Extraction et Gestion de Connaissance

Hammamet, 26 janvier 2010

7ème Atelier Fouille de Données Complexes dans un processus d'extraction de connaissances
Complexité liée aux données multiples

Responsables

Boutheina Ben Yaghlane (LARODEC, IHEC Carthage)

Guillaume Cleuziou (LIFO, Université d'Orléans)

Mustapha Lebbah (LIPN, Université Paris 13)

Arnaud Martin (ENSIETA - E3I2/EA3876)

I.H.E.C

7ème Atelier
Fouille de données complexes - complexité liée aux données multiples

Boutheina Ben Yaghlane*, Guillaume Cleuziou**, Mustapha Lebbah***, Arnaud Martin****

*LARODEC, IHEC Carthage, Carthage Présidence 2016, Tunisie
boutheina.yaghlane@ihec.rnu.tn
**LIFO, Université d'Orléans
guillaume.cleuziou@univ-orleans.fr
***LIPN, Université Paris 13
mustapha.lebbah@univ-paris13.fr
****ENSIETA, E^{3}I^2, EA3876, 2 rue François verny, 29806 Brest cedex 9
Arnaud.Martin@ensieta.fr

Résumé. L'atelier sur la fouille de données complexes dans un processus d'extraction de connaissances est proposé à l'instigation du groupe de travail EGC "Fouille de données complexes". Chaque année les organisateurs proposent une thématique de recherche qui suscite l'intérêt des chercheurs et des industriels. Cette année la septième édition a porté en priorité sur une problématique transversale : la complexité liée aux données multiples (multi-sources, multi-vues, tableaux multiples, séquentielles, etc.).

1 Présentation

La septième édition de l'atelier sur la fouille de données complexes dans un processus d'extraction de connaissances est faite à l'instigation du groupe de travail EGC "Fouille de données complexes" http ://eric.univ-lyon2.fr/ gt-fdc.

Cet atelier est devenu un lieu privilégié de rencontre où chercheurs et industriels viennent partager leurs expériences et expertises dans le domaine de la fouille de données complexes (i.e. des données non structurées comme c'est le cas dans le web, les séquences vidéo, etc.).

Quelques mots sur la fouille de données complexes

Dans tous les domaines tels que le multimédia, la télédétection, l'imagerie médicale, les bases de données, le web sémantique, la bio informatique et bien d'autres, les données à traiter pour y extraire de la connaissance sont de plus en plus complexes et volumineuses.
Nous sommes ainsi conduits à manipuler des données souvent non structurées :

- issues de diverses provenances comme des capteurs ou sources physiques d'informations variées ;

- représentant la même information à des dates différentes ;
- regroupant différents types d'informations (images, textes) ou encore de natures différentes (logs, contenu de documents, ontologies, etc.).
- ayant des distributions différentes et déséquilibrées. Actuellement ça devient la norme et non l'exception

Aussi la fouille de données complexes ne doit plus être considérée comme un processus isolé mais davantage comme une des étapes du processus plus général d'extraction de connaissances dans les bases de données (ECBD). En effet, avant d'appliquer des techniques de fouille de données, les données complexes ont besoin de mise en forme et de structuration. De plus anticiper, dès la phase de prétraitement, l'étape de fouille de données ainsi que la notion d'utilité des motifs extraits est également un thème visé par cet atelier.

Cette année l'atelier a porté en priorité sur une problématique transversale : la complexité liée aux données multiples (multi-sources, multi-vues, tableaux multiples, séquentielles, etc.). Dans ce contexte, la complexité peut concerner les processus (acquisition, structuration, extraction de connaissance) ou les données elles-mêmes (données manquantes, floues, incertaines, ...). Une liste de thèmes est donnée ci-dessous et reste ouverte et non limitative :

- Pré traitement, structuration et organisation des données complexes
- Processus et méthodes de fouille de données complexes
- Classification et fusion de données multi-sources et distribuées
- Les apports mutuels des méthodes de fouille de données et d'apprentissage et plus particulièrement les conditions qui justifient de faire appel aux unes et aux autres, les améliorations respectivement apportées.
- Retours d'expériences d'extraction de connaissances à partir de données complexes
- Rôle des connaissances en fouille de données complexes
- Fouille de données imprécises et/ou incertaines

2 Comité de programme

- Hanane Azzag (LIPN, Université Paris 13)
- Boutheina Ben Yaghlane (LARODEC, IHEC Carthage)
- Khalid Benabdeslem (LIESP, Université de Lyon 1)
- Omar Boussaïd (ERIC, Université de Lyon 2)
- Hend Bouziri (LARODEC-ISG, Tunisie)
- Martine Cadot (LORIA, Nancy)
- Guillaume Cleuziou (LIFO, Université d'Orléans)
- Sylvie Despres (LIM&BIO, Université Paris 13)
- Cyril De Runz (CReSTIC, LERI IUT de Reims-Châlons-Charleville)
- Gaël Harry Dias (University of Beira Interior, Portugal)
- Mounir Dhibi (ISSAT, Gafsa)
- Zied Elouedi (ISG, Université de Tunis)
- Rim Faiz (IHEC, Université de 7 Novembre de Carthage)
- Sami Faiz (INSAT, Université de 7 Novembre de Carthage)
- Pierre Gançarski (LSIIT-AFD, Université de Strasbourg)
- Lamia Hadrich (Laboratoire MIRACL, Université de Sfax)
- Mustapha Lebbah (LIPN, Université Paris 13)
- Eric Lefèvre (LGI2A, Université d'Artois)
- Arnaud Martin (ENSIETA, Brest)
- Florent Masséglia (AxIS-Inria Sophia Antipolis)
- Christophe Osswald (ENSIETA, Brest)
- Sébastien Régis (Université des Antilles et de Guyane)
- Brigitte Trousse (AxIS-Inria Sophia Antipolis)
- Cédric Wemmert (LSIIT-AFD, Univesité de Strasbourg)
- Djamel Zighed (ERIC, Université de Lyon 2)

3 Remerciements

Nous tenons à remercier les auteurs pour la qualité de leurs contributions, les membres du comité de programme et plus généralement tous les relecteurs de cet atelier pour le travail accompli et pour la qualité de leurs prestations.

Nous remercions également les responsables des ateliers pour EGC 2010, Amel Borgi, Tarek Hamrouni et Vasile-Marian Scuturici.

Enfin nous remercions vivement les présidents Jean-Marc Petit président du comité de programme et Sadok Ben Yahia, président du comité d'organisation d'EGC 2010.

4 Programme

8h45 Accueil

Session Fusion de données complexes

9h Fusion de segmentation et classification automatique d'images sonar
Julien Lengrand-Lambert, Arnaud Martin, Hicham Laanaya, Romain Courtis

9h30 Modélisation du conflit dans les bases de données évidentielles
Mouna Chebbah, Arnaud Martin, Boutheina Ben Yaghlane

10h Recalage et fusion d'images sonar multivues : utilisation du conflit
Cedric Rominger, Arnaud Martin

10h30-11h Pause

Session Classification non-supervisée

11h Approche graphique pour l'agrégation de classifications non supervisées
Fatma Hamdi, Haytham Elghazel, Khalid Benabdeslem

11h30 Etude d'opérateurs d'agrégation pour l'ordonnancement de clusters dans des images numériques de plantes
Jimmy Nagau, Sébastien Régis, Jean-Luc Henry

12h Visualisation de données spatiotemporelles imprécises : application en archéologie
Cyril de Runz, Frédéric Blanchard, Philippe Vautrot, Eric Desjardin

12h30-14h Repas

Session Visualisation des données

14h Etude de données multisources par simulation de capteurs et clustering collaboratif
Germain Forestier, Cédric Wemmert, Pierre Gançarski

14h30 Graphes multidimensionnels : Approche Coopérative
Lydia Boudjeloud-Assala, Hanane Azzag

15h Un langage et un générateur pour représenter les résumés visuels de bases de données géographiques
Ibtissem Cherni, Karla Lopez, Robert Laurini, Sami Faiz

15h30-16h Pause

Session Recherche et exploitation des données

16h Liaisons complexes entre variables : les repérer, les valider. Application à l'économie du mariage
Martine Cadot, Dhouha El Haj Ali

16h30 Les multi-sources dans un contexte d'Intelligence économique
Anass El Haddadi, Bernard Dousset, Ilham Berrada, Iloïse Loubier

17h Event Annotation based on Machine Learning
Aymen Elkhlifi, Rim Faiz

17h30 Réunion : Positionnement du groupe de travail Fouille de Données Complexes

17h45 Clôture

Summary

The workshop on mining complex data for extraction of knowledge is done at the incitement of the work group EGC "Complex data mining". Each year the organizers propose a topic that interest the researchers and companies. This year the seventh edition is focused on transverse field: complexity associated with multiple data (multi-source, multi-views, sequential, etc..).

Fusion de segmentation et classification automatique d'images sonar

Julien Lengrand-Lambert*, Arnaud Martin*
Hicham Laanaya**, Romain Courtis***

*ENSIETA, E^{3}I^2, EA3876, 2 rue François verny, 29806 Brest cedex 9
lengraju@ensieta.fr, Arnaud.Martin@ensieta.fr,
**HEUDIASYC, Université de Technologie de Compiègne (UTC),
Centre de recherche Royallieu, BP 20529 - 60205 Compiègne Cedex - France
hicham.laanaya@utc.fr
***GESMA/SDP/GDM, BP 42 - 29240 BREST ARMEES - France
Romain.Courtis@dga.defense.gouv.fr

Résumé. Cet article aborde la problématique de la classification et de la segmentation des images sonar du fond marin et de la fusion des deux approches afin d'en extraire de manière automatique des informations de classes et de frontières entre régions. Ce travail s'inscrit dans une démarche de cartographie automatique des fonds marins. La segmentation manuelle par expert est en effet une démarche coûteuse en temps et en argent et se traduit de plus par une variabilité des résultats. Deux experts n'auront en effet jamais exactement la même interprétation d'un fond marin. Ce travail s'appuie sur un précédent logiciel réalisé par Artigues et Billard (2005), qui permet de classifier automatiquement une image de fond marin à l'aide d'une base d'apprentissage. Afin de le compléter, nous nous sommes concentrés sur une approche par régions, puis avons cherché une méthode de fusion de ces deux approches. Ceci permet alors d'effectuer un premier bilan quand aux possibilités que pourraient offrir ces méthodes et l'utilité de continuer dans cette voie.

1 Introduction

La segmentation automatique des images sonar est un thème de recherche qui mobilise de nombreuses ressources à travers le monde aujourd'hui (*cf.* par exemple Leblond et al. (2005); Le Chenadec et Boucher (2005)). En effet, il n'existe aucune méthode générale qui traite automatiquement les données sonar et celles-ci sont de manière usuelle segmentées à la main par un expert du domaine. Ce traitement manuel, coûteux en temps et en moyens financiers, possède de plus une part d'incertitude car deux experts ne donneront jamais la même interprétation des images qu'ils ont traitées (voir figure 1). Enfin, il est important de noter que l'imagerie sonar est un domaine où s'expriment de fortes contraintes : hostilité et méconnaissance du lieu d'études, pertes importantes dues au domaine de propagation, et surtout changement constant de la typographie des fonds marins. Les données sonar sont ainsi complexes à traiter.

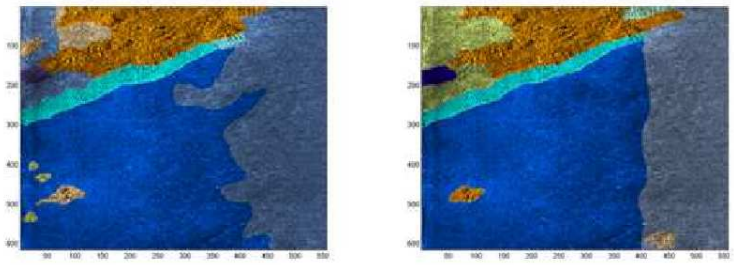


FIG. 1 – *Segmentation effectue par deux experts différents*

Devant l'intérêt croissant porté au monde sous-marin dans le monde industriel et militaire, à travers des applications telles que la chasse aux mines, on comprend aisément l'enjeu du développement d'un logiciel intégrant une chaîne complète de correction et segmentation des images sonar. Une première version présentée par Laanaya (2007), intègre une classification automatique des images fondée sur la caractérisation de la texture. Cependant, comme toute méthode automatique, elle est loin d'être parfaite et les erreurs de reconnaissance sont nombreuses. L'objectif de ce papier est donc de proposer une méthode de fusion de classification et de segmentation automatiques.

Il est important de distinguer clairement les deux méthodes qui ont été employées pour traiter le fond marin : la classification automatique, puis la segmentation automatique. Toutes deux ont la même image sonar en entrée mais diffèrent totalement dans leurs résultats, ce qui justifie une approche de fusion. La méthode de classification a pour but une division de l'espace de l'image en différentes classes que l'on choisit au départ. Ainsi, on utilise dans notre cas une classification supervisée. L'image de sortie est donc divisée en zones de types de fonds différents (sable, cailloux, roche, vase, ...). La segmentation quant à elle, utilise la même image, mais sans *a priori*. Le principe choisi pour cette segmentation repose sur une recherche des plages de niveaux. Cela signifie que les frontières obtenues en sortie correspondent à une forte variation de niveau dans l'image. Cette méthode ne présume absolument pas du type de fond de chaque classe de l'image et ne recherche que des frontières. C'est pour cela qu'on la nomme 'approche région'. Au vu de ces explications, on peut comprendre l'intérêt d'une fusion de ces deux résultats. Joindre la performance en frontière à une prévision des classes peut permettre d'obtenir l'objectif recherché.

Dans cet article, nous allons donc tout d'abord nous attacher à présenter rapidement les deux méthodes de segmentation et classification utilisées. Nous traitons ensuite la fusion, et la démarche suivie pour l'obtention des résultats, présentées dans une dernière section.

2 Classification des images sonar

Une première méthode de classification automatique a été mise au point par Laanaya (2007). Cette approche permet, à partir d'images sonar qui ont été au préalable débruitées, de fournir une image de sortie séparant l'image selon la classe de fond marin. L'utilisation d'une

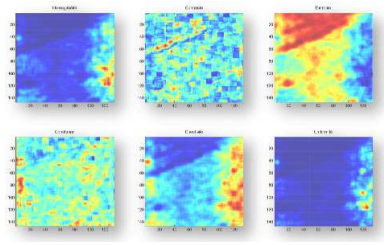


FIG. 2 – *Exemple de matrice des paramètres d'une image sonar*

classification supervisée peut s'expliquer par le fait que nous possédons des images d'expert segmentées manuellement et donc les différents types de sédiment cherchés.

Extraction de texture Les différents type de sédiments se caractérisent sur l'image sonar au travers de leur texture. Trois classes sont ici distinguées : rides, sable et roche. Pour représenter la texture, les images sonar sont divisées en petites imagettes pour lesquelles on calcule un vecteur de paramètres de texture. Différentes approches sont envisageables Martin et al. (2004), l'approche retenue ici se fonde sur les matrices de confusion proposée par Haralick (1979). Six paramètres de texture sont utilisées : homogénéité, contraste, entropie, corrélation, directivité et uniformité de l'énergie Russ (2002). Une image est donc représentée par six images de texture (*cf.* figure 2).

Classification d'images Laanaya (2007) a proposé différentes approches de classification supervisée pour les images sonar telles que les machines à vecteurs de support ou encore des approches plus simples telles que les k-plus proches voisins. Cette dernière approche montre des résultats déjà assez cohérents en elle-même. Les classes sont assez bien représentées pour peu que la base d'apprentissage soit fournie et que l'image ne soit pas de fond trop changeant. Cependant, elle manque de précision en raison du fait que l'on travaille sur des imagettes. On a donc des valeurs moyennes qui ne sont assignées qu'à un nombre plus faible de pixels ce qui conduit à une perte de résolution spatiale. De plus, une approche de classification ne permet pas de définir précisément les frontières entre les classes.

3 Segmentation automatique des images sonar

Afin de tenter d'améliorer ces résultats, nous avons appliqué une méthode originale de segmentation du fond marin dans Lengrand-Lambert (2009). Cette méthode est indépendante de la classification et se fonde sur une approche par régions. Cette dernière n'apporte donc pas d'informations sur le type de fond marin de l'image testée mais sur ses frontières. Cela devrait donc permettre d'être plus précis au niveau des frontières et avoir un gain d'informations face à la méthode de classification seule. Cette segmentation automatique a pour cœur la méthode des plages de niveaux proposée par Sumenger (2005), qui a été choisi dans l'objectif d'un travail préléminaire. L'algorithme de segmentation peut être divisé en trois étapes principales :

- **Paramétrage de la méthode :** Il s'agit ici de choisir les paramètres qui seront utilisés. C'est la seule interaction de l'opérateur dans la chaîne de traitement.
- **Prétraitement des données :** La première phase utilise les images crées à partir des imagettes, c'est-à-dire des images contenant les paramètres de Haralick calculés. Ces entrées sont modifiées afin de les rendre utilisables par la suite. On n'extraira notamment que certains des paramètres afin d'améliorer les résultats.
- **Segmentation :** La phase principale de la méthode prend en entrée les matrices de chaque paramètre de Haralick choisi par l'utilisateur et leur applique l'algorithme des plages de niveaux. Différents traitements seront encore effectués jusqu'à obtenir en sortie une matrice des régions de l'image sonar de base de même taille que l'image classifiée automatiquement.

Voici le détail des deux dernières étapes.

3.1 Prétraitement des données

Cette étape est importante car elle va conditionner la suite de l'approche. Si les données sont mal prétraitées, elles ne pourront en effet pas être de bonne qualité en sortie, et les résultats obtenus seront donc mauvais. Le prétraitement des données se décompose en plusieurs parties distinctes, qui sont appliquées de manière séquentielle.

Récupération des paramètres choisis Le but principal de la classification est de déterminer une nouvelle méthode qui permette de compléter l'existant à l'aide des mêmes données d'entrée. Cette partie consiste donc à examiner les données afin de les utiliser de la meilleure manière possible. L'implémentation de l'algorithme est assez ouverte pour permettre l'insertion de nouveaux paramètres par la suite ou encore d'insérer des méthodes de choix 'dynamiques'. Les constatations suivantes sont ainsi valables pour les images de notre base de données, et pourraient être approfondies dans une étude complémentaire. La première opération à effectuer est un choix des paramètres de Haralick à utiliser pour le reste de l'étude. Ce choix est resté le même lors de toute notre étude et est donc considéré comme appartenant à la phase de prétraitement. Il correspond à une phase de calibrage qui va dépendre de la base de données d'apprentissage. Tous les résultats obtenus ne sont pas de bonne qualité, et donc pas forcément utilisables. De manière assez générale sur nos images, nous observons que la corrélation et le contraste sont assez bruités (*cf.* figure 2). Ces deux paramètres n'ont donc pas été pris en compte pour la segmentation. De plus pour n'importe quelle image, l'uniformité et l'homogénéité donnent les mêmes informations en terme de segmentation, à ceci prêt que l'uniformité semble être atténuée par rapport à l'homogénéité. Afin d'optimiser le temps de calcul, ce paramètre n'est donc pas pris en compte. Les trois paramètres de Haralick choisis pour la segmentation sont donc : l'homogénéité, l'entropie et la directivité. On peut bien sûr tenir compte d'un grand nombre d'autres paramètres Martin et al. (2004).

Segmentation de chaque matrice On effectue une segmentation de Fisher sur chaque image des trois paramètres de texture retenus, afin de passer d'une matrice à niveaux continus à une matrice contenant 4 niveaux (*cf.* Russ (2002)). Cette segmentation permet de réduire de manière significative le nombre de plages de niveaux obtenues lors de la segmentation, et donc d'être plus précis dans les zones sélectionnées en sortie d'algorithme. La méthode de

Fisher est une méthode de segmentation qui se fonde sur un calcul à partir de l'histogramme d'une image. Quatre niveaux d'intensité sont discriminés avec pour critère de calcul celui de la minimisation de la somme des inerties de chacune des régions. Cette méthode de calcul recherche donc à partir de l'histogramme de départ et du nombre de régions à optimiser, la partition de l'algorithme en identifiant les séparateurs entre les régions.

