodologie sont présentées dans la section 3. Enfin, nous concluons en section 4.

2 Méthodologie

Dans un premier temps, nous présentons brièvement le modèle des histogrammes de forces. Ensuite, nous décrivons notre méthodologie d'extraction des régions d'intérêt d'un objet dans une image en couleurs. Nous présentons ensuite le descripteur de formes et de relations spatiales basé sur le calcul d'un histogramme de forces entre chaque couple de régions issues de l'étape d'extraction. Enfin, nous proposons différentes stratégies de comparaison et de mise en correspondance de tels descripteurs pour leur utilisation dans des processus de reconnaissance et de classification d'objets.

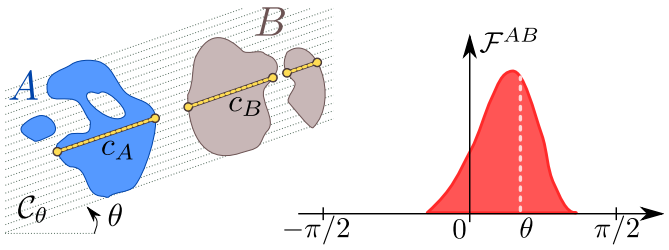


FIG. 1 – *Illustration du calcul d'un histogramme de forces. La force d'attraction $\mathcal{F}_{AB}(\theta)$ entre deux objets A et B dans une direction θ, est calculée en considérant toutes les coupes longitudinales $\mathcal{C}_A$ et $\mathcal{C}_B$ qui balayent les deux objets dans cette direction.*

2.1 Histogrammes de forces

Un histogramme de forces permet d'évaluer les relations spatiales directionnelles existant entre deux objets binaires dans une image. Il s'agit d'un histogramme circulaire le long des directions d'angles $\theta \in [0, 2\pi[$. Ainsi, pour deux objets binaires A et B, chaque valeur de l'histogramme correspond à un angle θ, et mesure la valeur de vérité de la proposition *« A est dans la direction θ de B »*, qui peut être vue comme une relation binaire floue $\mathcal{R}_\theta(A, B)$. L'ensemble de l'histogramme résume de manière globale le positionnement spatial de A par rapport à B, dans toutes les directions. Le modèle des histogrammes de forces est naturellement invariant par translation car les objets sont manipulés indépendamment de leur localisation dans l'image. Une illustration du calcul d'un histogrammes de forces entre deux objets A et B est présentée dans la Figure 1. Pour plus d'informations sur le fonctionnement et les propriétés du modèle des histogrammes de forces, voir Matsakis et Wendling (1999).

2.2 Extraction des régions d'intérêt des objets

Afin de décrire la structure spatiale interne d'un objet complexe présent dans une image, il est tout d'abord nécessaire de le décomposer en différentes sous-parties ou régions d'intérêt. Nous effectuons cette décomposition à l'aide d'un algorithme de segmentation, qui consiste à partitionner une image en différentes régions connexes selon différents critères. Cependant, dans notre contexte, les différentes sous-parties d'intérêt d'un objet peuvent être elles-mêmes composées de plusieurs régions déconnectées. Par conséquent, nous associons notre stratégie de segmentation à un algorithme de classification non-supervisée permettant de reconstruire de manière pertinente les différentes sous-parties de l'objet étudié.

2.2.1 Segmentation de l'objet en sous-parties

La segmentation d'images constitue un vaste champ de recherche à part entière et l'efficacité des méthodes existantes reste bien souvent fortement dépendante de l'application envisagée. Dans ces travaux, nous avons choisi d'utiliser l'algorithme de segmentation Mean Shift (Comaniciu et Meer, 2002) qui est adapté à la manipulation d'objets complexes dans des images en couleurs, où les frontières entre les régions ne sont pas toujours bien définies. Le

(a) Image requête I. (b) Segmentation Mean Shift. (c) Reconstruction par K-Means.

FIG. 2 – *Illustration de la stratégie d'extraction des régions d'intérêt des objets .*

Mean Shift est un algorithme issu de l'apprentissage automatique, qui envisage l'espace des caractéristiques comme une densité de probabilité, et qui cherche à en estimer les modes. Nous résumons ici son fonctionnement pour la segmentation d'images en couleurs. Celles-ci sont tout d'abord converties dans l'espace L*u*v* (luminance et chrominance) qui permet de répartir les couleurs de manière plus uniforme par rapport à la perception humaine. Pour chaque pixel, l'algorithme construit un voisinage dont il calcule la valeur moyenne. L'algorithme se déplace alors vers cette valeur moyenne et ce processus est itéré jusqu'à convergence, c'est-à-dire jusqu'à ce que tous les pixels soient assignés aux valeurs moyennes obtenues. Le Mean Shift produit une partition de l'image en R régions présentant des propriétés colorimétriques homogènes, et supposées délimiter les différentes régions d'intérêt composant l'image. Un résultat de segmentation d'un objet dans une image par l'algorithme Mean Shift est présenté en Figure 2 (b).

2.2.2 Reconstruction des sous-parties de l'objet

Le résultat de segmentation obtenu après application de l'algorithme Mean Shift produit un nombre de régions qui n'est pas connu par avance. De plus, toutes les régions sont composées uniquement de pixels connexes. Par exemple, en Figure 2 (b), la zone orange située sur les ailes du papillon est répartie en deux régions déconnectées, alors que d'un point de vue sémantique, nous souhaiterions qu'elles forment une seule et même sous-partie de l'objet. En ce sens, nous appliquons alors l'algorithme de classification non-supervisée (MacQueen, 1967) sur les pixels de l'image auxquels nous avons affecté la valeur d'intensité moyenne obtenue à l'issue de la segmentation Mean Shift. L'algorithme est initialisé avec N graines afin de produire un regroupement des régions de l'image en N couches structurelles, correspondant bien à la reconstruction des régions segmentées déconnectées. Dans la Figure 2 (c), les régions formées par l'étape de segmentation ont été reconstruites afin d'obtenir un total de 3 couches structurelles de l'objet.

2.3 Décomposition en histogrammes de forces

Une fois que l'objet représenté dans une image en couleurs a été décomposé en un ensemble de N régions d'intérêt $C_1, \ldots, C_N$, nous proposons de calculer un descripteur spatial

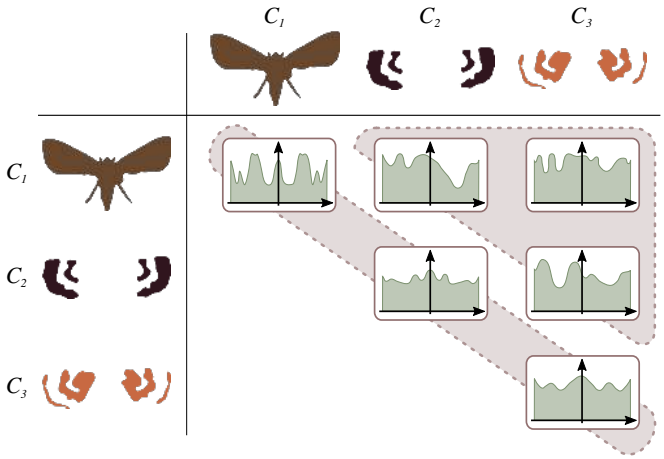


FIG. 3 – *Illustration du descripteur FHD : un histogramme de forces est calculé entre chaque couple de sous-parties (C_i, C_j) issues de l'étape de décomposition, regroupant des descripteurs de formes et de relations spatiales.*

conformément à ce qui a été exposé dans (Garnier et al., 2012). Le principe de ce descripteur consiste à calculer un histogramme de forces $\mathcal{F}^{C_iC_j}$ entre chaque couple de régions (C_i, C_j) issues de l'étape d'extraction des régions d'intérêt. Lorsque $i = j$, l'histogramme de forces calculé correspond à une région avec elle-même : il s'agit d'une description de la forme de la région. Les histogrammes de forces calculés pour $i < j$ correspondent à une information sur le positionnement spatial des régions, les unes par rapport aux autres. L'ensemble de ces histogrammes de forces constitue alors ce que nous appelons un descripteur *FHD* (pour *Force Histogram Decomposition*) :

$$\underbrace{\left\{\mathcal{F}^{C_iC_i}\right\}_{\forall i\in\{1..N\}}}_{\textit{formes}} \cup \underbrace{\left\{\mathcal{F}^{C_iC_j}\right\}_{\forall (i,j)\in\{1..N\}^2,\, i<j}}_{\textit{relations spatiales}}. \tag{1}$$

Un descripteur FHD est donc composé de $N(N+1)/2$ histogrammes de forces, et peut être représenté de manière matricielle, comme l'illustre la Figure 3. Cette matrice triangulaire est composée de N descripteurs de formes (sur la diagonale) et de $N(N-1)/2$ descripteurs de relations spatiales (sur le triangle supérieur). Dans la suite, on dira d'un descripteur FHD composé de N régions qu'il est *de taille N*.

Compte tenu des propriétés d'invariance des histogrammes de forces, les descripteurs FHD sont naturellement invariants face à la *translation*, ainsi qu'à la *mise à l'échelle* si les histogrammes de forces sont normalisés. Une estimation de la *rotation* peut être obtenue en effectuant des décalages circulaires des histogrammes de forces. De plus, ils proposent des propriétés symétriques intéressantes, ce qui permet notamment de ne pas avoir à calculer les histogrammes de forces pour le cas $i > j$ (triangle inférieur).

2.4 Comparaison des descripteurs FHD

Dans le but d'utiliser les descripteurs FHD dans un processus de reconnaissance d'objets ou de classification, il est important de définir une stratégie de comparaison de ces descripteurs. Il s'agit donc de construire une mesure de distance adaptée pour de tels descripteurs.

2.4.1 Mesure de distance

Étant donné deux descripteurs FHD de même taille N (correspondant au nombre de régions d'intérêt pour chaque objet), la manière la plus naïve de les comparer est de calculer la distance entre chacun de leurs histogrammes de forces pris deux-à-deux. Il existe de nombreuses mesures de distance entre histogrammes (Cha et Srihari, 2002). En se basant sur les résultats de (Garnier et al., 2012), nous avons choisi d'utiliser la distance du χ^2 pour comparer les histogrammes de forces deux-à-deux. La distance entre deux histogrammes de forces $\mathcal{F}_A$ et $\mathcal{F}_B$ le long de θ_{max} directions est donnée par :

$$d_{\chi^2}(\mathcal{F}_A, \mathcal{F}_B) = \sum_{i=0}^{\theta_{max}} \frac{\left(\mathcal{F}_A(i) - \mathcal{F}_B(i)\right)^2}{\mathcal{F}_A(i) + \mathcal{F}_B(i)}. \tag{2}$$

Comme indiqué précédemment, le descripteur matriciel FHD comprend à la fois des descripteurs de formes (diagonale) et des descripteurs de relations spatiales (triangle supérieur). Il paraît donc méthodologiquement pertinent de considérer séparément la distance entre ces deux types d'information. Ainsi, la distance globale entre deux descripteurs FHD Q et T peut être calculée de la manière suivante :

$$\mathcal{D}_\alpha(Q,T) = \alpha \times \mathcal{D}_{shape}(Q,T) + (1-\alpha) \times \mathcal{D}_{spatial}(Q,T), \tag{3}$$

où α est un nombre réel entre 0 et 1 permettant d'affecter un poids variable à la distance $\mathcal{D}_{shape}$ entre les descripteurs de formes et à la distance $\mathcal{D}_{spatial}$ entre les descripteurs de relations spatiales, ces distances étant définies de cette façon :

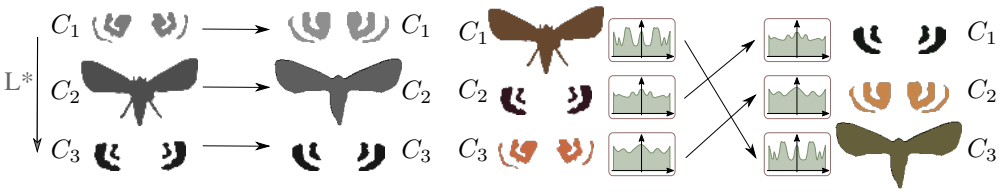
$$\mathcal{D}_{shape}(Q,T) = \frac{1}{N}\sum_{i=1}^{N} d_{\chi^2}\left(\mathcal{F}^{C_iC_i}(Q), \mathcal{F}^{C_iC_i}(T)\right), \tag{4}$$

$$\mathcal{D}_{spatial}(Q,T) = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N} d_{\chi^2}\left(\mathcal{F}^{C_iC_j}(Q), \mathcal{F}^{C_iC_j}(T)\right). \tag{5}$$

Une telle pondération des distances $\mathcal{D}_{spatial}$ et $\mathcal{D}_{shape}$ permet notamment d'éviter que les descripteurs de relations spatiales ne deviennent prédominants par rapport aux descripteurs de formes lorsque $N > 3$ (*i.e.*, il y a alors plus d'histogrammes de forces pour les relations spatiales que pour les formes).

2.4.2 Mise en correspondance

La mesure de distance présentée ci-avant n'a de sens que si les régions d'intérêt structurant les deux objets comparés sont correctement alignées dans les deux matrices. En effet, il est possible que les régions issues de l'étape de décomposition pour deux objets ne soient



(a) Alignement par luminance. (b) Appariement selon la forme.

FIG. 4 – *Illustration des deux stratégies de mise en correspondance proposées.*

pas ordonnées de la même façon, même si ceux-ci sont très similaires. Ce cas est notamment susceptible d'apparaître du fait que nous manipulons des images en couleurs : il n'y a alors pas d'ordre naturel entre les sous-parties des objets, celles-ci étant associées à des valeurs d'intensité dans l'espace RVB (rouge, vert, bleu). De ce fait, la distance entre les deux descripteurs risque d'être anormalement élevée, car calculée entre des histogrammes de forces qui ne correspondent pas deux à deux. Par conséquent, pour comparer des descripteurs FHD, il est nécessaire d'avoir recours à une stratégie de mise en correspondance des régions composant les objets, en amont du calcul de distance. Ce problème est de nature combinatoire : pour deux objets décomposés en N régions, il existe $N!$ possibilités d'appariement de leurs régions. Nous proposons ici d'employer des heuristiques se basant sur des hypothèses concernant les régions des objets pour réduire la complexité du problème.

Alignement par luminance La première stratégie consiste à partir de l'hypothèse que les régions correspondantes entre deux objets seront similaires en termes de colorimétrie. Afin de disposer d'un ordre total entre les couleurs associées aux sous-parties des objets, nous calculons leurs valeurs de luminance respectives dans l'espace L*u*v* (luminance et chrominance). Pour chaque objet, les régions sont alors triées par valeur de luminance. La Figure 4 (a) illustre le principe de cette stratégie.

Appariement selon la forme La seconde stratégie part de l'hypothèse que les régions correspondantes entre deux objets seront similaires du point de vue de leurs formes. La Figure 4 (b) illustre ce principe : chaque sous-partie est appareillée en fonction des histogrammes de forces décrivant la forme (c'est-à-dire ceux situés sur la diagonale du descripteur). Pour mettre en place cette stratégie, il est possible de l'assimiler à un problème d'optimisation dont l'objectif est de minimiser la distance globale $\mathcal{D}_{shape}(Q, T)$ entre les descripteurs de formes respectifs des deux objets. Nous proposons alors deux approches. L'une consiste à énumérer les $N!$ possibilités d'appariement entre les régions des deux objets : il s'agit d'une solution *« optimale »*, qui est cependant inappropriée pour des descripteurs FHD composés d'un nombre arbitraire de régions. Une approximation de cette solution consiste à mettre en correspondance les régions successivement en faisant à chaque fois le meilleur choix possible : il s'agit d'une stratégie *« gloutonne »*. D'un point de vue algorithmique, cette approximation peut être considérée comme un bon compromis nécessitant une complexité quadratique.

FIG. 5 – *Quelques images de papillons issues du jeu de données* PEALE.

Le choix d'une stratégie de mise en correspondance est fortement dépendant de l'application envisagée et chacune présente des avantages et des inconvénients. D'un côté, la stratégie d'alignement par luminance et relativement simple et peut être considérée comme une étape de pré-traitement, cependant, elle est vouée à échouer si la luminance n'est pas une caractéristique discriminative (par exemple, si deux objets ont la même structure mais des couleurs différentes). D'un autre côté, la stratégie d'appariement selon la forme paraît plus appropriée, mais implique d'être appliquée à chaque comparaison de deux descripteurs. De plus, cette stratégie est également plus disposée à commettre des erreurs, notamment lorsque certaines régions peuvent être assimilées à du bruit.

3 Validations expérimentales

3.1 Jeu de données

Dans le cadre de nos expérimentations, nous avons utilisé une base d'images de papillons nommée PEALE [1] et composée de 318 images présentant chacune un papillon vu du dessus. Les papillons appartiennent à des espèces différentes, divisant le jeu de données en un total de 28 classes non équiréparties, et constituant notre vérité terrain pour les tâches de reconnaissance et de classification. Les papillons sont un cas typique d'objets où la forme ainsi que la disposition spatiale des motifs sur les ailes constituent des caractéristiques particulièrement discriminantes pour distinguer les espèces. Quelques images issues du jeu de données PEALE sont présentées dans la Figure 5.

3.2 Protocole expérimental

Une classe (ou label) est assignée à chaque image du jeu de données PEALE et constitue notre vérité terrain. Afin d'éviter un quelconque biais dans nos expérimentations, nous utilisons un protocole de validation croisée de type *« leave-one-out »* : chaque image d'objet (échantillon de test) est comparée au reste du jeu de données (échantillon d'apprentissage), ceci successivement pour toutes les images. La classification de l'échantillon de test se fait par recherche du plus proche voisin : on assigne à l'image requête la classe de l'image la plus proche dans l'espace des caractéristiques engendré par les descripteurs FHD de l'échantillon d'apprentissage. Nous calculons alors le taux de reconnaissance global T_R.

Les histogrammes de forces sont calculés avec une force d'attraction constante et sur un total de 180 directions, balayant l'intervalle $[0, 2\pi[$ par pas réguliers de 2 degrés. Dans l'équa-

1. Academy of Natural Sciences, Philadelphia. `http://clade.ansp.org/entomology`

TAB. 1 – *Taux de reconnaissance global T_R pour le jeu de données* PEALE *pour différentes méthodes de décomposition des objets, en faisant varier le nombre N de sous-parties. Les triplets (h_s, h_r, M) correspondent aux paramètres utilisés pour la segmentation Mean Shift.*

Décomposition / N	2	3	4	5	6
Niveaux de gris	26,1	30,7	31,9	37,4	40,8
Mean Shift $(h_s, h_r, M) = (8, 4, 100)$	42,8	**45,0**	44,0	42,8	**45,0**
Mean Shift $(h_s, h_r, M) = (8, 8, 100)$	**47,5**	35,8	42,1	38,4	35,5
Mean Shift $(h_s, h_r, M) = (8, 12, 100)$	**39,9**	32,4	36,2	31,1	25,2

tion 3, nous avons fixé $\alpha = 5$ afin de donner un poids équivalent aux descripteurs de formes et aux descripteurs de relations spatiales.

3.3 Résultats expérimentaux

Taux de reconnaissance Le Tableau 1 présente les taux de reconnaissance obtenus pour la classification du jeu de données PEALE selon le protocole présenté précédemment. Les résultats sont présentés pour différentes méthodes de décomposition des objets, en faisant varier le nombre N de sous-parties des objets. La première ligne présente les résultats pour la version originale des FHD introduite par (Garnier et al., 2012) où les objets sont décomposés selon leurs niveaux de gris. Les lignes suivantes montrent les résultats obtenus pour notre extension des descripteurs FHD où les objets sont issus d'images en couleurs. Différents paramètres sont utilisés pour l'algorithme de segmentation Mean Shift. Le paramètre h_s correspond à la fenêtre spatiale utilisée par algorithme, et a été fixé empiriquement à $h_s = 8$ en fonction de la taille des images. Le paramètre h_r correspond à l'intervalle d'intensité parmi lequel les pixels d'un voisinage sont considérés comme appartenant à une même région. Ainsi, une valeur faible de ce paramètre, ici $h_r = 4$, produit une sur-segmentation de l'image, conservant ainsi les détails de texture. À l'inverse, une valeur plus élevée, $h_r = 12$ produit des régions plus grandes et plus hétérogènes. Enfin, le paramètre M correspond à la taille minimum des régions, fixée à $M = 100$ selon le même raisonnement.

On constate que les résultats obtenus pour l'extension des FHD aux images en couleurs sont dans l'ensemble supérieurs à ceux obtenus pour la décomposition en niveaux de gris. Cela confirme notre hypothèse selon laquelle une méthode de décomposition prenant en compte l'information spatiale et colorimétrique dans les images (par le biais d'une approche de segmentation) permet de construire des régions d'intérêt plus représentative de la structure des objets de ce jeu de données. Par ailleurs, nous pouvons également constater que les résultats semblent plus stables en adoptant une stratégie de sur-segmentation (*i.e.*, valeur de h_r faible) même en faisant varier le nombre N de sous-parties des objets.

Mise en correspondance Le Tableau 2 présente les taux de reconnaissance obtenus pour les différentes stratégies de mise en correspondance proposées en Section 2.4.2. Les paramètres de l'algorithme Mean Shift ont été fixés à $(h_s, h_r, M) = (8, 4, 100)$, afin d'obtenir une sur-segmentation des objets. La première ligne du tableau montre les résultats obtenus pour la

TAB. 2 – *Taux de reconnaissance global T_R pour le jeu de données* PEALE *pour les différentes stratégies d'appariement des descripteurs.*

Appariement / N	2	3	4	5	6
Luminance	42,8	**45,0**	44,0	42,8	**45,0**
Formes (glouton)	41,2	**45,0**	34,3	34,3	30,8
Formes (optimal)	**42,1**	40,6	39,3	35,2	33,0

stratégie d'alignement des régions par luminance. Les deux lignes suivantes présentent les résultats obtenus pour la stratégie de mise en correspondance en fonction des formes des régions, la première par appariement glouton, et la seconde par appariement optimal.

Nous pouvons constater que la stratégie d'alignement des régions par luminance fournit de meilleurs résultats globaux pour toutes les valeurs de N. Cela peut être dû au fait que pour la plupart des espèces de papillons du jeu de données, la couleur des régions d'intérêt constitue la caractéristique la plus discriminante. Nous pouvons également constater que pour la stratégie d'appariement sur la forme, l'approche gloutonne produit des résultats comparables à ceux de l'approche optimale mais avec un coût de calcul beaucoup plus faible. En plus de ces résultats globaux, nous avons étudié au niveau de la reconnaissance par classe le comportement des différentes stratégies de mise en correspondance.

La Figure 6 présente quelques résultats représentatifs de recherche d'objets similaires en utilisant la stratégie de mise en correspondance gloutonne basée sur la forme (pour $N = 3$ sous-parties). D'après ces résultats visuels, il apparaît que cette stratégie d'appariement permet de reconnaître de manière pertinente des objets où la forme des régions est une caractéristique plus discriminante.

4 Conclusion

Dans cet article, nous avons introduit une nouvelle approche pour la reconnaissance d'objets à partir d'images en couleurs. La principale originalité de cette approche repose sur une extension d'un descripteur d'objets qui intègre dans une représentation homogène des informations de formes et de relations spatiales entre les différentes régions d'intérêt composant les objets. Les premiers résultats obtenus montrent que de telles caractéristiques sont adaptées à la reconnaissance d'objets structurés au sein d'images en couleurs.

Ces travaux ont ouvert la voie à différentes perspectives. Une perspective à court terme consistera à envisager une décomposition des objets en un nombre variable de sous-parties. Ceci pourrait notamment être effectué par le biais d'une approche de segmentation hiérarchique des images. Ce type de segmentation adaptative permet de former des régions d'intérêt à différents niveaux d'échelle, résultant en une représentation des images sous la forme d'arbres de partition. Par ailleurs, une telle décomposition impliquerait alors d'être capable de comparer des descripteurs composés d'un nombre variable d'histogrammes de forces. Enfin, des travaux plus théoriques sur la modélisation et l'enrichissement du pouvoir de description des histogrammes de forces sont également envisagés.

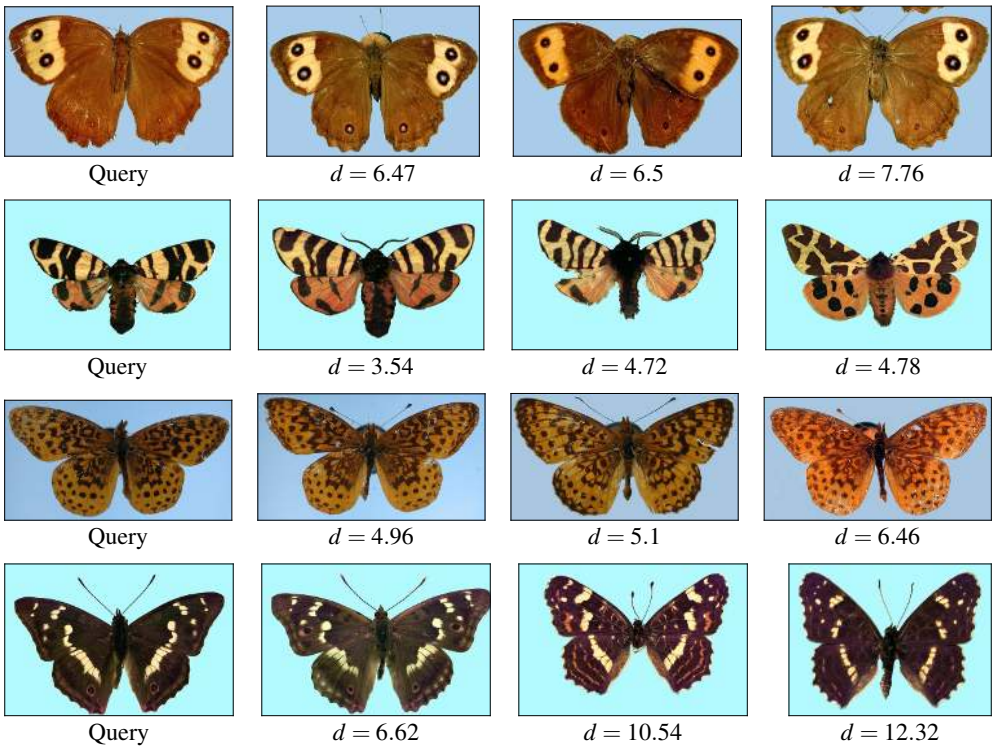


FIG. 6 – *Quelques résultats représentatifs de recherche d'objets similaires pour la base d'image* PEALE. *Pour chaque ligne, l'image de gauche représente l'image requête, par ordre décroissant de similarité.*

Références

Andreopoulos, A. et J. K. Tsotsos (2013). 50 Years of object recognition : Directions forward. *Computer Vision and Image Understanding 117*(8), 827–891.

Bloch, I. et A. L. Ralescu (2003). Directional relative position between objects in image processing : A comparison between fuzzy approaches. *Pattern Recognition 36*(7), 1563–1582.

Cha, S.-H. et S. N. Srihari (2002). On measuring the distance between histograms. *Pattern Recognition 35*(6), 1355–1370.

Comaniciu, D. et P. Meer (2002). Mean shift : A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*(5), 603–619.

Delaye, A. et E. Anquetil (2011). Fuzzy relative positioning templates for symbol recognition. In *Proceedings of the IEEE International Conference on Document Analysis and Recognition – ICDAR 2011*, pp. 1220–1224.

Egenhofer, M. J. (1989). A formal definition of binary topological relationships. In *Foundations of Data Organization and Algorithms*, Volume 367 of *Lecture Notes in Computer Science*, pp. 457–472.

Freeman, J. (1975). The modelling of spatial relations. *Computer Graphics and Image Processing 4*(2), 156–171.

Garnier, M., T. Hurtut, et L. Wendling (2012). Object description based on spatial relations between level-sets. In *Proceedings of the IEEE International Conference on Digital Image Computing Techniques and Applications – DICTA 2012*, pp. 1–7.