Mise en forme des matrices Pour appliquer la suite de la segmentation, il convient de faire en sorte que toutes les matrices aient des contours de fortes amplitudes dans les mêmes zones. En effet, pour les paramètres choisis au départ ; les matrices obtenues ont des variations qui se situent globalement dans les mêmes zones. Cependant, les matrices d'homogénéité et contraste ont des niveaux bas là où la matrice d'entropie possède des niveaux hauts. Il faut donc effectuer une modification (une inversion) des trois matrices obtenues afin que toutes possèdent des **maxima** pour les mêmes zones. Ceci est indispensable pour la suite de l'algorithme, qui va rechercher à récupérer des contours précis ; en se fondant sur les zones de l'image de plus fortes valeurs.

3.2 Segmentation

Cette partie est la plus importante de notre démarche de segmentation automatique des images. C'est elle qui va permettre d'obtenir les résultats finaux, et qui encapsule les algorithmes de plages de niveaux. Chaque étape de cette partie est effectuée à la suite des autres de manière séquentielle.

Application de l'algorithme des plages de niveaux Les plages de niveaux sont une méthode de calcul utilisée dans les cas où l'on recherche à observer l'évolution de force au sein d'une image, et notamment la recherche des discontinuités. Pour un contour fermé donné, les plages de niveaux vont permettre la recherche de la zone de discontinuité minimale ; et donc tendre à se refermer sur ce contour comme le montre Sumenger (2005). En supposant que ce dernier délimite deux zones d'évolution différentes plusieurs forces peuvent alors exister : dans la direction normale à la courbe, venant d'un vecteur extérieur ou fondée sur le contour lui-même. Cela se traduit par l'équation différentielle suivante :

$$\frac{\partial f}{\partial t} + \overrightarrow{S}.div(f) + V_n.|div(f)| = V_n.|div(f)|$$

où le premier terme est la force de type vectorielle, le seconde terme la force de direction normale et la dernière celle qui correspond à une force f appliquée sur la courbe. L'algorithme présenté ici est appliqué uniquement dans la direction de la courbe ce qui signifie que les zones recherchées sont du type :

$$\frac{\partial f}{\partial t} = V_n.|div(f)|$$

Cela équivaut à rechercher les zones de plus fort gradient dans l'image.

La méthode a été très peu modifiée lors de l'étude : elle prend en entrée l'une des matrices de la cellule de départ, ainsi qu'un nombre d'itérations, qui va jouer fortement sur la qualité finale du rendu. En effet, plus le nombre d'itérations sera grand, plus les contours obtenus

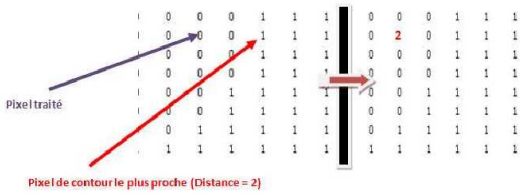


FIG. 3 – *Schéma explicatif de la création de matrice des distances gaussiennes*

seront lissés et fins. Il convient donc de modérer le nombre d'itérations choisis. Cette étape joue donc le rôle de filtre, et va permettre de discriminer rapidement les zones de fort changement dans l'image. Ce lissage sera ensuite utilisé pour obtenir les frontières recherchées.

Création de la matrice des distance gaussiennes Après passage de l'algorithme sur chacune des trois matrices, il convient de rechercher maintenant les contours lissés des images d'entrée. Une recherche de contours est appliquée puis les résultats sont discriminés pour ne garder que les contours les plus forts. Ces trois matrices sont ensuite fusionnées. La méthode utilisée est la création d'une matrice des distances gaussiennes (voir figure 3). Le principe est simple : il suffit de calculer pour chaque point d'une matrice la distance qui la sépare du point de contour le plus proche, pour obtenir une 'matrice des distances'. Une fois tous les pixels traités, il reste à créer la matrice 'gaussienne' de la précédente et à sommer les trois résultats. De cette manière, on peut observer des zones de recouvrement qui correspondent à des zones de corrélation des frontières pour les différents paramètres de Haralick.

Récupération de la matrice des régions Ceci est la dernière étape de la méthode de segmentation automatique. La matrice de distances gaussiennes contient en effet les frontières finales. Il ne reste maintenant plus qu'à mettre l'image sous la bonne forme pour pouvoir la réutiliser plus tard. C'est l'objectif de cette partie. Pour ce faire, nous passons d'une matrice des frontières à une matrice des régions. La seule barrière qu'il reste à franchir pour terminer cette segmentation est l'obtention de frontières exactes. En effet, les frontières sont encore larges de plusieurs pixels en raison de la création de la matrice des distances gaussiennes.

Cette méthode de segmentation est totalement décorrélée de la classification automatique et permet d'apporter de nouvelles informations sur l'image de fonds marins traitée. Au vu des résultats, les frontières obtenues en fin de segmentation sont de meilleure qualité que celles de la classification. Cependant, nous n'avons aucune information quand au type de fond des régions trouvées. Après vérification manuelle, les résultats obtenus sont de meilleure qualité pour les fonds rocheux et de sable, mais les zones de rides ont tendance à être occultée par l'algorithme.

Maintenant que nous parvenons d'une part à segmenter puis d'autre part à classifier une image sonar de manière automatique, la recherche d'une méthode de fusion qui permette de lier les résultats de ces deux méthodes paraît se justifier. C'est l'objet de la partie suivante.

4 Algorithme et méthode de la fusion

Cette partie présente une méthode dont le but est de fusionner les résultats obtenus jusqu'à maintenant afin d'avoir une classification automatique des images sonar. Celle-ci contiendrait à la fois les informations d'une approche 'classe' et d'une approche 'région'. Durant cette partie, les deux premières étapes de segmentation et classification sont déjà effectuées. L'objectif est alors d'obtenir une image prévisionnelle la plus proche possible de la réalité à partir des deux images obtenues précédemment. La démarche présentée est en deux étapes principales : récupération des contours finaux et détermination de la classe de chaque région obtenue.

4.1 Récupération des contours finaux

Il s'agit de la première étape de la fusion de données. Les deux matrices d'entrée sont la matrice des régions obtenue par la segmentation, ainsi que celle des régions de la classification. Ces régions et leurs frontières sont différentes, et le but de la fusion va être de les unifier du mieux possible. Une fois les frontières trouvées, on pourra alors rechercher les classes de fond de chaque région obtenue.

Une méthode similaire à celle de la segmentation automatique des trois paragraphes précédents est utilisée pour fusionner les frontières entre les deux images d'entrée. Ainsi, nous appliquons un seuillage à l'aide de l'algorithme de Fisher puis des plages de niveaux. Cette étape est suivie d'un passage de la matrice obtenue en matrice des distances gaussiennes. Cela permet donc d'avoir une matrice qui correspond à la somme gaussienne des frontières de chacune des deux matrices prises en entrée. Les frontières de cette image de somme gaussienne de sortie peuvent donc être récupérées en utilisant les zones de recouvrement. La frontière de deux régions sera donc déterminée par le maximum de recouvrement obtenue pour la matrice. Il ne reste plus alors qu'à passer d'une approche par contour à une approche par région.Chaque zone de l'image sera donc considérée en sortie comme une région. De cette manière, l'utilisation des outils déjà développés précédemment permet de sélectionner plusieurs zones de l'image qui représentent les différentes présences de classes.

4.2 Détermination de la classe de chaque région

L'objectif de cette partie est de réussir à attribuer une classe de terrain à chacune des régions de l'image qui vient d'être créée. Pour cela, les paramètres de Haralick vont encore une fois permettre de discriminer les différentes classes, en s'appuyant sur une classification supervisée. L'algorithme comprend deux étapes centrales, la recherche d'un vecteur de paramètres pour les régions de la matrice à classifier et la récupération de la classe de la région sélectionnée.

Recherche d'un vecteur de paramètres pour les régions de la matrice à classifier Chaque pixel d'une région étant représenté par un vecteur de paramètres de Haralick, il faut calculer le vecteur moyen des paramètres. Ce vecteur moyen est ensuite pris comme représentatif de la texture de la région homogène.

Classification des régions Pour classifier la région à partir du vecteur de paramètres, plusieurs approches sont envisageables. Nous présentons ici : *i* la méthode des k plus proches voi-

sins, qui cherche la proximité d'un élément avec un ensemble d'apprentissage, *ii* la méthode des prototypes, dérivée de la méthode précédente en moyennant les vecteurs d'apprentissage.

La première étape de test est effectuée avec une méthode simple, la méthode des k plus proches voisins (knn) (*cf.* Pasini et Grandgeorge (2003)). Il s'avère en fait que la méthode des knn a donné les résultats les plus précis et reproductibles. Ainsi, les k vecteurs les plus proches de chaque vecteur de région sont recherchées dans la base d'apprentissage. Le paramètre k est choisi entre 3 et 5 durant nos expériences. Les distances utilisées ici sont des distances euclidiennes qui conviennent bien dans ce type d'espace. À l'issue de cette recherche, il suffit enfin d'assigner la région à la classe majoritairement représentée dans l'échantillon. Nous supposons donc que l'échantillon choisi fait partie de l'ensemble le plus représenté autour de lui.

La méthode des prototypes est assez proche de la méthode des knn. À partir des vecteurs de la base d'apprentissage, les vecteurs sont moyennés pour obtenir en sortie un seul vecteur caractéristique par classe qui sera le 'prototype' de celle-ci. Notre base de connaissance se présente donc maintenant sous la forme de trois vecteurs, soit un par classe. Il reste alors à rechercher lequel de ces trois vecteurs est le plus proche de chacun des vecteurs de région et d'assigner la région à la classe correspondante.

Avant de présenter les résultats finaux, il reste à ajouter que le programme total est relativement rapide si l'on omet la création de la base d'apprentissage lors de la première utilisation. En effet, la totalité de l'algorithme de fusion est effectif en moins de 30 secondes sur une machine de bureau standard (2 Coeurs à 2.1 Ghz et 3 Gigas de mémoire) et pour une image au préalable débruitée de 600x3000 pixels. Cela permet d'imaginer dans le futur une approche opérationnelle de fusion en sortie directe de sonar. Les résultats obtenus pour chacune des deux approches de fusion, ainsi que les paramètres nous permettant de juger de leur qualité, sont présentés dans la section suivante.

5 Premiers résultats

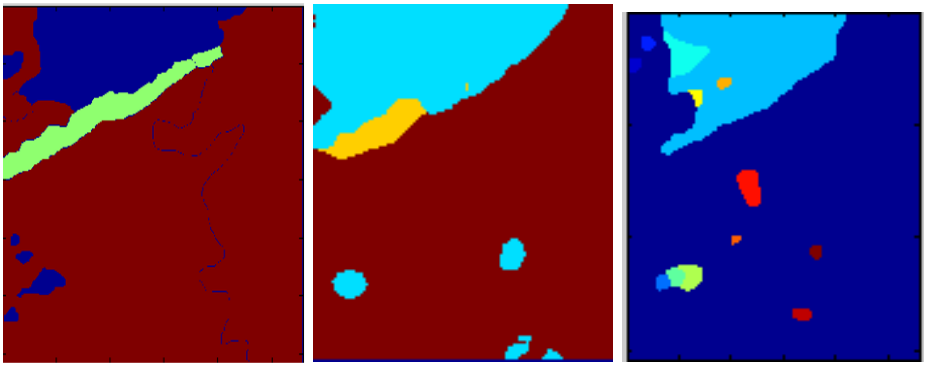
Dans cette dernière section, nous allons présenter tour à tour les résultats que nous avons pu obtenir avec les différentes méthodes. Le jugement de la qualité des résultats et leur exploitation s'appuiera sur le calcul des matrices de confusion comparant l'image de résultat avec l'image d'expert. Les 42 images sonar fournies par le GESMA (Groupe des Études Sous-Marines de l'Atlantique) proviennent d'une campagne de données effectuée au large des côtes finistériennes. Elles sont issues d'un sonar de type Klein 5400 et ont une résolution de 20 à 30 cm en azimut et 3 cm en range. La profondeur des fonds se situe entre 15 et 40 m.

L'utilisation de classifications supervisées lors de l'étude impose de scinder la base de données en deux parties : la base d'apprentissage qui permet d'obtenir les ressources nécessaires à appliquer les algorithmes et la base de test contenant les images sonar qui seront classifiées. La base d'apprentissage actuelle comporte 39 images sonar segmentées manuellement ce qui permet d'obtenir plus de 6700 imagettes. Les imagettes sont carrées et ont une dimension de 32 pixels de côté. De plus, le milieu naturel ne possède pas un équilibre parfait des différents types de sédiments qui se traduit par un nombre d'imagettes changeant en fonction des types de terrain. Afin de régler ce problème, la base contiendra autant de vecteurs pour chacune des trois classes de terrain à discriminer pour contenir finalement 2304 vecteurs de chaque classe (rides, sable et roche). Cette base devrait être suffisamment fournie pour obtenir de bons résul-

tats en sortie. Pour les images suivantes, nous utiliserons la couleur bleue pour la **classe roche**, la couleur verte pour la **classe ride** et la couleur rouge pour la **classe sable**.

5.1 Segmentation et Classification

La première étape de la phase de test passe par la validation de la classification et segmentation automatiques.



(a) Image segmentée par expert (b) Classification (c) Segmentation

FIG. 4 – *résultats obtenus pour la classification puis la segmentation automatiques*

Classification automatique Après application de la méthode de classification automatique détaillée, nous observons les résultats présentés sur la figure 4. Il existe une similitude apparente entre l'image segmentée par expert et l'image classifiée automatiquement. Les zones présentes concordent et les 3 classes sont bien présentes en sortie. La qualité de ces résultats se vérifie avec la matrice de confusion correspondante (*cf.* tableau 1). Les pourcentages de bonne détection obtenus dans le cas des classes roche et sable sont en effet très bons (supérieurs à 80%). Cependant, nous observons également que le taux d'erreur dans le cas de la classe ride est supérieur à 60%. La detection est donc peu fiable dans le cas de la détection de rides. Les résultats semblent cependant assez cohérents pour continuer notre démarche de fusion. Nous obtenons ainsi 84% de bonne détection pour la **classe roche**, 64% d'erreur pour la **classe ride** (la zone de ride est écourtée par rapport à l'image de base) et 82.9% de bonne détection pour la **classe sable**.

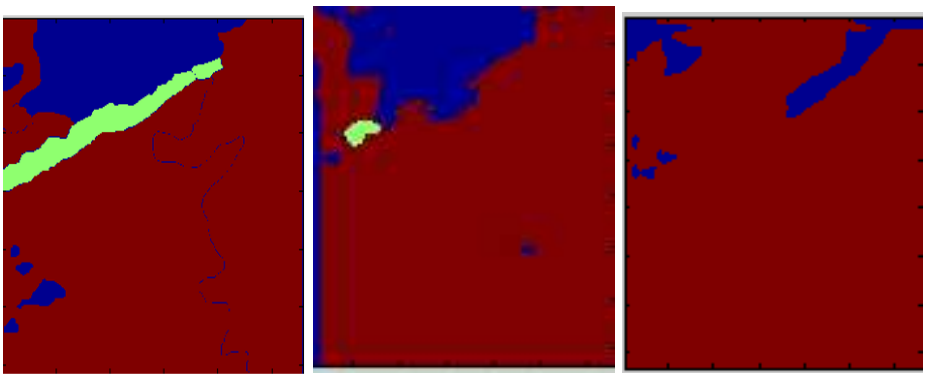
	ride	roche	sable
ride	1424	2243	298
cailloux	298	13336	2243
sable	414	11158	56075

TAB. 1 – *Matrice de confusion pour la classification automatique*

Segmentation automatique Avant d'effectuer cette fusion, il reste à bien vérifier que la segmentation automatique produit elle aussi des résultats convenables. Nous ne pourrons toutefois pas nous appuyer sur des résultats statistiques cette fois pour valider notre démarche. L'approche par régions ne permet en effet que de discriminer des ensembles et ne donne aucune information quand aux classes de types de fond. Une approche d'évaluation de la segmentation a cependant été proposée par Martin et al. (2006). L'image finale obtenue est représentée sur la figure 5. Chaque couleur sur l'image correspond une région différente. Les frontières apparaissent moins régulières que pour l'approche de classification. De plus, elles semblent se conformer à l'allure de la segmentation de l'expert. La démarche de fusion se justifie donc.

Dans l'idéal, la fusion de ces deux méthodes pourrait permettre de déterminer à la fois les types de fonds tout en gardant la précision des contours de l'approche région.

5.2 Fusion de la classification et segmentation



(a) Image segmentée par expert (b) k plus proches voisins (c) prototypes

FIG. 5 – *résultats obtenus pour la fusion par les deux méthodes utilisées*

Méthode des k plus proches voisins À l'issue de cette méthode, l'image obtenue est présentée sur la figure 5(b). L'allure générale de la matrice de fusion reste la même que la matrice d'expert. Les trois classes sont présentes et apparaissent globalement bien placées. Des différences significatives sont pourtant remarquables. La classe roche est très bien représentée, avec l'apparition de la zone bleue dans la partie supérieure de l'image. Cependant, les zones rocheuses dans la partie basse de l'image sont absentes dans les résultats. L'algorithme n'a pas permis de les détecter. De même, la zone de rides de sable est toujours présente en sortie, mais seule une petite partie apparaît. La perte d'informations est relativement élevée en ce qui concerne cette zone de rides, qui est remplacée par du sable simple (en rouge).

La matrice de confusion correspondant aux résultats permet de quantifier ces différences (*cf.* tableau 2). Nous obtenons ainsi 69.2% de bonne détection pour la **classe roche**, 62% d'erreur pour la **classe ride** (la zone de ride est écourtée par rapport à l'image de base) et 95.4% de bonne détection pour la **classe sable**.

	ride	roche	sable
ride	374	560	52
cailloux	4	2489	1101
sable	30	639	14023

TAB. 2 – *Matrice de confusion pour l'algorithme des knn*

Méthode des prototypes À l'issue de cette méthode, nous obtenons la segmentation présentée sur la figure 5. Cette méthode donne des résultats totalement incohérents. Cela est dû à la forte variation des paramètres de Haralick pour des images de même classe. Faire la moyenne de tous ces éléments implique une perte de l'information de dispersion des ensembles dans l'espace. De fait, les résultats obtenus sont totalement faux et inutilisables. Il suffit pour s'en convaincre de regarder la matrice de confusion correspondant à l'image (*cf.* figure 5).

	ride	roche	sable
ride	0	392	3202
cailloux	106	134	746
sable	220	332	14140

TAB. 3 – *Matrice de confusion pour l'algorithme des prototypes*

Nous obtenons ainsi 24.3% de bonne détection pour la **classe roche**, 100% d'erreur pour la **classe ride** (la zone de ride est écourtée par rapport à l'image de base) et 96.2% de bonne détection pour la **classe sable**.

Après visualisation des différents résultats, la méthode des k plus proches voisins semble la plus adaptée. En effet, les résultats sont cohérents et paraissent être assez proches de la réalité. La matrice de confusion obtenue par cette méthode est en effet de loin meilleure que celle de la méthode par prototypes. Enfin, si les classes roches et sable sont bien représentées, la classe ride est dans tous les cas fortement réduite par rapport à la réalité. Il serait opportun de concentrer les efforts de recherche sur cette partie dans un projet futur.

6 Conclusion et perspectives

Nous avons présenté d'une part une approche de segmentation des images sonar puis une méthode originale de fusion de la classification et de la segmentation d'images. Les résultats préliminaires montrent l'intérêt de l'approche. Les fonds marins peuvent donc ainsi être classifiés de façon autonome. De plus, cette méthode pourrait facilement être implémentée du fait de sa rapidité d'exécution. Même si ce travail est encourageant, la qualité des résultats obtenus reste modeste et est encore loin d'être utilisable directement. Nous pouvons cependant observer une bonne synergie entre les méthodes de classification et de segmentation, dont la fusion permet de garder un pourcentage de bonne détection acceptable tout en affinant les frontières.

Ce travail donne un aperçu d'une première méthode qui pourrait être réellement efficace, et ouvre la porte à une recherche plus approfondie dans le domaine. Nous pouvons ainsi ima-

giner l'implémentation d'approches plus complexes afin de combiner les résultats. La théorie des fonctions de croyance est par exemple un cadre envisageable. Enfin, il est important de garder à l'esprit que la présence d'un expert est toujours indispensable pour la phase d'apprentissage. Les experts indiquant leur degré de certitude, ce dernier pourrait être employé pour un meilleur apprentissage des approches supervisées. L'utilisation de méthodes de classification non-supervisées pourrait également permettre d'obtenir une segmentation de l'image.

Références

Artigues, S. et C. Billard (2005). Logiciel de segmentation et de classification automatique de sédiments marins. Technical report, Projet industriel, ENSIETA.

Haralick, R. (1979). Statistical and textural approaches to textures. *Proceedings of the IEEE 67*(5), 786–804.

Laanaya, H. (2007). *Classification en environnement Incertain : Application la Caractérisation de Sédiments Marins*. Ph. D. thesis, Université de Bretagne Occidentale, ENSIETA, Brest.

Le Chenadec, G. et J.-M. Boucher (2005). Sonar image segmentation using the angular dependence of backscattering distributions. In *IEEE Oceans'05 Europe*, Brest, France.

Leblond, I., M. Legris, et B. Solaiman (2005). Use of classification and segmentation of sidescan sonar images for long term registration. In *IEEE Oceans'05 Europe*, Brest, France.

Lengrand-Lambert, J. (2009). Fusion de la segmentation et de la classification automatique des images sonar. Technical report, Projet industriel, ENSIETA.

Martin, A., H. Laanaya, et A. Arnold-Bos (2006). Evaluation for uncertainty image classification and segmentation. *Pattern Recognition 39*(11), 1987–1995.

Martin, A., G. Sévellec, et I. Leblond (2004). Characteristics vs decision fusion for sea-bottom characterization. In *Journée d'Acoustique Sous-Marine*, Brest, France.

Pasini, S. et B. Grandgeorge (2003). Image segmentation. *Projet en Digital Photography - Image Segmentation 1*, 5–9.

Russ, J. C. (2002). *The image processing handbook*. Cleveland: CRC Press.

Sumenger, B. (2005). Level set method presentation. Technical report, Vision Research Lab, UCSB.

Summary

This issue handles the ability of processing sonar images in order to automatically draw a map of undersea borders and classes. This study takes part of the project of automatic undersea cartography. Indeed, the current use of experts to analyse those images is very long and costs money. In addition, two experts have almost always different points of view on the same image. This work is based on an older software capable of automatically classify images using a learning database. In order to enhance its results, we worked on a different way of segmentation using the 'level-sets method'. Then, we tried to fuse the two results. Finally, this issue allows us to draw a conclusion of such a method and to explore deeper in this way.

Modélisation du conflit dans les bases de données évidentielles

Mouna Chebbah*, Arnaud Martin**
Boutheina Ben Yaghlane***

*LARODEC, ISG Tunis, 41 Rue de la Liberté, Cité Bouchoucha 2000 Le Bardo, Tunisie
Mouna.Chebbah@gnet.tn
**E^3I^2, EA3876, ENSIETA, 2 rue François Verny, 29806 Brest Cedex 9
Arnaud.Martin@ensieta.fr
***LARODEC, IHEC Carthage, Carthage Présidence 2016, Tunisie
boutheina.yaghlane@ihec.rnu.tn

Résumé. La combinaison de différentes sources imparfaites fait inévitablement apparaître du conflit. Dans le cadre de la théorie des fonctions de croyance, la résolution du conflit se fait avant ou pendant l'étape de combinaison. Lors de la combinaison, plusieurs règles permettent d'éliminer le conflit en le redistribuant sur les informations disponibles de différentes manières. Par contre la gestion du conflit avant la combinaison revient à affaiblir les informations avant de les combiner ce qui nécessite une connaissance préalable des degrés de fiabilité des sources. Dans cet article, nous proposons une nouvelle méthode d'estimation de la fiabilité d'une source à partir de toutes les informations disponibles dans une base de données évidentielle. Les expérimentions sur des données radar réelles ont montré une amélioration remarquable des fiabilités des sources après affaiblissement.

1 Introduction

Les bases de données permettent de stocker une grande quantité d'information qui sont, la plupart du temps, incertaines ou imprécises. Pour aborder ce problème, des bases de données évidentielles ont été proposées par Hewawasam et al. (2005) et Bach Tobji et al. (2008). La fusion d'informations permet, d'une part, de réduire la quantité d'informations disponibles dans les bases de données évidentielles, et d'autre part, d'aider les utilisateurs qu'ils soient humains ou logiciels à la prise de décision en résumant les degrés de confiances en un seul facilement interprétable. Un degré de confiance associé à chaque information incertaine doit alors être défini en vu d'être stocké dans une base de données évidentielle.

Fusionner revient à combiner différentes informations, pouvant être imparfaites, en provenance de diverses sources. Pour ce faire la *théorie des fonctions de croyance*, introduite par Dempster (1967) et Shafer (1976), offre plusieurs avantages. En effet, cette dernière est utilisée pour sa robustesse en terme de représentation de données incertaines et de combinaison. La prise en considération de plusieurs sources hétérogènes lors de la combinaison peut induire l'apparition d'un conflit dû à une contradiction au niveau des informations fournies.

L'existence du conflit a favorisé l'apparition de plusieurs méthodes visant à le résoudre dans le cadre de la théorie des fonctions de croyance. Certaines méthodes proposent la résolution du conflit lors de la combinaison en proposant des règles adéquates. Ces règles permettent à la fois de combiner et de redistribuer le conflit de manières différentes sur les informations disponibles.

Une autre façon de gérer le conflit est de le réduire avant de combiner en affaiblissant les informations fournies par une source avec son degré de fiabilité. Cette méthode permet de différencier les informations fournies par une source fiable de celles fournies par une source de fiabilité moindre. Cette méthode nécessite une connaissance préalable des degrés de fiabilité des sources. Il est donc nécessaire de pouvoir estimer la fiabilité d'une source. Martin et al. (2008a) proposent une approche sans *a priori* fondée sur les informations fournies par les sources.

Dans cet article, nous proposons une méthode d'estimation de la fiabilité d'une source à partir de toutes les fonctions de masse qu'elle fournit et non pas à partir d'une seule fonction. Notre méthode est particulièrement applicable sur les bases de données évidentielles du moment où ces dernières permettent de stocker toutes les fonctions de masse fournies par une source. Nous proposons également d'améliorer le niveau de fiabilité des différentes sources.

Le reste de cet article est organisé comme suit : dans la deuxième section nous présentons brièvement les notions de base de la théorie des fonctions de croyance ensuite nous définissons les bases de données évidentielles dans la troisième section. La quatrième section présente notre méthode d'estimation du conflit et de la fiabilité d'une source et enfin nous présentons les résultats expérimentaux sur des données radar dans la cinquième et dernière section.

2 La théorie des fonctions de croyance

2.1 Formalisme

La théorie des fonctions de croyance, appelée aussi théorie de l'évidence initiée par Dempster (1967) et Shafer (1976), est un outil robuste pour la représentation des données imparfaites (imprécises et/ou incertaines). Nous présentons ici quelques concepts de base de cette théorie.

Soit $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ un ensemble fini de toutes les hypothèses possibles ω_i pour un problème donné. L'ensemble Ω représente *l'ensemble de discernement* ou *l'univers de discours* du problème en question.

Une fonction de masse est définie sur l'ensemble de tous les sous-ensembles possibles de Ω, noté 2^Ω et affecte à chaque sous-ensemble une valeur entre 0 et 1 représentant sa masse de croyance élémentaire. Une fonction de masse m est donc telle que :

$$m : 2^\Omega \mapsto [0, 1] \tag{1}$$

Un sous-ensemble ayant une masse de croyance non-nulle est *un élément focal*. Une fonction de masse doit satisfaire les conditions suivantes :

$$\sum_{X \subseteq \Omega} m(X) = 1 \tag{2}$$

On impose aussi en général $m(\emptyset) = 0$ qui permet de rester en monde fermé.

Une fonction de masse permet de représenter les connaissances incertaines et imprécises fournies par un expert (une source, un classifieur, ...). La masse affectée à un élément focal X représente le degré de croyance élémentaire d'une source à ce que la valeur réelle de l'attribut en question soit incluse ou égale à X, c'est donc son degré de croyance élémentaire en X.

La fonction de croyance (ou de crédibilité) *bel* représente le degré de croyance minimal affecté à un sous-ensemble de 2^Ω justifié par les informations disponibles. *bel(A)* mesure le degré auquel les informations données par une source ($B \subseteq A$) soutiennent A.

$$bel : 2^\Omega \rightarrow [0,1] \tag{3}$$

$$A \mapsto \sum_{B \subseteq A, B \neq \emptyset} m(B) \tag{4}$$

La fonction de plausibilité *pl* représente le degré de croyance alloué aux propositions non contradictoires avec A. C'est le degré de croyance maximal en A. La fonction de plausibilité est la fonction duale de la fonction de crédibilité.

$$\begin{array}{ll} \text{pl :} & 2^\Omega \rightarrow [0,1] \\ \text{A}\mapsto & pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \end{array} \tag{5}$$

2.2 La règle conjonctive normalisée

La théorie des fonctions de croyance représente un outil robuste de combinaison de différentes fonctions de masse définies sur le même ensemble de discernement et fournies par différentes sources. La combinaison permet de confronter différents avis d'experts afin d'en obtenir un seul en vue de la décision.

La règle conjonctive de combinaison, présentée dans le modèle de croyances transférables par Smets et Kennes (1994), permet de combiner deux fonctions de masse distinctes m_1^Ω et m_2^Ω. Elle est définie comme suit :

$$m_{1 \textcircled{\cap} 2}^\Omega(A) = (m_1^\Omega \textcircled{\cap} m_2^\Omega)(A) = \sum_{B \cap C = A} m_1^\Omega(B) \times m_2^\Omega(C) \tag{6}$$

Cette règle est non normalisée du moment qu'elle autorise l'affectation d'une masse nulle à l'ensemble vide après la combinaison. Elle est applicable sous l'hypothèse du monde ouvert dans lequel la masse attribuée à l'ensemble vide représente le degré de croyance en une hypothèse inconnue et non énumérée dans Ω. Une normalisation de cette règle est présentée par Dempster (1967).

La fonction de masse $m_{1 \oplus 2}^\Omega$ résultat de la combinaison de m_1^Ω et m_2^Ω par la règle de Dempster (1967) est obtenue comme suit :

$$m_{1 \oplus 2}^\Omega(A) = (m_1^\Omega \oplus m_2^\Omega)(A) = \begin{cases} \frac{m_{1 \textcircled{\cap} 2}^\Omega(A)}{1 - m_{1 \textcircled{\cap} 2}^\Omega(\emptyset)} & , \forall A \subseteq \Omega \text{ si } A \neq \emptyset \\ 0 & , \text{si } A = \emptyset \end{cases} \tag{7}$$

Cette règle est normalisée et elle est applicable avec l'hypothèse du monde fermé dans lequel on suppose que toutes les valeurs possibles que peuvent prendre un attribut sont énumérées dans l'ensemble de discernement.

Il existe toute une panoplie de règles de combinaison par exemple, la règle de combinaison de Yager (1987) et la règle de combinaison de Dubois et Prade (1988).

2.3 L'affaiblissement

La majorité des règles de combinaison ne font pas la distinction entre le conflit existant entre les sources et l'auto-conflit dû à la non-idempotence de la règle de combinaison utilisée. Une des origines du conflit est la non fiabilité d'au moins une des sources. La non fiabilité d'une source peut être réglée par l'affaiblissement des fonctions de masse avant la combinaison. Du moment qu'on arrive à quantifier les fiabilités α de chaque source, on peut affaiblir les fonctions de masse associées comme suit :

$$\begin{cases} m^{\alpha}(A) = \alpha\, m^{\Omega}(A), \text{ pour } A \subset \Omega \\ m^{\alpha}(\Omega) = (1-\alpha) + \alpha\, m^{\Omega}(\Omega) \end{cases} \tag{8}$$

avec α le degré de fiabilité de la source.

3 Les bases de données évidentielles

Une base de données évidentielle est une base de données qui contient des données parfaites, imparfaites ou même des données manquantes. L'imperfection (incertitude et/ou imprécision) dans les bases de données évidentielles est représentée grâce à la théorie des fonctions de croyance précédemment décrite.

Formellement, une base de données évidentielle est une base de données ayant X attributs (colonnes) et Y enregistrements (lignes), chaque attribut j ($1 \leq j \leq X$) possède un domaine D_j représentant toutes les valeurs que peut prendre cet attribut : *C'est son ensemble de discernement* tel que défini par Bach Tobji et al. (2008). Une base de données évidentielle doit contenir au moins un attribut évidentiel qui prendra des valeurs évidentielles décrite par une fonction de masse au lieu d'une valeur certaine et précise. Une valeur évidentielle V_{ij} de l'enregistrement i ($1 \leq i \leq Y$) pour l'attribut j ($1 \leq j \leq X$) est une fonction de masse m_{ij} telle que :

$$\begin{array}{l} m_{ij} : 2^{D_j} \rightarrow [0,1] \text{ avec :} \\ m_{ij}(\emptyset)\text{=0 et } \sum_{x \subseteq D_j} m_{ij}(\text{x})\text{=1} \end{array} \tag{9}$$

Les bases de données évidentielles sont utilisées dans différents domaines notamment pour le stockage des fonctions de masse de différents classifieurs présenté par Hewawasam et al. (2005).

4 Estimation du conflit et de la fiabilité d'une source

Une base de données évidentielle stocke différentes fonctions de masse fournies par une source. Avec la présence de plusieurs sources, experts ou classifieurs il y aura autant de bases de données évidentielles que de sources donc une quantité énorme de données à traiter. L'intégration des différentes bases de données en une seule permet de réduire la quantité d'information disponible afin de faciliter les requêtes sur la base et la prise de décision éventuelle. Lors

de l'intégration de différentes bases de données évidentielles, le principal problème rencontré réside dans l'intégration des valeurs évidentielles surtout quand elles sont conflictuelles.

Dans cet article, nous nous concentrons sur la fusion de plusieurs fonctions de masse stockées dans différentes bases de données. Nous proposons aussi d'enrichir ces bases de données par des informations sur le niveau de fiabilité des sources et de fiabilité des combinaisons. Ces fiabilités seront des indicateurs importants sur le degré de confiance des résultats des requêtes effectuées sur la base. L'utilisation d'une règle de combinaison paraît une solution simple permettant à la fois de combiner plusieurs fonctions de masse et de résoudre le conflit. Cependant, la résolution du conflit se fait alors par la redistribution de la masse affectée à l'ensemble vide sur les éléments focaux de différentes manières selon la règle utilisée sans prendre en considération les fiabilités des sources et le degré de véracité des fonctions de masse fournies.

Le degré de fiabilité de la source doit être pris en considération avant la combinaison pour prévenir le conflit au maximum. La difficulté réside dans la quantification de la fiabilité d'une source afin d'en tenir compte avant la combinaison en affaiblissant ses fonctions de masse. Dans une base de données évidentielle, plusieurs fonctions de masse relatives à une source y sont stockées, d'où la nécessité d'attribuer un degré de fiabilité global à la source qui prend en considération toutes les fonctions de masse qu'elle fournies.

4.1 La mesure du conflit

Martin et al. (2008b) considèrent que plus deux fonctions de masse sont éloignées plus les sources sont en conflit. Ainsi, une mesure de distance entre différentes fonctions de masse permet de quantifier le conflit entre leurs sources.

La distance de Jousselme et al. (2001) est utilisée dans cet article parce qu'elle permet de tenir compte des spécificités des fonctions de croyance. Elle est donnée par :

$$d(m_1, m_2) = \sqrt{\frac{1}{2}(m_1 - m_2)^t \underline{\underline{D}}(m_1 - m_2)} \tag{10}$$

avec :

$$D(A, B) = \begin{cases} \text{1, si A=B=}\emptyset \\ \frac{|A \cap B|}{|A \cup B|}, \forall A, B \in 2^{\Omega} \end{cases} \tag{11}$$

Le conflit entre deux sources S_1 et S_2 n'est autre que la distance séparant leurs fonctions de masse m_1 et m_2 :

$$Conf(S_1, S_2) = d(m_1, m_2) \tag{12}$$

Avec la présence de plusieurs sources, la mesure de conflit correspond à la distance entre la distribution de masse fournie par une source donnée et les autres distributions. Cette mesure de conflit peut être calculée de deux manières différentes. La première méthode consiste à calculer la moyenne des distances entre la fonction de masse de la source S_j en question avec les $s-1$ autres fonctions de masse ϵ :

$$Conf(j, \epsilon) = \frac{1}{s-1} \sum_{j=1, j' \neq j}^{s} Conf(j, j') \tag{13}$$

La seconde correspond au calcul de la distance entre la fonction de masse fournie par la source S_j et la fonction de masse m_{s-1} résultat de la combinaison des fonctions de masse des

$s-1$ différentes sources autre que la source en question :

$$Conf(j,s) = d(m_j, m_{s-1}) \tag{14}$$

4.2 Estimation de la fiabilité relative d'une source

L'affaiblissement peut être utilisé afin de prendre en considération la non-fiabilité relative d'une source avant la combinaison pour éliminer ou réduire le conflit qui pourra apparaître après. Nous faisons ici l'hypothèse que le conflit est issu de la non-fiabilité des sources.

Pour pouvoir affaiblir une fonction de masse, l'estimation de la fiabilité est nécessaire. La méthode d'estimation de fiabilité proposée par Martin et al. (2008b) est fondée sur la mesure du conflit. La *fiabilité relative* α_j d'une source S_j est une fonction décroissante de son conflit avec les autres sources telle que :

$$\alpha_j = (1 - Conf(j,s)^{\lambda})^{\frac{1}{\lambda}} \tag{15}$$

avec λ un réel quelconque.

4.3 Estimation de la fiabilité absolue d'une source

Pour s sources, l'estimation de la fiabilité relative revient à calculer pour chacune le conflit de sa fonction de masse par rapport aux autres qui servira à estimer sa fiabilité, cette fiabilité est utilisée pour affaiblir la fonction de masse correspondante avant la combinaison.

Dans une base de données évidentielle, différentes fonctions de masse fournies par la même source sont stockées. Il faut alors tenir compte de toutes les fonctions de masse pour estimer la fiabilité absolue de la source que nous introduisons ici. La fiabilité relative ne prend en compte qu'une seule fonction de masse alors que la *fiabilité absolue* reflète le niveau général de fiabilité de toutes les fonctions de masse fournies par une source.

Ces informations concernant les fiabilités relatives et la fiabilité absolue d'une source pourront servir à l'enrichissement de la base de données évidentielle en indiquant le niveau de fiabilité de chaque source.

À partir de s bases de données évidentielles concernant s sources, chacune ayant X attributs évidentiels et Y enregistrements, chaque base de données évidentielle stocke $X \times Y$ différentes fonctions de masse par source. À partir de chaque fonction de masse, l'estimation des fiabilités relatives pourra être faite ; il y aura donc $X \times Y$ degrés de fiabilités par source.

Dans cet article nous proposons de prendre la moyenne des différentes fiabilités relatives comme fiabilité absolue de la source. Le choix de la moyenne se justifie par le fait que la fiabilité d'une source est fixe, bien qu'elle puisse se tromper parfois donc sa fiabilité peut augmenter ou diminuer légèrement mais en moyenne elle garde le même niveau.

En effet, affaiblir avec une fiabilité relative minimale réduit la fonction de masse au maximum ce qui peut mener à l'ignorance totale. De plus, l'affaiblissement avec une fiabilité relative maximale ne permettra pas dans certains cas de réduire le conflit du moment où une source peut être au moins une fois complètement fiable alors qu'elle ne l'est pas en général. Affaiblir avec la fiabilité moyenne permet de réduire le conflit tout en gardant l'intégrité de la fonction de masse. Le choix de la fiabilité moyenne permettra ainsi d'éviter les valeurs aberrantes.