Inglada, J. et J. Michel (2009). Qualitative spatial reasoning for high-resolution remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing 47*(2), 599–612.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability – BSMSP 1967*, pp. 281–297.

Matsakis, P. et L. Wendling (1999). A new way to represent the relative position between areal objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence 21*(7), 634–643.

Summary

Classical object recognition methods from images usually rely on a statistical or structural description of the object content, summarizing different visual features such as contours, color or texture. Such features can then be used for the classification or the recognition of similar images. In this paper, we present a descriptor for complex objects represented in color images. The originality of this descriptor lies in an homogeneous representation of both shape attributes and spatial relations between structural regions extracted from objects using a robust decomposition method. We also propose different comparison strategies for this descriptor, based on specific matchings between structural regions of the objects. Results obtained on a dataset of color images suggest that the spatial relations between regions composing complex objects constitute interesting features for their description.

Exploration et Visualisation de sous-espaces pour la détection d'outliers dans un réseau informatique

Ibrahim Louhi[*,**] Lydia Boudjeloud-Assala[*]
Thomas Tamisier[**]

[*]Université de Lorraine,
Laboratoire d'Informatique Théorique et Appliquée, LITA-EA 3097,
Metz, F-57045, France
{ibrahim.louhi, lydia.boudjeloud-assala}@univ-lorraine.fr
[**]Centre de Recherche Public - Gabriel Lippmann,
41, rue du Brill, L-4422 Belvaux, Luxembourg
{louhi, tamisier}@lippmann.lu

Résumé. Nous proposons une approche basée sur le clustering pour la détection et la visualisation d'outliers dans un réseau informatique. L'enregistrement des évènements d'un trafic réseau génère un flux de données complexes. Pour traiter ces données, nous proposons de diviser le flux en plusieurs fenêtres. Dans chaque fenêtre, les données sont réparties sur des sous-espaces (sous-ensembles de dimensions). Nous appliquons un algorithme de clustering en exploitant les relations existantes entre ces sous-espaces. Les clusters obtenus peuvent représenter des partitions sur un autre sous-espace. Des tests sont effectués sur des données complexes constituées de fichiers logs d'un pare-feu. Notre approche a été testée sur deux sous espaces. Nous visualisons les résultats avec des graphes de voisinage sur chaque fenêtre. Les comparaisons avec l'algorithme *MCOD* sont prometteuses, les outliers sont identifiés visuellement; nous pouvons ainsi observer leur évolution dans le temps.

1 Introduction

Les données disponibles de nos jours sont de plus en plus complexes. Les données complexes peuvent se caractériser par un grand nombre de dimensions (sous-espaces) et par l'évolutivité dans le temps (Boussaid et al., 2005). Les réseaux informatiques font partie des domaines qui génèrent des séquences de données complexes sous la forme d'un flux. Afin d'assurer la sécurité et le bon fonctionnement d'un réseau informatique, il est indispensable de détecter les évènements atypiques (outliers). Ces évènements atypiques peuvent indiquer un changement dans le comportement du système, des tentatives d'intrusions ou des erreurs. Leur identification peut résoudre des problèmes, corriger des erreurs ou simplement alerter les utilisateurs d'un comportement inhabituel (Hodge et Austin, 2004).

Les approches de clustering peuvent être utilisées pour la détection des outliers. Cependant, l'identification d'outliers n'est pas une chose aisée dans le trafic réseau. Lorsque les

données sont temporelles, il est plus raisonnable de penser à adapter les approches classiques de clustering afin qu'elles puissent gérer l'évolution des flux.

Dans cet article, nous proposons une approche basée sur un algorithme de clustering pour la détection d'outliers dans un réseau informatique. L'algorithme de clustering est appliqué sur des sous-espaces (un sous-ensemble de dimensions) des données. Il est d'abord appliqué sur un premier sous-espace, les clusters obtenus sont considérés comme des représentants de partitions sur un autre sous-espace. Le clustering est ensuite appliqué sur les éléments appartenant à chaque partition du deuxième sous-espace séparément. Le processus peut être répété sur n sous-espaces. Le flux de données est divisé en plusieurs fenêtres pour permettre la détection des changements dans le temps. Les résultats du clustering sur chaque fenêtre sont visualisés par des graphes de voisinage. Le but est de détecter les outliers sur chaque fenêtre et de suivre leur évolution dans le temps. Les résultats sont comparés avec ceux de l'algorithme *MCOD* (Kontaki et al., 2011) à l'aide de l'outil Cadral (Pinheiro et al., 2014). Cadral est une plateforme logicielle qui comporte un module analytique visuel pour l'exploration des données et l'extraction des connaissances. Les tests sont effectués sur les données publiées pour le VAST challenge 2 (VAST, 2012) contenant les fichiers log du réseau de *la Banque Monétaire*.

Cet article est organisé de la façon suivante : nous présentons tout d'abord un bref état de l'art puis nous décrivons notre approche pour la détection des outliers dans un flux de données complexes avant de présenter les résultats expérimentaux. Nous terminons par une conclusion.

2 Etat de l'art

Un outlier est une observation qui semble incompatible avec les autres éléments de l'ensemble de données, ce qui laisse à supposer qu'elle est générée par un mécanisme différent (Hawkins, 1980) (Johnson et Wichern, 1992). Dans la littérature, les méthodes de la détection d'outliers se divisent en trois différentes approches (Chandola et al., 2009) : Les approches supervisées où un ensemble d'apprentissage, contenant des observations considérées comme étant normales et quelques outliers, est utilisé. Les approches semi-supervisées où l'ensemble d'apprentissage contient uniquement des observations normales. Enfin, les approches non-supervisées qui sont utilisées lorsqu'aucun ensemble d'apprentissage et aucune information préalable sur les données n'existent.

Nous nous intéressons aux approches non-supervisées où aucune information complémentaire sur les données n'est disponible. Parmi les approches-non supervisées, il existe les méthodes basées sur la distance (Knorr et al., 2000). Ces méthodes essayent généralement de trouver la distance locale entre les objets et elles peuvent traiter des ensembles de données de grande taille (Angiulli et Pizzuti, 2002) (Ramaswamy et al., 2000). Les outliers sont supposés être loin de leurs voisins les plus proches (Chandola et al., 2009). Ces méthodes mesurent la distance entre chaque élément et le centre du cluster le plus proche. Cependant, elles sont inefficaces face aux données contenant des régions denses et des régions à faible granulité (Ramaswamy et al., 2000). Regrouper les éléments homogènes dans des groupes (clusters) peut aider à identifier les outliers (Ramaswamy et al., 2000). Breuning propose une approche de détection d'outliers basée sur la densité (Breunig et al., 2000). Un outlier est désigné selon le degré d'isolation d'un élément par rapport à ses voisins. Le *Local Outlier Factor (LOF)* est utilisé pour obtenir la densité locale de chaque élément et calculer son degré d'isolement dans son voisinage.

Pour faire face à l'évolution du flux, les approches classiques de la détection d'outliers doivent être adaptées pour traiter les données temporelles. Parmi les approches de clustering des flux de données, il existe les méthodes basées sur des représentants de partitions. Les nouveaux éléments sont assignés au cluster représentatif le plus proche. L'algorithme *Stream* utilise un clustering basé sur les k-médianes (Guha et al., 2000). Le flux de données est divisé en plusieurs fenêtres qui contiennent un nombre prédéfini d'éléments. L'algorithme essaye pour chaque fenêtre de trouver un ensemble de k-médianes et d'associer chaque élément à la médiane la plus proche. Le but est de minimiser la somme carrée des distances entre les éléments et leur médiane associée. Les médianes sont les représentants des éléments. Lorsqu'un nombre prédéfini de représentants est atteint, un clustering hiérarchique est appliqué sur l'ensemble des représentants. La limite de l'algorithme *Stream* est son insensibilité à l'évolution du flux (Aggarwal, 2013).

L'algorithme *MCOD* (Micro-cluster based Continuous Outliers Detection) (Kontaki et al., 2011) est une approche basée sur les événements pour la détection des outliers dans un flux de données. Pour chaque nouvel élément, l'algorithme fixe un point temporel sur une des fenêtres suivantes pour vérifier si l'élément est devenu un outlier au lieu de vérifier chaque élément en permanence. Un élément est considéré comme outlier si le nombre de ses voisins est inférieur à un seuil prédéfini.

Dans notre approche, Nous appliquons un algorithme de clustering sur un premier sous-espace (sous-ensemble de dimensions). Nous considérons que les clusters obtenus sur ce premier sous-espace peuvent représenter des partitions sur un autre sous-espace. Chaque élément sur le premier sous-espace peut représenter plusieurs éléments sur le deuxième sous-espace en se basant sur les relations (liens) existantes entre les sous-espaces.

3 Approche proposée

L'approche que nous proposons est un algorithme de détection d'outliers utilisant un clustering basé sur le voisinage. Nous proposons d'utiliser les relations pouvant exister entre les sous-espaces pour traiter un flux de données. Pour traiter le flux, les données sont divisées en fenêtres avec une taille prédéfinie.

En premier lieu, nous pré-traitons les données pour extraire la table des distances entre les éléments sur un premier sous-espace. Nous appliquons ensuite le clustering sur l'ensemble des éléments. En utilisant les relations existantes entre les sous-espaces, nous croisons le premier sous-espace avec un autre sous-espace en remplaçant chaque élément du premier sous-espace par les éléments qu'il représente sur le deuxième sous-espace. Les clusters obtenus sur le premier sous-espace sont des représentants de partitions sur le deuxième sous-espace. Nous extrayons le tableau des distances entre les éléments de chaque partition, nous appliquons ensuite le clustering sur les éléments de chaque partition séparément. Le processus peut être itéré sur plusieurs sous-espaces (les clusters du deuxième sous-espace peuvent représenter des partitions sur un troisième sous-espace,...). Les clusters qui contiennent un seul élément sont considérés comme outliers, et les clusters qui contiennent peu éléments sont considérés comme suspects.

Algorithm 1 Detection d'outliers

ENTRÉES : Table de Distance : $(DisSubSpace_i)$;
SORTIES : Clusters.
DEBUT
Pour $i = 1$ to $n - 1$ **Faire**
Construire un graphe de voisinage $G = (V, E, p)$ à partir de $DisSubSpace_i$
Pour Chaque cluster de $SubSpace_i$ **Faire**
remplacer chaque élément par les éléments correspondants sur $SubSpace_{i+1}$ en se basant sur les relations existantes entre $SubSpace_i$ et $SubSpace_{i+1}$
Fin pour
Pour Chaque partition de $SubSpace_{i+1}$ **Faire**
Extraire la table de distance $(DisSubSpace_{i+1})$
Construire un graphe de voisinage $G' = (V', E', p')$ à partir de $DisSubSpace_{i+1}$
Fin pour
Fin pour
FIN.

Algorithm 2 Graphe de voisinage

ENTRÉES : Table de distance : $(DisSubSpace)$; Seuil.
SORTIES : Clusters.
DEBUT
Construire un graphe complet à partir de l'ensemble des éléments.
Tantque Il existe un noeud non-visité **Faire**
Trier les arêtes du graphe avec un ordre croissant.
Supprimer les arêtes avec un poids supérieur du seuil.
Supprimer l'arête avec le plus petit poids du graphe (ou du sous-graphe).
Pour Chaque noeud été connectée avec l'arête supprimée **Faire**
Marquer le noeud comme étant "visité" et trouver ses plus proches voisins.
Pour Chaque voisin **Faire**
Trouver ses plus proches voisins.
Fin pour
Fin pour
Fin tantque
FIN.

Le principe de cette approche est d'utiliser les relations pouvant exister entre les sous-espaces pour permettre aux clusters obtenus sur un sous-espace de représenter des partitions sur un autre sous-espace, au lieu de comparer chaque paire d'éléments dans l'espace (toutes les dimensions).

4 Expérimentation

Pour nos expérimentions, notre choix s'est porté sur les fichiers log publiés pour le VAST challenge 2 (VAST, 2012). Les fichiers log sont des fichiers texte où chaque ligne représente un événement enregistré sur le réseau. Les fichiers log contiennent un très grand nombre d'évènements qui sont représentés par un grand ensemble de dimensions (sous-espaces).

Nous choisissons de diviser le flux en dix fenêtres. Pour chaque fenêtre, nous choisissons d'appliquer notre approche sur deux sous-espaces : *Ports destinations* (le premier sous-espace) et *adresses IP* sources (le deuxième sous-espace). Les outliers sur le deuxième sous-espace (les *adresses IP*) représentent les machines avec un comportement atypique.

Nous appliquons le clustering basé sur le voisinage sur le premier sous-espace (*Ports)*. Comme il est illustré dans la partie gauche des figures ci-dessous (figures : 1 et 2), pour chaque fenêtre, nous obtenons des clusters de *Ports*. Nous choisissons la fréquence d'utilisation de chaque *port* par chaque *adresse IP* comme étant la relation entre les deux sous-espaces. Une *adresse IP* est représentée par son *port* le plus utilisé. En d'autres termes, chaque élément du premier sous-espace peut représenter plusieurs éléments du deuxième sous-espace. Les clusters des *Ports* sont les représentants des partitions sur le sous-espace des *adresses IP*. Chaque *Port* est remplacé par les *adresses IP* qu'il représente. Après avoir remplacé les clusters par les éléments correspondants sur le deuxième sous-espace, nous pouvons appliquer le clustering sur chaque partition des *adresses IP* séparément, les résultats obtenus sont affichés dans la partie droite des figures ci-dessous (figures : 1 et 2).

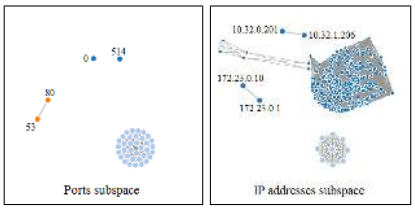


FIG. 1 – *La première fenêtre*

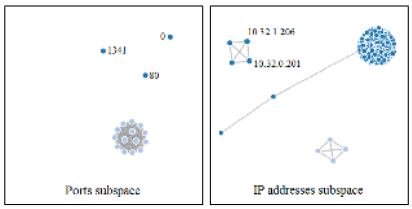


FIG. 2 – *La deuxième fenêtre*

Sur le sous-espace des *Ports* dans la figure 1, nous obtenons un cluster avec un grand nombre d'éléments et quatre éléments qui sont soupçonnés d'être des outliers : *80, 53, 514 et 0*. Sur le sous-espace des *adresses IP*, les *Ports 80, 53* ne représentent aucun élément. Nous remarquons que quatre éléments sont considérés comme des outliers. Les *Ports 80, 53* ne représentent aucun élément parce qu'ils ne sont pas utilisés comme des *Ports* destinations sur cette fenêtre. Les *adresses IP* détectées comme outliers ont des valeurs atypiques par rapport aux autres *adresses IP* figurant dans la même partition.

Dans la figure 2, sur le sous-espace des *Ports*, les *ports : 53 et 514* ne sont pas utilisés dans cette fenêtre. Cependant, les *ports : 0 et 80* sont toujours considérés comme outliers en plus d'un nouveau *port : 1341* qui est utilisé pour la première fois dans le flux de données. Sur le sous-espace des *adresses IP 10.32.0.201* et *10.32.1.206* sont rassemblés dans un même cluster avec deux autres *adresses IP*. Deux des *adresses IP* détectées comme outliers dans la figure 1 disparaissent dans cette fenêtre, ceci peut s'expliquer par le fait que ces éléments ne sont pas véritablement des outliers.

De la même façon, nous pouvons remarquer qu'il y a d'anciens outliers qui fusionnent avec d'autres éléments dans le même cluster. De nouveaux outliers peuvent aussi apparaître ou disparaître dans le temps.

Pour valider les résultats obtenus, nous appliquons l'algorithme *MCOD* sur les mêmes fenêtres. Nous avons également appliqué notre approche et l'algorithme *MCOD* sur tout l'ensemble de données non-divisé en fenêtres. Les outliers détectés par notre approche sont sur le tableau ci-dessous (tableau 1), les outliers en caractère gras sont ceux détectés aussi par l'algorithme *MCOD*.

Fenêtres	Ports	Adresses IP
1	**514, 53, 0, 80**	**172.23.0.1, 172.23.0.10**, 10.32.0.201, 10.32.1.206
2	**80, 0, 1341**	
3	**80, 0**, 1341	
4	**80, 53**, 0, 1333	10.32.0.100, 172.23.0.10
5	**1761, 80**, 0, 53	172.23.0.10
6	**1429, 80**, 0	172.28.29.9, 10.32.1.204, 10.32.0.202
7	**1772, 80**, 0	10.32.0.205
8	**80, 0, 1890**	10.32.0.100, 10.32.1.202
9	**80, 0, 1914**	
10	**80, 0**, 1914	10.32.1.204, 10.32.1.201
Flux	**514, 53, 0, 80**	**172.23.0.1, 172.23.0.10**

TAB. 1 – *Comparaison entre les résultats obtenus par notre approche et l'algorithme* MCOD

Nous remarquons que notre approche détecte les mêmes outliers détectés par l'algorithme *MCOD* sur tout l'ensemble de données. Toutefois, lorsque l'algorithme *MCOD* est appliqué sur les fenêtres, les résultats sont différents. Sur la première, la deuxième et la huitième fenêtre, les mêmes outliers sont détectés sur le sous-espace des *ports*. Sur les autres fenêtres, l'algorithme de *MCOD* détecte presque les mêmes outliers que ceux détectés par notre approche, sauf quelques différences. Sur le sous-espace des *adresses IP*, les résultats sont complètement différents. Ceci peut s'expliquer par le fait que *MCOD* ne prend pas en compte les partitions, à l'inverse de notre approche qui détecte les outliers parmi les éléments qui ont les mêmes représentants seulement.

Les résultats similaires obtenus sur l'ensemble des données par l'algorithme *MCOD* et par notre approche prouvent que notre approche peut détecter des outliers. De plus, notre approche permet de visualiser ces outliers et les changements des clusters dans le temps.

5 Conclusion

Nous proposons une approche qui utilise un algorithme de clustering basé sur le voisinage pour la détection d'outliers dans un flux de données complexes (trafic réseau). Le flux de données est divisé en fenêtres et les données en sous-espaces. En se basant sur les relations entre les sous-espaces, un élément peut représenter plusieurs éléments sur un autre sous-espace.

En premier lieu, le clustering est appliqué sur un seul sous-espace, les clusters obtenus sont considérés comme des représentants de partitions sur un autre sous-espace. L'algorithme est ensuite appliqué sur le sous-espace partitionné, mais uniquement sur les éléments de chaque partition séparément. Nous pouvons itérer ce processus sur n sous-espaces. Les éléments qui n'appartiennent pas à un cluster sont considérés comme des outliers. Les éléments qui appartiennent à un cluster avec peu d'éléments sont considérés comme suspects.

Nos expérimentations sont réalisées sur un flux de données complexes (des fichiers log). Nous avons appliqué notre approche pour détecter les machines avec un comportement atypique dans un réseau. Nous avons choisi de diviser les données sur deux sous-espaces. Un clustering basé sur le voisinage est appliqué sur le premier sous-espace, les clusters obtenus sont des représentants de partitions sur le deuxième sous-espace. Ensuite, le clustering est appliqué sur les éléments appartenant aux mêmes partitions séparément. Nous avons pu identifier visuellement les mêmes outliers détectés par l'algorithme *MCOD* et suivre leur évolution dans le temps. Une vérification numérique sur l'ensemble de données nous montre que les outliers détectés ont vraiment un comportement atypique.

En perspective, nous avons l'intention d'utiliser une visualisation qui permet une interaction entre l'utilisateur et l'algorithme de clustering. Nous projetons aussi de tester notre approche avec de différents ensembles de données et d'itérer l'approche sur plusieurs sous-espaces choisis automatiquement.

Références

Aggarwal, C. (2013). A survey of stream clustering algorithms. In *Data Clustering : Algorithms and Applications*, pp. 229–253. CRC Press.

Angiulli, F. et C. Pizzuti (2002). Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '02, London, UK, UK, pp. 15–26. Springer-Verlag.

Boussaid, O., P. Gançarski, F. Masseglia, B. Trousse, G. Venturini, et D. A. Zighed (2005). *Fouille de données complexes*. Revue des Nouvelles Technologies de l'Information. Cépaduès.

Breunig, M., H.-P. Kriegel, T. NG, Raymond, et J. Sander (2000). Lof : Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, Dallas, Texas, United States, pp. 93–104. ACM.

Chandola, V., A. Banerjee, et V. Kumar (2009). Anomaly detection : A survey. *ACM Comput. Surv. 41*(3), 15 :1–15 :58.

Guha, S., N. Mishra, R. Motwani, et L. O'Callaghan (2000). Clustering data streams. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, FOCS '00, Washington, DC, USA, pp. 359–366. IEEE Computer Society.

Hawkins, D. M. (1980). *Identification of Outliers*. Monographs on applied probability and statistics. Chapman and Hall.

Hodge, V. et J. Austin (2004). A survey of outlier detection methodologies. *Artif. Intell. Rev. 22*(2), 85–126.

Johnson, R. A. et D. W. Wichern (1992). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.
English

Knorr, E. M., R. T. Ng, et V. Tucakov (2000). Distance-based outliers : algorithms and applications. *The VLDB Journal 8*(3-4), 237–253.

Kontaki, M., A. Gounaris, A. N. Papadopoulos, K. Tsichlas, et Y. Manolopoulos (2011). Continuous monitoring of distance-based outliers over data streams. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ICDE '11, Washington, DC, USA, pp. 135–146. IEEE Computer Society.

Pinheiro, P., Y. Didry, O. Parisot, et T. Tamisier (2014). Traitement visuel et interactif dans le logiciel cadral. In *Atelier visualisation d' ? ?informations, interaction et fouille de données (GT-VIF, EGC 2014)*, Rennes, France.

Ramaswamy, S., R. Rastogi, et K. Shim (2000). Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec. 29*(2), 427–438.

VAST (2012). Visual analytics community. http ://www.vacommunity.org/VAST+Challenge+2012.

Summary

We propose an outlier detection approach based on a clustering algorithm. Our approach is developed to be applied on a computer network. The computer networks traffic generates complex data streams that represent the network events. To deal with this data, we propose to split the streams into several windows. In each window, the data is divided on subspaces (dimensions subsets). A clustering algorithm is applied on each subspace. Based on the existing relations between the different subspaces, the obtained clusters on one subspace can represent partitions on another subspace. We perform tests on firewall logs data sets. We choose to test our approach with two subspaces and to visualize the results with neighborhood graphs in each window. A comparison is provided between the obtained results and the *MCOD* algorithm results. We can identify visually the outliers events and observe the evolution of the stream.

La découverte des règles d'association dans un contexte distribué avec des données manquantes : Décomposition tensorielle.

Isam El Ayyadi *, Mourad Ouiziri *, Salima benbernou *,Mohamad Younas **

*Université Paris Descartes, Sorbonne Paris Cité, France
{ isam.el-ayyadi,mourad.ouiziri,salima.benbernou }@parisdescartes.fr,
http://lipade.mi.parisdescartes.fr/
** Oxford Brookes University, Oxford, UK
m.younas@brookes.ac.uk
http://cms.brookes.ac.uk/staff/MYounas/

Résumé. Le recueil des flux de données à travers des capteurs est devenu essentiel dans notre vie de chaque jour, allant de la surveillance du trafic en temps réel aux interventions d'urgence et de surveillance de la santé. Le volume des données entrant est généralement trop élevé pour être stocké et les calculs sur le flux doivent être exécutés en temps réel pour détecter rapidement des événements intéressants (par exemple : la détection et la notification d'accident, le contrôle de congestion du réseau, gestion des pannes de réseau, détection d'intrusion). Cependant, plusieurs de ces événements isolés peuvent également être conjointement surveillés et corrélés afin d'adapter le comportement du système et de prendre les mesures appropriées lors d'une détection d'anomalie.

Dans cet article nous présentons une nouvelle technique permettant de découvrir les règles d'association manquantes dans un réseau multimodal. L'approche proposée est basée sur la décomposition d'un tenseur de confiance avec des valeurs manquantes. Elle est validée par des résultats expérimentaux qui montrent son importance et sa viabilité.

1 Introduction

Aujourd'hui, chaque organisation est confrontée à la manipulation d'une quantité importante de données qui proviennent de sources multiples : données météorologiques, les données des capteurs, les pagesWeb etc.

De nombreux événements intéressants peuvent être détectés par la fouille de ces données provenant de différentes sources distribuées et les analyser à des fins spécifiques Byung-Hoon Park (2002).

Prenant l'exemple de l'industrie automobile Kargupta (2012), les futurs systèmes d'assistance au conducteur devront découvrir, recueillir et analyser des informations dynamiques sur l'environnement de la voiture et de l'état du conducteur. Pour cela Les données seront recueillies à partir de différentes sources qui sont distribués à travers différents endroits.

Cependant la perte d'information et des erreurs dans le processus de collecte sont les deux principaux facteurs qui contribuent à des données manquantes. La conséquence est que certains jeux de données importants peuvent être jetés ou mal analysés produisant des informations incorrectes Kanishka Bhaduri (2011). Les questions mentionnées ci-dessus seront étudiées dans le présent document dans un nouveau type d'environnement distribué à savoir le cloud computing.

Cet article aborde la question de découverte et de prévoir les règles d'association manquantes à partir de données incomplètes sur un nœud de nuage, en les corrélant avec des données provenant d'autres nœuds.

L'approche proposée est basée sur la décomposition tensorielle Tamara G. Kolda (2009). Les décompositions sont appliquées à des tableaux de données pour l'extraction et l'explication de leurs propriétés. Les offres proposées avec un réseau multimodal où les règles d'association manquants sont détectés et leurs confidences sont estimés. Pour cela, les règles d'association à savoir leurs confidences seront représentés sous forme de tableaux dans chaque nœud, où les tableaux obtenus sont incomplets et les résultats de la corrélation entre l'association règles avec d'autres nœuds sont représentés par un tenseur. En d'autres termes, notre objectif à la première tentative est de capturer la structure latente des données via d'ordre supérieur de factorisation en présence de règles d'association. La deuxième tentative est de récupérer les entrées manquantes vers une corrélation distribuée des règles d'association sur le réseau.

Pour valider les résultats obtenus, l'approche distribuée est discuté avec des expériences numériques sur des ensembles de données simulées en présence de données incomplètes et manquantes.

2 État de l'art

Dans le domaine de la fouille des données, l'extraction des règles d'association est une méthode populaire pour découvrir des relations intéressantes entre les variables dans les grandes bases de données Fayyad (1996). Dans cette section, nous donnons quelques définitions utiles traitant des règles d'association et aussi que les tenseurs.

2.1 Itemsets fréquents

— **Définition 1 :** Le support d'un itemset I, noté $Supp(I)$, est égal au nombre d'objets le contenant $Supp(I) \in [1; |G|]$ (appelé aussi support absolu de I).
La fréquence d'un itemset I, notée $Freq(I)$, est égale à $\frac{Supp(I)}{|G|}$.

— **Définition 2 :** Itemset fréquent
Un itemset I est dit fréquent si son support, $Supp(I)$, est supérieur ou égal à un seuil minimal d'objets, noté $minsupp$, fixé par l'utilisateur.

— **Proposition :** L'ensemble des itemsets fréquents forme un idéal d'ordre dans $(2M; \subseteq)$ (par rapport à la contrainte de fréquence) :
Tout sous-ensemble d'un itemset fréquent est aussi fréquent, Tout sur-ensemble d'un itemset infréquent est aussi infréquent.

2.2 Règles d'association

— Les mesures les plus utilisées sont le support et la confiance.
— Pour une règle $R : X \implies Y$, la confiance mesure la probabilité qu'une transaction contenant X contienne aussi Y (càd $P(Y|X) = supp(\frac{X \cap Y}{X})$).
— R est donc valide si $Conf(R) \geq minconf$.

2.3 Tenseurs

Un tenseur $\mathcal{X}$ peut être vu comme une généralisation de la notion de vecteur ou de matrice à plusieurs dimensions Herman. et Mechelen (2001) Appellof et Davidson (1981). L'ordre d'un tenseur désigne le nombre de ces dimensions. La figure 2.3 illustre un tenseur d'ordre 3 dont les dimensions successives sont I, J et K. Par analogie avec les matrices, nous pouvons noter un élément de ce tenseur par $x_{i;j;k}$. Un tenseur est un tableau multidimensionnel. L'ordre d'un tenseur représente le nombre de ces dimensions (ou modes).
Les coupes d'un tenseur d'ordre trois sont : les coupes horizontales $\mathbf{X}_{i::}$, les coupes latérales $\mathbf{X}_{:j:}$ et les coupes frontales $\mathbf{X}_{::k}$.
Les fibre d'un tenseur d'ordre 3 sont : les fibres du premier mode $\boldsymbol{x}_{:jk}$, du deuxième mode $\boldsymbol{x}_{i:k}$ et ceux du troisième mode $\boldsymbol{x}_{ij:}$.

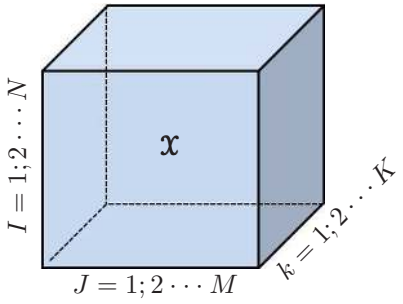


FIG. 1 – *Tenseur* $\mathcal{X} \in \mathbb{R}^{N \times M \times K}$

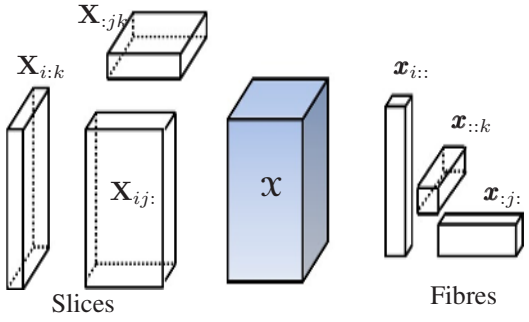


FIG. 2 – *Les coupes d'un tenseur.*

Le produit scalaire et la norme d'un tenseur sont définis par :

$$\mathcal{X} \cdot \mathcal{R} = \sum_{ijk} x_{ijk} x_{ijk}, \qquad \|\mathcal{X}\|^2 = \mathcal{X} \cdot \mathcal{X} = \sum_{ijk} x_{ijk}^2$$

3 Modèle et algorithme de résolution

3.1 Modèle

Dans cette section, on présentera le modèle proposé. L'objectif est de prévoir des règles d'association dans l'environnement de Cloud Computing. dans un réseau multimodal, les données sont réparties entre les différents nœuds $N_1, N_2 ... N_R$ (ou des systèmes informatiques) qui sont reliés par des réseaux. Nous traitons la situation où les données découvertes et extraites

de différents nœuds sont incomplètes.

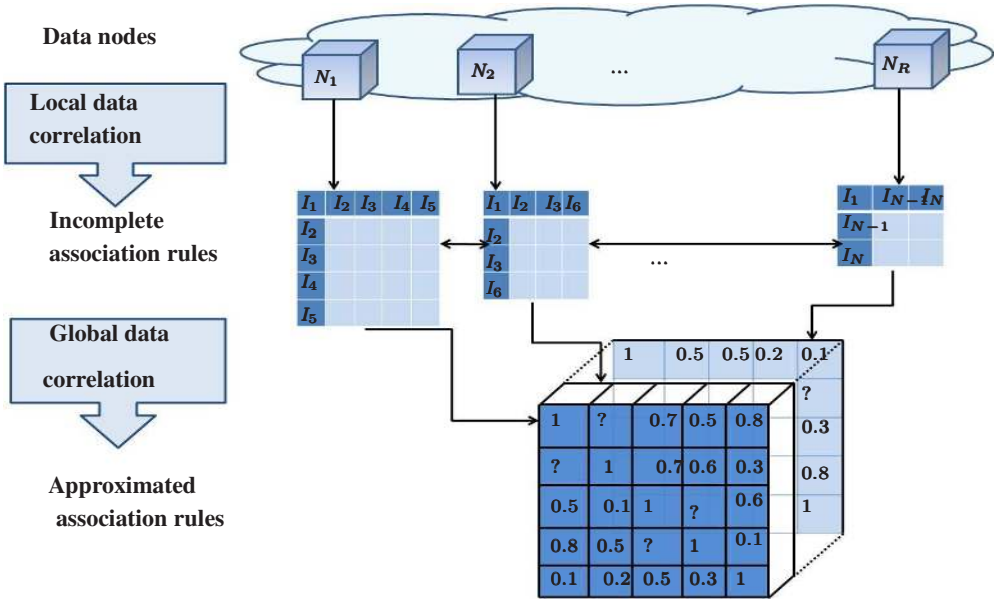


FIG. 3 – *Framework de la corrélation des données distribuées dans le cloud*

3.2 Algorithme

Pour l'extraction locale des relations d'associations, on applique l'algorithme Apriori introduit par Agrawal (1994) qui utilise une méthode bottom-up dans laquelle, à chaque étape, les sous-ensembles fréquents sont élargis d'un item. L'idée de base d'Apriori est qu'un itemset est fréquent si tous ses sous-ensembles sont fréquents.

Le modèle de corrélation 3.2sera représenté par le tenseur $\mathcal{R}$, on notera par ? les valeurs des confiances inconnues. Un fibre $\boldsymbol{r}_{ij:}$ représente les confiances entre les itemsets i et j dans tous les noeuds.

On définit la matrice $\mathcal{W}$ par :

$$\mathcal{W}_{ij} = \begin{cases} 1 & \text{si } X_{i,j,k} \text{ est connu} \\ 0 & \text{si } X_{i,j,k} \text{ est manquant} \end{cases}$$

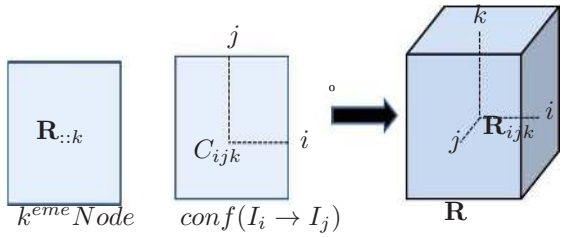


FIG. 4 – *Tenseur de confiance*

Le problème de l'approximation des valeurs manquantes dans le tenseur de confiance $\mathcal{R}$ est équivalent à la recherche d'un tenseur $\mathcal{X}$ qui minimise la forme quadratique suivante :

$$\mathtt{f}(\mathcal{X}) = \sum_{i,j=1}^{N} \mathcal{W}_{ij} \|\mathcal{R}_{ij:} - \mathcal{X}_{ij:}\|^2$$

On applique la décomposition ParaCand decomposition (**CP**)au tenseur $\mathcal{X}$: La forme L de-

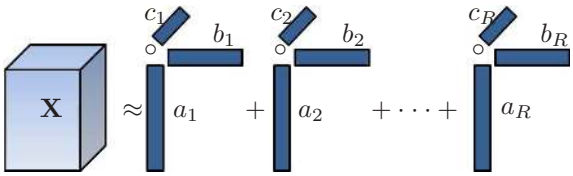


FIG. 5 – *CP*

vient :

$$\mathtt{L}(\mathcal{X}) = \sum_{i,j=1}^{N} \sum_{r=1}^{R} \mathcal{W}_{ij} (\mathcal{R}_{IJR} - \sum_{k=1}^{K} A_{ik} B_{jk} C_{rk})^2$$

Nous appliquerons la méthode du gradient pour minimiser L. Les dérivées partielles de L suivant les trois directions sont données par :

$$\begin{aligned} \frac{\partial \mathtt{L}}{\partial X_{ik}} = & \sum_{j=1}^{N} \sum_{r=1}^{R} \mathcal{W}_{ij} (\sum_{k=1}^{K} X_{ik} Y_{jk} Z_{rk}) Y_{jk} Z_{rk} \\ & - \sum_{j=1}^{N} \sum_{r=1}^{R} \mathcal{W}_{ij} \mathcal{R}_{i,j,r} Y_{jk} Z_{rk} \end{aligned}$$

La découverte des règles d'association : Décomposition tensorielle.

$$\frac{\partial \mathtt{L}}{\partial Y_{jk}} = \sum_{i=1}^{N}\sum_{r=1}^{R}\mathcal{W}_{ij}(\sum_{k=1}^{K}X_{ik}Y_{jk}Z_{rk})X_{ik}Z_{rk} - \sum_{i=1}^{N}\sum_{r=1}^{R}\mathcal{W}_{ij}\mathfrak{R}_{i,j,r}X_{ik}Z_{rk}$$

$$\frac{\partial \mathtt{L}}{\partial Z_{rk}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{W}_{ij}(\sum_{k=1}^{K}X_{ik}Y_{jk}Z_{rk})X_{ik}Y_{jk} - \sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{W}_{ij}\mathfrak{R}_{i,j,r}X_{ik}Y_{jk}$$

4 Résultats

Dans cette section, nous simulons des données afin d'évaluer la performance de l'algorithme proposé en termes de sa capacité à trouver les valeurs de confiance manquantes.
Pour cela, nous générons une base de données D. Pour chaque nœud, on enlève une partie aléatoire de la base de données. Les bases de données qui en résultent sont : $\{D_1 \cdots D_{10}\}$. L'algorithme Apriori est appliqué à chaque nœud $\{N_i | i = 1 \cdots 10\}$.Le tableau **??** donne des détails sur les statistiques de notre base de données.

TAB. 1 – *Application des statistiques*

Nœuds	10
Nombre total d'Itemsets fréquents	55
Nombre total de confiances manquantes	2360
Tenseur de confiance $\mathcal{X} \in \mathbb{R}^{55\times55\times10}$	30250

Aprés l'application de notre algorithme, nous avons obtenu les résultats suivants :

TAB. 2 – *Model Results*

	N_1	N_2	N_3	N_4
Nombre des Itemsets fréquents	46	50	52	54
Erreur relative ≈	6%	7%	3%	5%

N_5	N_6	N_7	N_8	N_9	N_{10}
45	48	53	48	54	51
6%	7%	8%	7%	1.5%	5%

Globalement, les approximations pour les nœuds 3 et 9 sont intéressants car leurs données sont fermées. Toutefois, les données de nœud et les longerons 7 sont loin de celles des autres nouds, pour laquelle l'erreur d'approximation est la plus élevée.

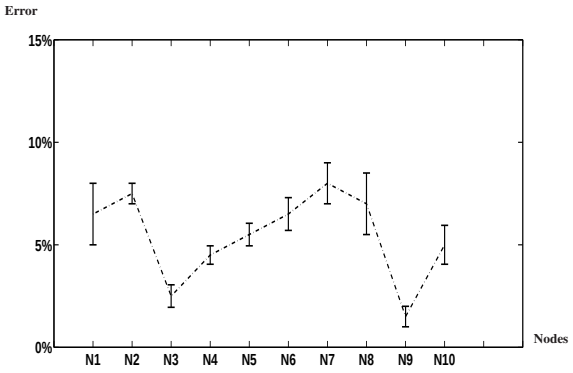


FIG. 6 – *L'erreur relative*

La figure 8 montre l'évolution de l'erreur moyenne en fonction de la quantité de données manquantes. La courbe obtenue montre une certaine stabilité dans notre modèle. En fait, l'effet de l'évolution est linéaire, avec une valeur maximale qui ne dépasse pas 15%.

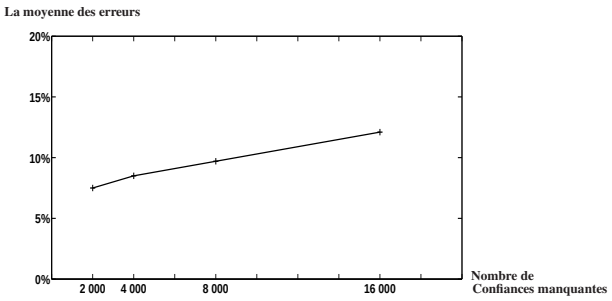


FIG. 7 – *L'évolution de l'erreur*

5 Conclusion

Dans cet article, on a abordé le problème de la découverte des règles d'association manquantes dans le cas où les données sont réparties entre différents nœuds du cloud et certaines données sont manquantes ou erronées. L'algoritme d'approximation est basé sur la décomposition tensorielle. Diverses expériences ont été menées et les résultats obtenus sont convaincants.

Dans l'approche actuelle, le tenseur de confiance considère que les données manquantes ayant des valeurs nulles est basé sur des hypothèses relativement simple. L'avenir comprend l'élargissement du cadre pour gérer (1) hypothèse plus complexe (2) découvrir et analyser les données d'applications en streaming.

Références

Agrawal, R. Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 487–499.

Appellof, C. J. et E. R. Davidson (1981). Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluent. *Anal. Chem*, 2053âĂŞ2056.

Byung-Hoon Park, H. K. (2002). Distributed data mining: Algorithms, systems, and applications. pp. 341–358.

Fayyad, U.M., P.-S. G. S. (1996). From data mining to knowledge discovery : An overview. *An Advances in Knowledge Discovery and Data Mining*, 1–34.

Herman., K. et I. V. Mechelen (2001). Three-way component analysis : Principles and illustrative application. *Psychological Methods 6*, 84âĂŞ110.

Kanishka Bhaduri, Kamalika Das, K. D. B. (2011). Scalable, asynchronous, distributed eigen monitoring of astronomy data streams. *Statistical Analysis and Data Mining 4*, 336–352.

Kargupta, H. (2012). Connected cars: How distributed data mining is changing the next generation of vehicle telematics products. *S-CUBE*, 73–74.

Tamara G. Kolda, B. W. B. (2009). Tensor decompositions and applications. *SIAM Review 51*, 455–500.

Summary

An increasing number of data applications such as monitoring weather data, data streaming, data web logs, and cloud data, are going online and are playing vital in our every-day life. The underlying data of such applications change very frequently, especially in the cloud environment. Many interesting events can be detected by discovering such data from different distributed sources and analyzing it for specific purposes (e.g., car accident detection or market analysis). However, several isolated events could be erroneous due to the fact that important data sets are either discarded or improperly analysed as they contain missing data. Such events therefore need to be monitored globally and be detected jointly in order to understand their patterns and correlated relationships. In the context of current cloud computing infrastructure, no solutions exist for enabling the correlations between multi-source events in the presence of missing data. This paper addresses the problem of capturing the underlying latent structure of the data with missing entries based on association rules.This necessitate to factorize the data set with missing data.

The paper proposes a novel model to handle high amount of data in cloud environment. It is a model of aggregated data that are confidences of associations rules. We first propose a method to discover the association rules locally on each node of a cloud in presence of missing rules. Afterward, we provide a tensor based model to perform a global correlation between all the local models of each node of the network.

The proposed approach based on tensor decomposition, deals with a multi modal network where missing association rules are detected and their confidences are approximated. The approach is scalable in terms of factorizing multi-way arrays (i.e. tensor) in the presence of

missing association rules. It is validated through experimental results which show its significance and viability in terms of detecting missing rules.

Comparison of Crosslingual Similarity Measures for Multilingual Documents Clustering

Manuela Yapomo[*,**], Delphine Bernhard[*], Pierre Gançarski[**]

[*] LiLPa - Linguistique, Langues, Parole, EA 1339
[**] ICube - Laboratoire des sciences de l'Ingénieur,
de l'Informatique et de l'Imagerie, UMR 7357
Université de Strasbourg

Résumé. This paper compares the performance of one thesaurus-based approach against three lexicon-based techniques to the measurement of the crosslingual similarity of domain-specific texts. These methods are applied to an unstructured manually annotated corpus of texts in three languages: French, English and German. We investigate the correlation between these measures and human judgement as well as their ability to detect subtle (in the same topic) and broader (in related topics) differences in comparability. Additional experiments aim to determine the extent to which terminology helps improving measures of similarity in a specialised domain. Results suggest that injecting domain-specific knowledge, when available, is a good alternative to more shallow techniques.

1 Introduction

Text mining applications are increasingly targeting the extraction of information from unstructured multilingual corpora. Processing multilingual documents is more complex than dealing with documents in one language, due to the language gap. For instance, to perform multilingual document clustering, it is necessary to convert the documents into a common representation space, thus bridging the language gap. The challenge is then to obtain a faithful representation of the original documents from which similarity can be computed for clustering.

Two different kinds of multilingual corpora are distinguished : parallel corpora and comparable corpora. Parallel corpora are made up of source documents and their translations (McEnery et Xiao, 2007) ; comparable corpora can be defined as collections of multilingual documents that have similar contents. Comparable documents are easier to collect than exclusively parallel documents. Few studies have been devoted to the notion of "comparability", even though it is essential for the construction of high-quality comparable corpora. Ideally, comparable corpora to be used for text mining applications should contain highly comparable documents, with only few unrelated documents.

In this study, we compare the performance of one thesaurus-based measure to three lexicon-based measures for calculating the crosslingual similarity of texts. The underlying hypothesis is that the similarity measures should correlate with human judgement. The goal is also to assess whether existing measures can distinguish between different degrees of comparability in multilingual and topically related documents.

We detail the main trends in the state-of-the-art in section 2. Section 3 describes the corpus collected for the evaluation of comparability measures covered in section 4. Evaluation results are reported and discussed in section 5. Section 6 is a conclusion of our work.

2 State of the Art

2.1 Comparability

Corpus comparability has a large influence on the performance of systems for crosslingual terminology extraction (Morin et Prochasson, 2011), machine translation (MT) (Bin et al., 2010), crosslingual information retrieval (Talvensaari et al., 2007) among others. Su et Babych (2012) define the comparability of documents as their capacity to allow the extraction of parallel segments for machine translation systems. In the same vein, Li et Gaussier (2010) consider documents to be comparable if a significant amount of translation pairs can be extracted from these documents. According to them, the main property of comparable documents is to share parallel segments of texts.

<table>
<tr><th></th><th>Bekavac et al. (2004)</th><th>Skadiņa et al. (2010b)</th><th>Braschler & Schäuble (1998)</th><th>Pouliquen et al. (2004)</th></tr>
<tr><td rowspan="5">linguistic & extra-linguistic criteria</td><td></td><td>(1) parallel</td><td></td><td></td></tr>
<tr><td rowspan="4">(1) hard comparability</td><td rowspan="2">(2) strongly comparable</td><td>(1) same story</td><td>(1) same news story</td></tr>
<tr><td>(2) related story</td><td>(2) interlinked news story</td></tr>
<tr><td rowspan="2">(3) weakly comparable</td><td>(3) shared aspects</td><td rowspan="2">(3) loosely connected story</td></tr>
<tr><td>(4) common terminology</td></tr>
<tr><td>extra-linguistic criteria (only)</td><td>(2) light comparability</td><td>(4) non-comparable</td><td>(5) unrelated</td><td>(4) wrong link</td></tr>
</table>

TAB. 1 – *Some comparability scales for multilingual corpora*

Several relevance scales exist in the literature to account for the different degrees of comparability in multilingual corpora (Braschler et Schäuble, 1998; Bekavac et al., 2004; Pouliquen et al., 2004; Skadiņa et al., 2010) (see Table 1). These scales usually serve as guidelines for human comparability judgements on topically diverse collections of documents. This is the case with *TREC-like* (Braschler et Schäuble, 1998) or newspaper (Pouliquen et al., 2004) collections. In such collections, human judgement has a high probability to strongly correlate with automatic similarity, given the fact that documents found in such datasets usually refer to different topics. In the case of multilingual documents with the same or similar topic(s), a high correlation between human and automatic similarity is more difficult to achieve.

In this work, we investigate the ability of existing measures to grasp fine differences of comparability or to distinguish different degrees of comparability in less heterogeneous multilingual corpora, that is, within the same specific domain.

2.2 Approaches to Crosslingual Similarity Measurement

In this section, we present the traditional approaches for computing the similarity of multilingual texts. The main challenge is to bridge the language barrier, and thus to convert the texts to a common representation space that reflects their semantic content. These approaches can either be classified as language-dependent (translation based on bilingual lexicons or machine translation) or language-independent (concept space).

2.2.1 Bilingual Lexicons and Machine Translation

The simplest approach consists in translating all the documents to a common target language and then use monolingual text relatedness measures. Translation is usually performed either with a bilingual lexicon containing word translation pairs or with a machine translation system (Evans et al., 2004). The first approach necessitates fewer resources, but is also more likely to generate erroneous translations, given the ambiguity of natural language and the sparseness of cross-lingual lexicons which might not contain technical or recent terms.

Crosslingual lexical overlap and information retrieval methods can also be used to compute the similarity of multilingual texts. The first technique consists in comparing documents based on the amount of translation equivalents they share (Li et Gaussier, 2010; Su et Babych, 2012). Different from the lexical overlap methods, crosslingual information retrieval methods do not take into account all the content words in texts for similarity computation. Only keywords, obtained by means of term weighting measures are used. Similarity is measured by comparing documents through their keywords lists translated into a target language (Su et Babych, 2012). Another way is to run translated source keywords as queries against the target collection in order to retrieve similar texts (Baradaran Hashemi et al., 2010).

These techniques only exploit shallow form-based features but raise several problems, due to ambiguity or the limited coverage of lexicons/machine translation systems.

2.2.2 Concept Space

Multilingual thesauri and ontologies describe language-independent concepts and provide concept labels in several languages. For instance, the Eurovoc thesaurus contains 22 languages of the European Union. In order to obtain a representation of textual data in the space of concepts listed in a thesaurus or an ontology, the most straightforward method consists in looking for the concept names in the text itself. For instance, Pouliquen et al. (2003) describe a statistical method trained on manually annotated documents to identify concepts from the Eurovoc thesaurus in multilingual documents.

2.2.3 Combination of Methods

For many tasks, better results can be obtained by combining the output of several methods. A natural way of dealing with different representations of data is to try and combine them in

order to improve the results obtained by each form of representation on its own. For instance, Pouliquen et al. (2008) build independent interlingual representations based on (a) the EUROVOC thesaurus, (b) person and organisation names, (c) direct or indirect references to countries and (d) cognates. The overall comparability between documents is the linear combination of the similarities obtained for the four representations.

In this paper, we compare the performances of one thesaurus-based and three lexicon-based methods to crosslingual similarity of texts on a manually collected corpus.

3 Corpus

In this section, we describe the compilation of the corpus used for evaluating the methods covered in section 4. We provide statistics and properties of the initial annotated collection and additional collections exploited.

3.1 Initial Corpus

Corpus collection Documents on the topic of "biogas" were harvested and annotated by Master students of the University of Strasbourg, who study computational linguistics and translation. Different sources were used for the collection of documents :

- Articles in French, German and English were collected on the Web ;
- Some websites providing articles or links to articles in the three target languages, such as Presseurop or Linguee, were also used.

The corpus mainly contains general and specialised press articles as well as scientific articles. Documents in the corpus can be grouped in three subcorpora according to the languages involved. Table 2 reports statistics of the corpus :

Collections	French	English	German	Total
# of doc.	20	17	18	55
# of words	13,746	15,492	12,926	42,164

TAB. 2 – *Initial corpus statistics.*

For evaluation purposes, a human annotation of the data with comparability levels was performed.

Comparability Judgements In total, seventeen annotators assigned comparability judgements to texts in the corpus. All annotators worked on two languages depending on their linguistic competence.

As regards annotation guidelines, we adapted Braschler et Schäuble (1998)'s relevance scheme to suit our corpus. This scale takes into account comparable and unrelated texts while our data is also made up of parallel documents belonging to the topic of biogas. Table 3 illustrates our modification of Braschler and Schäuble's comparability scale.

Our classes	Braschler and Schäuble (1998)	Scores	Comments
(1) Parallel texts		5	translations
(2) Same story	(1) Same story	4	same event or topic from similar viewpoint
(3) Related story	(2) Related story	3	Same or similar event(s) from different viewpoint
(4) Shared aspect	(3) Shared aspect	2	Related events
(5) Common terminology	(4) Common terminology	1	Events may not be related but share terminology
(5) Unrelated	(5) Unrelated	0	

TAB. 3 – *Modification of Braschler and Schäuble's guidelines for comparability annotation*

Our tests are limited to comparability in two language directions : from French to English and from French to German. Judgements were not assigned for the English-German language pair because of the small number of assessors. Two judges assessed each pair of texts. The percentage of agreement, reflecting the number of times identical classes were assigned to pairs, is 34%. This weak score shows that even for human judges, assigning comparability classes from a five-level scale is not a simple task. We also computed the weighted percentage of agreement (the code can be found in the annotation task of the nltk metrics package in Python). It is based on a linear distance whose value increases as the gap between annotations gets larger. The weighted percentage of agreement is 79%. Finally, we computed the agreement based on a two-point scale (Paramita et al., 2012) by gathering results for classes 1-3 and 4-5 respectively in high comparability and low comparability classes. We get a very high percentage of agreement of 92% and a weighted Kappa of 0.363.

Based on the annotations, the pairs were organised in three classes ranging from less comparable (with scores of 0-1) to parallel (a score of 5), through very comparable (with scores from 2-4). The initial number of 200 annotated pairs was reduced to 134. This was done by discarding pairs whose scores belong to different classes. Fourteen parallel pairs of documents for which we do not need judgements with the other texts were added. Comparability judgments were simply induced from the the fact that these pairs are parallel. Table 4 below shows statistics of the final set of judgement pairs.

Classes	FR-EN	FR-DE	# of doc. pairs
Parallel	9	9	18
Very comp.	9	7	16
Less comp.	61	53	114
Total	79	69	148

TAB. 4 – *Judgment pairs by classes*

The assessment of the similarity measures in section 4 will rely on these human comparability judgements and their organisation in three groups : *parallel*, *very comparable* and *less comparable*, shown in Table 4.

3.2 Out-of-domain Corpus

In order to asses how well the comparability measures discriminate in-domain from out-of-domain texts, we used the TTC corpus [1], which is made up of texts belonging to the topic of "wind energy". Table 5 shows statistics of the out-of-domain corpus.

Collections	English	German	Total
# of doc.	26	27	52
# of words	70,780	51,184	121,964

TAB. 5 – *Out-of-domain corpus statistics*

We want to randomly select English and German languages' subsets of the TTC corpus to be similar both in terms of size and number of documents to the French part of the initial corpus. However, texts from the TTC corpus have 2,354 words in average while those from the initial corpus have 766 words. To compensate this size difference, we add to the French part of the initial corpus 9 texts from Wikipedia, on the topic of biogas, with a total of 36,683 words. This gives us a French corpus of 29 texts and 50,429 words.

To evaluate the ability of measures to discern documents of a topic (biogas) from those of another (wind), we choose to pair each English and German document from the TTC corpus with two randomly selected French documents in the initial and Wikipedia corpus.

4 Crosslingual Similarity of Texts

In this section, we describe the different measures of crosslingual document similarity that are tested in this work.

4.1 Lexicon-based Similarity

This approach basically uses a multilingual lexicon to identify common vocabulary in documents. The more translation equivalents two documents have in common, the higher their comparability.

Li et Gaussier (2010) propose measuring the comparability of corpora using a bilingual lexicon. We adapt this method to the comparability of single documents. It relies on a function that determines the number of times a translation from the translation set T_w of a word w from the source document D_s is found in the vocabulary of the target document D_t as follows :

1. http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html

$$\sigma(w, D_s) = \begin{cases} 1 & \text{if } T_w \cap D_t \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

The comparability measure M for this pair of documents is :

$$M(D_s, D_t) = \frac{\sum_{w \in D_s \cap L_s} \sigma(w, D_t) + \sum_{w \in D_t \cap L_t} \sigma(w, D_s)}{\#w(D_s \cap L_s) + \#w(D_t \cap L_t)}$$

where L is a bilingual lexicon made of L_s and L_t which are respectively its source and target language parts. In this method, no particular document preprocessing is performed.