La *fiabilité absolue* α_j^a d'une source S_j est donc la moyenne de ses Y fiabilités relatives α_{yj} :

$$\alpha_j^a = \frac{1}{Y} \sum_{y=1}^{Y} (\alpha_{yj}) \tag{16}$$

4.4 Estimation de la fiabilité de la combinaison

Un degré de fiabilité relative α_j est attribué à chaque fonction de masse, qui pourra servir à estimer la fiabilité de la combinaison.

Si on a s sources fournissant chacune une distribution de masse différente, l'estimation de la fiabilité de chaque source se fera en calculant la distance entre sa distribution de masse et le reste des distributions ce qui représente son conflit relatif. La fiabilité relative est calculée à partir de ce degré de conflit relatif. La fiabilité de la combinaison est la moyenne des s fiabilités relatives propres aux différentes sources concernant les fonctions de masse à combiner pour une observation donnée. La fiabilité d'une combinaison représente le degré de confiance moyen attribué à la fonction de masse résultante.

$$\alpha_c = \frac{1}{s} \times \sum_{j=1}^{s} (\alpha_j) \tag{17}$$

Ces fiabilités associées aux combinaisons pourront enrichir la base de données évidentielle en indiquant à l'utilisateur à quel point il pourra faire confiance à la fonction de masse fournie. Ces fiabilités ne sont utiles que pour l'utilisateur lors de la prise de décision.

Les équations (16) et (17) sont différentes : la première équation permet d'obtenir la fiabilité moyenne d'une source à partir de toutes ses fiabilités relatives correspondant à ses fonctions de masse, tandis que la seconde équation permet d'obtenir la fiabilité de la combinaison à partir des fiabilités relatives de toutes les fonctions de masse des différentes sources combinées.

5 Expérimentation

Afin de pouvoir tester la méthode précédemment décrite, nous avons considéré une base de données radar. Ces données ont été recueillies dans la chambre anéchoïque de l'ENSIETA en plaçant une cible (maquette d'avion) et un capteur radar pouvant détecter la cible sous différents points angulaires. Le système d'acquisition est présenté par Martin et Radoi (2004). Une base de données a été proposée pour l'acquisition et le stockage des signaux par Toumi (2007). Nous considérons ainsi cinq cibles radar différentes (Mirage, F14, Rafale, Tornado, Harrier). Chaque table contient 250 représentations fréquentielles obtenues dans un domaine angulaire d'environ 60° et utilisant une bande de fréquence d'environ 6 GHZ. Pour caractériser les cibles, et donc renseigner la bases de données, nous avons utilisé trois classifieurs différents : le k-plus proche voisin flou, le k-plus proche voisin crédibiliste et les réseaux de neurones. Ces trois classifieurs sont considérés comme des sources, sur lesquelles on déduit des fonctions de masse tel que présenté par Martin et Radoi (2004). Ils ont donc fourni 250 fonctions de masse stockées dans trois tables différentes et permettant de classifier les cinq cibles radar différentes.

Notre but est d'intégrer ces trois tables en combinant les 250 fonctions de masse fournies par chaque source (classifieur) pour obtenir une seule table facilitant les requêtes sur la base et ainsi aider à la prise de décision.

La première étape consiste à estimer les conflits relatifs à chaque source pour chaque fonction de masse, donc chaque source aura 250 conflits relatifs. Le conflit absolu d'une source n'est autre que la moyenne de ses 250 conflits relatifs. Afin de calculer les conflits relatifs, nous avons utilisé deux types de méthode de calcul de distance :

- **Distance type1 :** correspond au calcul du conflit donné par l'équation (13), *i.e.* à la moyenne des distances séparant une fonction de masse fournie par une source et les autres fonctions de masse sans utiliser une règle de combinaison.
- **Distance type2 :** correspond au calcul du conflit donné par l'équation (14), *i.e.* à la distance séparant la fonction de masse fournie par une source et la fonction de masse combinée relative aux autres sources. Il existe plusieurs règles de combinaison pouvant être utilisées pour la combinaison des fonctions de masse telles que rappelées par Smets (2007) et Martin et Osswald (2007), mais dans cet article nous avons utilisé uniquement la moyenne des fonctions de masse et la règle de combinaison de Dempster donnée par l'équation (7). Cette dernière a un comportement conjonctif telle que la règle de Yager (1987) et la règle de Dubois et Prade (1988) mais distribue le conflit de façon différente.

Le conflit absolu initial propre à chaque source et la variance des conflits relatifs sont donnés dans le tableau 1.

Experts	**Type de distance**	**Règle de combinaison**	**Conflit initial**	**Variance**
K-ppv flou	Type1	-	0.19582	0.01370547
K-ppv flou	Type2	Règle de Dempster	0.1250316	0.02006961
K-ppv flou	Type2	Moyenne	0.147562	0.01127087
K-ppv crédibiliste	Type1	-	0.2232652	0.01772268
K-ppv crédibiliste	Type2	Règle de Dempster	0.0856512	0.02090321
K-ppv crédibiliste	Type2	Moyenne	0.2137072	0.01685384
Réseau de neurones	Type1	-	0.301354	0.03387129
Réseau de neurones	Type2	Règle de Dempster	0.3341664	0.04060376
Réseau de neurones	Type2	Moyenne	0.2922108	0.03484972

TAB. 1 – *Conflits absolus initiaux des sources et les variances associées*

Nous avons utilisé les conflits absolus initiaux pour affaiblir les fonctions de masse des différentes sources avec différentes valeurs de λ (le paramètre de calcul de la fiabilité à partir du conflit de l'équation (15)). Après affaiblissement, nous avons recalculé les conflits absolus pour chaque valeur de λ. Disposant des valeurs des conflits absolus initiaux (indépendantes de λ) et des valeurs des conflits absolus après affaiblissement en fonction de λ, les pourcentages de diminution des valeurs des conflits sont présentés dans la figure 1.

Nous remarquons que dans certains cas, le taux de diminution est négatif ce qui signifie que le conflit absolu après affaiblissement est supérieur au conflit absolu initial.

Le but de l'affaiblissement est ou bien de réduire le conflit absolu ou bien de garder le même niveau de conflit absolu en diminuant les conflits relatifs. La diminution de la variance des conflits relatifs indique une amélioration des conflits relatifs.

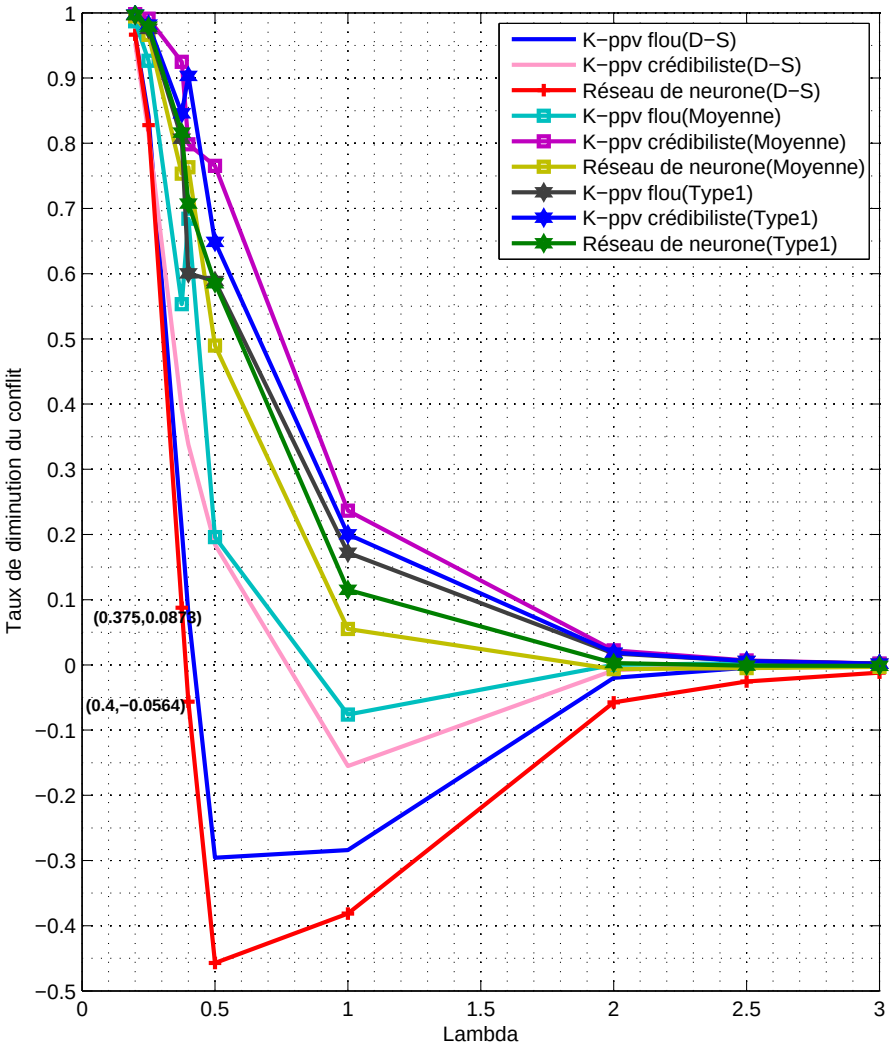


FIG. 1 – *Distributions du conflit avant et après affaiblissement en fonction de* λ

Le choix de λ est conditionné par un pourcentage de diminution positif. D'après la figure 1, pour $\lambda \geq 0.4$ le réseau de neurones a un taux de diminution négatif et pour $\lambda \leq 0.375$ son taux de diminution est positif, c'est pour cela que nous choisissons $\lambda \leq 0.375$ qui garantit un taux de diminution positif pour les deux types de distance.

Pour une valeur $0.375 \leq \lambda_0 \leq 0.4$, le pourcentage de diminution devient nul pour le réseau de neurones avec la distance type 2 en utilisant la règle de combinaison de Dempster et positif pour les autres classifieurs indépendamment de la mesure de distance. Il est difficile de déterminer λ_0 dû à la taille réduite de l'intervalle auquel elle appartient ainsi que l'impossibilité d'estimer la fonction liant le pourcentage de diminution à λ sur ces données réelles.

La fiabilité est une fonction croissante de λ (voir les distributions des fiabilités initiales en fonction de λ dans la figure 2), plus λ est petite plus l'affaiblissement est important d'où le choix de $\lambda = 0.375$ et non pas $\lambda < 0.375$. $\lambda = 0.375$ garantie un pourcentage de diminution positif pour les trois sources indépendamment du type de la distance et de la règle de combinaison tout en affaiblissant le moins possible.

Les conflits absolus propres à chaque source après affaiblissement avec un λ=0.375 sont donnés dans le tableau 2.

Experts	**Type de distance**	**Règle de combinaison**	**Conflit absolu**	**Variance**
K-ppv flou	Type1	-	0.0378216	0.000039045
K-ppv flou	Type2	Règle de Dempster	0.0990748	0.00032168
K-ppv flou	Type2	Moyenne	0.0659344	0.000052011
K-ppv crédibiliste	Type1	-	0.034408	0.000074342
K-ppv crédibiliste	Type2	Règle de Dempster	0.0520692	0.00072428
K-ppv crédibiliste	Type2	Moyenne	0.0160376	0.00020475
Réseau de neurones	Type1	-	0.0553784	0.00010062
Réseau de neurones	Type2	Règle de Dempster	0.304864	0.00044402
Réseau de neurones	Type2	Moyenne	0.072086	0.00012238

TAB. 2 – *Conflits absolus et variances après affaiblissment*

Nous remarquons bien l'amélioration au niveau des conflits après affaiblissement ainsi qu'au niveau des variances des conflits relatifs. Prenons l'exemple de l'utilisation de la distance type 2 avec la moyenne comme règle de combinaison, le conflit initial du réseau de neurones était de 0.29 avec une variance des conflits relatifs de 0.034. Après affaiblissement, le conflit absolu est passé à 0.072 et la variance des conflits relatifs est réduite à un taux de 0.00012.

D'après la figure 2, il est clair que pour $\lambda = 0.375$, les fiabilités après affaiblissement sont supérieures aux fiabilités initiales ce qui signifie une amélioration considérable des fiabilités. Nous remarquons également que les distributions de fiabilités après affaiblissement sont des fonctions croissantes qui tendent vers les distributions des fiabilités initiales. En augmentant la valeur de λ, les fiabilités initiales et les fiabilités après affaiblissement se rejoignent pour atteindre la valeur 1.

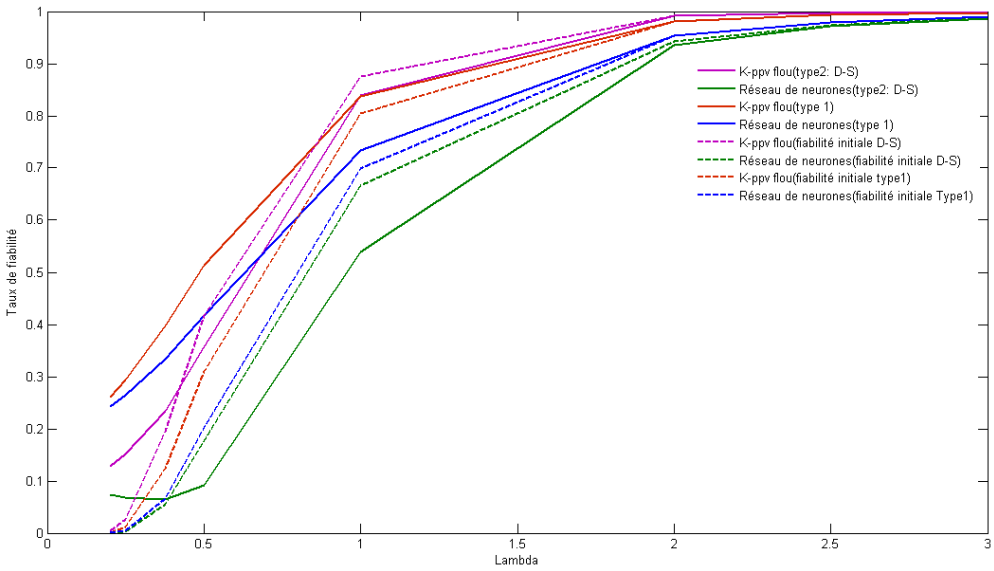


FIG. 2 – *Distributions des fiabilités avant et après affaiblissement en fonction de* λ

6 Conclusion

Dans cet article, nous avons proposé une méthode permettant d'estimer la fiabilité d'une source à partir de toutes les fonctions de masse qu'elle fournit. Cette méthode a pour objectif d'utiliser les fiabilités estimées pour affaiblir les fonctions de masse stockées dans une base de données évidentielle afin de fusionner ces fonctions avec d'autres stockées dans différentes bases de données évidentielles. Cette fusion permettra d'aider l'utilisateur à la prise de décision en réduisant la quantité d'information à traiter et en lui indiquant les degrés de fiabilité des sources et des informations combinées.

Comme perspective à ce travail, la proposition d'une méthode de transfert de masse autre que l'affaiblissement des masses permettrait de modifier l'ensemble des éléments focaux ce qui pourrait améliorer la qualité de classification des fonctions de masse, résultats de la combinaison.

Références

Bach Tobji, M.-A., B. Ben Yaghlane, et K. Mellouli (2008). A new algorithm for mining frequent itemsets from evidential databases. In *Information Processing and Management of Uncertainty*, Malaga, Spain, pp. 1535–1542.

Dempster, A. P. (1967). Upper and Lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics 38*, 325–339.

Dubois, D. et H. Prade (1988). Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence 4*, 244–264.

Hewawasam, K., K. Premaratne, S. Subasingha, et M.-L. Shyu (2005). Rule mining and classification in imperfect databases. In *International Conference on Information Fusion*, Philadelphia, USA, pp. 661–668.

Jousselme, A.-L., D. Grenier, et E. Bossé (2001). A new distance between two bodies of evidence. *Information Fusion 2*, 91–101.

Martin, A., A.-L. Jousselme, et C. Osswald (2008a). Conflict measure for the discounting operation on belief functions. In *International Conference on Information Fusion*, Cologne, Germany.

Martin, A. et C. Osswald (2007). Toward a combination rule to deal with partial conflict and specificity in belief functions theory. In *International Conference on Information Fusion*, Québec, Canada.

Martin, A., C. Osswald, J. Dezert, et F. Smarandache (2008b). General combination rules for qualitative and quantitative beliefs. *Journal of Advances in Information Fusion 3*(2), 67–82.

Martin, A. et E. Radoi (2004). Effective ATR Algorithms Using Information Fusion Models. In *International Conference on Information Fusion*, Stockholm, Sweden.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Smets, P. (2007). Analyzing the combination of conflicting belief functions. *Information Fusion 8*, 387–412.

Smets, P. et R. Kennes (1994). The Transferable Belief Model. *Artificial Intelligent 66*, 191–234.

Toumi, A. (2007). *Intégration des bases de connaissances dans les systèmes d'aide à la décision : Application à l'aide à la reconnaissance de cibles radar non-coopératives*. Ph. D. thesis, Université de Bretagne Occidentale, ENSIETA, Brest.

Yager, R. R. (1987). On the Dempster-Shafer Framework and New Combination Rules. *Informations Sciences 41*, 93–137.

Summary

The conflict appearing while combining several uncertain informations reflects the degree of conflict between their sources. This conflict can be managed before the combination step by discounting belief functions using sources' reliability. In this paper, we propose a generalization method for sources' reliability estimation taking into account all its belief functions stored in an evidential database. This method is evaluated on real radar data and supplied good results in terms of sources' reliability improvement.

Recalage et fusion d'images sonar multivues : utilisation du conflit

Cedric Rominger*, Arnaud Martin*

*ENSIETA E^{3}I^2 EA3876
2 rue François Verny 29806 Brest Cedex 9
{Cedric.Rominger,Arnaud.Martin}@ensieta.fr,
http://www.ensieta.fr/e3i2/index.php

Résumé. Ce papier présente une application pour le recalage et la fusion d'images sonar classifiées. Nous adaptons ici la méthode présentée dans un précédent papier à des données multivues. Pour la caractérisation de fond marin, nous avons besoin de fusionner des images sonar multivues afin d'améliorer les résultats. Néanmoins, avant de pouvoir fusionner ces images, il faut les recaler. Notre approche de recalage s'appuie sur une critère de dissimilarité calculé à partir du conflit issu de la combinaison des fonctions de croyance. L'utilisation de la théorie des fonctions de croyance offre un cadre théorique unique qui permet une bonne modélisation des imperfections, et qui a déjà prouvé son intérêt pour la fusion de classifieurs en traitement d'images.

1 Introduction

Le domaine de l'imagerie sous marine s'appuie principalement sur des données issues de capteurs acoustiques. Les sonars présentent l'intérêt de pouvoir imager les fonds marins à des distances bien plus importantes que les capteurs optiques (comme la vidéo), et de couvrir des surfaces à des cadences pouvant aller jusqu'a plusieurs miles nautiques carrés par jour.

Le traitement des images sonar présente plusieurs difficultés. En effet, les mouvements du sonar peuvent altérer la géométrie des objets qui reposent sur le fond marin. De plus le signal peut suivre plusieurs chemins, en fonction de la réflexion sur le fond ou d'autres objets, la faune ou la flore conduisant à des interférences sur l'intensité résultante. À ces incertitudes et imprécisions s'ajoutent un autre élément de complexité : la variation de l'information suivant l'angle de prise de vue. Cette multiplicité de l'information peut conduire à des informations conflictuelles.

Un aspect du traitement des images sonar est la caractérisation du fond marin. Comme le montrent Martin et al. (2006), la nature même des images rend cette caractérisation difficile, même pour des experts humains, qui par exemple pourront reconnaître le même sédiment, mais être en désaccord sur sa frontière. De plus, ils ne peuvent humainement pas traiter l'énorme masse de données disponible. Dhibi et al. (2008) et Williams (2009) ont montré que les techniques issues du domaine de la fusion peuvent apporter une réponse à ce problème en fusionnant des données provenant de plusieurs points de vues.