Su et Babych (2012) also make use of bilingual lexicons to have source and target documents in a unique language. Monolingual similarity can then be applied. In contrast to Li et Gaussier (2010), some preprocessing steps are performed : part-of-speech tagging and stemming of content words, which help disambiguate some words and identify different word forms as a unique stem. The similarity of documents represented as index vectors is then computed with cosine.

In our experiments, we used lexicons built from the Europarl corpus (Koehn, 2005) with Anymalign (Lardilleux et al., 2012). To get the bilingual lexicons, we ran Anymalign for 3 hours. It resulted in 1,658,870 English-French and 1,632,782 and German-French alignments respectively. Multiword units and alignments with a number of occurrences under 10 were discarded. The final numbers are of 25,945 English-French and 19,694 German-French lexicon entries. For the sake of uniformity, we take into account translation candidates with a probability above 0.3 for all experiments as Su et Babych (2012). A second set of experiments with the injection of biogas terminology into the lexicons is carried out. This terminology comes from the initial corpus and was provided by the annotators. 51 term pairs were added to the English-French lexicon and 56 to the German-French lexicon.

4.2 Machine Translation-based Similarity

In a second method, Su et Babych (2012) extract features from machine-translated documents to derive comparability. The assumption is that machine translation works better as it partly solves the problem of ambiguity occurring with lexicons. Four features are taken into account to measure the similarity of the part-of-speech tagged and stemmed documents :

- Lexical mapping : vector representations of documents
- Structure : number of content words and sentences of documents
- Keywords : keywords list of documents built with TF-IDF
- Named entities : number of named entities

Cosine measures similarity for these four features separately. The four different similarities are combined in one global measure by summing their weighted values. The weight assigned to each similarity feature reflects its importance in the overall comparability. Lexical mapping has a weight of 0.5, structure and keywords 0.2 and named entities, 0.1.

Su et Babych (2012)'s lexical overlap and machine translation measures are some of the tools released under the ACCURAT project [2].

2. http://www.accurat-project.eu/

4.3 Thesaurus Indexing

We automatically annotated the documents using the topic-indexing tool Maui [3] (Medelyan et Witten, 2008). Maui combines the use of a thesaurus, to identify keyphrase candidates, and machine learning, to select the most significant candidates. As we did not have annotated data for the machine learning part, we only used the Maui candidate extraction tool to identify terms from the Agrovoc thesaurus.

The Agrovoc thesaurus [4] is provided by the FAO (Food and Agriculture Organisation of the United Nations) and covers the domains of agriculture, environment, etc. It contains over 30,000 concepts, with labels in up to 22 languages, including French, English and German. It is therefore particularly adapted to index our corpus.

In order to map document phrases to Agrovoc descriptors, Maui uses several preprocessing steps, including stopword removal and stemming. However, we did not allow for term reordering, which aligns candidates with descriptors if they contain the same tokens but in a different order.

The average number of Agrovoc descriptors automatically assigned to documents in the biogas domain are as follows : 112 for the German documents, 247 for the English documents and 129 for the French documents.

In the wind energy domain, there are on average 71 descriptors for the German documents, 125 for the English documents and 106 for the French documents.

Once the documents are indexed with Agrovoc descriptors, we compute the similarity of pairs of documents d_1 and d_2 with the Jaccard index :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where A is the set of Agrovoc descriptors in d_1 and B is the set of Agrovoc descriptors in d_2.

5 Results and Discussion

5.1 Performance of Similarity Measures

We present results in terms of average similarity at each comparability level taken into account in this study. The Pearson correlation between human judgement and automatic similarity is also reported. These results are reported in Table 6. We also compute the ratios between the average similarity scores of two close classes (see Table 7).

5.2 Discussion

We notice that for the initial corpus, the average similarities of pairs of documents in the classes *parallel* (*Par*) are the highest, which is coherent. However, the average similarity scores in the classes *very comparable* (*VC*) and *less comparable* (*LC*) are close although higher for *VC* class. This is also consistent. To further analyse these results, we compute the ratio between the averages of two close classes (see 7).

3. http://code.google.com/p/maui-indexer/
4. http://aims.fao.org/standards/agrovoc/about

Methods	Lexicon-based measures				MT-based measure		Thesaurus indexing	
	Li et Gaussier (2010)		Su et Babych (2012)					
Language pair	DE-FR	EN-FR	DE-FR	EN-FR	DE-FR	EN-FR	DE-FR	EN-FR
Parallel	0.473	0.573	0.533	0.616	0.540	0.614	0.176	0.268
Very comparable	0.166	0.196	0.376	0.327	0.266	0.333	0.072	0.099
Less comparable	0.141	0.169	0.328	0.289	0.208	0.245	0.049	0.090
Out-of/In-domain	0.073	0.110	0.198	0.118	0.114	0.095	0.034	0.054
Correlation	0.467	0.625	0.276	0.582	0.445	0.678	0.437	0.535

TAB. 6 – *Average comparability scores with correlation values for each method*

Methods	Lexicon-based measures				MT-based measure		Thesaurus Indexing	
	Li et Gaussier (2010)		Su et Babych (2012)					
language pair	DE-FR	EN-FR	DE-FR	EN-FR	DE-FR	EN-FR	DE-FR	EN-FR
VC/Par	0.350	0.342	0.705	0.530	0.492	0.542	0.409	0.369
LC/VC	0.849	0.862	0.872	0.883	0.781	0.735	0.680	0.909

TAB. 7 – *Ratios between average values of close classes*

The higher the ratio of C_1/C_2 – here VC/Par or LC/VC –, the less discriminant the similarity measure between a pair of documents of class C_1 and another pair of class C_2. For example, if this value is 1, the comparability of two documents from C_1 is identical to that of two documents of C_2. We observe that the thesaurus indexing measure performs better than the measures of Su et Babych (2012) to distinguish documents of classes VC and Par (minimum ratio values on all measures). It is nevertheless less effective than the lexicon-based measure of Li et Gaussier (2010). As regards differentiation of classes VC and LC, this measure yields the best performance for the language pair German-French and the lowest for the French-English pair. Although this measure does not realise the best performance as a whole, the majority of results reflects equivalent or better performances than those of the other measure.

To judge the measures by their ability to discern pairs of documents on the same topic (biogas) from those in dissimilar topics (wind energy vs. biogas), we pair each English/German document of the TTC corpus with two randomly selected French documents to compute their similarity. The row "Out-of/In-domain corpus" of Table 6 shows the average similarity scores of all pairs of unrelated documents formed that way. The measures are able to distinguish these pairs from others by assigning them lower scores, which is perfectly consistent.

Finally, to assess the reliability of these measures, we compute the Pearson coefficient, which indicates the correlation between automatic similarity and manual annotation. The low Pearson values, which indicate low correlation between automatic similarity and human judgement, is not surprising given the low inter-annotator agreements obtained for the manual annotation task. Only Li et Gaussier (2010)'s and Su et Babych (2012)'s MT-based measures achieve scores higher than 0.6.

As reported in section 4.3, the thesaurus-based method is quite simple, using a small similarity space by just counting concepts occurring in documents. This is not the case of the lexicon-based measures which take into account all content words for similarity computation. However, the thesaurus-based method performs similarly to the lexicon-based methods and even better than Su et Babych (2012)'s lexicon-based measure as regards correlation with human judgement.

5.3 Enhancing Lexicons with Terminology

In this section, we aim to measure the impact of terminology on comparability in a specialised domain. This is done by comparing average similarity values per level of comparability with and without domain-specific terms. Terms specific to the biogas domain were added to lexicons built from the Europarl corpus (section 4.1) used in the experiments. The results, detailed in Table 8, show that there is an overall increase of average similarity values. Nevertheless, the use of domain-specific terms yields a slight reduction of the average value for English-French pairs with Li et Gaussier (2010)'s method. The increase is more prominent in results from Su et Babych (2012)'s measure.

Methods	**Li et Gaussier (2010)**				**Su et Babych (2012)**			
	without terms		**with terms**		**without terms**		**with terms**	
Language pair	**DE-FR**	**EN-FR**	**DE-FR**	**EN-FR**	**DE-FR**	**EN-FR**	**DE-FR**	**EN-FR**
Parallel	0.473	0.573	0.481	0.577	0.533	0.616	0.582	0.693
Very comparable	0.166	0.196	0.172	0.195	0.376	0.327	0.420	0.450
Less comparable	0.141	0.169	0.143	0.170	0.328	0.289	0.364	0.402
Out-of/In-domain	0.073	0.110	0.076	0.111	0.198	0.118	0.203	0.140
Correlation	0.467	0.625	0.485	0.623	0.276	0.573	0.372	0.517

TAB. 8 – *Results of lexicon-based measures with and without additional terminology*

We observe in Table 8 that comparability scores of unrelated (Out-of/In-domain row) document pairs increase less than in other classes with Li et Gaussier (2010)'s measure. The exploitation of terminology has a greater impact for the German-French language pair. This pair is also the one having the smallest bilingual lexicon. This indicates that using a domain-specific lexicon, even small, can play a relevant role in similarity measurement in a specialised domain.

6 Conclusion

In this paper, we presented a comparison of several cross-lingual similarity methods. To this aim, we collected and annotated a domain-specific corpus in three languages. The manual annotation of fine-grained comparability levels in topically related documents is a difficult task. Likewise, automatic measures assign scores which follow the comparability scale, from unrelated to parallel, but with little difference for less comparable and very comparable documents. The results also show that even a simple thesaurus indexing approach, without any form of disambiguation or concept weighting, is good enough to distinguish parallel documents from comparable documents. The use of a thesaurus-based measure seems realistic in the context of a clustering task.

In future work, we aim at developping the thesaurus indexing-based similarity measure discussed in this article for performing cross-lingual document clustering, to build clusters of topically related multilingual documents. We would also like to perform a new annotation study, with fewer comparability levels, but with more diversified document topics.

Acknowledgements This work was funded through a PhD contract IdEx 2012 from the University of Strasbourg.

Références

Baradaran Hashemi, H., A. Shakery, et H. Faili (2010). Creating a Persian-English comparable corpus. In *Proceedings of the 2010 international conference on Multilingual and multimodal information access evaluation*, Padua, Italy, pp. 27–39.

Bekavac, B., P. Osenova, K. Simov, et M. Tadic (2004). Making monolingual corpora comparable : a case study of Bulgarian and Croatian. In *Proceedings of the 4th Language Resources and Evaluation Conference*, Lisbonne, pp. 1187–1190.

Bin, L., T. Jiang, K. Chow, et T. Benjamin K. (2010). Building a large english-chinese parallel corpus from comparable patents and its experimental application to SMT. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, Malta, pp. 42–49.

Braschler, M. et P. Schäuble (1998). Multilingual information retrieval based on document alignment techniques. *Research and Advanced Technology for Digital Libraries*, 513–513.

Evans, D. K., J. L. Judith L. Klavans, et K. R. McKeown (2004). Columbia Newsblaster : multilingual news summarization on the web. In *Demonstration Papers at HLT-NAACL 2004*, pp. 1–4.

Koehn, P. (2005). Europarl : A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, Volume 5.

Lardilleux, A., F. Yvon, et Y. Lepage (2012). Hierarchical Sub-sentential Alignment with Anymalign. In *Proceedings of the 16th annual conference of the European Association for Machine Translation (EAMT 2012)*, pp. 279–286.

Li, B. et E. Gaussier (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 644–652.

McEnery, A. M. et R. Z. Xiao (2007). Parallel and comparable corpora : What are they up to ? In *Incorporating Corpora : Translation and the Linguist* (Multilingual Matters ed.). Clevedon, UK : Anderman, G. & Rogers, M.

Medelyan, O. et I. H. Witten (2008). Domain-Independent Automatic Keyphrase Indexing with Small Training Sets. *Journal of the American Society for Information Science and Technology 59*(7), 1026–1040.

Morin, E. et E. Prochasson (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, pp. 27–34.

Paramita, M., P. Clough, A. Aker, et R. Gaizauskas (2012). Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, pp. 790–797.

Pouliquen, B., R. Steinberger, et O. Deguernel (2008). Story tracking : linking similar news over time and across languages. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, pp. 49–56.

Pouliquen, B., R. Steinberger, et C. Ignat (2003). Automatic annotation of multilingual text collections with a conceptual thesaurus. In *Proceedings of the EUROLAN workshop on Ontologies and Information Extraction*.

Pouliquen, B., R. Steinberger, C. Ignat, E. Käsper, et I. Temnikova (2004). Multilingual and cross-lingual news topic tracking. In *Proceedings of the 20th International Conference on Computational Linguistics*, Genève, Suisse.

Skadiņa, I., A. Aker, V. Giouli, D. Tufiş, R. Gaizauskas, M. Mieriņa, et N. Mastropavlos (2010). A collection of comparable corpora for under-resourced languages. In *Proceedings of the Fourth International Conference Baltic HLT*, pp. 161–168.

Su, F. et B. Babych (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-) parallel translation equivalents. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, pp. 10–19.

Talvensaari, T., J. Laurikkala, K. Järvelin, M. Juhola, et H. Keskustalo (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM TOIS 25*(1), 4.

Summary

Cet article compare la performance d'une approche basée sur un thésaurus à celle de trois techniques basées sur des équivalents de traduction pour le calcul de la similarité translingue de textes spécifiques à un domaine. Ces méthodes sont appliquées à un corpus non structuré et annoté de textes français, anglais et allemand. Nous étudions essentiellement la corrélation entre ces mesures et le jugement humain et leur capacité à détecter des différences subtiles (dans un même thème) et plus larges (en thèmes connexes) de comparabilité. Des expériences supplémentaires nous permettent d'observer le rôle de la terminologie dans l'amélioration des mesures automatiques de comparabilité dans un domaine spécialisé. Les résultats suggèrent que l'injection de connaissances spécifiques au domaine (ici à partir d'un thésaurus), lorsqu'elles sont disponibles, est une bonne alternative à des techniques peu profondes.

Gestion de l'incertitude dans le cadre d'une extraction des connaissances à partir de texte

Fadhela Kerdjoudj[*,**] Olivier Curé[*]

[*]Université Paris-Est Marne-La-Vallée, LIGM, CNRS UMR 8049, France
fadhela.kerdjoudj@univ-mlv.fr, ocure@univ-mlv.fr
[**]GEOLSemantics 12 rue Raspail, 94250, Gentilly

Résumé. Le domaine de l'extraction de connaissances à partir de texte nécessite des méthodes permettant de détecter et de manipuler l'incertitude. En effet, de nombreux textes contiennent des informations dont la véracité peut être remise en cause. Il convient alors de gérer de manière efficace ces informations afin de représenter les connaissances de manière explicite. Une première démarche consiste à identifier les différentes formes d'incertitudes pouvant intervenir durant un processus d'extraction. Puis, nous proposons une représentation RDF basée sur une ontologie développée destinée à modéliser l'incertitude.

1 Introduction

La multiplication de sources textuelles sur le Web offre un champ pour l'extraction de bases de connaissances. Dernièrement, de nombreux travaux dans ce domaine sont apparus ou se sont intensifiés, Dong et al. (2014), Niu et al. (2012). Dans le domaine de la construction automatique de bases de connaissances, il est nécessaire de faire collaborer des approches linguistiques, pour extraire certains concepts relatifs aux entités nommées, aspects temporels et spatiaux, à des méthodes issues des traitements sémantiques afin de faire ressortir la pertinence et la précision de l'information véhiculée. Pour présenter un intérêt à l'échelle du Web, les traitements linguistiques doivent être multi-sources et inter-lingues.

GEOLSemantics est une entreprise qui s'appuie sur une expérience cumulée de plusieurs années dans le monde de la linguistique et de la sémantique. Cette société propose une solution logicielle de traitement linguistique basée sur une analyse linguistique profonde. Le but est d'extraire automatiquement, d'un ensemble de textes, des connaissances structurées, localisées dans le temps et l'espace, des concepts, des relations et des événements impliquant des Entités Nommées. Pour représenter les connaissances extraites du texte, nous avons opté pour les technologies du web sémantique. Nous représentons donc nos extractions sous forme de triplets RDF et exploitons une ontologie que nous avons spécifiquement développées pour couvrir certains domaines. Cette approche permet de relier les résultats de nos extractions à des connaissances externes contenues dans des bases de références du Linked Open Data, Bizer et al. (2008), tels que Dbpedia et Geonames, et ainsi d'effectuer un certain nombre de vérifications. Les textes actuellement traités par GEOLSemantics sont issus d'articles de presse et traitent de sujets divers allant de la politique au sport en passant par des faits divers.

Gestion de l'incertitude.

Lors de l'analyse linguistique, il arrive que l'information traitée contienne des imperfections. En effet, l'information peut être bruitée, biaisée, implicite, imprécise, incohérente ou incertaine. Dans notre travail, nous accordons un intérêt particulier à l'incertitude. Il s'agit de l'une des imperfections les plus courantes et qui apporte une importante information à la connaissance acquise. Notre première contribution porte sur une catégorisation de l'incertitude lors des différentes phases d'extraction. Notre seconde contribution se situe au niveau de la représentation de l'incertitude dans le graphe RDF.

Cet article est organisé comme suit : Dans la Section 2, nous introduisons les grandes lignes du système d'extraction de GEOLSemantics. La Section 3 détaille notre catégorisation de l'incertitude. Ensuite, nous motivons et présentons notre représentation de l'incertitude dans un graphe RDF. Nous concluons et présentons nos travaux futurs en Section 5.

2 Acquisition de l'information et représentation des connaissances

L'acquisition de l'information comporte plusieurs étapes distinctes allant du simple découpage du texte en mots à la représentation de son contenu. Ces étapes consécutives consistent en :

- *L'analyse morpho-syntaxique :* il s'agit de la mise en évidence des structures d'agencement des catégories grammaticales (nom, verbe, adjectif, etc.), afin d'en découvrir les relations formelles ou fonctionnelles (Exemple : sujet, verbe et complément).
- *L'analyse sémantique :* il s'agit de l'étude linguistique du sens. L'objectif principal de cette analyse est de déterminer le sens de chaque mot dans la phrase.
- *L'extraction de connaissances :* permet de mettre en évidence des entités nommées et des relations relatives à un concept particulier. Dans cette étape, nous nous basons sur la notion de marqueur ou déclencheur. Il s'agit d'un terme ou une expression indiquant la présence d'un concept donné. Exemple : les termes *partir, aller, voyager* sont des déclencheurs indiquant un *déplacement*. Ces déclencheurs indiquent qu'une relation relative à un concept est présente et peut être extraite. A chaque concept de l'ontologie correspondra une liste de déclencheurs possibles, ceci permet de sélectionner les règles d'extraction de connaissances à appliquer.
- *La mise en cohérence :* permet de consolider les connaissances extraites, notamment le regroupement des entités nommées et la résolution des dates relatives.
- *L'enrichissement :* permet de compléter l'information à partir des données du Linked Open Data.

A l'issue de ces traitements, grâce à la représentation RDF, Cyganiak et al. (2013), nous disposons d'un ensemble de triplets qui permettra d'utiliser par la suite la connaissance exprimée dans le texte étudié. Le format RDF présente l'avantage d'une syntaxe simple sous forme de triplets (sujet-prédicat-objet). Il repose sur des URI (Uniform Resource Identifier) qui permettent d'identifier de manière unique chaque ressource et de faciliter la distribution et la publication des données sémantiques sur le web à travers le Linked Open Data [1]. Néanmoins, ce traitement suppose que les données fournies sont toutes fiables et sûres. Ceci n'est pas toujours garanti. Le but de notre travail est de pondérer la connaissance extraite en fonction de la fiabilité de

1. `http://linkeddata.org/`

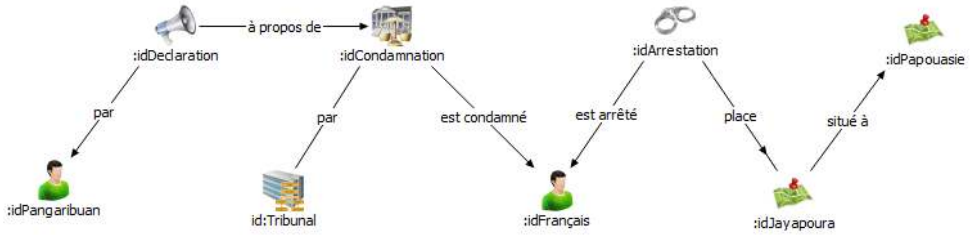


FIG. 1 – *Représentation graphique des triplets RDF après extraction.*

l'information initiale. En effet, celle-ci peut être remise en cause suivant différents critères que nous catégorisons dans la section suivante. L'exemple 2.1 permettra d'illustrer nos propos tout au long de cet article :

Exemple 2.1 *Le figaro, Par Jeanne Fremin Du Sartel Publié le 02/09/2014 à 16 :43*
***Détenus** dans un **centre** de rétention à Jayapura, **capitale** de la Papouasie, les **Français** vont très probablement devoir **comparaître** devant un **tribunal** pour avoir triché sur la nature de leur visa, prévient leur **avocat** Aristo Pangaribuan.*

Dans cet exemple, nous remarquons la présence de plusieurs déclencheurs que nous avons mis en gras :

- détenus : permet d'identifier une arrestation.
- centre, capitale : annonceurs d'entités nommées de type Lieu.
- comparaître : permet d'identifier un jugement.
- avocat : entité nommée de type Personne
- tribunal : entité nommée de type Organisation

La figure 1 propose une représentation graphique de l'extraction textuelle de notre exemple. Celle-ci décrit la déclaration de l'avocat concernant l'arrestation en Papouasie puis la condamnation des français. Les noeuds représentent les instances de concepts de l'ontologie, alors que les arcs formalisent les liens entre ces concepts.
Cependant, la condamnation est annoncée comme étant "probable", il est donc nécessaire d'introduire une incertitude concernant cet événement. Dans la suite de cet article nous décrirons les étapes qui nous permettent d'identifier et d'inclure cette incertitude dans notre graphe RDF.

3 Catégorisation de l'incertitude

La fiabilité de l'information est très souvent remise en cause. En effet, l'imprécision, l'incertitude ou encore l'incomplétude sont des problèmes récurrents dans le traitement de l'information. Dans cet article, nous accordons un intérêt particulier à l'aspect incertain de l'information acquise. Notre démarche est de considérer les modalités d'acquisition et d'expression de l'information qui constituera la connaissance, ainsi que le traitement de cette dernière jusqu'à la génération d'un graphe RDF permettant de la stocker dans des bases de connaissances afin de pouvoir la réutiliser par la suite. Nous considérons que le degré de confiance accordé à une

information est influencé par différentes phases du traitement. Pour cela, nous identifions trois niveaux où peut intervenir de l'incertitude dans le traitement de l'information.

3.1 Incertitude pré-extraction

Dans cette partie, nous considérons la source du texte ainsi que les métadonnées qui lui sont associées. Ces métadonnées peuvent faire référence à l'auteur de l'article, l'organisation chargée de le publier, le contexte abordé... Ces métadonnées sont obtenues lors de la récupération du texte. L'information peut provenir de sources variées avec différents niveaux de fiabilité, il s'agira donc lors de cette étape de qualifier la fiabilité de la source utilisée. En effet, la cotation d'informations est une tâche qui vise à mesurer la qualité d'une information, Mombrun et al. (2010), et en particulier la confiance qu'on peut lui accorder. Cette confiance peut varier en fonction des informations qu'elles produisent, la compétence des auteurs, la certitude qu'elles expriment, la vraisemblance du contenu ou l'existence de confirmations ou d'infirmations...

Pour modéliser ces métadonnées nous nous basons sur l'ontologie PROV-O [2], un standard RDF certifié par le W3C. PROV-O est une ontologie qui permet de modéliser les informations liées à la source des données. Elle décrit les entités, les activités et les agents impliqués dans la production d'informations, ainsi que la qualité, la fiabilité et la confiance associée, Lebo et al. (2013). Dans Missier et al. (2013), les auteurs proposent un tutoriel qui permet de modéliser les métadonnées ainsi que la confiance associée à la source.

Par ailleurs, nous disposons d'une base de connaissances sur la confiance accordée aux sources. À chaque source sera associé un degré de confiance évaluant la fiabilité des informations qu'elle fournit. Ce degré de confiance dépend également de l'utilisateur. En effet, pour une même source, la confiance accordée peut varier suivant l'utilisateur.
Nous avons alors créé une base de connaissances (que nous nommerons Trust) permettant de stocker la confiance associée à une source pour chaque utilisateur. L'ontologie de cette base comporte trois classes : *Source, User et Trustworthiness*.
La classe *Source* correspond à la super-classe d'une hiérarchie de concepts comportant des classes décrivant des auteurs et éditeurs, e.g., bloggeur, journaliste, chaîne TV, journal. La classe *User* désigne l'utilisateur du programme. Enfin, la classe *Trustworthiness* décrit le degré de confiance qu'attribue l'utilisateur à la source considérée.
Le degré de confiance permet d'évaluer la fiabilité de la connaissance véhiculée. Nous avons décidé de représenter le degré de confiance par un nombre réel compris entre 0 et 1. La représentation en nombres continus (intervalle) poserait par la suite plus de problèmes, lors du raisonnement.

Exemple 3.1 *Un utilisateur donné peut croire le journal "Le figaro" avec un degré de confiance de* 0.8 *alors qu'un autre utilisateur décidera de croire cette source à* 0.6.

Les informations fournies dans cet exemple nous permettent d'ajouter les triplets suivants dans notre base :
: $idTrustworthiness1 - hasSource - : idLeFigaro$;
: $idTrustworthiness1 - hasUser - : idUserX$;
: $idTrustworthiness1 - hasTrust - 0.8$;
: $idTrustworthiness2 - hasSource - : idLeFigaro$;

2. http://www.w3.org/TR/prov-o/

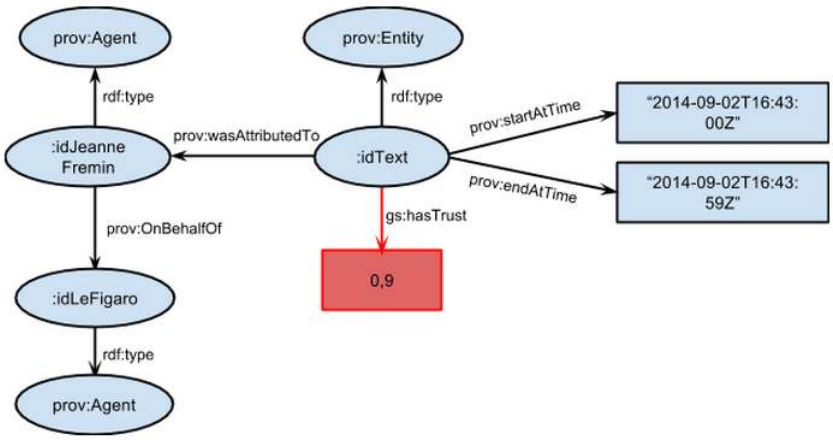


FIG. 2 – *Représentation de l'exemple 2.1 avec Prov-o.*

: $idTrustworthiness2 - hasUser -$: $idUserY$;
: $idTrustworthiness2 - hasTrust - 0.6$;

La figure 2 décrit le graphe RDF associé à la représentation des métadonnées de l'exemple 2.1, en utilisant la représentation recommandée dans Prov. La propriété *wasAttributedTo* permet de lier le contenu du texte à l'auteur de la publication. Cet auteur est, quand à lui, lié à l'organisme de publication par la propriété *onBehalfOf*. Enfin, les proprétés *startAtTime* et *endAtTime* permettent respectivement d'indiquer la date de début et de fin de la publication. Le degré de confiance *hasTrust* est attribué après interrogation de la base de connaissances Trust. Les URI des objets des triplets dont le prédicat est *prov :wasAttributedTo* et/ou *prov :OnBehalfTo* permettent d'identifier la source. L'utilisateur avant d'évaluer la certitude des connaissances du texte, devra en premier vérifier dans Trust si le degré de confiance qu'il attribue à cette source lui convient ou pas.

3.2 Pendant l'extraction

Le deuxième niveau où une incertitude peut être exprimée concerne le contenu même du texte. En effet, durant l'analyse du texte, quelques imperfections de l'information peuvent être identifiées. De plus, les règles d'analyse et d'extraction de connaissances peuvent elles aussi être incertaines. Un degré d'incertitude devient alors nécessaire pour évaluer la certitude et la qualité de la connaissance extraite.

3.2.1 Imperfection au niveau de l'information

Le but dans cette partie est de considérer le caractère même de l'information. Celle-ci peut être objective (provenant d'une machine tel qu'un capteur ou radar) ou bien subjective (provenant d'une déclaration faite par un agent). L'information objective ne peut être remise en cause alors que l'information subjective peut être considérée de manière différente suivant

les personnes. L'information peut contenir de l'incertitude, des imprécisions ou encore être incomplètes... Dans Smets (1997), l'auteur distingue trois variantes d'imperfections

1. *L'incertitude* liée à la relation observée entre la donnée et l'univers pris en compte indiquant lorsque la donnée est possible.
   Exemple : "L'ancien président, Nicolas Sarkozy **pourrait** se présenter aux présidentielles 2017".
2. *L'imprécision*, causée par des données statistiques imprécises. Elle diffère d'une interprétation à une autre. Elle est généralement décrite grâce à des termes linguistiques décrivant le monde réel.
   Exemple : "Pierre a l'air **jeune**"
   "Jeune" étant une définition abstraite, l'âge de la personne peut varier de 18 à 25 ans par exemple.
   L'espace temporel peut lui aussi être exprimé de manière floue. Dans ce cas nous représenterons la date avec un intervalle temporel comprenant cette date.
   Exemple : " Pierre est né en **2003** à Paris".
   La date de naissance n'étant pas précise, un intervalle d'incertitude permettra alors de modéliser cette connaissance.
3. *Ignorance totale ou partielle*, il arrive qu'une information soit livrée sans que tous les détails la concernant ne soient décrits. L'information devient alors incomplète.
   Exemple : "Pierre est allé en **Allemagne**".
   Dans cet exemple, le lieu d'arrivée n'est pas bien précisé et les autres informations telles que la date ou encore le lieu de départ ne sont pas précisés. Il s'agit donc ici d'une information partielle.

Dans notre travail, nous avons mis l'accent sur la première forme d'imperfection, à savoir l'incertitude.
L'incertitude qualifie la connaissance de l'auteur sur l'information fournie. Cette dernière est soit vraie ou fausse à un moment donné, mais si l'auteur peut ne pas avoir connaissance de cet état ni de sa véracité, il exprimera alors une certaine incertitude lors de son récit. L'incertitude peut être employée pour exprimer une intention, une volonté, une supposition, une éventualité, un doute, une hésitation, une indécision, une croyance, une préférence, une émotion...
Comme pour l'extraction de connaissances, nous nous basons sur la notion de marqueurs telle que nous l'avons décrite dans la section 2. Les déclencheurs permettent de repérer l'incertitude que peut exprimer l'auteur lors de son récit. Une liste d'indicateurs d'incertitude est établie. Nous distinguons

- Les verbes d'opinion : croire, penser, douter...
- Les verbes impersonnels : il parait que, il semble que...
- Les adjectifs : douteux, incertain, possible...
- Les adverbes : peut-être, apparemment, probablement...
- Les locutions adverbiales : éventuellement, hypothétiquement..
- Les expressions : selon lui, à mon avis, il se peut...

Chaque marqueur d'incertitude est associé à un degré qui permettra de quantifier la fiabilité de l'information véhiculée.

Le marqueur peut également être accompagné d'une valeur numérique qui facilitera alors la quantification de l'incertitude. Exemple : *"La **probabilité** qu'il pleuve demain est de **60%**."*

La difficulté rencontrée à cette étape est la portée de cette incertitude. Il faut savoir si par exemple l'incertitude porte sur une propriété d'un concept donné ou bien sur le concept en entier. Le plus souvent, l'utilisation d'une structure emphatique telle que *"c'est ...qui/que..."* permet d'identifier le premier cas.

Exemple 3.2 *Je pense que **c'est** Paul qui emmène Julie à Paris.*

Dans cet exemple, le voyage de Julie n'est pas remis en cause, mais c'est le fait que ce soit Paul qui l'emmène est incertain.

3.2.2 Application des règles d'extraction

Il arrive quelques fois que l'ambiguité soit très importante à tel point que le système n'arrive pas à distinguer la meilleure solution, étant donné que même un humain peut mal interprêter ces textes.

Exemple 3.3 *Le livre de Paul.*

Dans cet exemple, nous ne pouvons distinguer entre la relation d'appartenance ou bien d'auteur. En effet, cette phrase peut exprimer le fait que Paul soit le propriétaire du livre ou bien le fait que Paul soit l'auteur du livre. En général, les systèmes d'analyse linguistique choisissent soit une règle d'extraction au hasard ou la plus probable. Dans cet exemple, la relation la plus probable est la relation d'appartenance, mais une incertitude subsiste quand même. Avec une démarche qui prend en compte l'incertitude, un poids est associé aux règles d'extraction afin d'extraire tous les triplets possibles. L'utilisateur décidera par la suite quelle connaissance il souhaite garder et ajouter à la base de connaissances.

3.3 Incertitude post-extraction

3.3.1 Mise en cohérence

Après l'extraction des connaissances brutes, il est nécessaire d'ajouter un traitement supplémentaire qui consiste à mettre en cohérence l'ensemble des connaissances extraites. L'extraction de connaissances traite le texte phrase par phrase, sans prendre en considération de lien sémantique qui peut exister entre les phrases. Le résultat de l'extraction des connaissances est un graphe RDF faisant référence aux concepts et propriétés issus de l'ontologie intégrée dans le système. L'étape de mise en cohérence correspond aux opérations de consolidation, résolution d'ambigüités et enrichissement de ce graphe. Les principaux traitements effectués lors de cette étape sont :

- Le regroupement des entités nommées : ceci consiste à regrouper les co-occurrences de chaque entité (Personne, Organisation, Lieu) citée dans le texte.
- La résolution des dates relatives : une date non absolue n'a pas d'intérêt à être stockée dans une base de connaissances car elle ne réfère à aucune date précise. Il est alors nécessaire de calculer la date relative en fonction de la date de référence. Cette dernière peut être la date d'émission de l'article, ou bien d'une date citée à l'intérieur du texte.
- Le regroupement d'événements : un événement peut être décrit avec différents déclencheurs ce qui créera plusieurs événements au lieu d'un seul et même évènement au sein de la phrase. La règle de regroupement suppose alors que si rien n'indique le contraire

-aucune contradiction dans la description de ces deux évènements ou encore qu'aucun adverbe de temps n'a été introduit- alors il s'agit d'un seul événement.

– Inférence au niveau de la phrase : de nouvelles connaissances peuvent être déduites du contexte. Ces connaissances étant exprimées implicitement il est impossible de les extraire lors de l'analyse linguistique. Dans l'exemple 2.1, le lieu où se situe le tribunal n'est pas mentionné mais nous pouvons supposer qu'il se trouve à Jayapura, avec cependant un certain degré de réticence.

Chacun de ces traitements repose sur des règles spécifiques. Le risque d'incertitude est évalué à l'application des règles. Pour le regroupement des entités nommées, nous nous basons sur des algorithmes d'Entity Matching, Köpcke et Rahm (2010). Ces méthodes de regroupement reposent principalement sur des algorithmes de comparaison sémantique de deux entités, ou de calcul de mesures de similarités, qui permettent d'évaluer le degré de ressemblance de deux entités, telles que : *Jaccard, Levenshtein, TF-IDF, etc.* Le calcul du niveau de ressemblance permet alors de quantifier le niveau d'incertitude associé à ce regroupement.
Pour ce qui est de la résolution des dates relatives, nous avons proposé une représentation sous forme d'intervalle de toutes les dates citées dans le texte. Ainsi une date qui réfère à la semaine dernière, par exemple, aura pour date de début le lundi de la semaine passée et pour date de fin le dimanche de cette même semaine. Cependant, aucun degré d'incertitude supplémentaire n'est associé à cette extraction.
Enfin, en ce qui concerne le regroupement des évènements et l'inférence intra-phrase, le degré dépendra également de la règle appliquée. Par exemple, l'inférence d'une entité nommée telle que la date ou le lieu aura un degré de possibilité plus élevé que l'ajout d'autres propriétés telles que la cause, ou les agents intervenants dans la description d'un évènement.

3.3.2 Enrichissement et vérification avec les bases de références

La dernière étape de notre système est la vérification et/ou l'enrichissement des connaissances extraites. La vérification est réalisée afin de s'assurer de la cohérence et de la véracité des connaissances extraites avant de les ajouter à la base de connaissances. Ceci consiste à interroger les jeux de données du Linked Open Data (LOD) tels que DBpedia ou Geonames. Cependant, il arrive que, dans des articles de presse, les entités nommées telles que les personnes ou les organisations soient orthographiées de manières différentes. Grâce aux pages d'homonymie disponibles sur le web noil devient possible de résoudre ce problème.
Enfin, pour ce qui concerne l'enrichissement, il s'agit de pouvoir améliorer la pertinence de la connaissance extraite en complétant l'information à partir des datasets du LOD. Ceci est une démarche pour gérer l'information incomplète. Cependant, les données du LOD peuvent elles aussi être incertaines. En effet, la qualité de ces données est parfois remise en cause. Quelques travaux ont été menés dans ce cadre afin d'évaluer la qualité de ces données : Bizer et Cyganiak (2009); Hartig (2008); Flemming (2010). Zaveri et al. (2013) considère que la qualité peut varier suivant six dimensions : le contexte abordé, la confiance accordée, le caractère intrinsèque de l'information, l'accessibilité des données, la pérennité et la représentation. Malheureusement, aucun dataset n'offre un score d'une précision de 100%. Il est donc nécessaire d'évaluer la qualité des données avant de compléter la connaissance extraite. Ce n'est pas toujours le cas. Par exemple, DBpedia comporte une grande quantité d'erreurs qui remettent en cause son utilisation Knuth et al. (2012).

4 Représentation de l'incertitude

Dans cette section, nous présentons le formalisme adopté pour la représentation de l'incertitude dans les graphes issus de nos extractions textuelles. Comme tout formalisme de représentation des connaissances, celui-ci comporte un aspect syntaxique et un aspect sémantique. Au niveau syntaxique, nous nous intéressons essentiellement à la manière d'intégrer les valeurs d'incertitude dans nos graphes. Nous dédions une sous-section à la notion de réification qui dans un premier temps semblait être une solution évidente puis introduisons notre approche.

4.1 Aspect sémantique

L'interprétation de nos graphes RDF se base sur la sémantique standard des graphes RDF qui sont associés aux ontologies OWL telle qu'elles ont été évoquées précédemment, e.g., l'ontologie de l'extraction, de l'incertitude et PROV-O. La sémantique qui est attribuée à la notion d'incertitude n'est pas développée dans cet article. Nous détaillerons cet aspect dans un prochain article lorsque nous approfondirons les aspects requêtage et raisonnement sur nos graphes. Nous pouvons simplement mentionner que notre représentation est compatible avec la sémantique standard des logiques possibiliste et probabiliste.

4.2 Ascpect syntaxique

4.2.1 Réification

La réification est une recommandation RDF du W3C, Semantics et al. (2004). Elle permet de décrire des informations concernant les triplets, telles que les métadonnées par exemple. Le principe de la réification est de diviser le triplet en quatre sous-triplets ayant la même ressource en sujet. Cette ressource est désignée par un rdf :nodeID. Le premier triplet permet d'identifier le sujet, le deuxième le prédicat, le troisième l'objet et enfin le quatrième permet d'indiquer que le nodeID décrit un triplet (Statement). Ce nodeID pourra par la suite être utilisé pour ajouter des triplets supplémentaires qui permettront de décrire les informations à ajouter au triplet initial. La figure 4 illustre l'exemple 4.1, et permet de décrire l'incertitude exprimée dans la phrase.

Exemple 4.1 *John est probablement marié avec Mary.*

A première vue cela semble intéressant car l'incertitude peut être considérée comme une information supplémentaire à ajouter à un triplet. Cependant, vu que ce triplet est divisé, il devient difficile de le relier aux autres triplets dans le graphe. Ceci pose un sérieux problème lors de l'interrogation car la syntaxe habituelle (s,p,o) n'est pas respectée. De plus, le fait de diviser le triplet en quatre augmente la taille du graphe RDF, et rend son parcours plus long, ce qui ralentit le temps de réponses aux requêtes. Dans (Hartig et Thompson, 2014), les auteurs proposent une alternative à la réification telle que l'a définie le W3C, en créant une nouvelle syntaxe qui admet de prendre en sujet un triplet et non pas une ressource. A cette syntaxe RDF*, ils définissent un langage alternatif à SPARQL, SPARQL* qui interroge des graphes RDF*. Cependant, à notre connaissance, aucun raisonneur ne permet d'inférer de nouvelles connaissances à partir des triplets RDF*.

Gestion de l'incertitude.

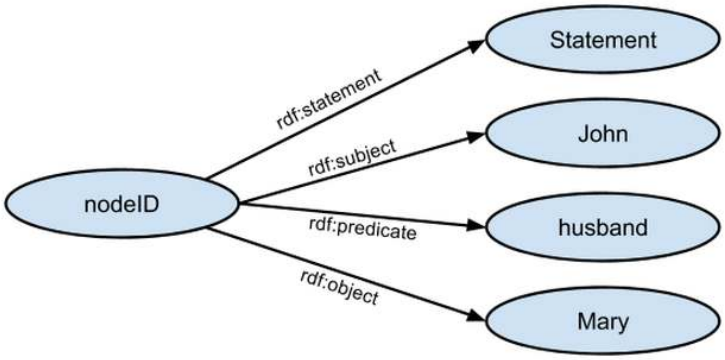


FIG. 3 – *représentation de l'incertitude avec reification.*

4.2.2 Notre approche

Notre approche consiste à considérer l'incertitude comme une connaissance à part entière et non pas comme une simple métadonnée à ajouter au triplet. Pour cela, nous avons décidé de créer dans l'ontologie une classe, nommée *Uncertainty*, pour modéliser cette incertitude. Elle servira à décrire ce qui est incertain dans le texte. Cette classe est décrite par trois propriétés :

- weight : une propriété littérale pour quantifier l'incertitude identifiée.
- isUncertain : propriété object qui aura pour co-domain le top-concept, cela veut dire que tout concept de l'ontologie pourra être visé par une incertitude.
- hasUncertainProp : une propriété object qui servira d'intermédiaire entre le domaine initial de la propriété et la propriété en question

La figure 4 permet de modéliser l'incertitude exprimée dans l'exemple 2.1. L'auteur de l'article indique que la condamnation des accusés est probable donc incertaine. Pour cela nous avons créé une instance de Uncertainty qui nous permet de qualifier l'information incertaine (l'ensemble de la condamnation) ainsi que de la quantifier grâce au poids associé à cette incertitude grâce au terme marqueur "probable". D'un autre coté, nous avons déduit que le tribunal pourrait peut être se situer en Papouasie, nous l'avons donc ajouté avec un degré d'incertitude adéquat.

L'ontologie que nous avons développé est indépendante de tout domaine d'application. Dès lors, elle peut être ajoutée à toute autre ontologie voulant prendre en compte l'incertitude.

À noter également que le degré de confiance associée (cf 3.1) à la source sera répercuté sur le reste des connaissances extraites du texte.

5 Conclusion et perspectives

Dans cet article nous nous sommes intéressés au traitement de l'information incertaine dans le cadre d'une extraction de connaissances à partir de texte. Le traitement repose sur les technologies du web sémantique pour permettre de faire le lien avec les données du Linked Open Data.

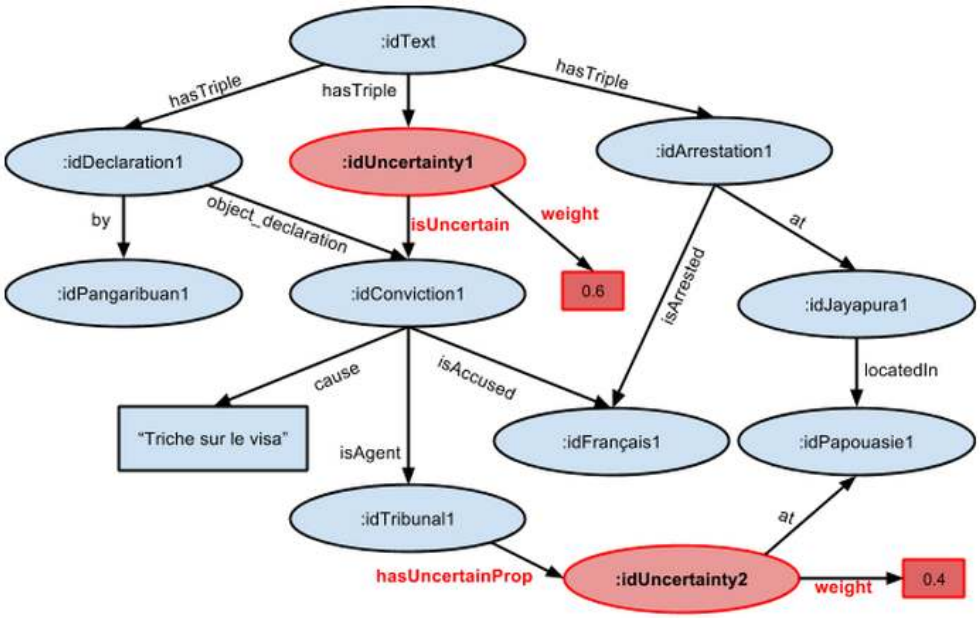


FIG. 4 – *Représentation RDF de l'incertitude.*

Notre démarche consiste à identifier les différentes situations où une incertitude remettant en cause la validité de l'information peut subsister. Nous proposons une ontologie pour modéliser l'information incertaine et la représenter au format RDF.
Nous travaillons actuellement sur développement d'un ensemble de patterns pouvant faciliter l'interrogation du graphe RDF prenant en compte notre représentation de l'incertitude. Nous prévoyons par la suite de développer un raisonneur basé sur le formalisme des logiques possibilistes afin de permettre l'inférence sur les données incertaines.

Références

Bizer, C. et R. Cyganiak (2009). Quality-driven information filtering using the wiqa policy framework. *Web Semantics : Science, Services and Agents on the World Wide Web 7*(1), 1–10.

Bizer, C., T. Heath, K. Idehen, et T. Berners-Lee (2008). Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pp. 1265–1266. ACM.

Cyganiak, R., D. Wood, et M. Lanthaler (2013). Rdf 1.1 concepts and abstract syntax. *World Wide Web Consortium, Working Draft WD-rdf11-concepts-20130723*.

Dong, X., E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, et W. Zhang (2014). Knowledge vault : a web-scale approach to probabilistic knowledge

fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pp. 601–610.

Flemming, A. (2010). Quality characteristics of linked data publishing datasources. *Master's thesis, Humboldt-Universität of Berlin*.

Hartig, O. (2008). Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*. Citeseer.

Hartig, O. et B. Thompson (2014). Foundations of an alternative approach to reification in rdf. *arXiv preprint arXiv :1406.3399*.

Knuth, M., J. Hercher, et H. Sack (2012). Collaboratively patching linked data. *arXiv preprint arXiv :1204.2715*.

Köpcke, H. et E. Rahm (2010). Frameworks for entity matching : A comparison. *Data & Knowledge Engineering 69*(2), 197–210.

Lebo, T., S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, et J. Zhao (2013). Prov-o : The prov ontology. *W3C Recommendation, 30th April*.

Missier, P., K. Belhajjame, et J. Cheney (2013). The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pp. 773–776. ACM.

Mombrun, Y., A. Pauchet, B. Grilhères, S. Canu, et al. (2010). Collecte, analyse et évaluation d'informations en sources ouvertes. In *Atelier COTA des 21es Journées francophones d'Ingénierie des Connaissances*.

Niu, F., C. Zhang, C. Re, et J. W. Shavlik (2012). Deepdive : Web-scale knowledge-base construction using statistical learning and inference. In *Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012*, pp. 25–28.

Semantics, R., P. Hayes, W. W. W. Consortium, et al. (2004). W3c recommendation. *Reification, Containers, Collections and rdf : value*.

Smets, P. (1997). Imperfect information : Imprecision and uncertainty. In *Uncertainty Management in Information Systems*, pp. 225–254. Springer.

Zaveri, A., A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, et P. Hitzler (2013). Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal*.

Summary

The knowledge representation area needs some methods that allow to detect and handle uncertainty. Indeed, a lot of text hold information whose the veracity can be called into question. These information should be managed efficiently in order to represent the knowledge in an explicit way. As first step, we have identified the different forms of uncertainty during a knowledge extraction process, then we have introduce an RDF representation for these kind of knowledge based on an ontologie that we developped for this issue.

Using Clustering for Type Discovery in the Semantic Web

Kenza Kellou-Menouer*, Zoubida Kedad*

*PRISM - University of Versailles Saint-Quentin-en-Yvelines,
45 avenue des Etats-Unis , Versailles, France
kenza.menouer@prism.uvsq.fr, zoubida.kedad@prism.uvsq.fr

Abstract. The Web has witnessed a multiplication of RDF(S)/OWL data sources referred to as linked data, making a huge amount of information available to users and applications. This has enabled interconnecting, enriching and querying Web data sources. One important feature in this context is that data sources are not organized according to any predefined schema, as they are structureless by nature. This lack of schema limits their use to express queries and to understand their content.

Our work is a contribution towards the inference of the structure of RDF(S)/OWL data sources. We present an approach to infer the types describing the entities of a data set. This information can be defined for some entities, but is often missing. In this paper, we present a clustering-based approach for type inference along with some preliminary experimentation results performed on real data sets to demonstrate the feasibility of our approach and its effectiveness to extract types from possibly incomplete and noisy data.

Keywords: Type Extraction, Semantic Web, Noisy Data, Clustering.

1 Introduction

The Web has recently witnessed a proliferation of RDF(S)/OWL [1] [2] [3] data sources, and its content has evolved from documents connected through hypertext links to a set of interconnected data sources, so-called linked data. This has made a huge amount of data and knowledge available to users and applications, but has risen new challenges related to interconnecting, enriching and querying this data.

Querying and exploiting RDF(S)/OWL data sources requires information about their content, i.e. their resources and properties. Without a description of the data set, it is difficult to target the relevant properties and resources, and browsing the data sets in order to understand their content can be a tedious process.

One important feature of RDF(S)/OWL data sources is that they are not structured according to any predefined schema. They are structureless by nature and the languages used

1. http://www.w3.org/TR/RDF
2. http://www.w3.org/TR/rdf-schema
3. http://www.w3.org/OWL

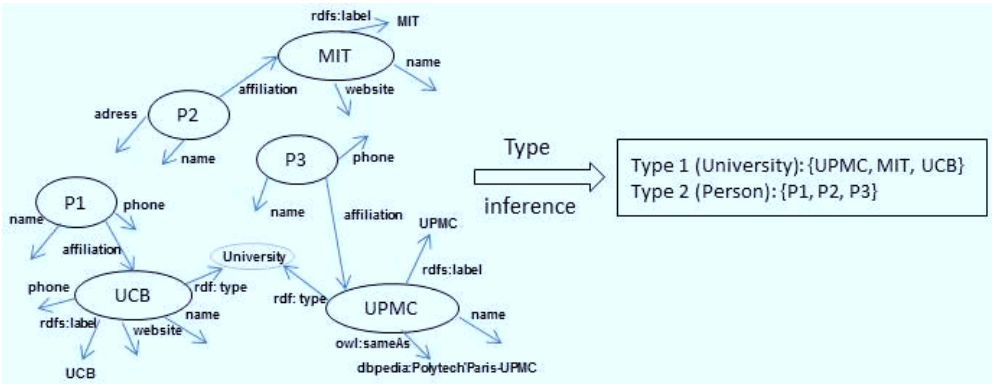


FIG. 1 – *Type inference from RDF(S)/OWL data sets.*

to describe data on the Web do not impose any constraints or restrictions on the properties describing resources. Some primitive properties in the RDF(S)/OWL vocabularies provide schema-related information, such as *rdf:type*, *rdfs:subClasseOf*, *rdfs:domain* and *rdfs:range*, but they are not always provided. Besides, resources of the same type may have different properties, making it difficult for reasoners to infer type information when missing. Our goal is to provide an approach to help understanding the content of RDF(S)/OWL data sources and our work is a contribution towards the inference of a schema describing their content. We are interested in discovering type definitions, or classes, in a data set where this information is missing.

In this paper, we present a clustering-based approach for type inference, which relies on the similarity between resources based on their properties to identify groups of resources having the same type. We provide some preliminary experimentation results performed on real data sets to demonstrate the feasibility of our approach and its effectiveness to extract types from possibly incomplete and noisy data.

The paper is organized as follows. The problem addressed in our work is presented in section 2. Section 3 describes our approach for type inference. In section 4, we present the methodology used for evaluating our approach along with some experiments. Related works are provided in section 5, and finally, section 6, concludes the paper.

2 Problem statement

Consider the sets R, B, P and L representing resources, blank nodes (anonymous resource), properties and literals respectively. A data set described in RDF(S)/OWL is defined as a set of triples $T \subseteq (R \cup B) \times P \times (R \cup B \cup L)$. A property $p \in P$ is either user-defined or defined in the RDF(S)/OWL vocabularies and in this case will be referred to as a primitive property. Graphically, a data set is represented by a labeled directed graph G, where each node is a resource, a blank node or a literal and where each edge from a node r_1 to another node r_2 labeled with the property p represents the triple (r_1, p, r_2) of the data set T.

In such RDF(S)/OWL graph, we define an entity as a node corresponding to either a resource or a blank node, that is, any node apart from the ones corresponding to literals.

Figure 1 shows an example of such data set, related to universities. We can see that some entities are described by the property *rdf:type*, defining the classes to which the entities belong, which is the case for entities "UPMC" and "UCB". For other entities, such as "MIT" and "P1", this information is missing. The reason may simply be that no one has entered this data. Even when linked data sets are automatically extracted from a controlled source, type information can be missing: in DBpedia (Auer et al. (2007)), which is extracted automatically from Wikipedia, the experiments reported by Paulheim and Bizer (2013) show that at most 63.7 % of the data have type declaration. Type definitions are provided for at most 53.3 % of resources in YAGO (Suchanek et al. (2007)). This is generally due to either a missing infobox, or to a very general or ambiguous semantic. Two entities having the same type are not necessarily described by the same properties, as we can see for "UPMC" and "UCB" in our example, which are both associated to the "University" type, but unlike "UPMC", "UCB" has a "phone" and a "website" properties.

This paper is a step towards the definition of the schema describing RDF(S)/OWL data sets, and our problem can be stated as follows: given a data set with missing type information, how to group entities so as to infer type definitions from this incomplete and structureless data? As a results, our goal is to provide an extracted schema defined as follows:

The extracted schema S of a data set T is composed of a set of classes $\{C_1, ..., C_p\}$, each class corresponds to a set of entities in T and defines their type.

We can infer this schema description by grouping the entities according to the similarity of their structure, i.e. to the similarity of their incoming and outgoing properties. Depending on this similarity, it is possible that type information corresponding to some entities can not be inferred. Obviously, the classes and their number is not known in advance. In the example given in figure 1, two type definitions would be inferred: "University" grouping the entities "MIT", "UCB" and "UPMC", and "Person", grouping the entities "P1", "P2" and "P3". In the next section, we present the principles underlying our approach for type inference.

3 Type discovery

In order to infer missing type definitions in a RDF(S)/OWL data set, our approach relies on grouping entities according to their similarity. A group of similar entities corresponds to a type definition. The similarity between two given entities is evaluated considering their respective sets of properties.