Cette caractérisation permet de produire des points de repère utiles pour la navigation sous marine. Quand un robot sous marin autonome (AUV) navigue, il peut déterminer sa position par différents instruments (comme une centrale inertielle) qui peuvent subir des dérives ou des imprécisions. L'utilisation de points de repère (ou amers) issus de la caractérisation du fond peut aider l'AUV à déterminer sa position.

La production de carte de fonds marin s'appuie sur des techniques de recalage d'images appliquées aux images sonar. Une fois déterminée la transformation la plus adaptée pour mettre en correspondance les deux images, celles-ci sont fusionnées pour en produire une plus fiable. En recalant toutes les images sonar d'une même campagne, une carte est ainsi produite. Une fois caractérisée, cette carte peut être utilisée par un AUV. Le processus de recalage d'images sonar peut être amélioré en utilisant des images préalablement classifiées (Leblond et al. (2005); Leblond (2006)), et la phase finale du processus, la génération de la mosaïque, peut alors être traité comme un simple problème de fusion.

Nous avons proposé l'utilisation de la théorie des fonctions de croyance pour la fusion, et le recalage d'image. Les fonctions de croyance permettent de prendre en compte l'incertitude et l'imprécision des images sonar. En associant à chaque pixel des images classifiées une fonction de croyance, nous définissons un critère de dissimilarité pour le recalage fondé sur le conflit issu de la combinaison conjonctive (Rominger et al. (2009)). De plus la génération de la mosaïque peut s'appuyer sur la combinaison des fonctions de masse, directement à l'aide des résultats obtenus pendant le calcul du critère de dissimilarité.

Dans la suite de ce papier, nous commençons par présenter le principe du recalage d'images. Puis, nous introduisons les éléments de la théorie des fonctions de croyance dont nous avons besoin. La section 4 présente le processus de recalage proposé. Enfin, nous présentons les résultats obtenus sur des images sonar multivues.

2 Recalage d'images

Le recalage d'images est un processus visant à déterminer quelle est la meilleure transformation qui permet d'aligner deux prises de vues. La figure 1 illustre la problématique pour deux images I_1 et I_2 de taille et angle différents, où l'on cherche à recaler l'image I_2 sur l'image I_1 prise comme image de référence. Le problème étant symétrique pour deux images ce choix importe peu *a priori*. Classiquement, comme présenté par Zitova et Flusser (2003), les méthodes de recalage sont organisées en deux familles :

- Les méthodes géométriques sont fondées sur une extraction d'un ensemble de caractéristiques des images (points, contours, formes), et leur mise en correspondance pour déterminer la meilleure transformation à appliquer.
- Les méthodes iconiques prennent en compte l'intégralité des pixels des images, et comparent directement leur intensité, ou une fonction de cette intensité.

En environnements naturels et incertains, il existe peu de formes géométriques simples et comparables d'une image à l'autre. Ces images peuvent de plus être fortement déformées selon l'angle de vue par exemple avec l'apparition d'ombre. Ainsi dans cet article notre choix s'est porté sur une méthode iconique.

Les méthodes de recalage cherchent à déterminer la meilleure transformation au sens d'un critère de similarité. Cette transformation appartient à un ensemble de transformations caractérisé par différents types que Maintz et Viergever (1998) énumèrent :

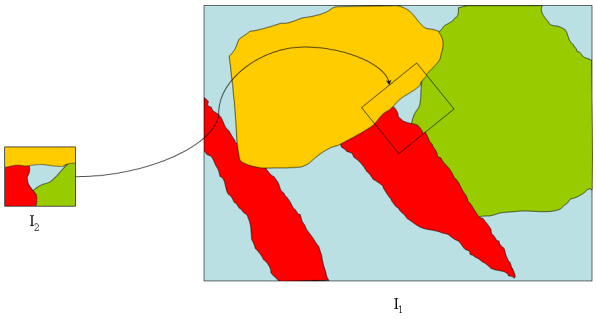


FIG. 1 – *Technique du recalage, image I_2 recalée sur l'image I_1.*

– Rigide : uniquement translation et rotation.
– Affine : transforme des lignes parallèles en lignes parallèles.
– Projective : met en correspondance des lignes non parallèles.
– Élastique : transforme des lignes droites en courbes.

De plus, les modèles de transformation peuvent être appliqués à tout ou partie de l'image. On parle alors respectivement de modèles globaux ou locaux.

Par nature, le meilleur moyen de mettre en correspondance deux images sonar est de s'appuyer sur un modèle de transformation projectif. Les modèles élastiques restent les plus intéressants puisqu'ils permettent de recaler le plus finement possible deux images. Néanmoins, traiter de tels modèles conduit à devoir traiter un nombre très important de paramètres de transformation. Dans le but de réduire cette complexité, nous nous intéressons aux transformations rigides locales que nous considérons comme une approximation d'un modèle de transformation élastique global.

Dans le cadre des méthodes de recalage iconique, nous cherchons à mesurer la similarité (ou dissimilarité) s entre l'intensité des pixels d'une image I_1 et d'une image ayant subit une transformation $t(I_2)$. Suivant la nature de ce lien, il existe différentes mesures qui sont plus ou moins bien adaptées. Une classification de ces mesures est d'ailleurs suggérée par Roche (2001). Le choix du critère de similarité dépend de la relation supposée entre les intensités de pixels des images.

L'idée la plus simple est que l'intensité est stationnaire d'une image à l'autre. On parle alors de relation *identité*, et on peut utiliser une mesure de corrélation (corrélation croisée, somme absolue des différences des intensités, variance des différences, etc, telles que présentées par Chambon (2005)). L'utilisation de ces mesures permet de déterminer rapidement la meilleure transformation, mais elles sont sensibles aux valeurs aberrantes.

Dans la pratique la relation identité est rarement vérifiée, car la valeur des intensités dépend de l'instrument de mesure. Il convient donc de prendre en compte une remise à l'échelle des intensités grâce à une relation affine de type $j = \alpha i + \beta$. Hill et Hawkes (2000) montrent que le coefficient de corrélation affine permet de prendre en compte cette relation.

Bien que donnant généralement de bons résultats dans le cadre du recalage monomodal, l'hypothèse de relation affine n'est plus adaptée dès qu'il s'agit de recalage multimodale. On généralise alors à une relation fonctionnelle de type $j = f(i)$, avec l'hypothèse que pour une

intensité d'une image donnée on peut associer une unique intensité de l'autre image. On trouve dans cette catégorie de mesures le critère de Woods et al. (1993) et le rapport de corrélation listés par Roche (2001).

Il est possible de considérer une relation moins restrictive que la relation fonctionnelle en considérant les images comme des réalisations de variables aléatoires dont on cherche à caractériser la dépendance. L'outil alors utilisé est l'histogramme conjoint. Et les critères tels que l'information mutuelle utilisée par Chailloux (2007) sur des images sonar, permettent de mesurer la dispersion de l'histogramme conjoint, avec l'idée que plus la dispersion est faible, plus forte est la dépendance entre les deux images.

Dans le cas d'images classifiées, l'intensité de chaque pixel ne représente plus une mesure physique, mais son appartenance à une classe. Il faut donc chercher à modéliser la relation entre ces appartenances. De plus en environnements incertains, ces appartenances ne sont pas strictes. Il est donc raisonnable d'envisager une telle relation, dont les précédents critères de similarité ne tiennent pas compte.

L'étape de décision consiste donc à déterminer la transformation t à appliquer à I_2, parmi l'ensemble des transformations T considérées, donnant la dissimilarité d la plus faible (ou la similarité d la plus forte). Nous décidons donc de la transformation t_d donnée par :

$$t_d = \underset{t \in T}{\operatorname{argmin}}\, d(I_1, t(I_2)) \tag{1}$$

ou

$$t_d = \underset{t \in T}{\operatorname{argmax}}\, s(I_1, t(I_2)) \tag{2}$$

Dans la suite, nous allons voir comment le critère de dissimilarité peut s'appuyer sur la théorie des fonctions de croyance.

3 Fonctions de croyance

La théorie des fonctions de croyance est issue des travaux de A. Dempster (1967), et du formalisme de G. Shafer (1976) sous le nom de *theory of evidence*. Les fonctions de croyance ont trouvé de nombreuses applications en traitement d'images (Bloch et Maitre (1994); Vannoorenberghe et al. (2003)), telles que la segmentation d'images (Taleb-Ahmed et al. (2002); Vannoorenberghe et al. (1999)), ou la fusion de classifieurs sur des images (Milisavljevic et al. (2003); Martin (2005); Dhibi et al. (2008)). Dans ces deux dernières applications, les images sonar sont supposées déjà recalées. Si ce n'est le cas, il peut être intéressant de chercher à les recaler et les fusionner en conservant le même formalisme de la théorie des fonctions de croyance que nous décrivons ci-dessous.

La théorie des fonctions de croyance repose sur la manipulation de fonctions de masse. À la différence des probabilités, les fonctions de masse sont définies sur l'ensemble de toutes les disjonctions possibles des classes C_k noté $2^\Theta = \{\emptyset, \{C_1\}, \{C_2\}, \{C_1 \cup C_2\}, \ldots, \Theta\}$. Une fonction de masse m est donc définie sur 2^Θ à valeurs dans $[0, 1]$, et vérifie la propriété de normalisation suivante :

$$\sum_{A \in 2^\Theta} m(A) = 1. \tag{3}$$

La fonction de masse modélise le degré de croyance élémentaire que l'on accorde à chaque proposition A de 2^Θ. Ce degré croyance est indépendant de la croyance que l'on peut accorder aux éventuels sous-ensembles et sur-ensembles de A.

Lorsqu'on suppose l'exhaustivité des classes de notre cadre de discernement, on se place en monde fermé (*i.e.* $m(\emptyset) = 0$). Par opposition dans l'hypothèse du monde ouvert de Smets et Kennes (1994) (que nous ferons ici), nous admettons que l'on puisse avoir $m(\emptyset) > 0$.

Une fois définies les fonctions de masse pour chaque classifieur S_i, différents opérateurs de combinaison sont envisageables. La règle conjonctive non normalisée proposée par Smets (1990) est définie pour deux fonctions de masse m_1 et m_2 et pour tout $A \in 2^\Theta$ par :

$$m_{\text{Conj}}(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \tag{4}$$

Cet opérateur est associatif et commutatif, mais non-idempotent. La masse affectée sur l'ensemble vide $m_{\text{Conj}}(\emptyset)$ s'interprète généralement comme une mesure de conflit. Si cette masse représente dans une certaine mesure le conflit entre les classifieurs, une partie provient également de la non-idempotence. Ce conflit peut résulter d'un manque d'exhaustivité des sources, d'un manque de fiabilité des classifieurs, ou encore lorsque les classifieurs ne représentent pas la même chose comme le souligne Appriou (2002). Dans le premier cas, le fait d'avoir une masse non nulle sur l'ensemble vide (cas d'un monde ouvert) est concevable. Si l'on souhaite rester en monde fermé, de nombreuses règles de combinaison répartissant le conflit ont été proposées dans la littérature. Initialement, Dempster, repris par Shafer, a proposé une règle orthogonale normalisée, répartissant le conflit de manière uniforme, donnée pour tout $X \in 2^\Theta$, $X \neq \emptyset$ par :

$$m_{\text{DS}}(X) = \frac{m_{\text{Conj}}(X)}{1 - m_{\text{Conj}}(\emptyset)}. \tag{5}$$

La dernière étape de la fusion de classifieurs concerne la décision de la classe C_k sur l'image ou partie de l'image observée. Il est mal aisé de réaliser cette décision directement sur les fonctions de masse, ainsi plusieurs fonctions de croyance (telles que plausibilité ou crédibilité, qui peuvent être vues comme une croyance supérieure ou inférieure), ou probabilités (telle que la probabilité pignistique) ont été définies.

Dans le cadre de notre application, nous nous appuyons sur la probabilité pignistique pour décider de la classe du pixel issue de la fusion. La probabilité pignistique est définie pour tout $X \in \Theta$ avec $X \neq \emptyset$ par :

$$betP(X) = \sum_{Y \in 2^\Theta, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m(Y)}{1 - m(\emptyset)} \tag{6}$$

où $|X|$ est le cardinal de X.

La classe du pixel x_i décidée par la fusion est alors déterminée par :

$$cl_{x_i} = max_{X \in \Theta}(BetP(X)) \tag{7}$$

4 Recalage d'images à partir des fonctions de croyance

Nous nous plaçons dans ce papier dans les cas où nous possédons des images classifiées I_i. Ces images sont donc les sorties des classifieurs, notés S_i, et sont composées pour chaque

pixel, d'information symbolique correspondant au type de classe. Les approches fondées sur des opérations arithmétiques de ces images symboliques, telles que proposées par Olteanu (2007), sont donc peu envisageables.

Dans le but d'approcher un modèle de transformations élastiques global, nous traitons plusieurs transformations rigides localement. Pour ce faire, nous déterminons un ensemble de zone d'intérêt, dans lesquelles nous savons qu'il y a une correction à déterminer, et dont nous extrayons des imagettes I_i^j que nous recalons au travers d'un modèle de transformation rigide global (au sens des imagettes) noté T.

Les images à recaler ont été classifiées et donc implicitement segmentées par des classifieurs possédant le même cadre de discernement Θ selon n classes C_k. De fait, chaque pixel de chacune des deux images (celle de référence et celle à recaler) appartient à une (unique) classe. L'ensemble Θ des n classes C_k est donc le cadre de discernement de nos images I_i (et de nos imagettes I_i^j).

Chaque pixel x_i de chaque image I_i (images qui peuvent être de taille différente) ayant déjà été affecté à une classe C_k par un classifieur, nous pouvons utiliser une fonction à support simple pour définir les masses élémentaires de chaque pixel. Pour une image I_j composée des pixels x_i, une fonction de masse à support simple pour chaque pixel x_i est définie ainsi :

$$\begin{cases} m_{x_i}(C_k) = \alpha_{ik} \text{ si } x_i \in C_k \\ m_{x_i}(\Theta) = 1 - \alpha_{ik} \\ m_{x_i}(A) = 0 \text{ si } A \in 2^\Theta \setminus \{C_k, \Theta\} \end{cases} \tag{8}$$

où α_{ik} est la fiabilité du classifieur ayant produit I_i pour la classe C_k. Il a été proposé par Martin (2005) d'estimer cette fiabilité à partir du taux de bonne classification par classe. Lorsque les performances du classifieur sont sensiblement les mêmes pour chaque classe, la fiabilité peut être donnée par α_i, le taux de bonne classification global du classifieur.

Dans le cas de deux images, nous cherchons donc à combiner l'image de référence I_1^j avec le résultat de la transformation $t(I_2^j)$ de l'image à recaler I_2^j, pour une transformation donnée $t \in T$. Cependant, rien ne nous garantit que le résultat $t(I_2^j)$ et I_1^j représente la même chose, puisque justement nous cherchons également cette transformation. Nous sommes alors dans le cas où les sources à combiner ne représentent pas la même chose, et il ne faut donc pas les combiner, ou en l'occurrence recaler $t(I_2^j)$ sur I_1^j. Le problème du recalage par les fonctions de croyance se résume alors par le choix de la transformation t la plus crédible pour la combinaison de I_1^j et $t(I_2^j)$.

La théorie des fonctions de croyance permet de gérer le conflit entre deux sources d'information. Suivant sa nature (qui peut être multiple), différentes réponses peuvent être apportées pour minimiser son impact sur la masse combinée **?**. Comme nous utilisons des fonctions de masse à support simple, nous sommes certain que le conflit est une mesure du désaccord entre deux sources, ou dans notre cas de la classe à laquelle a été affectée un pixel. L'utilisation de fonctions de masses plus complexes nous obligerait à étudier plus précisément ce conflit pour déterminer la part de désaccord qu'il contient.

Le conflit qui apparaît si l'on combine $t(I_2^j)$ avec I_1^j à tort est donc une bonne mesure de dissimilarité de nos deux images I_1^j et $t(I_2^j)$. Nous considérons que cette mesure de conflit est directement donnée par la masse $m_{\text{Conj}}(\emptyset)$ transférée sur l'ensemble vide lors de la combinaison conjonctive non normalisée de l'équation (4).

Formellement, nous cherchons donc à combiner la fonction de masse m_{x_1} associée à un pixel x_1 de l'image I_1^j avec la fonction de masse $m_{t(x_2)}$ associée à $t(x_2)$ où x_2 est un pixel de l'image I_2^j, telle que $t(x_2) = x_1$. Le conflit associé à la combinaison des fonctions de masse de ces deux pixels est donc nul si $C_{x_1} = C_{x_2}$ et sinon est donné par :

$$m_{(x_1,t(x_2))}(\emptyset) = m_{x_1}(C_{x_1})m_{x_2}(C_{x_2}), \tag{9}$$

où C_{x_i} est la classe du pixel x_i et avec $x_1 = t(x_2)$, $x_1 \in I_1^j$, $x_2 \in I_2^j$. Le conflit associé à la transformation $t \in T$ est alors donné par :

$$m_t(\emptyset) = \sum_{x_1 \in I_1} m_{(x_1,t(x_2))}(\emptyset). \tag{10}$$

Dans le cas des transformations rigides, il y a bijectivité des fonctions t et l'équation précédente s'écrit aussi :

$$m_t(\emptyset) = \sum_{x_2 \in I_2} m_{(x_1,t(x_2))}(\emptyset). \tag{11}$$

Nous avons donc intérêt à considérer l'image de taille la plus petite.

Enfin pour décider de la transformation t la plus crédible dans l'ensemble T, il suffit de considérer la transformation minimisant ce conflit $m_t(\emptyset)$. Nous choisissons donc, après une recherche exhaustive sur T, la transformation t_r^j donnée par :

$$t_r^j = \underset{t \in T}{\operatorname{argmin}}\, m_t(\emptyset). \tag{12}$$

L'étape finale de notre application est la génération d'une mosaïque par fusion des différentes imagettes I_i^j au travers des transformations optimales t_r^j. La fusion de chaque paire d'imagettes I_1^j et $t_r^j(I_2^j)$ a déjà été réalisée lors du calcul de la valeur du critère de dissimilarité pour cette transformation. Nous pouvons donc reprendre immédiatement les fonctions de masse générées pour les futurs pixels de la mosaïque. La classe de chacun de ces pixels est ainsi déterminée en prenant le maximum de la probabilité pignistique (équations (6) et (7)).

5 Application au recalage d'images sonar segmentées

Nous travaillons à partir de données sonar Klein 5500B acquises sur la « Grande Vaille » (83) par la société SEMANTIC-TS et le GESMA (Groupe d'Études Sous-Marines de l'Atlantique) dans le cadre du contrat REI (Recherche Exploratoire et Innovation) n° 05.34.011.00.470.75.65 mis en place par la DGA/D4S/MRIS et intitulé "cartographie de la couverture du fond marin par fusion multi-capteurs".

Nous disposons de sept passes sonar mesurant jusqu'à 800m de long pour 130m de large, avec une résolution finale de 10cm sur les images. Ces relevés portent sur une zone côtière peu profonde (environ 15m).

Les passes sonar sont caractérisées automatiquement par un classifieur de type k-plus proches voisins (k-ppv) crédibiliste développé par Denoeux (1995). Un exemple de classification est présenté par la figure 2. Les images sont classifiées selon quatre classes : sable (jaune pale), rides de sable (jaune foncé), roche/posidonie (vert), et vase (gris).