All the properties don't have the same importance for type inference. Some of them, such as *rdfs:label*, could apply to any entity, and therefore they will not be considered as important as others when evaluating the similarity between entities. Some properties provide information about the type of an entity, such as *rdf:type*, which relates the type "University" of the entity "UPMC" in figure 1. Other properties such as *owl:sameAs* can be used to validated the results of the type inference approach: if two entities r_1 and r_2 are related by the primitive property *owl:sameAs*, then we can check that the type inferred for r_1 and r_2 is the same.

Our requirements for type inference are the followings: firstly, the number of types is not known in advance, and secondly, the data sets are evolving, large and may contain noise. The most suitable grouping approach is density-based clustering, introduced by Ester et al. (1996), because it is robust to noise, deterministic and it finds classes of arbitrary shape, which is useful for data sets where resources are described with heterogeneous property sets. In

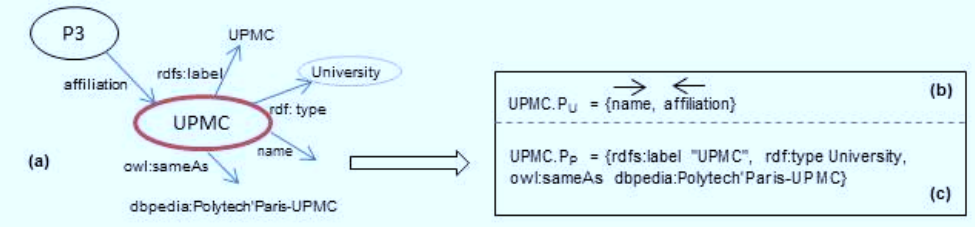


FIG. 2 – *Example of Entity Description.*

addition, unlike the algorithms based on k-means, described by Jain (2010), and k-medoid, introduced by Kaufman and Rousseeuw (1987), the number of classes is not required.

In this section, we first give a description of the entities according to their different types of properties, then we present our approach for type discovery.

3.1 Entity description

In our context, an entity is described by different kinds of properties. Some of them are part of RDF(S)/OWL vocabulary, and we will refer to them as primitive properties, and others are user-defined. The set of primitives properties is denoted P_P, and the set of user-defined properties is denoted P_U. The set P of properties describing entities is such that $P = P_U \cup P_P$. In order to infer type definitions, entities are compared according to their structure. We consider the following description of an entity, defined as follows.

Definition. An entity r in the data set T is described by:

1. An user-defined properties set $r.P_U$ composed of oriented properties p_u from P_U as follows:

– if $(\exists(r, p_u, r_1) \in T)$ then $(\overrightarrow{p_u} \in r.P_U)$;

– if $(\exists(r_2, p_u, r) \in T)$ then $(\overleftarrow{p_u} \in r.P_U)$.

2. A primitives properties set $r.P_P$ composed of p_p from P_P with their values. Each primitive property p_p of r is defined as follows:

– $\exists(r, p_p, r_1) \in T$ and $p_p \in P_P$.

Figure 2 shows an example of entity description. The density-based clustering algorithm takes as input the set $D_U = \{r_i.P_U\}$ where $r_i.P_U$ represents the set of user-defined properties describing the entity r_i. The set $D_P = \{r_i.P_P\}$ where $r_i.P_P$ represents the set of primitive properties describing the entity r_i can be used to validated the results of the approach and to specify type checking rules.

3.2 Clustering

To discover missing type information in a RDF data set, we have used and adapted the density-based algorithm described by Ester et al. (1996). We apply our algorithm to the set D_U (see subsection 3.1). In order to measure the similarity between two property sets $r_i.P_U$ and $r_j.P_U$ describing two entities r_i and r_j, we use the Jaccard similarity distance defined as follows:

$$Similarity(r_i, r_j) = \frac{|r_i.P_U \cap r_j.P_U|}{|r_i.P_U \cup r_j.P_U|}$$

The maximum radius of neighborhood ε represents the minimum similarity value for two entities to be considered as neighbors, in other words, it is the maximum distance between two entities r_i, r_j such that $distance(r_i, r_j)$ = 1- $similarity(r_i, r_j)$.

The minimum number of neighbors for an entity, denoted $MinPts$ is the minimum number of their similar entities required to be considered as a core (Ester et al. (1996)); it allows to exclude the outliers and the noise. An entity is not assigned to a class if it is considered as noise, i.e. if it is neither a core itself nor the neighbor of a core.

In order to speed up the clustering process, and especially to perform successive executions with different parameters values, we perform once and for all the calculation of the nearest neighbors of each entity. To this end, we index the data and we order the entities according to their similarity. We store a $|D_U|$ X $|D_U|$ neighborhood matrix containing for each entity the ordered list of its neighbors, as well as the distance between this entity list (see figure 3) and the number of the line representing the index of an entity. It is then straightforward to find the nearest neighbors for an entity at a distance lower than ε.

$$M_n(|D_U|, |D_U|) = \begin{pmatrix} (i, 0.05) & (j, 0.1) & (k, 0.25) & ... \\ (b, 0.1) & (d, 0.3) & ... & ... \\ . & & & \\ . & & & \\ ... & (i, 0.05) & (h, 0.9) & (d, 0.93) \end{pmatrix}$$

FIG. 3 – *Neighborhood matrix.*

For the neighborhood matrix described in figure 3 and considering a 0.2 value for ε, the neighbors of entities having an index value of 1 are the first entities from the first line of the matrix for which the distance is less than 0.2, which are the entities corresponding to i and j in our example. Thanks to the indexing of the data and the ordering of the neighbors according to the distance, it is not necessary to go through the entire matrix to find the neighbors of a given entity.

Thanks to the use of the neighborhood matrix, the complexity of the approach is linear o(n). Indeed, finding the neighbors of an entity r simply consists in returning the previously sorted entities having a distance lower than ε in the row corresponding to r in the matrix (see figure 3).

4 Evaluation

This section presents some experimentation results using our approach. We have evaluated both the quality of the results provided by our approach using well-known information retrieval metrics, and their performance. We have used different real data sets, described in the following section.

4.1 data sets

In order to evaluate the quality of our approach we have first used the Conference [4] and DBpedia [5] data sets. The Conference data set exposes papers, presentations and people for several semantic Web related conferences and workshops. The data set contains 1430 triples, 322 entities and the following type definitions: Presentation, Person, Organization, InProceedings, ProgrammeCommitteeMember, TutorialEvent, ConferenceEvent, Chair, KeynoteTalk, Point. The different entities are described by very similar sets of properties in the Conferences data set, unlike in DBpedia, where entities of the same type are heterogeneous in their characteristics. Another reason for choosing DBpedia is that the number of properties of an entity is important, about 150 on average, which enables to test the quality of the approach in the case of high-dimensional data. For testing our approach we have randomly selected 100 instances of the types Politician, SoccerPlayer, Museum, Movie, Book, Country. We have set ε to 0,5 so that two entities are considered as neighbors if the number of their shared properties represents at least the half of the total number of the properties characterizing them. We have set $MinPts$ to 1 so that an entity is considered as noise if it has no neighbors.

For performance evaluation, we have used the DBpedia and Conference data sets, and we have also used Timbl [6], a large real-world data set for performance evaluation taken from Billion Triple Challenge Dataset (BTC) 2012 [7]. We have tested the performance of the approach on three of their subsets (Timbl/data-0, Timbl/data-1 and Timbl/data-2) having different sizes. The tests have been performed on an Intel(R) Xeon(R) machine, CPU of 2.80 GHz, 64-bit with 4 GB of RAM.

4.2 Metrics

In order to evaluate the quality of the results provided by our approach, we have extracted the existing type definitions from our data sets and considered them as a gold standard. Then we have run our approach on the data sets without the type definitions and evaluated for each of the inferred classes precision, recall and F-Score presented in (Larsen and Aone (1999)). We have annotated each inferred class C_i with the most frequent type definition of its entities. For each type label T_r of size n_r in the data set and each class C_i of size n_i inferred by our approach, such that T_r is the label of C_i, the quality metric are evaluated for each C_i with respect to T_r as follows: n_{ri} being the number of instance in the class C_i that belong to T_r, the precision $P(T_r, C_i)$ is defined as n_{ri}/n_i, the recall $R(T_r, C_i)$ is defined as n_{ri}/n_r and F-Score is defined as:

$$F(T_r, C_i) = \frac{2 \times P(T_r, C_i) \times R(T_r, C_i)}{P(T_r, C_i) + R(T_r, C_i)}$$

If most of the entities in a generated class are untyped, it is impossible to associate this class to a type defined in the data set. In such case, we have manually labeled it with an appropriate type. Precision has been evaluated considering the number of entities of this type in the class, and recall has been evaluated considering the number of entities of this type outside the class.

4. http://data.semanticweb.org/dumps/conferences/dc-2010-complete.rdf
5. http://dbpedia.org
6. http://km.aifb.kit.edu/projects/btc-2012/
7. http://challenge.semanticweb.org

4.3 Results

In this section, we describe the results of our experiments, regarding both the quality of the results and the performances.

4.3.1 Quality

The results of the quality of the approach for the Conference data set are shown in figure 4 (a). The approach gives good results and can detect types not declared in the data set, which have been annotated as follows: classes 5, 8, 9 and 10 are labeled "AuthorList", "PublicationPage", "HomePage" and "City" respectively. It is necessary to consider the incoming properties in order to be able to group them because they don't have outgoing properties, especially containers such as "AuthorList". Classes 1 and 6 don't have a good precision because they contain entities typed differently in the data set. Indeed, class 1 annotated "Presentation" corresponds to three types in the data set: "Presentation", "Tutorial" and "ProgrammeCommitteeMember"; and Class 7 annotated "TutorialEvent" corresponds to two types: "TutorialEvent" and "KeynoteTalk". Types grouped in classes 1 and 7 have the same structure, it is therefore not possible to make a distinction between them considering the structure only.

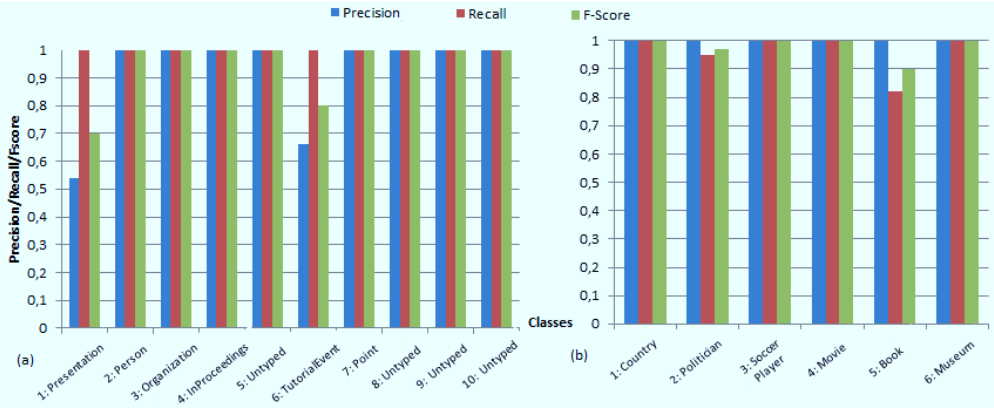


FIG. 4 – *Precision, Recall and F-Score quality of the data sets Conference (a) and DBpedia (b).*

The results for the DBpedia data set (see figure 4 (b)) show that the assignment of types to entities has achieved good precision and recall, and that the approach gives good results even when the data set is heterogeneous. The recall for types "Book" and "Politician" is not maximum, because the approach has detected noisy instances for these types. Entities of the two types "Politician" and "SoccerPlayer" have not been grouped together despite having similar property sets.

4.3.2 Performance

Table 1 shows the execution time for the computation of the neighborhood matrix and the clustering process on five data sets of different size, expressed in terms of number of triples

and number of entities.

TAB. 1 – *Performance of the approach (in ms) for different data sets.*

Data set	Nb triples	Nb entities	neighborhood matrix	Clustering
Conference	1430	322	31	0.4
DBpedia	19696	100	37	0.2
Timbl/data-0	89	33	0.8	0.02
Timbl/data-1	15616	3630	16130	2200
Timbl/data-2	87250	19944	1698015 (28.3mn)	154481 (2.57mn)

The calculation of the neighborhood matrix saves considerable time for clustering, which is very useful as it may be performed iteratively with different parameter values. The number of entities and the number of triples affects the calculation of the neighborhood matrix because the similarity between each pair of entities is calculated according to their properties and the number of entities affects the ordering of the neighbors because the entities are sorted according to their similarity to every other entity. Clustering is also affected by the number of entities because it brings together entities according to their similarity.

5 Related works

Some works in the literature have addressed the problem of inferring the schema of a semi-structured data set. Wang et al. (2000) propose an approximate DataGuide based on an incremental hierarchical clustering algorithm, (COBWEB Fisher (1987)), for grouping similar nodes having the same incoming/outgoing edges. The proposed method considers both types of edge at the same level as a property of a node, as it is proposed for semi-structured data described in OEM (Papakonstantinou et al. (1995)). It doesn't differentiate between the domain and the range of properties which could affect the result of the process if applied to RDF(S)/OWL data sets. The method uses COBWEB which is not deterministic, expensive and not very suitable for large data sets. Nestorov et al. (1997) propose a method that infers the inherent structures and types of semi-structured data. It distinguishes between incoming and outgoing edges: the incoming ones are considered as roles (candidate labels for the type). This is applicable to the OEM model but not to RDF, where incoming edges don't necessarily reflect the type. The proposed algorithm requires to set the threshold of the jump which is not easy to define, because it depends on the regularity of the data. In addition, the construction of the graph is expensive, which is not suitable for a large data set. An extension of this approach is proposed in Nestorov et al. (1998), using bottom-up grouping: initially, each object is a cluster, then, the similar objects are merged. The method infers classes, but it requires a threshold of similarity, and the number of classes. To solve the problem of determining the number of cluster, the authors propose to use k-means algorithm with different k values, representing the number of clusters, but this is an expensive process. Christodoulou et al. (2013) use standard ascending hierarchical clustering to deduce structural summaries of linked data. Unlike our

approach, only outgoing properties are considered, which could be a problem for entities having only incoming properties such as containers. A hierarchical clustering algorithm is applied on instances to form classes. The choice of the algorithm is due to the fact that the number of classes is not known in advance. To determine the best partition of the hierarchical algorithm, the approach tries to maximize the average dissimilarity between each instance and instances of other clusters, and minimize the average dissimilarity inside a cluster. Hierarchical clustering is expensive, and may not be suitable for large RDF graphs. In addition, the method traverses the hierarchical clustering tree to assess the best cutoff level, which increases the cost of the process.

Other approaches enrich data sets by adding more structure-relatd information through RDF(S)/OWL properties. SDType (Paulheim and Bizer (2013)) is a statistical method that enriches an entity by several types using RDFS inference rules, and computes the confidence of a type for an entity. The contribution is mainly the evaluation of the relevance of inferred types for an entity rather than finding types, as RDFS inference rules are used for this. Indeed, the *rdfs:domain*, *rdfs:range* and *rdfs:subClassOf* properties are required for type inference, which limits the approach. Moreover, it can not introduce new types apart from those already in the data set. The approach allows to assign a weight to the properties according to their relevance for type inference, but it requires part of the data set with type information as a training set. Works in Nuzzolese et al. (2012) and Gangemi et al. (2012) infer types for the DBpedia data set only: Nuzzolese et al. (2012) uses K-NN to finds the type of untyped entities and Gangemi et al. (2012) finds the most appropriate type based on descriptions from Wikipedia and links with WordNet [8] and Dolce Ultra Lite ontology [9].

6 Conclusion and perspectives

In this paper, we have proposed an approach for type discovery relying on a density-based clustering algorithm and using the similarity between sets of properties describing entities. We have performed some experiments on existing data sets to assess the quality of the inferred types. In addition to achieve good precision and recall even when the sets of properties describing entities are very heterogeneous, the approach enabled to infer type definitions which were not specified in the data set. In some cases, the analysis of the structure of entities alone is not sufficient to infer the types, and we could consider other information such as the domain and the range of properties for entities of the same class.

One problem we plan to address in future works is the annotation of the extracted classes. It is currently done by identifying the most frequent type in the class, but we could also use external knowledge bases. Beside the classes describing a data set, other information could be very useful in order to extract a complete schema, such as the links between the classes, or some information related to class inclusion. We also plan to explore the possible ways of extracting such information.

8. http://wordnet.princeton.edu/
9. http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

References

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives (2007). *Dbpedia: A nucleus for a web of open data*. Springer.

Christodoulou, K., N. W. Paton, and A. A. Fernandes (2013). Structure inference for linked data sources using clustering. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, pp. 60–67. ACM.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Volume 96, pp. 226–231.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning 2*(2), 139–172.

Gangemi, A., A. G. Nuzzolese, V. Presutti, F. Draicchio, A. Musetti, and P. Ciancarini (2012). Automatic typing of dbpedia entities. In *The Semantic Web–ISWC 2012*, pp. 65–81. Springer.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31*(8), 651–666.

Kaufman, L. and P. Rousseeuw (1987). *Clustering by Means of Medoids*. Reports of the Faculty of Mathematics and Informatics. Faculty of Mathematics and Informatics.

Larsen, B. and C. Aone (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22. ACM.

Nestorov, S., S. Abiteboul, and R. Motwani (1997). Inferring structure in semistructured data.

Nestorov, S., S. Abiteboul, and R. Motwani (1998). Extracting schema from semistructured data. In *ACM SIGMOD Record*, Volume 27, pp. 295–306. ACM.

Nuzzolese, A. G., A. Gangemi, V. Presutti, and P. Ciancarini (2012). Type inference through the analysis of wikipedia links. In *LDOW*.

Papakonstantinou, Y., H. Garcia-Molina, and J. Widom (1995). Object exchange across heterogeneous information sources. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pp. 251–260. IEEE.

Paulheim, H. and C. Bizer (2013). Type inference on noisy rdf data. In *The Semantic Web–ISWC 2013*, pp. 510–525. Springer.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706. ACM.

Wang, Q. Y., J. X. Yu, and K.-F. Wong (2000). Approximate graph schema extraction for semi-structured data. In *Advances in Database Technology-EDBT 2000*, pp. 302–316. Springer.

Résumé

Un nombre croissant de sources de données au format RDF(S)/OWL sont publiées sur le Web. Ces sources interconnectées rendent une masse très importantes de données et de connaissances disponibles pour les utilisateurs et les applications en permettant l'interconnexion, l'en-

richissement et l'interrogation de ces données. Une caractéristique importante de ces sources réside dans le fait qu'elles ne sont pas décrites en suivant un schéma particulier, ce qui peut limiter les possibilités de les interroger et rend difficile la compréhension de leur contenu.

Notre travail est une contribution à la découverte de la structure des sources de données RDF(S)/OWL, et nous présentons une approche permettant de découvrir les types décrivant les entités d'une source de données. Ces types peuvent parfois être explicitement définis, mais sont souvent manquants. Dans cet article, nous présentons une approche utilisant un algorithme de clustering, ainsi que des résultats d'expérimentations sur des sources réelles pour illustrer la faisabilité de l'approche et la qualité des résultats produits dans le cas de données incomplètes et éventuellement bruitées.

Keywords: Extraction de Type, Web Sémantique, Données Bruitées, Clustering.

Extraction de la Valeur des données du Big Data par classification multi-label hiérarchique sémantique

Thomas Hassan*, Rafael Peixoto**
Christophe Cruz*, Aurélie Bertaux*, Nuno Silva**

*Université de Bourgogne
thomas.hassan@u-bourgogne.fr
christophe.cruz@u-bourgogne.fr
aurelie.bertaux@iut-dijon.u-bourgogne.fr
**Polytechnic of Porto
rafpp@isep.ipp.pt
nps@isep.ipp.pt

Résumé. Dans le cadre d'une veille stratégique, analyser les données du Big Data permet aux entreprises d'améliorer leur compétitivité. Dans cette optique, cet article présente une nouvelle approche d'extraction d'informations pertinentes à partir de données du Big Data par une Classification Hiérarchique Multi-label (HMC) Sémantique. Le processus proposé consiste en l'apprentissage non supervisé d'une ontologie par des méthodes de machine learning, et en la classification dynamique d'items par un raisonneur basé sur des règles d'inférence. L'architecture est composée de 5 étapes pouvant individuellement passer l'échelle dans un contexte Big Data.

1 Introduction

De nos jours, découvrir des connaissances et des informations de Valeur à partir de données du web est une tâche majeure pour les entreprises. Cependant, déterminer la Valeur d'une information est une tâche complexe qui demande de recourir au domaine de la fouille de données (Witten et Frank, 2005). Dans un contexte Big Data, cette tâche est d'autant plus difficile de par les caractéristiques associées, définies autour de 3 V (Chen et al., 2014)(Hitzler et Janowicz, 2013) : Volume, Variété et Vélocité. Le Volume définit la quantité croissante de données, générée et stockée au fil du temps par les réseaux sociaux, les données de senseurs, etc (Chen et al., 2014). La Vélocité concerne la vitesse importante de production des données, et par conséquent le besoin de traitement rapide des données. La Variété représente la grande hétérogénéité des formats du Big Data. En particulier, les données non-structurées et semi-structurées nécessitent des traitements importants, et représentent 90% du contenu du Big Data (pages web, documents en langage naturel, audio, etc)(Syed et al., 2013).

D'autres V émergent comme la Véracité, la Visualisation ou la Valeur. Contrairement aux 3V du Big Data, il ne s'agit plus d'indicateurs de performance et de robustesse

face aux données. En revanche, la Valeur définit la pertinence d'une information pour un "utilisateur" (par exemple une entreprise) du Big Data, c'est donc une caractéristique primordiale.

L'évaluation de la Valeur des données du web a été étudiée dans la littérature. Comme dans la plupart des systèmes de recherche d'information, on distingue les approches basées sur les données de celles basées sur l'utilisation de connaissances extérieures [1]. Les premières sont généralement les plus simples et les plus performantes en terme de temps de calcul, mais fournissent des résultats de moindre qualité. IDC(Gantz et Reinsel, 2011) propose d'extraire la Valeur d'un grand volume de données hétérogènes, par découverte, capture et analyse à haute vélocité, en se basant sur des méthodes statistiques. Sheth(Sheth, 2014) propose d'utiliser les technologies du web sémantique et les métadonnées disponibles dans les données (semi) structurées afin d'extraire la Valeur tout en considérant les pré-requis de performance du Big Data.

Notre objectif est d'évaluer la Valeur des grands volumes de données du Big Data par une nouvelle approche d'apprentissage d'une ontologie basée sur la classification hiérarchique multi-label appelée *HMC Sémantique*. Cette approche hybride tire parti à la fois de méthodes statistiques et des technologies du web sémantique. La Classification Hiérarchique Multi-label (HMC) est la combinaison de la classification multi-label et de la classification hiérarchique : elle consiste à associer un ou plusieurs labels (concepts) à des items (ou objets) qui peuvent être de sources et formats différents (textes, fichiers de logs, pages web, images, vidéos...). Dans cette approche, plusieurs chemins de la hiérarchie sont attribués à chaque item, qui peut donc appartenir à différentes classes d'un même niveau(Bi et Kwok, 2011) .L'ontologie (Studer et al., 1998) joue un rôle déterminant dans la définition des termes utilisés pour représenter la connaissance, rapprochant ainsi le modèle généré du langage de l'utilisateur.

Ce papier n'a pas pour objectif de compléter l'état de l'art de la Classification Hiérarchique Multi-label ou de la génération de taxonomie. Nous proposons un processus d'apprentissage d'une ontologie pouvant passer l'échelle dans un contexte Big Data, basé sur des méthodes de machine learning, et un raisonnement sur les règles d'inférence (Urbani et al., 2011a) afin de classer les items, extrayant ainsi de la Valeur. Les contributions de ce travail sont les suivantes : un processus d'apprentissage d'une ontologie basé sur le machine learning, et l'analyse du Big Data par la *HMC Sémantique*. Ce papier est divisé en 3 sections. La seconde section décrit comment utiliser l'approche de *HMC Sémantique* pour évaluer de la Valeur du Big Data. La section 3 décrit le processus de *HMC Sémantique* et les 5 phases associées, ainsi que les méthodes permettant de passer l'échelle pour chaque phase. La dernière section conclue sur les travaux en cours et suggère les pistes de réflexion futures.

2 Extraire la Valeur depuis les données du Big Data

La Valeur des données du Big Data est dépendante du cas d'utilisation et des objectifs de l'entreprise qui possède ces données. Il peut par exemple s'agir de rechercher des motifs récurrents dans un flux de données continu, ou de détecter des pics d'intérêt

1. data-driven / knowledge-driven

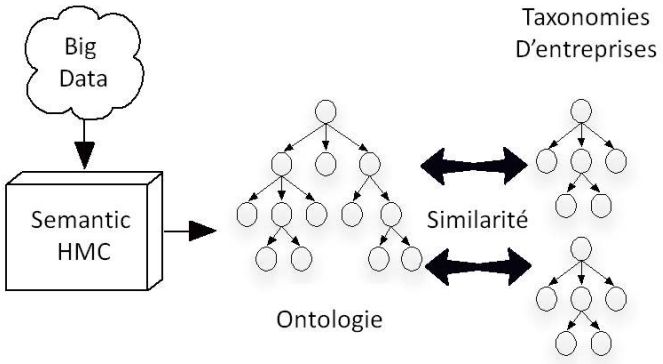


FIG. 1 – *Extraction de Valeur pour les entreprises*

dans un laps de temps donné. Quelle que soit l'application, extraire la Valeur consiste à utiliser les grands Volumes de données comme une nouvelle source d'informations inédites et potenciellement stratégiques pour l'entreprise. Notre approche consiste à déterminer la Valeur de larges volumes de données générés en continu, en utilisant une classification basée sur la sémantique. Le processus de *HMC Sémantique* génère la partie Tbox [2] d'une ontologie, i.e. la taxonomie et les règles à partir des données en se basant sur la fréquence des termes. Une fois cette phase d'apprentissage terminée, le système classifie les items arrivant au niveau Abox [3], en tant qu'individus du modèle précédemment appris. Le niveau Tbox de l'ontologie est modifié par incréments en répercussion à la classification effectuée. Le système prend ainsi en compte l'aspect de Vélocité des données. Le résultat de ce processus *HMC Sémantique* est une ontologie riche, basée sur le contenu des items, qui sont associés aux concepts correspondants dans la taxonomie. Les entreprises utilisent de façon récurrente des taxonomies de domaine, afin de représenter leur connaissance métier. Ils formalisent ainsi un modèle associant une Valeur aux données (Lambe, 2014). Notre approche a pour but d'utiliser des taxonomies issues des entreprises et de les comparer à l'ontologie générée, afin de valider cette dernière (Fig 1).

Une similarité importante entre les deux hiérarchies de concepts suggère un meilleur alignement entre les résultats de la classification et les attentes de l'entreprise pour valoriser les données. Par conséquent, les items classés sous des concepts clés de l'entreprise auront plus de Valeur. Afin de comparer les concepts des deux hiérarchies, il est nécessaire d'utiliser une méthode de distance sémantique. La comparaison des deux hiérarchies sera l'objet de travaux ultérieurs.

2. Partie conceptuelle de l'ontologie (modèle)
3. Assertions (faits) en accord avec le modèle

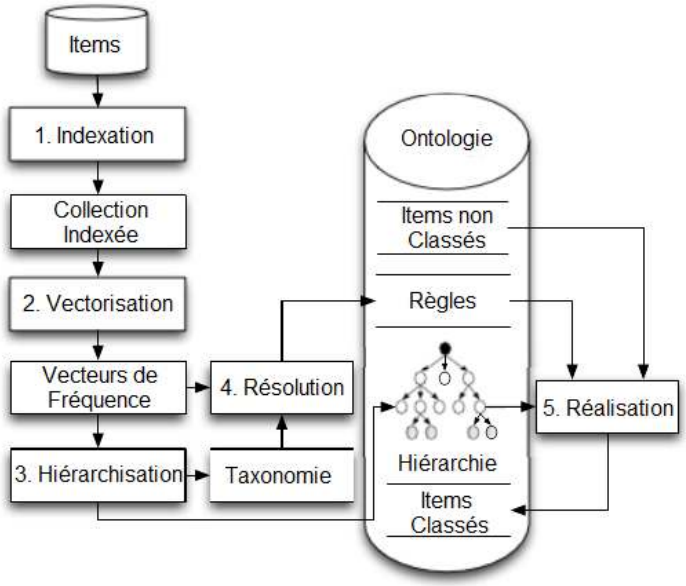


FIG. 2 – *Processus de "HMC Sémantique"*

3 Processus de "HMC Sémantique"

Notre processus de *HMC Sémantique* est générique, et est applicable à une grande Variété de données (section 3.1). L'apprentissage effectué est non supervisé : aucun exemple préétabli ne sert à construire la taxonomie initiale, ou à établir des règles de classification. Le concept le plus spécifique (ainsi que tous les concepts "parents") est inféré pour chaque nouvel item. Nous utilisons un raisonneur basé sur des règles d'inférence : un ensemble de règles est appliqué aux triplets correspondants aux items afin de déterminer leur classification (Urbani et al., 2011a). Cette approche basée sur des règles est distribuable sur des clusters de machines (section 3.4). La montée en charge peut ainsi s'effectuer, notamment via l'utilisation du framework MapReduce(Dean et Ghemawat, 2008). Des raisonneurs à échelle du web (Urbani, 2013) utilisent déjà un raisonnement basé sur des règles, et permettent un passage à l'échelle par parallélisation et distribution de la charge de calcul. Notre processus comprend 5 étapes qui passent individuellement l'échelle (figure 2) :

— *L'indexation* parse les items et crée un index inversé des items.
— *La vectorisation* calcule les vecteurs de fréquence de termes pour chaque item.
— *La hiérarchisation* crée une taxonomie basée sur les vecteurs de fréquence.
— *La résolution* génère un ensemble de règles de classification permettant de faire correspondre les items aux concepts pertinents.
— *La réalisation* peuple l'ontologie avec les items et détermine pour chaque item les concepts les plus spécifiques correspondants.

3.1 Indexation

L'indexation parse et crée un index des items. L'indexation doit être optimisée afin de filtrer des termes non pertinents, en particulier dans le cas ou les items sont des documents non structurés (textes en langage naturel par exemple). Chaque type d'item requiert un parser spécifique afin de récupérer des informations pertinentes et réduire ainsi *l'Analyse de Contenu Limitée* (Lops et al., 2011)(Bobadilla et al., 2013). *L'analyse de Contenu Limitée* répond à la difficulté d'extraire automatiquement des informations fiables suivant le format des items (textes, fichiers de logs, pages web, images, vidéos...). L'indexation est une étape obligatoire, qui permet de palier *l'Analyse de Contenu Limitée*, et ainsi de gérer une plus grande Variété de données. Le résultat de l'indexation est un corpus d'items défini par un index inversé des termes correspondants aux items.

3.1.1 Passage à l'échelle

Le passage à l'échelle de cette phase dépend du stockage des données. En l'occurrence, l'utilisation du système de fichiers HDFS de l'écosystème Hadoop est une solution appropriée. Les moteurs de recherche tels que ElasticSearch et Solr permettent de créer un index distribué sur un système HDFS. Pour des requêtes portant sur un index plusieurs millions de documents, ces moteurs affichent d'excellentes performances (temps de réponse inférieur à la seconde). L'indexation de différents flux de données provenant de sources variées (logs, crawler web...) est ainsi effectuée, les traitements ultérieurs bénéficient des avantages de la distribution des données sur un cluster HDFS [4].

3.2 Vectorisation

3.2.1 Définition

La Vectorisation crée deux types de vecteurs de fréquence de termes pour les items indexés. Un vecteur de fréquences des termes pour l'ensemble du corpus (DF) est d'abord généré, contenant l'ensemble des unigrams et n-grams du corpus d'items. La matrice de fréquence des termes est ensuite générée : chaque ligne correspond à un unigram/n-gram. Cette matrice permet de faire des appareillements de termes / n-grams dont la co-occurence est élevée. Les étapes suivantes utilisent cette matrice de termes afin de générer la taxonomie (hiérarchisation) ainsi que les règles de classification (résolution).

3.2.2 Passage à l'échelle

La création du vecteur de fréquence (des unigrams/n-grams) sur l'ensemble des documents correspond a un ensemble d'opérations de comptage. Des implémentations existent déjà : la librairie Mahout propose une implémentation distribuée de l'algorithme de calcul de collocations(n-grams).

La création de la matrice de fréquences consiste à définir chaque ligne individuellement. L'ensemble des items correspondant à chaque ligne est récupéré depuis l'index. A

4. `http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html`

partir de cet ensemble, un comptage parallélisé des termes est effectué via l'utilisation du framework MapReduce. Le comptage parallélisé est un cas d'utilisation typique du framework : la répartition des tâches consiste à reporter l'apparition de chaque terme dans une table (map), puis à regrouper et sommer les différentes occurrences du même terme (reduce) [5].

3.3 Hiérarchisation

3.3.1 Définition

La hiérarchisation sélectionne les termes pertinents à partir des vecteurs de termes et définit ainsi les concepts qui appartiennent à la hiérarchie. Deux méthodes pour construire automatiquement une hiérarchie sont utilisées dans la littérature (de Knijff et al., 2013)(Meijer et al., 2014) : la méthode de subsomption qui construit les relations de généralisation - spécialisation en se basant sur la co-occurrence des concepts, et le regroupement hiérarchique qui consiste à regrouper les concepts les plus proches les uns des autres. La méthode de subsomption est en l'occurrence adaptée à notre solution : on utilise en entrée les vecteurs créés lors de la vectorisation.

3.3.2 Passage à l'échelle

Cette étape permet également de tirer parti du framework MapReduce afin de distribuer les calculs. La création de la hiérarchie par subsomption peut être parallélisée de la façon suivante : pour chaque concept choisi comme pertinent, la recherche de candidats pour les concepts parents correspond à une opération Map. L'opération Reduce correspond à la sélection du meilleur candidat.

L'ensemble des concepts et la fréquence des couples sont récupérés depuis la matrice de fréquence générée lors de la vectorisation. Un mapper est exécuté pour chaque concept : chaque mapper reçoit en entrée un couple <X, freq(Y)> correspondant à la fréquence d'occurence du concept Y dans l'ensemble des items où le concept X apparait. Chaque mapper effectue une correspondance avec le couple opposé afin de calculer la probabilité de parenté entre les deux concepts. Les couples parents/enfants possibles sont renvoyés en sortie avec la probabilité de subsomption de chaque couple. Les concepts parents candidats respectent la relation suivante (de Knijff et al., 2013) :

$$P(p|x) \geq t, P(x|p) < t$$

où P(p|x) est le pourcentage d'occurence du concept parent p dans les items du concept x. La fonction reduce détermine le parent de chaque concept, i.e. le parent avec le score maximum. Le score de parenté pour les parents potenciels p de x est défini comme le ratio [6] :

$$score(p, x) = \frac{P(p|x)}{P(x|p)}$$

5. Concrètement, il s'agit d'une application du "wordcount", un exemple de MapReduce couremment utilisé : `http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html`

6. Simplification du score de (de Knijff et al., 2013) ou le poids des ancêtres du parent potenciel p est pris en compte dans le calcul du score

Si aucun parent n'est trouvé, le parent correspondra à la racine de la hiérarchie. La figure 3 décrit les opérations map et reduce de la subsomption :

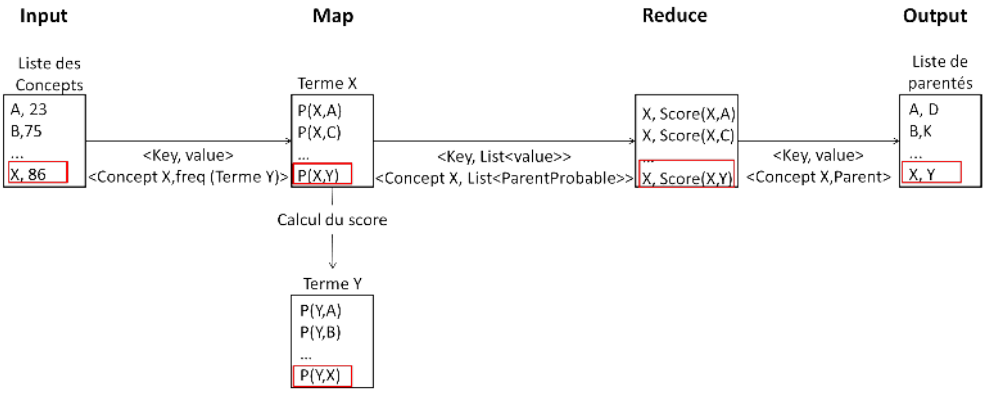


FIG. 3 – *Algorithme de subsomption (MapReduce)*

3.4 Résolution

3.4.1 Définition

La résolution crée les règles de l'ontologie utilisées pour la classification des items par rapport à la taxonomie, en se basant sur les vecteurs de fréquence de termes. Le processus de création des règles utilise une méthode de seuils (Werner et al., 2014) pour sélectionner les termes les plus pertinents pour chaque concept de la hiérarchie, et inclue ces termes dans la définition de la règle. La principale différence avec (Werner et al., 2014) est que plutôt que de traduire les règles en contraintes logiques de l'ontologie en logique de description, les règles sont transcrites au format SWRL (Semantic Web Rule Language). Le principal intérêt des règles SWRL est de réduire la charge de calcul du raisonneur, donc d'améliorer la performance du système. Un ensemble conséquent de règles SWRL simples (i.e. courtes) est généré lors de cette phase : chaque concept est associé à un ensemble de règles de classification, permettant de classer les nouveaux items en fonction des règles.

3.4.2 Passage à l'échelle

La définition des règles de classification pour un concept de l'ontologie est indépendante des autres concepts, ce qui permet de répartir la génération des règles par concept (approche de type "diviser pour régner").

3.5 Réalisation

3.5.1 Définition

La réalisation peuple l'ontologie avec les items, en accord avec la taxonomie et les règles de classification. L'ontologie est peuplée avec les items au niveau Abox (assertion). Les règles SWRL générées lors de la Résolution sont utilisées afin d'associer des étiquettes aux items. Le moteur d'inférence basé sur les règles utilise ensuite les règles et la hiérarchie pour inférer les concepts les plus spécifiques pour chaque item. La classification des items est donc multi-label et hiérarchique, basée sur une ontologie et un ensemble de règles (*Semantic HMC*).

3.5.2 Passage à l'échelle

Le raisonnement est le point critique du processus, sa parallélisation est donc primordiale. L'application des règles de classification par raisonnement sur l'ontologie lors de la requête (query time) est coûteux en temps de calcul. Cette étape est dépendante du triple store donnant accès aux à l'ontologie et aux individus générés précédemment. Deux méthodes de raisonnement sont implémentées dans les triple stores tels que AllegroGraphe, Stardog ou Virtuoso :

— *Le forward chaining* : la phase de raisonnement est effectuée une seule fois, et les inférences sont ensuite stockées (matérialisation) ce qui réduit considérablement le temps de calcul lors de la requête. Cette méthode est peu adaptée dans un système où l'ontologie change fréquemment (le raisonnement doit alors être réexécuté entièrement).
— *Le backward-chaining* : le raisonnement est effectué lors de la requête. L'inconvénient est que les requêtes demanderont un temps d'exécution relativement long, indépendamment de leur fréquence ou de leur complexité. Plusieurs approches effectuent en conséquence une ré-écriture des requêtes, afin de restreindre les axiomes et par conséquent le temps de calcul.

Dans notre cas d'utilisation, les règles sont modifiées de façon dynamique, lors de l'ajout de nouveaux items. La solution appropriée est d'utiliser un raisonnement par backward-chaining. Afin de palier le désavantage de performance du backward-chaining, des travaux récents s'intéressent au passage à l'échelle des raisonneurs par optimisation des requêtes (dans le cas du backward-chaining) d'une part et par utilisation du framework MapReduce d'autre part(Urbani et al., 2011b). Ces différentes techniques ont un intérêt important pour notre application et sont envisagées pour les phases de développement ultérieures.

4 Conclusion

Dans ce papier nous présentons notre approche pour extraire de la Valeur depuis le Big Data en utilisant un procédé de *HMC Sémantique*. Nous proposons un processus passant à l'échelle basé sur 5 étapes distinctes pour créer un modèle de classification

et classer les items par rapport au modèle. Nous utilisons le machine learning afin de générer l'ontologie ainsi que les règles de classification, et l'écosystème Hadoop (HDFS, MapReduce) afin de distribuer la charge de calcul dans un cadre Big Data pour chacune des phases du processus. Nous présentons par ailleurs une version parallélisée de l'algorithme de subsomption pour la création automatisée de la hiérarchie.

L'extraction de la Valeur est appuyée par l'intégration de connaissances métier précises dans le processus de classification. Le prototype complet de *HMC Sémantique* est en développement, nous espérons montrer l'implémentation ainsi que les résultats obtenus dans une prochaine publication. Notre travail actuel consiste en l'évaluation de l'ontologie générée, en considérant 3 aspects : la performance du système (en terme de passage à l'échelle), la qualité de la hiérarchie, et la qualité du processus de classification (« concept-tagging » des items).

Références

Bi, W. et J. Kwok (2011). Multi-label classification on tree-and DAG-structured hierarchies. *Yeast*, 1–8.

Bobadilla, J., F. Ortega, A. Hernando, et A. Gutiérrez (2013). Recommender systems survey. *Knowledge-Based Systems*, 109–132.

Chen, M., S. Mao, et Y. Liu (2014). Big Data : A Survey. *Mobile Networks and Applications*, 171–209.

de Knijff, J., F. Frasincar, et F. Hogenboom (2013). Domain taxonomy learning from text : The subsumption method versus hierarchical clustering. *Data & Knowledge Engineering*, 54–69.

Dean, J. et S. Ghemawat (2008). Mapreduce : simplified data processing on large clusters. *Communications of the ACM 51*(1), 107–113.

Gantz, J. et D. Reinsel (2011). Extracting value from chaos. *IDC iview*, 1–12.

Hitzler, P. et K. Janowicz (2013). Linked data, big data, and the 4th paradigm. *Semantic Web 4*, 233–235.

Lambe, P. (2014). *Organising knowledge : taxonomies, knowledge and organisational effectiveness*. Elsevier.

Lops, P., M. de Gemmis, et G. Semeraro (2011). Content-based recommender systems : State of the art and trends. In *Recommender Systems Handbook*, pp. 73–105. Springer.

Meijer, K., F. Frasincar, et F. Hogenboom (2014). A Semantic Approach for Extracting Domain Taxonomies from Text. *Decision Support Systems*.

Sheth, A. (2014). Transforming Big Data into Smart Data. *2014 IEEE 30th International Conference on Data Engineering*, 2–2.

Studer, R., V. R. Benjamins, et D. Fensel (1998). Knowledge engineering : Principles and methods. *Data & Knowledge Engineering*, 161–197.

Syed, A., K. Gillela, et C. Venugopal (2013). The Future Revolution on Big Data. *Future 2*, 2446–2451.

Urbani, J. (2013). Three Laws Learned from Web-scale Reasoning. In *2013 AAAI Fall Symposium Series*.

Urbani, J., F. van Harmelen, S. Schlobach, et H. Bal (2011a). QueryPIE : Backward Reasoning for OWL Horst over Very Large Knowledge Bases. In *ISWC'11*, pp. 730–745. Springer-Verlag.

Urbani, J., F. Van Harmelen, S. Schlobach, et H. Bal (2011b). Querypie : Backward reasoning for owl horst over very large knowledge bases. In *The Semantic Web–ISWC 2011*, pp. 730–745. Springer.

Werner, D., N. Silva, et C. Cruz (2014). Using DL-Reasoner for Hierarchical Multilabel Classification applied to Economical e-News. In *Science and Information Conference*, pp. 8.

Witten, I. H. et E. Frank (2005). *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann.

Summary

Analyzing Big Data can help corporations to improve their efficiency. In this work we present a new vision to derive Value from Big Data using a Semantic Hierarchical Multi-label Classification called *Semantic HMC* based on a non-supervised Ontology learning process. We also propose a *Semantic HMC* process, using scalable machine learning techniques and Rule-based reasoning. The architecture consists of 5 individually scalable steps.

De la scène de crime aux connaissances : représentation d'évènements et peuplement d'ontologie appliqués au domaine de la criminalistique informatique

Yoan Chabot*,** Aurélie Bertaux**
Tahar Kechadi*, Christophe Nicolle**

*School of Computer Science and Informatics, University College Dublin, Ireland
**Equipe CheckSem, Laboratoire Le2i, UMR CNRS 6306,
Faculté des sciences Mirande, 21078 Dijon, France
yoan.chabot@hotmail.fr

Résumé. Avec la démocratisation des technologies, les enquêtes de criminalistique informatique impliquent des volumes de données toujours plus grands et hétérogènes. Pour faciliter le travail des enquêteurs, nos travaux ont pour objectif de reconstruire automatiquement les évènements liés à un incident numérique, tout en respectant les exigences légales. Pour cela, il est nécessaire d'introduire un modèle de représentation de connaissances permettant de structurer les informations recueillies sur une scène de crime dans le but de faciliter l'utilisation de processus d'analyse automatisés. Ce papier propose un état de l'art des modèles de représentations d'évènements pour le domaine de la criminalistique informatique et introduit ensuite une nouvelle représentation basée sur une ontologie. Un processus de peuplement automatique est ensuite présenté afin d'instancier l'ontologie à partir de données collectées durant une enquête.

1 Introduction

Les nouvelles technologies occupant désormais une place prédominante dans nos vies quotidiennes, il est courant de trouver sur une scène de crime des objets numériques qui sont autant de sources d'informations possibles pour aider les enquêteurs dans la résolution d'une affaire. Le domaine de la criminalistique informatique propose des méthodes d'investigation numériques visant à fournir à la justice des pièces à conviction afin de déterminer la culpabilité ou l'innocence de suspects. Ce domaine de recherche s'intéresse à la résolution de crimes où les technologies sont une cible (e.g attaques par déni de service, utilisation frauduleuse de cartes bancaires), un vecteur principal (e.g. approche d'une victime par un pédophile via les réseaux sociaux) ou un vecteur secondaire (e.g. échange de SMS entre deux complices d'un braquage). Durant une enquête, il est nécessaire d'analyser des grands volumes de données hétérogènes de part la multiplication des objets numériques et l'augmentation de leur capacité de stockage. Par exemple, la boîte à outils *Plaso* (utilisée pour produire des chronologies à partir d'images disques) peut identifier plusieurs milliers d'évènements générés par des sources variées (historiques web, journaux d'évènements Windows, etc.) à partir d'une image disque de quelques

giga-octets. Il est par conséquent nécessaire de développer des processus automatiques pour assister les enquêteurs dans le traitement et l'interprétation de ces données. Cependant, l'utilisation de représentation des données non structurées (format textuel Mactime (Farmer et Venema, 2004) par exemple, utilisé par un grand nombre d'outils) rend le développement de processus d'analyse complexe de part le manque d'informations sur la sémantique des données. Pour faire face à ces problèmes, le recours à une représentation ontologique permet d'une part de structurer les données et d'une autre part de standardiser la représentation des informations. Les objectifs d'une telle représentation sont la simplification du développement d'outils d'analyse ainsi que la mise à disposition des données, pour les investigateurs, sous une forme permettant une consultation intuitive de l'information. Dans nos travaux, nous introduisons une représentation des connaissances permettant de modéliser de manière précise un incident numérique et l'ensemble des étapes composant une enquête. Cette représentation est utilisée dans le cadre d'une architecture ayant pour objectif l'étude a posteriori de machines pour la construction et l'analyse de chronologies sémantiquement riches d'incidents numériques (Chabot et al., 2014b). L'utilisation d'une ontologie pour la reconstruction de scénarios d'incidents présentée dans ce papier est une approche novatrice permettant de combler les manques causés par l'utilisation de formats de données plus rudimentaires.
La Section 2 évalue les représentations d'évènements existantes au regard de quatre critères déterminants pour juger de la qualité d'un modèle. Une ontologie pour la représentation d'évènements composant des incidents numériques est ensuite présentée dans la Section 3. Pour conclure, la Section 4 introduit un processus d'extraction et de peuplement permettant d'instancier cette ontologie à partir de données extraites dans une scène de crime.

2 Étude des représentations d'évènements pour la criminalistique informatique

Cette section a pour objectif d'évaluer les solutions de représentation d'évènements au regard de quatre critères :
Complétude du modèle : un modèle doit proposer un vocabulaire suffisamment complet pour représenter de manière précise les entités (évènements, objets, processus, etc.) liées à un incident, leurs caractéristiques et les relations entre ces entités. Plusieurs formats de données (Bodyfile, Mactime (Farmer et Venema, 2004), TimeLiNe (Carvey, 2009)) existent pour représenter des chronologies d'évènements. Ces formats utilisant un faible nombre d'attributs, la représentation des évènements est imprécise. De plus, un autre inconvénient de ces formats est qu'ils ne permettent pas de représenter les relations entre entités. Des modèles de représentation plus évolués sont proposés tels que ECF (Chen et al., 2003) et FORE (Schatz et al., 2004). Ces modèles permettent de représenter des dimensions caractéristiques des évènements (temps, objets utilisés, participants impliqués dans un évènement, etc.). Toutefois et à l'instar des formats précédents, ils ne modélisent pas les relations entre les entités (e.g. il est possible de modéliser le fait qu'un évènement interagit avec un objet mais pas de spécifier la nature de cette interaction). (Mudholkar et Bharambe, 2013) proposent une ontologie incluant des dimensions également proposées dans notre modèle telles que le temps ou les protagonistes impliqués dans un incident. Cependant, l'ontologie proposée n'est pas suffisamment décrite

pour permettre son évaluation. Enfin, CybOX [1] est un ensemble de schémas XSD permettant la représentation d'entités (processus ou ressources) et d'évènements les affectant. L'une des spécificités de ce modèle est l'intégration de connaissances techniques à travers un ensemble d'objets (fichiers PDF, historiques Web, connexions réseau, etc.).

Traçabilité des informations : Le deuxième critère de cette étude est l'intégration de données dans le modèle assurant la traçabilité de l'information. Pour satisfaire les exigences légales, les pièces à conviction utilisées lors d'un procès doivent respecter plusieurs critères parmi lesquels la crédibilité des preuves et la reproductibilité de leur méthode de production (Baryamureeba et Tushabe, 2004). Un modèle doit donc permettre la modélisation de la provenance de chaque information produite durant une enquête, incluant des informations sur la nature de chaque tâche accomplie, sur les enquêteurs ayant contribué à chacune d'elles et sur les outils utilisés. Le modèle CybOX incorpore des éléments pour modéliser la provenance de l'information afin de mémoriser pour chaque entité la source d'information et les techniques utilisées ainsi que les contributeurs ayant participé à son identification. Un autre travail pertinent pour ce critère est la recommandation W3C PROV-O (Lebo et al., 2013) décrivant une ontologie composée de concepts et de relations permettant de définir une information ainsi que le processus utilisé pour la produire. Toutefois, cette ontologie n'étant pas appliquée à la criminalistique informatique, certaines caractéristiques spécifiques au domaine sont manquantes.

Automatisation des processus : Ce critère est lié au besoin de produire des outils automatisés capables de traiter de grands volumes de données. La conception de tels outils nécessite que les données soient représentées dans un format compréhensible par des machines. Le rôle de ce critère est d'évaluer le niveau de structuration des données ainsi que la mise à disposition de mécanismes pour faciliter l'utilisation des données par des processus d'analyse automatiques. L'approche ECF introduit une représentation à deux niveaux : un premier niveau contenant des informations génériques sur les évènements et un deuxième niveau contenant des informations spécifiques à chaque type d'évènement. La représentation canonique des évènements permet de les modéliser de manière uniforme indépendamment de la source à partir de laquelle ils sont extraits (facilitant l'analyse des informations) tout en préservant les spécificités de chaque évènement. Le recours à une ontologie, comme illustré dans l'approche FORE, facilite également l'automatisation de part la description formelle de la sémantique des entités permettant aux machines de comprendre la signification des données.

Utilisabilité du modèle : En complément de la mise à disposition des données pour des processus automatiques, les modèles doivent également permettre aux enquêteurs d'accéder à l'information et de comprendre les données (outils de recherche et de visualisation pour un accès intuitif et rapide aux données). Les formats textuels ne permettent pas aux enquêteurs de comprendre aisément les informations contenues dans une chronologie. Les ontologies sont plus à même de répondre à ce critère en proposant des visualisations sous forme de graphes permettant notamment d'identifier rapidement des connexions entre les entités.

En conclusion, aucun des modèles présents dans la littérature ne donne des réponses satisfaisantes à l'ensemble des critères énoncés. Dans la section suivante, nous introduisons un modèle de représentation des évènements apportant des réponses à ces quatre critères. L'ontologie proposée tire parti de l'ontologie PROV-O pour la représentation de la provenance des informations et du modèle CybOX pour garantir sa complétude.

1. https ://cybox.mitre.org/

3 Une ontologie pour la représentation d'évènements liés à des incidents numériques

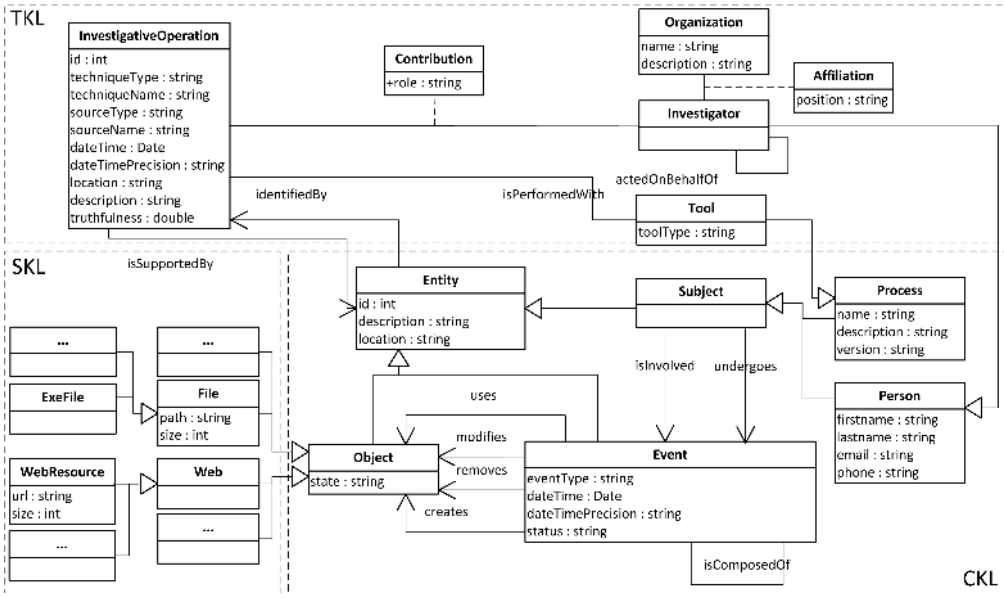


FIG. 1 – *Ontologie pour la représentation d'incidents numériques*

Pour satisfaire les critères énoncés dans la section précédente, une ontologie implémentée à l'aide du langage *OWL 2 RL* est utilisée. Une ontologie permet de représenter les connaissances d'un domaine donné en structurant ces informations sous forme d'entités, de relations et de contraintes logiques sur ces entités et relations. Les ontologies sont ainsi capables de représenter formellement les connaissances générées durant une enquête (connaissances sur les évènements, les processus et les personnes, etc.). Contrairement à des formats de données plus rudimentaires, elles permettent de représenter des relations entre entités ainsi que la logique sous-jacente aux données. La nature explicite et formelle des ontologies permet de faciliter la conception et l'emploi d'outils d'interprétation et d'analyse (déduction de nouvelles informations, vérification de la cohérence des connaissances, etc.) en complément des enquêteurs. Les ontologies sont également une structure facilement manipulable grâce à des outils tels que SPARQL, un langage d'interrogation conçu pour travailler sur des graphes de connaissances. Enfin, la structuration en triplets rend possible la visualisation sous forme de graphes, une représentation claire et intuitive pour les enquêteurs.
L'ontologie proposée dans ce papier est implémentée à l'aide du profil *OWL 2 RL* (sous ensemble de *OWL 2 DL*, un langage basé sur les logiques de descriptions *SHROIQ(D)*). Le choix de ce langage est motivé par plusieurs raisons, dont la mise à disposition d'une expressivité suffisante pour modéliser le domaine nous intéressant. *OWL 2 DL* permet notamment de définir des hiérarchies de classes et de propriétés, des restrictions ou encore d'établir des faits sur les individus tels que l'égalité d'instances. L'utilisation du profil *OWL 2 RL* permet de contraindre

sur certains aspects (expressions de classes notamment) le langage *OWL 2 DL* afin de garantir la décidabilité et la rapidité (complexité polynomiale) des raisonnements à base de règles (Motik et al., 2009). La nécessité d'opérer sur de grands volumes de données et la volonté de proposer aux enquêteurs des outils d'inférence et d'analyse puissants rendent le langage *OWL 2 RL* pertinent pour implémenter une ontologie pour la représentation de chronologies d'incidents. Pour répondre aux besoins de complétude et de traçabilité de l'information, l'ontologie proposée est divisée en trois couches ("Common Knowledge Layer", "Specialized Knowledge Layer" et "Traceability Knowledge Layer") illustrées dans la Figure 1. Afin de garantir la lisibilité, seules les classes, propriétés et attributs nécessaires à la compréhension du papier sont représentés. La classe centrale de l'ontologie est la classe *Entity*, notion abstraite subsumant les classes principales de l'ontologie. Chaque instance de *Entity* est définie par un identifiant unique, une description courte et une localisation. *Entity* est directement spécialisée par les classes *Event*, *Object* et *Subject* et indirectement spécialisée par les classes *Investigator*, *Tool* (TKL) et *Process*, *Person* (CKL).