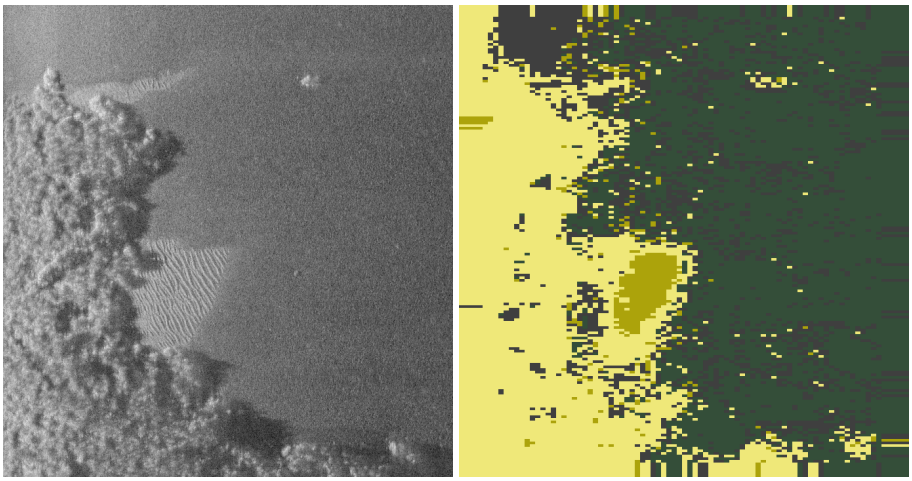


FIG. 2 – *Un exemple de trace sonar brute, et classifiée (extrait)*

Le classifieur a déjà été testé, et confronté à une classification par un expert humain, nous permettant de déterminer ses taux de bonne classification par classe. C'est ces taux que nous utilisons dans la définition des fonctions de masse des pixels par l'équation (8).

Associés aux relevés sonar, nous possédons les informations de géoréférencement du bateau tractant le sonar. Ces données sont moins bruitées que dans le cas d'une application avec un robot sous-marin, où ces données seraient issues de la centrale inertielle du robot.

Nous pouvons donc réaliser une projection de nos relevés sonar classifiées (voir figure 3). Néanmoins, ces informations ne prennent pas en compte les mouvements du sonar lui même, et même si nous pouvons mettre en correspondance deux traces, il reste encore quelques erreurs de recalage que nous cherchons à corriger.

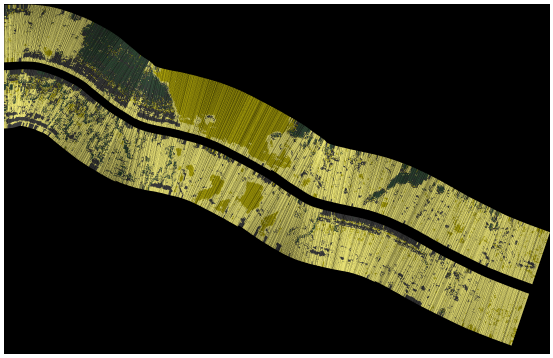


FIG. 3 – *Une trace sonar projetée selon ses données de géoréférencement.*

En s'appuyant sur les données de géoréférencement, nous pouvons mettre en correspon-

dance deux traces sonar, et les fusionner. Le résultat de cette fusion génère une certaine quantité de conflit. Les endroits où le conflit est maximum indiquent là où nous devons chercher à corriger des erreurs. Nous déterminons donc les zones d'intérêt, avec une taille fixée, sur nos deux traces qui englobent le maximum de conflit généré.

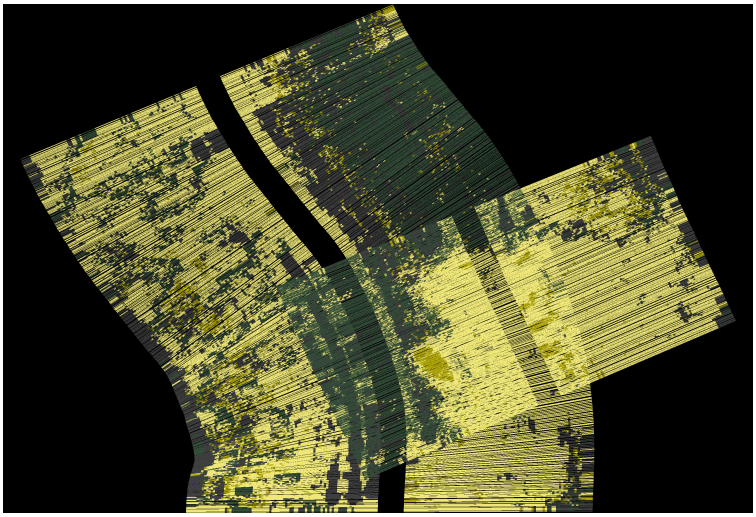


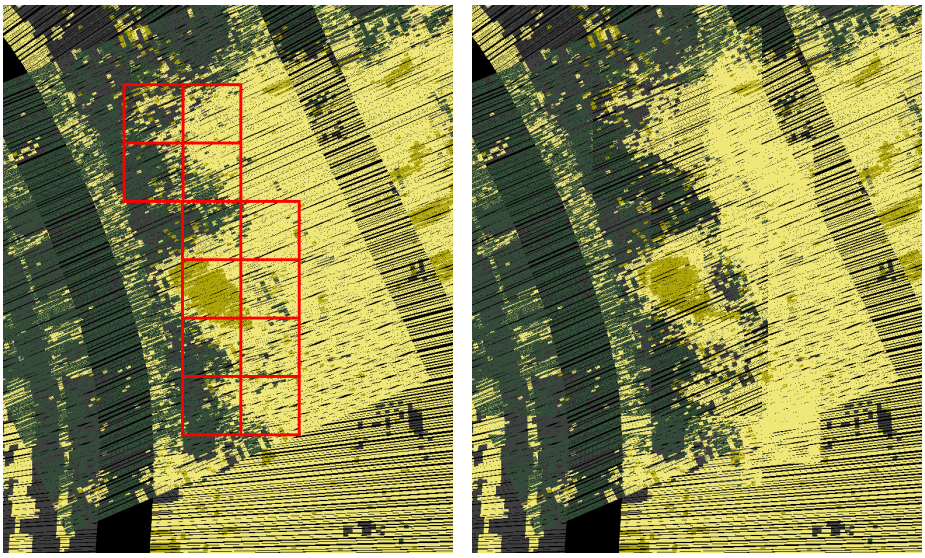
FIG. 4 – *Deux extraits de traces sonar recalés selon leurs données de géoreferecement, puis fusionnés.*

Après observation, nous considérons que les mouvements du sonar et l'imprécision d'autre paramètres (tels que la longueur du câble reliant le sonar au bateau) nous conduit à faire une erreur pouvant aller jusqu'à une dizaine de mètres sur le recalage par géoréférencement. Nous choisissons aléatoirement parmi les deux traces celle qui deviendra l'image de référence, et celle que nous cherchons à recaler. Au sein de la zone d'intérêt, nous extrayons de l'image de référence des imagettes carrées de 300 pixels de coté (soit 30 m), qui deviendront des imagettes de référence, et de l'image à recaler des imagettes de 100 pixels de coté (soit 10 m). Ceci nous permet de calculer la valeur de dissimilarité sur une zone de 10 m de coté, et de recaler sur une distance de 10 m autour de la position d'origine des imagettes.

Nous avons appliqué notre processus de recalage aux images présentées précédemment. La figure 5(a) présente les zones d'intérêt que nous allons chercher à corriger.

La figure 5(b) présente la mosaïque générée à la fin de notre processus de recalage. L'ensemble des transformations rigides locales déterminées sont cohérentes dans leur orientation (vers la droite) et leur distance (quelques mètres). Les transformations déterminées sont liées au fort conflit dans les zones d'intérêt. En effet, la passe de référence a été principalement classifiée avec de la posidonie alors que la passe à recaler l'a été avec du sable. Notre critère de dissimilarité a donc conduit à déterminer des transformations qui projette nos imagettes sur la zone de sable de la passe de référence.

Le coût calculatoire de notre algorithme est actuellement très élevé, en particulier à cause de la combinaison. La taille des données ne limite que la génération de la mozaïque (l'image



(a) Zones d'intérêt détectées (b) Mosaïque après recalage

FIG. 5 – *Détail de la zone de recalage*

finale de notre base de donnée actuelle tend vers 9000×6000 pixels) car nous nous ramenons à un problème de recalage sur des données limitées (300×300 pixels).

6 Conclusion

Nous présentons dans cet article une application de notre processus de recalage pour les images classifiées. Afin de tenir compte des imperfections des résultats de classification nous nous sommes tournés vers la théorie de fonction de croyance. Cette théorie a de plus été employée avec succès pour la fusion de classifieurs qui est la dernière étape du recalage. Ce choix nous permet donc réaliser deux étapes (le recalage et la fusion) en une seule opération.

L'approche proposée repose sur l'utilisation du conflit issu de la combinaison des fonctions de croyance comme critère de dissimilarité afin de déterminer la meilleure transformation au sein d'un algorithme de recalage. Nous nous sommes confrontés ici à des données réelles et multivues, et avons cherché a prendre en compte un modèle de transformation rigide local (en tant qu'approximation d'un modèle élastique global).

Ce travail peut être étendu en utilisant une définition plus complexe pour les fonctions de masses. Il faudra alors déterminer au mieux la nature du conflit généré lors de la combinaison, pour n'en considérer que la part qui représente le désaccord des sources. De plus la théorie des fonctions de croyance permet de combiner plus de deux sources, donc nous pourrions étendre le recalage et la génération de mosaïque à trois images sources ou plus en même temps.

Références

Appriou, A. (2002). Discrimination multisignal par la théorie de l'évidence. In R. Lengllé (Ed.), *Décision et Reconnaissance des formes en signal* (Lavoisier ed.)., Chapter 7, pp. 219–258. Hermes Science Publication.

Bloch, I. et H. Maitre (1994). Fusion de données en traitement d'images : modèles d'information et décisions. *Traitement du Signal 11*(6), 435–446.

Chailloux, C. (2007). *Recalage d'images sonar par apariement de régions. Applicaton à la génération d'une mosaïque*. Ph. D. thesis, ENST Bretagne.

Chambon, S. (2005). *Mise en correspondance stéréoscopique d'images couleur en présence d'occultations*. Ph. D. thesis, Université Paul Sabatier, Toulouse.

Dempster, A. P. (1967). Upper and Lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics 38*, 325–339.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems Man and Cybernetics 25*(5), 804–813.

Dhibi, M., R. Courtis, et A. Martin (2008). Multi-segmentation of sonar images using belief function theory. In *Acoustics*, Paris, France.

Hill, D. L. G. et D. J. Hawkes (2000). Across-modality registration using intensity-based cost functions. pp. 537–553.

Leblond, I. (2006). *Recalage à long terme d'images sonar par mise en correspondance de cartes de classification automatique des fonds*. Ph. D. thesis, ENSIETA.

Leblond, I., M. Legris, et B. Solaiman (2005). Use of classification and segmentation of sidescan sonar images for long term registration. *Oceans 2005-Europe 1*.

Maintz, J. et M. Viergever (1998). A survey of medical image registration. *Medical Image Analysis 2*(1), 1–36.

Martin, A. (2005). Comparative study of information fusion methods for sonar images classification. In *International Conference on Information Fusion*, Philadelphia, USA.

Martin, A., H. Laanaya, et A. Arnold-Bos (2006). Evaluation for uncertainty image classification and segmentation. *Pattern Recognition 39*(11), 1987–1995.

Milisavljevic, N., I. Bloch, S. Van Den Broek, et M. Acheroy (2003). Improving mine recognition through processing and dempster-shafer fusion of ground-penetrating radar data. *Pattern Recognition 36*(5), 1233–1250.

Olteanu, A. (2007). A multi-criteria fusion approach for geographical data matching. In *Proceedings of the Fifth Internat. Symp. on Spatial Data Quality (ISSDQ'07), June*.

Roche, A. (2001). *Recalage d'Images Médicales par Inférence Statistique*. Ph. D. thesis, Université de Nice Sophia Antipolis.

Rominger, C., A. Martin, A. Khenchaf, et H. Laanaya (2009). Sonar image registration based on conflict from the theory of belief functions. In *International Conference on Information Fusion*.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Smets, P. (1990). The Combination of Evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*(5), 447–458.

Smets, P. (2007). Analyzing the combination of conflicting belief functions. *Information Fusion 8*, 387–412.

Smets, P. et R. Kennes (1994). The Transferable Belief Model. *Artificial Intelligent 66*, 191–234.

Taleb-Ahmed, A., L. Gautier, et M. Rombaut (2002). Architecture de fusion de données basée sur la théorie de l'évidence pour la reconstruction d'un vertèbre. *Traitement du Signal 19*(4), 267–283.

Vannoorenberghe, P., O. Colot, et D. de Brucq (1999). Color image segmentation using Dempster-Shafer's theory. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, Volume 4.

Vannoorenberghe, P., E. Lefevre, et O. Colot (2003). Traitement d'images et théorie des fonctions de croyance. *Rencontres Francophones sur la Logique Floue et Ses Applications, LFA'2003*, 26–27.

Williams, D. P. (2009). Bayesian data fusion of multi-view synthetic aperture sonar imagery for seabed classification. *IEEE Signal and Image Processing 18*(6), 1239–1254.

Woods, R., J. Mazziotta, et S. Cherry (1993). MRI-PET registration with automated algorithm. *Journal of computer assisted tomography 17*(4), 536–546.

Zitova, B. et J. Flusser (2003). Image registration methods : a survey. *Image and Vision Computing 21*(11), 977–1000.

Summary

This paper presents an application for classified image registration and fusion. We extend here results developed on a previous paper to multiview images. For seabed characterization, we need to fuse the multiview of sonar images to increase performances. However, before fusion, we have to proceed to an image registration. The proposed approach is based on the use of the conflict due to the combination as a disimilarity measure in the classified images registration. The theory of belief functions allows an unique framework to model the imperfections and to fuse the classified images.

Approche graphique pour l'agrégation de classifications non supervisées

Fatma Hamdi*, Haytham Elghazel **
Khalid Benabdeslem **

* Université Paris13, LIPN, UMR CNRS 7030
99, Av Jean Baptiste Clément 93430 Villetaneuse
Fatma-hamdi@lipn.univ-paris13.fr
**Université Lyon1, LIESP EA 4125
43, Bd du 11 Novembre 1918 Villeurbanne Cedex
{kbenabde, elghazel}@bat710.univ-lyon1.fr

Résumé. Dans cet article nous proposons de traiter le problème de combinaison de plusieurs résultats de classification pour obtenir une partition consensus. Dans ce cadre nous proposons un algorithme ensembliste : *G-Cons* basée sur une technique connue en théorie des graphes appelée la coloration minimale permettant de retourner une seule partition finale et relativement stable. Les résultats expérimentaux ont montré des performances très encourageantes.

1 Introduction

L'extraction de connaissances par exploration de données fait souvent appel à des processus de classification automatique en mode non supervisé. Effectuer une classification, c'est mettre en évidence d'une part les relations entre les différentes observations et d'autre part les relations entre ces observations et leurs caractéristiques (leurs variables). A partir d'une certaine mesure de proximité ou de dissemblance, il s'agit de regrouper un ensemble de données en un ensemble de classes qui soient les plus hétérogènes possible. En outre, il existe une panoplie de méthodes de classification automatique ayant intéressé beaucoup de chercheurs dans la communauté scientifique, notamment, celle dites par partitions (K-moyenne, k-medoïde), hiérarchique (CAH, pyramidale....) et celles basées sur des modèles connexionnistes ou graphiques (SOM, b-coloration, …etc). Ceci dit, ces méthodes produisent généralement des partions peu ou très différentes sur un même ensemble de données. C'est le cas par exemple des techniques de classification hiérarchique où une coupure du dendrogramme fournit une partition de l'ensemble des données à classer. Il est à mentionner que plusieurs niveaux de troncature sont souvent retenus et les partitions qui s'en déduisent sont comparées afin de retenir la meilleure en termes de compacité et de séparabilité des classes. C'est le cas aussi des approches de classification par partitionnement où le nombre de classes doit être fixé à priori, une information non pas toujours disponible pour toutes les bases de données. Plusieurs valeurs pour ce nombre de classes peuvent donc

être fixées et les partitions correspondantes sont ensuite comparées. L'existence d'une telle variabilité intra et inter méthodes est souvent évaluée par des indices de qualité en mode concurrentiel pour justifier la préférence d'une partition ou d'une méthode par rapport à une autre. Dans le cadre de ce travail, nous proposons de contourner ce problème de variabilité dans les approches de classification automatique par une alliance des différentes partitions retournées. il s'agit de mettre en place une approche de *classification ensembliste* (*consensus*) basée sur une technique de la théorie des graphes baptisée la *coloration minimale* permettant de retourner une seule partition finale et relativement stable, par le biais de la combinaison de différentes partitions.

Plusieurs approches de consensus (Ghemi et al 2009)) ont été proposées dans la littérature dont on peut citer les approches de partitionnement d'un graphe, les approches par vote, les approches qui cherchent à optimiser l'information mutuelle comme fonction objectif et aussi les approches basées sur une matrice de co-association. Strehl et Ghosh (2002) se sont situés dans la famille des approches de partitionnement de graphe afin de développer trois algorithmes de combinaison de partitions CSPA[1], CSPA[2] et HGPA[3]. Ces algorithmes consistent à transposer le problème de consensus dans la classification non supervisée à un problème d'algorithmique dans la théorie des graphes. Pour cela ils proposent la modélisation des différentes partitions à combiner par un hypergraphe définit par un ensemble de sommets (les individus) et d'hyper-arêtes (une arête dans un graphe permet de connecter deux sommets alors qu'une hyper-arête dans un hypergraphe peut connecter un ensemble de sommets).

Les différents algorithmes proposés, cherchent à maximiser la même fonction objective dite l'information mutuelle. Pour un problème de consensus donné les auteurs proposent d'appliquer les trois algorithmes et de ne retenir que la partition retournée par l'algorithme offrant la meilleure qualité en termes d'information mutuelle. Cette partition est dite la ***Supra Consensus***.

.

2 Approche proposée : Consensus par coloration de graphes

Dans cette section, nous présentons une nouvelle méthodologie basée sur la théorie des graphes. Cette approche, utilisée pour résoudre le problème de combinaison des partitions, consiste à chercher une seule partition finale et relativement stable. Cette dernière provient d'un ensemble de partitions, obtenues suite à l'application de plusieurs algorithmes de classification non supervisée.

[1] Clusterd –based Similarity Partitionning Algorithm.
[2] Clusterd –based Similarity Partitionning Algorithm.
[3] HyperGraph-Partitionning Algorithm

2.1 Description générale du problème de combinaison des partitions

Partant d'un ensemble de partions ***r*** obtenu à partir de l'application de plusieurs techniques de classification (sur le même ensemble de données ***X***: la q-éme partition $\lambda^{(q)}$ contient $k^{(q)}$ classes), la résolution du problème de combinaison des partitions (consensus) revient à trouver la fonction consensus Γ. Cette fonction est définie comme une fonction $\mathbb{N}^{nxr}$ $\mathbb{N}^{n}$ et qui permet de donner la partition finale λ'', tel que :

$$\Gamma : \{ \lambda^{(q)} / q \in \{1, \ldots, r\} \} \longrightarrow \lambda'' \tag{1}$$

Les informations concernant les différentes partitions peuvent être représentées par une matrice de dissimilarité appelée ***D.*** Cette dernière indique le nombre de fois où chaque deux individus ne se sont pas apparus ensemble dans toutes les partitions.
Considérons dans la suite le problème du consensus comme étant un problème de classification automatique à base d'une matrice de proximité ***D.*** Nous cherchons à optimiser une fonction objective qui est l'information mutuelle dont le principe est le suivant :
Dans une première étape, il s'agit de définir la fonction $\varphi^{(NMI)}$, cette dernière permet de calculer l'information mutuelle entre deux partitions :

$$\varphi^{(NMI)}(\lambda^{(a)}, \lambda^{(b)}) = \frac{\sum_{h=1}^{k^{(a)}} \sum_{l-1}^{k^{(b)}} n_{h,l} \log(\frac{n - n_{h,l}}{n_h^{(a)} n_l^{(b)}})}{\sqrt{(\sum_{h=1}^{k^{(a)}} n_h^{(a)} \log \frac{n_h^{(a)}}{n})(\sum_{l=1}^{k^{(b)}} n_l^{(b)} \log \frac{n_l^{(b)}}{n})}} \tag{2}$$

Où $n_h^{(a)}$ le nombre d'individus dans la classe C_h de la partition $\lambda^{(a)}$, $n_l^{(b)}$ le nombre d'individus dans la classe C_l de la partition $\lambda^{(b)}$ et $n_{h,l}$ le nombre d'individus qui sont dans la classe *h* de la partition $\lambda^{(a)}$ et aussi dans la classe *l* de la partition $\lambda^{(b)}$.

Par la suite, nous cherchons à optimiser la fonction $\varphi^{(ANMI)}$ qui calcule l'information mutuelle entre la partition recherchée λ'' et les autres partitions à combiner. Notre approche ensembliste vise à trouver la partition qui définie la meilleure combinaison, i.e la plus grande valeur de $\varphi^{(ANMI)}$:

$$\varphi^{(ANMI)}(\Lambda,\lambda'') = \frac{1}{r}\sum_{q=1}^{r}\phi^{(NMI)}(\lambda'',\lambda^{(q)}) \quad (3)$$

2.2 Modélisation du problème par la théorie des graphes

Dans cette partie, nous allons modéliser les informations des différentes partitions à l'aide d'un graphe. Pour cela, nous considérons la représentation graphique des données à grouper comme un graphe complet, non orienté et pondéré. Dans ce graphe les sommets sont les individus à analyser et les arêtes les liens pondérés par les dissimilarités entre les paires de données.
Une définition classique suppose qu'une classe ou «cluster» est un ensemble d'éléments similaires ou semblables, et les éléments de différentes classes sont différents. En effet une classe devrait satisfaire les deux conditions suivantes : la première c'est que l'homogénéité interclasses doit être élevée ; la deuxième consiste à une hétérogénéité forte entre les éléments de classes différentes. Ces deux conditions s'élèvent à affirmer que les arêtes entre deux sommets de la même classes devraient avoir une forte similarité reflétant une faible pondération ; et ceux entre les sommets de classes différentes devraient avoir une faible similarité donc une pondération élevée.

2.2.1 Coloration minimale pour un consensus de classification non supervisée

Afin de définir notre algorithme, qui permet de trouver le meilleur compromis entre les différentes partitions à combiner, nous nous sommes basés sur le principe de la coloration minimale. L'approche fondée sur la coloration minimale permet de définir des partitions à faible diamètre (Hanssen et al (1978)) (un critère d'homogénéité *intraclasse*). Ceci répond exactement à notre objectif qui est de maximiser l'information mutuelle de la partition retenue, considérée lui aussi comme un critère d'homogénéité *intraclasse* (Strehl et al (2002)).
La représentation par graphe complet ne convient pas au problème de classification non supervisée. En effet, la coloration minimale du graphe retournerait la classification "*triviale*" où chaque *cluster* (*couleur*) contient un seul individu (singleton). La coloration minimale passe donc par la construction d'un *graphe seuil supérieur* défini comme le graphe *partiel* du graphe de départ. Un *graphe seuil supérieur* $G(V,E)$ est un graphe simple ayant pour ensemble de sommets les sommets du graphe d'origine $V=\{v_1,...,v_n\}$ et pour ensemble d'arêtes E les paires de sommets dont la dissimilarité est supérieure à un seuil θ choisi à partir de la table de dissimilarité des individus (*i.e.* $\forall v_i,v_j \in V$, *l'arête* (v_i,v_j) existe *ssi* $D(v_i,v_j) > \theta$ où $D(v_i,v_j)$ est la dissimilarité entre v_i *et* v_j).
Dans la suite de cet article, deux sommets sont *voisins* (*resp. non voisins*) s'ils sont "*adjacents*" (*resp.* "*non adjacents*"). Nous cherchons donc à établir une *coloration valide* du graphe $G(V, E_{>\theta})$, qui consiste à affecter une couleur c à chaque sommet v

tel que deux sommets adjacents n'ont pas la même couleur. Le nombre de couleur utilisé doit être minimal.

- Algorithme de coloration minimale

Plusieurs algorithmes ont été développés afin de résoudre le problème de la coloration minimale d'un graphe, le plus connu et le plus utilisé c'est l'algorithme *Largest First* (LF) développé par Welsh et Powell en 1967. Cet algorithme permet de ranger l'ensemble des sommets dans un ordre décroissant par rapport à leurs degrés, il vise à construire une partition de l'ensemble des données à classer sans donner de l'importance à la séparation entre les classes. Cet algorithme présente quelques limitations liées au choix de couleurs des sommets dans certain cas. En effet il n'attribut pas un critère de choix entre les sommets du moment où ils ont le même degré.

- Adaptation et/ou amélioration de LF : *G-Cons*

Dans un premier temps, nous avons adapté l'algorithme LF à la problématique d'agrégation des partitions, ce qui nous a permis d'obtenir des résultats encourageants. Néanmoins le LF présente certaines faiblesses. Par conséquent, nous avons proposé un algorithme ensembliste appelé *G-Cons*. Cet algorithme est une modification de *Largest First* dans le but d'améliorer la qualité de la partition retournée en apportant des solutions à ces problèmes.
Dans une première étape de notre algorithme nous proposons d'effectuer un prétraitement des individus, en regroupant ceux qui sont toujours ensemble dans toutes les partitions. Le prétraitement des individus va conduire à la construction d'une nouvelle matrice réduite ***D'*** de taille ***n'*** x ***n'*** (***n'***<<***n***).
De manière équivalente à la construction du graphe ***G*** à partir de la matrice de dissimilarité ***D,*** les informations de la matrice ***D'*** peuvent être aussi représentées par un graphe ***G'*** = ***G'(V', E')*** qui contient des sommets composés tel que :

$$V_i' \epsilon\, V' = \{x_l \epsilon\, V\}\ tel\ que\ \forall(x_l, x_j) \epsilon\, V^2, D(x_l, x_j) = 0$$

Cette étape est très importante dans la mesure où elle nous a permis de réduire remarquablement la taille de la matrice ***D***, ce qui permet de réduire la complexité de notre algorithme.
En deuxième étape, nous proposons d'utiliser un algorithme qui prend en entrée le graphe obtenue à partir de la matrice de dissimilarité ***D'*** et qui procède à la résolution du problème de *largest first* de la manière suivante:
Comme nous l'avons cité auparavant, le problème principal de LF consiste à trouver le sommet approprié quand il s'agit de choisir entre plusieurs ayant le même degré. Ce choix doit être contraint de la maximisation de l'homogénéité *intraclasse* et par conséquent la dissimilarité interclasse de la partition λ'' retournée par le consensus.
Comme solution à ce problème, nous proposons la stratégie de choix des sommets suivante:

Pour un sommet v_i sélectionné pour être coloré avec la première couleur ***c*** différente aux couleurs de ces sommets adjacents. Nous procédons comme suit : les sommets non encore colorés de même degré que v_i et qui ne sont pas adjacents à aucun sommet de couleur ***c***, seront aussi simultanément considérés par la coloration. Le sommet dont sa coloration donne une valeur de dissimilarité minimale, c'est-à-dire celui qui maximise notre fonction objective, sera choisi pour être coloré avec ***c*** et les autres seront traités après par l'algorithme.

Cependant la distance entre un sommet v_i et une couleur ***c***, est définie par la moyenne des dissimilarités entre un sommet v_i, non encore coloré, et tous les sommets de couleur ***c***. Cette distance est obtenue par la formule suivante :

$$d(v_i, c) = \frac{1}{|c|} \sum_{j=1}^{|c|} D(v_i, v_j) \tag{4}$$

Notre idée est présentée par l'algorithme *G-Cons*. Nous avons ainsi besoin d'introduire quelques notations que nous allons utilisés :

degré (v_i), c'est le degré du sommet v_i dans $\boldsymbol{G}_{>\theta}$.C'est le nombre de voisins du sommet v_i.

c(v_i) : la couleur du sommet v_i dans le graphe $\boldsymbol{G}_{>\theta}$.

$\boldsymbol{N}$ (v_i) : le voisinage du sommet v_i dans le graphe $\boldsymbol{G}_{>\theta}$.

$\boldsymbol{N_c}(v_i)$: le voisinage de couleur du sommet v_i dans le graphe $\boldsymbol{G}_{>\theta}$.

Nbcouleur : le nombre de couleurs dans le graphe $\boldsymbol{G}_{>\theta}$. Nbcouleur est initialisé à 0 (les sommets du graphe $\boldsymbol{G}$ ne sont pas encore colorés).

-Enlever (v, V) c'est la méthode qui permet de supprimer le sommet v de l'ensemble des sommets V.

- Mise à jour (d(v_j, c)) : c'est la méthode qui permet la mise à jour de la valeur de dissimilarité entre un sommet v_j et une couleur ***c*** lorsqu'un sommet v_i est récemment affecté à cette couleur, autrement la coloration de v_i avec la couleur ***c*** change la valeur de dissimilarité d(v_j, c) de chaque sommet $v_j$du graphe $\boldsymbol{G}_{>\theta}$.

Algorithme 1 : *G-Cons*

Début

Entrée : $G_{>\theta} = G(V, E_{>\theta})$ *// un graphe avec un ensemble de sommets et un ensemble d'arêtes.*

1. Choisir le sommet v de l'ensemble des sommets V avec le plus grand degré ;
2. c(v)=1 ;
3. Enlever (v, V);
4. Pour chaque sommet $v_j \in$ V faire
5. Mise à jour (d(v_j, c(v))
6. Fin pour
7. Nbcouleur :=1;
8. Répéter
9. Choisir v_i de V;
10. c:= min {h| 1 ≤ h ≤ Nbcouleur, h$\notin N_c(v_i)$} ;
11. Si (c ≤ Nbcouleur) alors
12. M := V \ $N(v_i)$;
13. H:= {v_h| $v_h \in$ M et degré(v_h)=degré(v_i) et c$\notin N_c$ (v_h)} ;
14. H := H $\cup$ { v_i};
15. v := argmin $_{v_h \in H}$ (d(v_h, c));
16. c(v) := c;
17. Enlever (v, V);
18. sinon
19. c(v_i) := c;
21. Nbcouleur=Nbcouleur+1;
21. Enlever (v, V);
22. Fin si
23. Pour chaque sommet $v_j \in$V faire
24. Mise à jour (d(v_j, c)) ;
25. Fin pour
26. Jusqu'à (V = ϕ)

Fin.

3 Expérimentations et validation

Nous avons considéré dans cette étude les jeux de données benchmark *Anneaux*, *Atom*, *Engytime*, *Tow-diamonds*, *target* obtenues à partir de la base de données UCI Blake et al (2007).

	Nombre d'individus	**Nombre de variables**	**Nombre de labels**
Anneaux	1000	3	2
Atom	800	3	2
Engytime	4096	2	2
Towdimonds	800	2	2
Target	770	2	6

Tab 1**.** Description des bases de données utilisées.

Notre méthodologie d'évaluation se base sur les deux principes suivants:
1) Dans un premier temps nous avons cherché à évaluer la performance de notre fonction du consensus *G-Cons* et celle de chacune des autres approches HGPA, MCLA et CSPA (Strehl et al (2002)). D'autre part, afin de mieux étudier l'apport des différentes modifications que nous avons apporté à l'algorithme de la coloration minimale, la fonction de consensus par coloration minimale originale **(que nous appelons *Ccm*)** sera également évaluée. La pertinence de ces différentes fonctions consensus est évaluée en terme de leurs valeurs finales de la fonction objective: l'information mutuelle.
2) Dans un deuxième temps, nous travaillons à comparer les partitions finales retournées par notre technique *G-Cons*, par la meilleure des trois techniques CSPA, HGPA et MCLA (dite *Supra Consensus* dans (Strehl et al (2002)) et les partitions originales à combiner. Pour une meilleure évaluation de ces partitions nous avons utilisé les critères de qualité internes et externes suivants : un critère interne basé sur la notion de dissimilarité qui est l'indice de *Davies-Bouldin* et une mesure statistique externe basés sur les étiquètes (labels) fournies dans les bases de données originale qui est l'*information mutuelle*.
Les différentes partitions en entrée de ces fonctions consensus ont été obtenues par l'application de l'algorithme des *k-moyennes* (Berkhin (2002)), la *classification ascendante hiérarchique* CAH (Diday et al) et les cartes topologiques de Kohonen optimisées par la CAH et aussi par les k-moyennes (Vesanto et al (2000)), Le nombre de classes fourni en entrée de chacune de ces quatre approches de classification non supervisée correspond à celui des labels originaux dans chaque jeu de données utilisé.

3.1 Etude comparative des fonctions consensus considérées

Le tableau ci-dessous résume les résultats obtenus, en termes d'information mutuelle, par l'application de différentes approches de consensus, *G-Cons*, *Ccm,* CSPA, MCLA et HGPA sur nos différents jeux de données considérés.

	CSPA	***MCLA***	***HGPA***	***Ccm***	***G-Cons***
Target	**0.6497**	**0.6640**	**0.6261**	**0.9057**	**0.9057**
Engytime	**0.7952**	**0.8055**	**4.7654^{e-} 005**	**0.8159**	**0.8191**
Atom	**0.3074**	**0.1236**	**1.4496^{e}- 004**	**0.6665**	**0.6665**
Anneaux	**0.6382**	**0.6260**	**9.7926^{e}- 004**	**0.6663**	**0.6663**
Towdimonds	**0.9842**	**0.9657**	**4.5091^{e}- 006**	**0.9842**	**0.9842**

Tab 2. Valeurs de la fonction objective pour les différents algorithmes d'agrégations et sur différents jeux de données.

Ces différents résultats montrent d'une part (1) l'apport de notre adaptation du principe de la coloration minimale au contexte du consensus de classification, qui a permis de mieux optimiser notre fonction objective qui est l'information mutuelle. Ceci confirme le fort lien entre l'information mutuelle et le critère d'homogénéité *intraclasse* (diamètre de partition minimum) souvent recherché par la coloration minimale et d'autre part (2) la pertinence des modifications que nous avons apporté à l'approche par coloration minimale *Ccm* sur un seul jeu de donnée *Engytime*. Ce dernier résultat étant attendu vu le nombre d'individus dans cette base. Pour les autres jeux de données (de tailles plus petites que *Engytime*), où les résultats sont similaires, l'étape de réduction de la matrice de dissimilarité décrite dans la section 2 a permis de bien diminuer le nombre d'individus (en occurrence les sommets dans le graphe) conduisant à des graphes de petites tailles où le problème de choix de couleurs pour les sommets est rare. En effet le tableau suivant permet de fournir une vision sur la réduction de sommets sur les différents jeux de données testés.

	Nb de sommets
Target	19
Engytime	50
Atom	7
Anneaux	11
Towdimonds	4

Tab 3. Taille des différentes matrices après réduction.

3.2 Etude comparative des résultats

Les différents indices de qualité mentionnés ci-dessus sont calculés pour les partitions obtenues par les algorithmes ensemblistes (*Supra Consensus*, *Ccm* et *G-Cons*) et les partitions à combiner fournies par les algorithmes de classification non supervisée.
Les différentes qualités des partitions à combiner influencent celle de la partition obtenue par l'application d'un algorithme de consensus. Nous proposons donc de comparer les indices de qualité obtenus pour chaque algorithme ensembliste avec ***MoyenneP*** qui représente la moyenne des valeurs des indices obtenus par les algorithmes utilisés afin de fournir les partitions à combiner.

- **Indice de *l'information mutuelle* :**

	MoyenneP	***Supra consensus***	***Ccm***	***G-Cons***
Target	0.650300	0.664800	0.657500	0.657500
Engytime	0.701300	0.722800	0.722000	0.727800
Atom	0,307375	0.03520	0.533100	0.533100
Anneaux	0.170250	0.070500	0.296000	0.297100
Towdimonds	0.984225	1.000000	1.000000	1.000000

Tab 4. Résultats de différents algorithmes en termes d'information *mutuelle*.

- **L'indice de *Davies-Bouldin*** (Bezdek et al (1998))

	MoyenneP	***Supra consensus***	***Ccm***	***G-Cons***
Target	0.832842	1.836464	0.842219	0.842219
Engytime	1.146891	1.134809	1.141622	1.134441
Atom	1.327712	1.708694	1.408429	1.408429
Anneaux	1.396211	1.406439	1.202712	1.202712
Towdimonds	1.008505	1.007958	1.007958	1.007958

Tab 5. Comparaison entre les résultats de différents algorithmes avec l'indice de *Davies-Bouldin*.

Les différents résultats obtenus pour les différent critères de qualités montrent que *G-Cons* donne une meilleure performance par rapport au Supra Consensus ainsi nous pouvons constatés que notre approche améliore dans la majorité des cas la qualité en terme de différents indices utilisés par rapport aux partitions à combiner. Cependant les résultats obtenus par notre approche sont similaire à *Ccm* dans les bases *Target, Atom, Anneaux* et *Twodimonds* et meilleur pour la base *Engytime (la plus grande base)* ce qui est attendu.
Ces différentes valeurs confirment le bon résultat obtenu par *G-Cons*, ce qui montre l'apport de notre adaptation du principe de la coloration minimale au contexte du consensus de classification, qui a permis de donner des meilleurs résultats par rapport au *Supra Consensus* et au *Ccm* lorsqu'il s'agit de grandes bases.

4 Conclusion et perspectives

Dans le cadre de ce travail, nous avons proposé une nouvelle approche graphique pour la combinaison de partitions issues de la classification non supervisée.
Cette approche n'est pas directement dédiée à un type de données complexes, mais elle y contribue de manière implicite à partir de partitions qui pourraient résulter de n'importe quel type de données.
De plus, de par son rôle d'agrégation, elle permet de traiter les des données de sources multiples en fournissant une partition finale et relativement stable à partir de plusieurs regroupements obtenus par plusieurs techniques connues en classification non supervisée.
Les résultats obtenus sont très encourageants et ont montré l'utilité d'une telle approche dans un domaine émergeant en fouille de données complexes.

Référence

A. Strehl et J. Ghosh. *Cluster ensembles – A knowledge reuse framework for combining multiple partitions*. Journal on Machine Learning Research (JMLR), 3:583–617, December 2002.

Bezdek J. C et Pal N. R. *Some new indexes of cluster validity*, IEEE Transactions on Systems, Man and Cybernetics, 28(3), pp. 301-315, 1998

Berkhin P. *Survey of clustering data mining techniques*, Accrue Software, 2002

Blake C. L. et Merz C. J. UCI repository of machine learning databases, Available from http://www.ics.uci.edu/~mlearn/MLRepository.html (Octobre 2007), University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

Diday E. Bertrand P. ***Une généralisation des arbres hiérarchiques: les représentations pyramidales***. *Revue de Statistique Appliquée*

G. Karpys et V. Kumar. *A fast and high quality multilevel scheme for portioning irregular graphs*. SIAM Journal on Scientific Computing, 20(1):359-392,1998.

Hassen P. and Delattre M. *Complete-link cluster Analysis by graph coloring*, Journal of the American Statistical Association, 73, pp. 397-403, 1978.

J. Vesanto et E. Alhoniemi. *Clustering of the self-organizing map*, IEEE Transactions on Neural Network, VOL. 11, NO, pp. 586-600, 3, May 2000.

R. Ghaemi, N. Sulaiman, H. Ibrahim, et N. Mustapha, A survey : Clustering ensembles techniques. In world Academy Science, Engeneering and Technologie.

Welsh D. J. A. & Powell M. B. (1967). *An upper bound for the chromatic number of a graph and its application to timetabling problems*. Computer Journal, 10(1),85–87.

Etude d'opérateurs d'agrégation pour l'ordonnancement de groupes de pixels dans des images numériques de plantes

Jimmy Nagau*, Sébastien Régis*
Jean-Luc Henry*

*Laboratoire GRIMAAG
Université des Antilles et de la Guyane Département Mathématiques et Informatique
B.P. 592, 97159 Pointe-à-Pitre Cedex
jnagau,sregis,jlhenry@univ-ag.fr,
http://grimaag.univ-ag.fr/

Résumé. Dans cet article, on utilise une méthode de segmentation appliquée à la reconnaissance de végétaux dans des images. La segmentation fournit des groupes de pixels, et nous proposons d'utiliser des opérateurs d'agrégation pour ordonner ces groupes afin d'améliorer les performances du système de reconnaissance. En particulier, on utilise un opérateur totalement renforcé appelé triple Π. Les premiers tests sont réalisés et montrent que l'utilisation du triple Π améliore les résultats pour la structuration des groupes de pixels.