3.1 Traçabilité des informations et reproductibilité des processus

La couche TKL, inspirée de l'ontologie PROV-O (Lebo et al., 2013), stocke des informations sur la manière dont l'enquête est menée (e.g. participants, étapes de l'enquête, informations en entrée/sortie de chaque étape, etc.). L'objectif de cette couche est de satisfaire les exigences légales en assurant d'une part la reproductibilité des résultats via la mémorisation de chaque action et d'une autre part la crédibilité des résultats en conservant le cheminement et les données utilisées pour produire les résultats. Chaque tâche (*InvestigativeOperation*) est caractérisée par un ensemble d'attributs permettant notamment de définir : le type de techniques utilisées (extraction à partir d'une source d'informations, déduction de nouvelles connaissances, corrélation d'évènements, etc.), les sources d'informations utilisées (archives de conversations, registre Windows, etc.), la date et le lieu où la tâche a été effectuée ou encore une valeur numérique quantifiant le degré de confiance du résultat (peu élevé par exemple, dans le cas d'une tâche utilisant des informations potentiellement corrompues par des assaillants). Les instances de *InvestigativeOperation* sont liées aux outils (*Tool*) utilisés et aux personnels (*Investigator*) impliqués en utilisant respectivement les propriétés d'objets *isPerformedWith* et *Contribution*. Chaque instance de *InvestigativeOperation* est utilisée pour augmenter la connaissance des enquêteurs sur les évènements survenus durant l'incident. Ainsi, chaque tâche de l'enquête est liée aux évènements ainsi qu'aux sujets et objets qu'elle a permis d'identifier. La propriété d'objet *identifiedBy* modélise le fait que toute entité est identifiée à l'aide d'une instance de *InvestigativeOperation* (e.g. une tâche d'extraction d'information à partir d'un historique web peut engendrer l'identification d'un évènement représentant la visite d'une page web). Pour certaines tâches, les enquêteurs doivent raisonner sur des informations déjà existantes pour produire de nouvelles connaissances. La propriété d'objet *isSupportedBy* modélise ce principe en liant les instances de *InvestigativeOperation* aux informations utilisées par celles-ci.

3.2 Connaissances génériques sur l'incident

La couche CKL, dérivée du modèle formel introduit dans (Chabot et al., 2014a), est utilisée pour stocker des connaissances génériques sur les évènements. Elle modélise notamment des connaissances temporelles, des informations sur les objets utilisés par chaque évènement et les

sujets participant à chacun d'eux. Son objectif est d'obtenir une représentation uniforme des évènements composant un incident afin de simplifier les tâches d'analyse en aval. La classe *Event* permet de modéliser tout évènement numérique survenant sur une machine. Chaque instance de *Event* est définie par un type (e.g. copie d'un fichier, suppression d'une clé de registre, etc.), un intervalle de temps représentant la durée ainsi qu'un statut (succès, échec, en cours, inconnu). La propriété d'objet *isComposedOf* est utilisée pour lier un évènement à un évènement le composant. Les classes *Subject*, *Process* et *Person* sont utilisées pour modéliser les protagonistes impliqués dans les évènements. Un sujet peut participer (*isInvolved*) à un évènement ou subir (*undergoes*) ce dernier. La classe *Object* représente les ressources utilisées (*uses*), modifiées (*modifies*), supprimées (*removes*) ou créées (*creates*) par les évènements.

3.3 Représentation de connaissances métiers

La couche SKL est utilisée pour stocker des connaissances spécialisées sur les évènements, et notamment les objets utilisés par ces derniers. Elle permet de modéliser des connaissances techniques sur tout objet numérique pouvant être identifié dans une scène de crime numérique. Les informations techniques sur les évènements (adresses IP, chemin et métadonnées de fichiers, etc.) stockées dans cette couche sont des informations de valeur durant la phase d'analyse. La couche SKL propose un panel important de classes permettant de représenter un grand nombre d'objets numériques. Cette couche inclut notamment des objets permettant de représenter :

- Des fichiers (*File*) : *OLECF*, *Link*, *ArchiveFile*, *ImageFile*, *PDFFile*, *ExeFile*.
- Des comptes d'utilisateurs (*Account*) : *UnixUserAccount*, *WinUserAccount*, *ComAccount*.
- Des objets spécifiques au Web et à son utilisation (*Web*) : *Webpage*, *WebResource*, *EmailMessage*, *Bookmark*, *Cookie*, etc.
- Des objets relatifs aux communications (*Communication*) : *MMS*, *SMS*, *Chat*, *Call*.
- Des clés de registre (*RegisterKey*).

4 Peuplement de l'ontologie à partir de traces extraites dans une scène de crime

Cette section a pour objectif d'illustrer le peuplement de l'ontologie à partir de données extraites dans une scène de crime. L'introduction de techniques de peuplement automatisées est primordiale pour permettre le traitement des grands volumes de données extraits lors d'une enquête. La méthode de peuplement utilisée dans notre approche est un processus séquentiel, illustré dans la Figure 2, débutant par la collecte des traces numériques trouvées sur une machine et se terminant par l'instanciation des concepts et des propriétés de l'ontologie.

4.1 Utilisation de la boîte à outils Plaso pour l'extraction d'information à partir de traces numériques

La première étape consiste à extraire l'ensemble des informations contenues dans les différentes sources d'évènements présentes dans l'image disque analysée (image disque des vo-

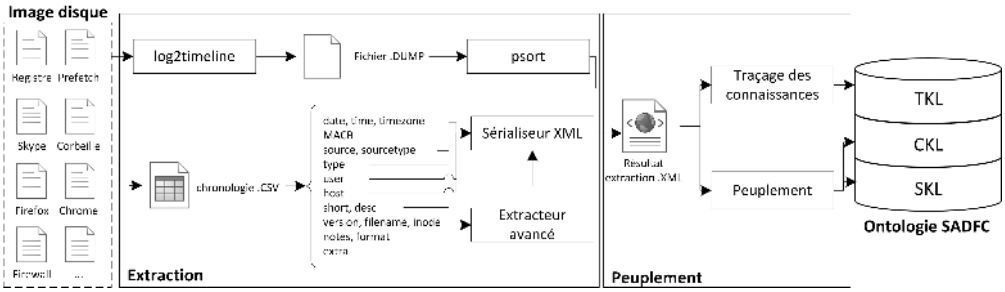


FIG. 2 – *Chaîne d'extraction et de peuplement*

lumes de la machine étudiée). Durant une investigation, de nombreuses sources peuvent être utilisées afin d'obtenir des informations sur les activités de l'utilisateur. Pour gérer l'ensemble de ces sources, l'outil *log2timeline* (Gudhjonsson, 2010), proposé dans la boîte à outils *Plaso*, est utilisé. Ce dernier collecte des informations à partir de nombreuses sources d'informations, parmi lesquelles : les sources inhérentes au système d'exploitation (e.g. base de registre, système de fichier, corbeille, journaux d'évènements) ; les historiques, cookies et fichiers de cache des navigateurs Web ; les fichiers et journaux inhérents à des logiciels divers tels que Skype, Google Drive, etc. Le résultat produit lors de cette étape est un fichier *.dump* contenant l'ensemble des informations extraites de l'image disque. Une transformation du résultat est ensuite nécessaire pour rendre les données utilisables par les processus en aval. Pour cela, l'outil *psort* de la boîte à outils *Plaso* est utilisé. Cet outil permet de sérialiser les données produites par *log2timeline* dans de nombreux formats parmi lesquels le format CSV. Un fragment d'exemple de résultat obtenu en sortie de l'outil *psort* est donné dans la figure 3. Chacune des lignes du fichier illustré dans l'exemple est une entrée décrivant une action survenue sur la machine étudiée. Le premier évènement extrait représente le téléchargement d'un fichier *.exe* à l'aide de Google Chrome. La deuxième entrée décrit l'exécution de ce même fichier *.exe* qui a pu être identifiée via les informations contenues dans le dossier Windows Prefetch. Enfin, la troisième entrée représente la suppression du fichier téléchargé (envoi dans la corbeille).

4.2 Extraction avancée et sérialisation des informations

La sérialisation CSV est structurée en dix-sept attributs donnant des informations temporelles (*date*, *time* et *timezone*), une description de la source d'informations via les champs *source* et *sourcetype* (e.g. historiques de Chrome, Windows Prefetch, corbeille, etc.), une description de l'évènement et de ses conséquences(*MACB*, *type*, *short*, *desc* et *extra*), des informations sur les outils utilisés pour l'extraction et les fichiers utilisés comme source d'informations (*version*, *filename*, *inode*, *notes* et *format*). Certains champs (*date*, *time*, etc.), ne nécessitent aucun traitement particulier et l'extraction des connaissances est aisée. D'autres champs tels que le champ *desc* nécessitent davantage de traitements car leur contenu dépend fortement du type et du déroulement de l'évènement. Ces champs présentent les mêmes inconvénients que les formats textuels présentés dans l'état de l'art et la variabilité de leur contenu rend leur manipulation complexe. Un deuxième problème posé par la chronologie produite par *log2timeline* et *psort* est celui du volume de données. En effet, une chronologie produite à l'aide de ces

```
date , time , timezone , MACB, source , sourcetype , type , user , host , short , desc , version ,
     filename , inode , notes , format , extra
11/24/2014,11:50:24 ,UTC,...B,WEBHIST,Chrome History ,File Downloaded,-,WIN-I51P7DIKOO0,C:\
     Users\User1\Downloads\Firefox Setup Stub 33.1.1.exe downloaded (244336 bytes),
     https://download-installer.cdn.mozilla.net/pub/firefox/releases/33.1.1/ Firefox%20
     Setup%20Stub%2033.1.1.exe. Received: 244336 bytes out of: 244336 bytes.,2,TSK:/Users
     /User1/AppData/Local/Google/Chrome/User Data/Default/History,43770,-,sqlite ,plugin:
     chrome_history
11/24/2014,11:51:19 ,UTC,.A..,LOG,WinPrefetch ,Last Time Executed,-,WIN-I51P7DIKOO0,FIREFOX
     .EXE was run 4 time(s),Prefetch [FIREFOX.EXE] was executed - run count 4 path: \
     USERS\USER1\DESKTOP\SOFTWARES\FIREFOX.EXE hash: 0x9336A096 volume: 1 [serial number:
      0x724766A7 device path: \DEVICE\HARDDISKVOLUME1],2,TSK:/Windows/Prefetch/FIREFOX.
     EXE-9336A096.pf,43408,-, prefetch ,number_of_volumes: 1 volume_device_paths: [u'\\
     DEVICE\\HARDDISKVOLUME1'] volume_serial_numbers: [1917281959L] version: 23
     prefetch_hash: 2469830806
11/24/2014,11:51:31 ,UTC,M...,RECBIN,Recycle Bin,Content Deletion Time,-,WIN-I51P7DIKOO0,
     Deleted file: C:\Users\User1\Downloads\Firefox Setup Stub 33.1.1.exe,C:\Users\User1\
     Downloads\Firefox Setup Stub 33.1.1.exe,2,TSK:/$Recycle.Bin/S
     -1-5-21-2714290424-3384145025-262107571-1000/ $IOP4Y0X.exe,50944,-,recycle_bin ,
     file_size: 244336
```

FIG. 3 – *Données produites par log2timeline et formatées à l'aide de psort*

outils, à partir de l'image disque d'une machine ayant fonctionné environ trente minutes avec une utilisation standard, est composée d'environ 300 000 entrées. L'étude du fichier est réalisée manuellement (e.g. grep, recherche par dates, etc.) par les enquêteurs et l'interprétation de la chronologie est par conséquent particulièrement laborieuse. Cet état de fait valide le choix de l'utilisation de modèles de représentation des données plus avancés tels que les ontologies.
Dans l'objectif de faciliter le peuplement de l'ontologie à partir des données produites par *Plaso*, une étape intermédiaire d'extraction de l'information et de sérialisation au format XML est introduite dans notre processus. Pour cela, les informations structurées telles que les informations temporelles (*date*, *time* et *timezone*), les informations sur l'utilisateur (*user*), les informations sur l'hôte (*host*) ainsi que des informations sur le type d'évènements (*source* et *sourcetype*) sont tout d'abord extraites. Les informations contenues dans les champs plus faiblement structurés, tel que le champ *desc*, sont ensuite extraites. Le contenu du champ *desc* dépendant de la source d'informations et du type d'évènements, un ensemble de motifs est défini afin d'extraire correctement les informations. Dans le cas de la première entrée de la Figure 3 correspondant au téléchargement d'un fichier à l'aide de Google Chrome, un motif est utilisé pour extraire dans le champ *desc* l'URL du fichier téléchargé et le chemin local utilisé pour son stockage ainsi que la taille du fichier. L'étape suivante consiste à filtrer les données collectées afin de conserver uniquement les données pertinentes pour alimenter notre modèle dans le but de réduire la quantité de données, améliorer la lisibilité du résultat final et optimiser les temps de traitements. Après le filtrage, les données sont ensuite sérialisées au format XML. La sérialisation de la première entrée de la Figure 3 est donnée dans la Figure 4. Cette figure montre notamment la décomposition à l'aide des motifs des informations contenues dans le champ *desc* au sein de l'élément XML *description*.

```
1  <footprint id="1">
2          <datetime>11/24/2014 11:50:24 UTC</datetime>
3          <type>Chrome History</type>
4          <subtype>Download of a file</subtype>
5          <location>WIN-I51P7DIKOO0</location>
6          <user>-</user>
7          <process>Google Chrome</process>
8          <description>
9          <url>https://download-installer.cdn.mozilla.net/pub/firefox/ releases/33.1.1/
               win32/fr/Firefox%20Setup%20Stub%2033.1.1.exe</url>
10         <localPath>C:\Users\User1\Downloads\Firefox Setup Stub 33.1.1.exe</localPath>
11         <receivedBytes>244336</receivedBytes>
12         <sizeFile>244336</sizeFile>
13         </description>
14         <extra>https://download-installer.cdn.mozilla.net/pub/firefox/releases/33.1.1/
               Firefox%20Setup%20Stub%2033.1.1.exe. Received: 244336 bytes out of: 244336
               bytes.</extra>
15 </footprint>
```

FIG. 4 – *Données XML en sortie du processus d'extraction*

4.3 Peuplement de l'ontologie et traitements sur les connaissances

La dernière étape consiste à peupler l'ontologie à partir du résultat de l'extraction. Pour chaque élément *footprint* composant le fichier XML, les couches CKL et SKL sont peuplées en créant des instances d'évènements, d'objets et de sujets et des liens entre les individus conformément aux propriétés formelles définies dans l'ontologie (Chabot et al., 2014a). Les relations liant un évènement à un objet ou à un sujet sont déduites en fonction du type de l'évènement (e.g. dans le cas du déplacement d'un fichier vers la corbeille, l'évènement est lié au fichier via la propriété *:removes*). La Figure 5 présente une sérialisation en Turtle des connaissances relatives à l'évènement de téléchargement du fichier à l'aide de Google Chrome utilisé tout au long du document. L'instance *:event1* (lignes 9 à 19) représente cet évènement et est liée aux instances *:webResource1* et *:exeFile1*. L'instance *:webResource1* (lignes 20 à 24) de la classe *:WebResource* représente la ressource distante téléchargée par l'utilisateur. Cette instance est liée par une relation d'utilisation *:uses* à l'évènement car le téléchargement est réalisé à partir de cette ressource distante. *:exeFile1* (lignes 25 à 29) est une instance de la classe *:ExeFile* représentant le fichier *.exe* local, téléchargé par l'utilisateur. Cette instance est reliée à l'évènement par une relation de création *:creates*. Le processus Google Chrome, utilisé pour mener à bien l'évènement, est représenté par l'instance *:googleChrome* (lignes 30 à 34) et lié à l'évènement via la propriété d'objet *:isInvolved*. Des connaissances décrivant la manière dont les connaissances précédentes ont été extraites sont ensuite ajoutées dans l'ontologie dans la couche TKL. L'instance *:investigativeOperation1* (lignes 42 à 51) représente la tâche d'extraction de l'information réalisée à l'aide de l'outil Plaso (lignes 52 à 57).
Au terme du peuplement de l'ontologie, chaque élément *footprint* en entrée est représenté par un graphe ontologique. Les étapes suivantes consistent à consolider les connaissances présentes dans l'ontologie ainsi qu'à traiter et analyser ces dernières. Ces étapes sont hors de la portée de cette publication mais sont toutefois décrites succinctement ici. La consolidation des connaissances est une étape permettant l'identification des connexions entre les différents graphes de connaissances dans le cas où les évènements interagissent avec des objets ou des

sujets identiques. La phase de consolidation comprend également une étape d'inférence de nouvelles connaissances afin de compléter les connaissances des enquêteurs sur l'incident. Après l'étape de consolidation, un graphe de connaissances de grande taille représentant les informations contenues dans le résultat produit par *Plaso* est obtenu. La structure de ce graphe est dictée par le schéma de notre ontologie. Ce graphe présente de nombreux avantages car il structure l'information et ainsi facilite sa compréhension par les enquêteurs et la mise en place de processus automatiques d'analyse. Le premier outil d'analyse proposé dans notre approche est un outil de corrélation d'évènements permettant de détecter des couples d'évènements liés (Chabot et al., 2014a). L'identification de tels couples est réalisée à l'aide de quatre critères : l'interaction des deux évènements avec des objets communs ou des sujets communs, la proximité temporelle et la validation ou non de règles métiers définies par les spécialistes. Par exemple, soit un évènement A représentant la création d'un marque page pour une page donnée et l'évènement B représentant la visite de cette même page, la valeur du score de corrélation entre les évènements A et B est augmentée par l'utilisation d'un objet commun (la page Web) et l'interaction avec un même processus (le navigateur Web). Le deuxième outil proposé est un algorithme de recherche de motifs permettant de détecter des actions illicites en identifiant des séquences d'évènements particulières (une action illicite peut être composée de plusieurs évènements autorisés d'où la nécessité d'utiliser un système à base de motifs pour détecter correctement les actions délictueuses).

5 Conclusion et travaux futurs

Durant une enquête de criminalistique informatique, les enquêteurs doivent faire face à plusieurs problèmes parmi lesquels le volume de données à traiter, l'hétérogénéité de ces données et les exigences légales. Pour servir de support au développement d'outils d'aide à la décision pour les enquêteurs, il est nécessaire d'introduire un modèle de représentation des informations permettant une modélisation précise des connaissances, l'intégration de données sur la traçabilité, l'automatisation des tâches en aval et une restitution des données intuitive et rapide. Afin de répondre à ces besoins, nous proposons un nouveau modèle de représentation des évènements basé sur une ontologie *OWL 2 RL* et couplé à un processus de peuplement automatisé. Cette ontologie, grâce à son expressivité, permet de répondre au besoin de complétude et à la nécessité de représenter des informations sur la provenance des résultats. De plus, la possibilité d'associer à cette ontologie des outils d'interrogation et de visualisation permet un accès intuitif et efficace aux connaissances et facilite la compréhension des données.
Les travaux futurs se concentreront tout d'abord sur l'extension du processus de peuplement à de nouvelles sources d'informations, afin de s'approcher d'une vision complète des évènements survenus durant un incident. Un autre objectif important est la validation de l'ontologie par des experts du domaine de la criminalistique informatique puis l'obtention d'un consensus au sein de la communauté pour l'adoption d'un modèle de représentation commun afin de faciliter l'interopérabilité des outils utilisés dans le domaine. Enfin, ces travaux soulèvent également des questions éthiques. Bien que notre approche s'applique au domaine de la criminalistique informatique, elle peut également être utilisée par des sociétés tiers (e.g. un fournisseur de services peut utiliser un modèle similaire pour profiler ses utilisateurs) au risque de porter atteinte à la vie privée.

Références

Baryamureeba, V. et F. Tushabe (2004). The enhanced digital investigation process model. In *Proceedings of the Fourth Digital Forensic Research Workshop*. Citeseer.

Carvey, H. (2009). Timeline analysis, pt iii, http ://windowsir.blogspot.fr/2009/02/timeline-analysis-pt-iii.html.

Chabot, Y., A. Bertaux, C. Nicolle, et T. Kechadi (2014a). A Complete Formalized Knowledge Representation Model for Advanced Digital Forensics Timeline Analysis. *Digital Investigation 11*(2), S95–S105.

Chabot, Y., A. Bertaux, C. Nicolle, et T. Kechadi (2014b). Automatic Timeline Construction and Analysis For Computer Forensics Purposes. In *IEEE Joint Intelligence & Security Informatics Conference 2014 (IEEE JISIC2014)*, La Haye, Netherlands, pp. 4.

Chen, K., A. Clark, O. De Vel, et G. Mohay (2003). Ecf-event correlation for forensics. In *First Australian Computer Network and Information Forensics Conference*, Perth, Australia, pp. 1–10. Edith Cowan University.

Farmer, D. et W. Venema (2004). The coroner's toolkit (tct), http ://www.porcupine.org/forensics/tct.html.

Gudhjonsson, K. (2010). Mastering the super timeline with log2timeline. *SANS Reading Room*.

Lebo, T., S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, et J. Zhao (2013). Prov-o : The prov ontology. *W3C Recommendation, 30th April*.

Motik, B., B. C. Grau, I. Horrocks, Z. Wu, A. Fokoue, et C. Lutz (2009). Owl 2 web ontology language : Profiles. *W3C recommendation 27*, 61.

Mudholkar, M. et U. Bharambe (2013). A study on significance of event ontology approach in web crime mining. *International Journal of Latest Trends in Engineering and Technology 2*.

Schatz, B., G. Mohay, et A. Clark (2004). Rich event representation for computer forensics'. *Proceedings of the Fifth Asia-Pacific Industrial Engineering and Management Systems Conference (APIEMS 2004) 2*(12), 1–16.

Summary

Due to the democratization of technologies, computer forensics investigators have to deal with volumes of data increasingly large and heterogeneous. To facilitate the work of investigators, our work aims at reconstructing automatically the events related to a digital incident, while respecting legal requirements. To reach this goal, it is necessary to introduce a knowledge representation model allowing to structure the information collected from a crime scene in order to facilitate the use of analysis processes. This paper first gives a comprehensive state of the art of event representation models for digital forensics and then proposes a new ontology. In addition, an automatic settlement process is then presented to instantiate the ontology using data collected on a machine seized during an investigation.

```
@prefix : <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@base <http://www.semanticweb.org/sadfc> .
<http://www.semanticweb.org/sadfc> rdf:type :Ontology ;
        :versionIRI <http://www.semanticweb.org/sadfc/1.0.0> ;
        :imports <http://www.w3.org/2006/time> .
:event1 rdf:type :Event ,
        :NamedIndividual ;
        :hasID 1 ;
        :uses :webResource1 ;
        :creates :exeFile1 ;
        :hasEventType "Chrome History"^^xsd:string ;
        :hasEventSubtype "Download of a file"^^xsd:string ;
        :isIdentifiedBy :investigativeOperation1 ;
        :hasLocation "WIN-I51P7DIKOO0"^^xsd:string ;
        :hasDateTime :interval1 ;
        :hasDateTimePrecision "sec"^^xsd:string .
:webResource1 rdf:type :WebResource ,
        :NamedIndividual ;
        :hasID 2 ;
        :hasSize 244336 ;
        :hasURL "https://download-installer.cdn.mozilla.net/pub/firefox/
            releases/33.1.1/Firefox%20Setup%20Stub%2033.1.1.exe"^^xsd:string .
:exeFile1 rdf:type :ExeFile ,
        :NamedIndividual ;
        :hasID 3 ;
        :hasSize 244336 ;
        :hasPath "C:\\Users\\User1\\Downloads\\Firefox Setup Stub
            33.1.1.exe"^^xsd:string.
:googleChrome rdf:type :Process ,
        :NamedIndividual ;
        :hasID 4 ;
        :hasName "Google Chrome"^^xsd:string ;
        :isInvolved :event1 .
:interval1 rdf:type :NamedIndividual ,
        <http://www.w3.org/2006/time#Interval> ;
        <http://www.w3.org/2006/time#hasEnd> :instant1 ;
        <http://www.w3.org/2006/time#hasBeginning> :instant1 .
:instant1 rdf:type :NamedIndividual ,
        <http://www.w3.org/2006/time#Instant> ;
        <http://www.w3.org/2006/time#inXSDDateTime> "2014-11-24T11:50:24"^^xsd:string .
:investigativeOperation1 rdf:type :InvestigativeOperation ,
        :NamedIndividual ;
        :hasID 5 ;
        :hasTruthfulness "100.0"^^xsd:double ;
        :hasTechniqueType "Information Source"^^xsd:string ;
        :hasTechniqueName "Extraction using Plaso"^^xsd:string ;
        :hasSourceName "Google Chrome History"^^xsd:string ;
        :isPerformedWith :plaso .
        :hasDateTime :interval2 .
        :hasDateTimePrecision "sec"^^xsd:string ;
:plaso rdf:type :Tool ,
        :NamedIndividual ;
        :hasID 6 ;
        :hasVersion "1.1.0"^^xsd:string ;
        :hasToolType "Digital Forensics"^^xsd:string ;
        :hasName "Plaso"^^xsd:string .
:interval2 rdf:type :NamedIndividual ,
        <http://www.w3.org/2006/time#Interval> ;
        <http://www.w3.org/2006/time#hasEnd> :instant2 ;
        <http://www.w3.org/2006/time#hasBeginning> :instant2 .
:instant2 rdf:type :NamedIndividual .
        <http://www.w3.org/2006/time#Instant> ;
        <http://www.w3.org/2006/time#inXSDDateTime> "2014-11-25T14:53:10"^^xsd:string .
```

FIG. 5 – *Sérialisation Turtle du résultat à l'issu du peuplement de l'ontologie*

Index