1 Introduction

La reconnaissance de plantes à partir d'images numériques est une thématique qui touche de nombreux domaines. L'enjeu est important, par exemple, pour les centres médicaux en cas d'intoxication, pour les observatoires botaniques pour les recensements ou encore pour les écoles par rapport à l'apprentissage de la Botanie. Dans cette démarche, on peut imaginer une chaîne de traitements qui reçoit une ou plusieurs images d'un végétal et propose le nom de ce végétal en s'appuyant sur un ensemble d'images contenu dans une base de données répondant au mieux à la requête d'un utilisateur. Un système de ce type se base dans un premier temps sur les éléments issus d'une opération de segmentation dont le but est de mettre ensemble chaque pixel représentant une forme dans une image.

Pour ce faire, nous utilisons une méthode de segmentation globale basée sur un Mean Shift D. Comaniciu (2002) avec changement d'échelle Nagau et Henry (2010) qui fournit des agrégats de pixels. Ces groupes correspondent en général à des parties spécifiques de la plante : branche, feuille, pétale, etc.

En fonction de l'espèce présente dans l'image, certaines parties (représentées par ces groupes de pixels) sont plus pertinentes pour la reconnaissance de la plante : suivant l'espèce, certaines parties sont suffisantes pour la reconnaissance alors que les autres parties sont secondaires BINET et BRUNEL (1967).

Dans cet article, nous proposons de fusionner des caractéristiques de ces groupes de pixels pour établir un ordre de priorité des groupes afin d'améliorer la reconnaissance des végétaux.

Pour la fusion de l'information, on utilisera des opérateurs d'agrégation. Dans la section 2, nous présentons plus en détail la problématique et la méthode utilisée pour analyser et traiter les images de végétaux. Dans le paragraphe 3, la fusion d'information et les opérateurs d'agrégation sont présentés. Dans la partie 4, nous présentons les premiers résultats expérimentaux avant de conclure.

2 Contexte et problématique

Les images numériques sont d'énormes sources d'informations, on observe dans chacune d'elle un ensemble de pixels qui, à travers des relations de connexité, permet de représenter des objets. Dans une photographie numérique de plante par exemple, chaque agrégat de pixels représente une partie de celle-ci comme ses feuilles, ses pétales, ses nervures, etc. Diverses approches ont été utilisées pour la reconnaissance de végétaux dans des images ; on peut citer entre autres, les outils de morphologie mathématique Wang et al. (2008), le changement d'espace colorimétrique Philipp et Rath (2002), les réseaux de neurones Aitkenhead et al. (2003) ou encore les SVM Camargoa et Smith (2009). Cependant, il faut noter que la plupart de ces méthodes sont utilisées pour la reconnaissance du végétal à partir d'une seule de ses parties (en général la feuille). Néanmoins, ces végétaux possèdent plusieurs éléments (tige, fleur, fruit, etc.) qui peuvent aider à la caractérisation. Cependant, tous ces éléments dans une image ne possèdent aucune structure pour une machine. On cherche à organiser les groupes de pixels dans un arbre afin d'améliorer le processus de reconnaissance et d'ajouter de la sémantique en combinant ces groupes à partir de l'arbre. Pour ce faire, on effectue une segmentation de l'image (voir figure 1). Après l'utilisation de la chaîne de traitement de la figure 2 où l'on effectue deux traitements qui ont pour but d'éliminer le bruit (utilisation de la théorie gestalt Wertheimer (1958)) et les objets flous (morphologie mathématique), on extrait des groupes de pixels, et l'on s'intéresse à deux de leurs caractéristiques.

La premère est liée aux conditions d'acquisition : il s'agit de la position spatiale de chaque groupe de pixels. On suppose que plus les éléments à traiter se rapprochent du rebord de l'image, moins ils deviennent pertinents pour une identification. Ce critère permet de simuler l'intérêt que le photographe accorde aux éléments appartenant à l'environnement capturé. Les groupes de pixels ayant une bonne évaluation de ce paramètre ont une forte probabilité de ne pas subir d'occlusion (éléments coupés par les rebords de l'image ou d'autres objets appartenant à la scène).

Le second critère est axé sur la nature formelle des groupes de pixels d'une image : on parle de l'homogénéité. En effet, le passage d'une forme à une autre, les variations lumineuses, les autres éléments de la scène, etc. créent parfois du bruit. Les agrégats formés à cause de toutes ces perturbations sont souvent filaires et hétérogènes : ils ne permettent pas une étude précise à cause du manque de densité de pixel.

2.1 Position spatiale des agrégats

Les formes utilisées sont celles placées aux alentours du centre de l'image. Cette contrainte permet d'obtenir de bonnes conditions de traitement et d'éviter tout phénomène d'occlusion.

FIG. 1 – *Résultat de la segmentation sur une photographie de Matricaria recutita*

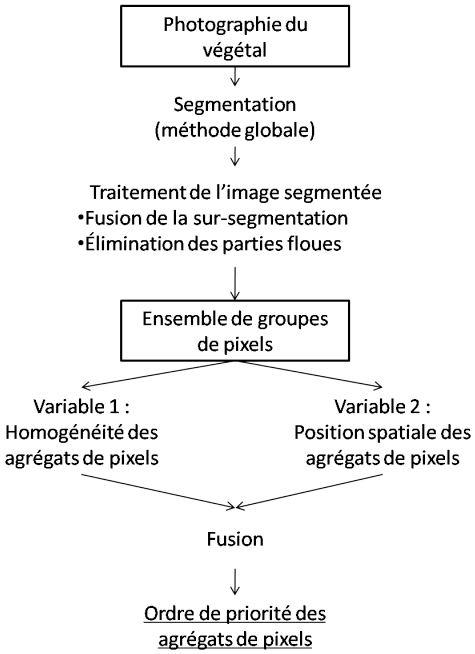


FIG. 2 – *Schéma de la chaîne de traitement, d'une image à l'ordonnancement d'une partition de ses pixels.*

Le système réagit comme le ferait un être humain qui n'accorde de l'importance dans un premier temps lorsqu'il analyse une image qu'aux objets occupant le centre de cette dernière. Un objet dans l'image est représenté par un agrégat de pixels, la position spatiale de cet objet est donc décrite par l'ensemble des pixels qui le composent.

La probabilité de pertinence d'un pixel de l'image est mesurée en fonction de sa position dans cette dernière. On observe une loi gaussienne qui attribue une valeur de 1 pour un pixel situé au plus près du centre de l'image, et une valeur proche de zéro pour un pixel en bordure d'image.

Le résultat pour une forme est la moyenne des valeurs de ses pixels. Chaque position de pixels a une valeur entre 0 et 1 calculée à partir de la formule 1,

$$PSF = \frac{1}{n}\sum_{i=0}^{n} C_n \tag{1}$$

$$C_n = e^{-\alpha \frac{d(F_1,P_n)+d(F_2,P_n)}{2A}}$$

$$A = \frac{l}{4} + \frac{1}{2}\sqrt{L^2 + \left(\frac{l}{2}\right)^2}$$

où PSF est le coefficient qui qualifie la position spatiale de l'objet, n est le nombre de ses pixels, α est un coefficient de paramétrage, F_1 et F_2 représentent les foyers de l'ellipse de rayon l (longueur de l'image) et L (largeur de l'image) et d est la distance euclidienne.

2.2 Homogénéité des agrégats

Les formes traitées doivent avoir leurs pixels qui décrivent un amas dense. Ce procédé permet d'éviter de traiter celles qui n'apportent aucune information sur la structure spatiale d'un objet. Les pixels matérialisant le passage d'une forme à une autre sont ainsi classés en faible priorité à l'aide de ce procédé. Ce critère est représenté par un scalaire compris entre zéro et un, la valeur zéro signifiant que nous sommes en présence d'un objet ne formant que très peu d'amas et fortement dispersé dans l'image traitée.

Nous utilisons l'uniformité d'un pixel dans sa zone, chaque pixel connexe à celui-ci et appartenant à sa classe lui rapporte un score de "+1", la moyenne est réalisée à partir des huit connexes. Pour un pixel ayant une forte connexité avec ses voisins, nous atteignons une valeur égale à un, et pour un pixel faiblement connexe le calcul renvoie une valeur équivalente à zéro. Le critère d'homogénéité est la moyenne des coefficients obtenue pour chaque pixel d'une forme donnée, on le calcul avec la formule 2,

$$HF = \frac{1}{n}\sum_{i=0}^{n} P_n \tag{2}$$

avec :

$$\delta_{P_{n_{x+a,y+b}}} \left\{ \begin{array}{c} 1\ si\ P_{x+a,y+b} \in C_{P_{x,y}} \\ 0\ sinon \end{array} \right\} \tag{3}$$

et

$$0 < \sqrt{a^2 + b^2} < 1$$

où HF représente le critère d'homogénéité, δ est un critère d'uniformité, $P_{n_{x,y}}$ est un pixel de coordonnées (x, y) appartenant à la n ème forme de l'image et C_P est la classe de la forme à laquelle appartient le pixel P.

3 La fusion d'informations et les opérateurs d'agrégation

3.1 Les opérateurs d'agrégation

Il existe un grand nombre d'opérateurs d'agrégation. Le choix et l'utilisation d'un opérateur dépendent de nombreux paramètres. Mais ce choix dépend surtout de la fusion elle-même. Avant d'aller plus loin, il est important de donner la définition de la fusion Bloch et (Eds) (2001) : *La fusion consiste à réunir ou agréger des informations provenant de différentes sources, et à exploiter cette information réunie ou agrégée, dans diverses applications comme la réponse à une question, la prise de décision, une estimation numérique, etc*.
Cette définition met l'accent sur deux éléments principaux. D'abord, elle met l'emphase sur la combinaison de l'information. Puis l'accent est mis sur l'objectif de la fusion. Pour définir cette combinaison et l'objectif sous-jacent, il convient de connaître le type de données que l'on cherche à fusionner. Nous reprenons succinctement les types de données proposés par Bloch, Hunter et al Bloch et (Eds) (2001) et repris par Dubois et Prade Dubois et Prade (2004) :

- *les observations*. Elles décrivent le monde d'un point de vue plus ou moins particulier. Il s'agit le plus souvent de données numériques fournies par des capteurs.
- *la connaissance*. Elle décrit la façon dont le monde est *en général*. Il s'agit de données plus subjectives que les observations, et qui sont souvent issues de personnes plutôt que de capteurs.
- *la préférence*. Ce sont des informations subjectives qui décrivent comment *on aimerait* que le monde soit. Il s'agit là aussi de données issues de personnes.
- *les régulations*. Il s'agit d'informations génériques qui sont la plupart du temps énoncées sous forme de lois.

Pour notre application, nous travaillons essentiellement sur des *observations*. Notons que les opérateurs d'agrégation doivent vérifier au moins la condition de monotonie, c'est-à-dire que si les valeurs marginales à fusionner augmentent, l'opérateur de fusion doit augmenter aussi, ou du moins ne pas diminuer Dubois et Prade (2004).
Les opérateurs d'agrégation peuvent être divisés en 4 groupes par rapport à leurs propriétés relatives au minimum (min) et au maximum (max) :

- Les opérateurs *conjonctifs* qui vérifient $f \leq min$. Les normes triangulaires (t-normes) et les copulas Nelsen (1999) appartiennent à cette classe.
- les opérateurs *disjonctifs* qui vérifient $f \geq max$. Les conormes triangulaires appartiennent à ce groupe.
- Les *moyennes* qui vérifient $min \leq f \leq max$. Par exemple les *OWA* Yager (1988)Yager (2004)Zarghami et al. (2008) appartiennent à cette classe.
- Les opérateurs d'agrégation *hybrides*. Ce groupe représente tous les autres opérateurs qui ne peuvent être comparés avec le min ou le max. On peut citer par exemple les

connecteurs mixtes Zimmerman et Zynso (1980) Piera-Carreté et al. (1988), les uninormes Yager et Rybalov (1996), les nullnormes Calvo et al. (2001) et les sommes symétriques Silvert (1979)..

Diverses études s'intéressent aux multiples propriétés des opérateurs comme la non-contradiction, les éléments neutres etc. Dans cet article nous nous intéresserons plus particulièrement à la propriété de renforcement (notion expliquée ci-dessous) de certains opérateurs.

3.2 La notion de renforcement, triple Π et moyenne triple Π

Supposons que pour une classe donnée, un objet présente des degrés d'appartenance marginaux élevés pour tous les attributs considérés. Dans le raisonnement humain, une agrégation de toutes ces informations marginales sera plus élevée que chacun des degrés d'appartenance pris séparément Yager et Rybalov (1998). Dans ce type de raisonnement, les degrés d'appartenance marginaux forts se renforcent mutuellement. Ce comportement est appelé *renforcement positif*.
De façon analogue, si pour une classe donnée, un objet a des degrés d'appartenance marginaux faibles quelque soit l'attribut considéré, alors l'agrégation sera plus faible que la plus faible des valeurs d'appartenance marginales. On parle dans ce cas de *renforcement négatif*.
Le renforcement total est une propriété qui traduit certains aspects du raisonnement humain. L'utilisation d'opérateur ayant cette propriété peut donc être intéressante dans la mesure où l'on cherche un système proche de ce type de raisonnement.

Définition
Un opérateur d'agrégation L dont les arguments sont dans l'intervalle $[0, 1]$, a la propriété de renforcement positif si lorsque tous ses attributs sont affirmatifs (i.e. suprieur ou égaux à 0,5) il vérifie :

$$L(x_1, ..., x_n) \geq \max_i[L(x_i)] \tag{4}$$

De façon similaire, un opérateur d'agrégation L dont les arguments sont dans l'intervalle $[0, 1]$, a la propriété de renforcement négatif si lorsque tous ses attributs sont non-affirmatifs (i.e. inférieur ou égaux à 0,5), il vérifie :

$$L(x_1, ..., x_n) \leq \min_i[L(x_i)] \tag{5}$$

Un opérateur qui possède les deux propriétés est défini comme étant *totalement renforcé* (*full reinforced*).

Les t-normes sont des opérateurs à renforcement négatif ($T(x_1, ..., x_n) \leq \min_i[T(x_i)]$) mais elles ne sont pas à renforcement positif. Par ailleurs les t-conormes sont à renforcement positif ($C(x_1, ..., x_n) \geq \max_i[C(x_i)]$) mais ne possèdent pas de renforcement négatif. On pourrait espérer que des combinaisons de t-normes et t-conormes (comme les connectifs mixtes) soit totalement renforcées mais Yager et Rybalov ont trouvé des contre-exemples qui prouvent que ce n'est pas le cas Yager et Rybalov (1998).
Les moyennes ne sont ni renforcées positivement ni renforcées négativement par définition. En effet, on a pour une moyenne $M(x_1, ..., x_n)$:

$$\min_i(x_i) \leq M(x_1, ..., x_n) \leq \max_i(x_i)$$

Le seul opérateur qui soit, à notre connaissance, totalement renforcé est le triple Π défini par Yager et Rybalov Yager et Rybalov (1998).
Le triple Π est défini comme suit :

$$PI(x_1, ..., x_n) = \frac{\Pi_{j=1}^{n} x_j}{\Pi_{j=1}^{n} x_j + \Pi_{j=1}^{n}(1 - x_j)} \tag{6}$$

Notons que cet opérateur est aussi une somme symétrique Silvert (1979).

La propriété du renforcement total est donc particulièrement intéressante car elle permet d'avoir une bonne modélisation du comportement humain, ce qui est souvent l'objectif de nombreux systèmes de fusion d'informations.

Par ailleurs, la moyenne triple Π est un opérateur moyenne créé à partir de l'opérateur triple Π mais qui est de type moyenne. La moyenne triple Π est une moyenne définie comme suit :

$$MPI(x_1, ..., x_n) = \tag{7}$$

$$= \frac{\Pi_{j=1}^{n}(x_j)^{(1/n)}}{\Pi_{j=1}^{n}(x_j)^{(1/n)} + \Pi_{j=1}^{n}(1 - x_j)^{(1/n)}} \tag{8}$$

$$= \frac{1}{1 + \Pi_{j=1}^{n}\left[\frac{1-x_j}{x_j}\right]^{1/n}}$$

Elle possède des caractéristiques qui présentent une certaine analogie avec celles du triple Π. Ainsi, la moyenne triple Π possède la caractéristique d'être *moyennement renforcée* Emilion et al. (2004)Doncescu et al. (2007) ; on considère la moyenne arithmétique classique, $\frac{1}{n}\sum_{j=1}^{n} x_j$, alors :
Si $\forall j \in 1, ..., n$, on a $x_j \geq 0,5$, on a :

$$MPI(x_1, ..., x_n) \geq \frac{1}{n}\sum_{j=1}^{n} x_j \tag{9}$$

Si $\forall j \in 1, ..., n$, on a $x_j \leq 0,5$, on a :

$$MPI(x_1, ..., x_n) \leq \frac{1}{n}\sum_{j=1}^{n} x_j \tag{10}$$

La moyenne triple Π tente de "concilier" les propriétés d'une moyenne et d'un opérateur renforcé.
Ces deux opérateurs seront testés pour notre application car leur analogie avec le raisonnement humain peut conduire à des résultats pertinents pour l'ordonnancement des groupes de pixels.

4 Résultats expérimentaux

L'étude est réalisée sur une dizaine d'images extraites au hasard d'une base de données. Après avoir réalisé la segmentation, on réalise la fusion d'informations (par rapport aux deux

critères : position spatiale et homogénéité) pour évaluer l'ordre d'importance des groupes de pixels. Pour chacune des dix images, il existe entre trente et quarante groupes réalisés à partir de la propriété colorimétrique, mais nous ne nous intéresserons qu'aux cinq premiers groupes les plus importants. Cinq opérateurs d'agrégation ont été utilisés : la moyenne arithmétique, le maximum, le minimum, le triple Π et la moyenne triple Π. Les résultats fournis par ces cinq opérateurs ont été comparés à l'ordonnancement des groupes proposé par un humain expert en biologie végétale. L'ordonnancement des groupes de pixels proposé par cet expert sera considéré comme la référence pour déterminer la qualité des résultats de chaque opérateur. Les résultats sont présentés dans les tableaux un et deux correspondant à deux images prises au hasard parmi les dix :

Ordre	triple π		Max		Min		Moyenne		moyenne triple π		Humain
	n°	val.	n°	val.	n°	val.	n°	val.	n°	val.	
1	1	0.96	1	0.98	1	0.46	1	0.72	1	0.86	3
2	3	0.86	3	0.97	17	0.45	3	0.70	3	0.84	1
3	2	0.85	5	0.93	3	0.43	2	0.64	2	0.72	2
4	0	0.84	0	0.92	10	0.42	0	0.63	0	0.71	0
5	5	0.65	2	0.84	11	0.40	5	0.60	5	0.70	4

TAB. 1 – *Photographie de Matricaria recutita*

Ordre	triple π		Max		Min		Moyenne		moyenne triple π		Humain
	n°	val.	n°	val.	n°	val.	n°	val.	n°	val.	
1	0	0.98	0	0.98	2	0.55	0	0.74	0	0.87	0
2	1	0.96	1	0.97	0	0.55	1	0.73	1	0.84	1
3	2	0.94	2	0.93	3	0.54	2	0.73	2	0.80	2
4	3	0.90	3	0.91	6	0.54	3	0.71	3	0.75	3
5	4	0.89	4	0.89	11	0.54	6	0.70	4	0.75	4

TAB. 2 – *Photographie de Acalypha alopecuroides*

Critères	Sélection des classes	Ordonnancement des classes
Triple π	74%	34%
Maximum	72%	20%
Minimum	42%	6%
Moyenne	76%	30%
Moyenne Triple π	74%	34%

TAB. 3 – *Résumé de l'étude des critères de fusion sur les images de plantes utilisées*

Les tableaux 1 et 2 montrent deux exemples de résultats pour les fusions des paramètres "position spatiale" et "homogénéité" pour les plantes *Matricaria recutita* et *Acalypha alopecuroides*. Les colonnes "n°" et "val." indiquent respectivement le numéro identifiant d'un

groupe de pixels parmi les cinq premiers agrégats les plus pertinents et la valeur renvoyée par l'opérateur de fusion. Dans le tableau 1, aucune des méthodes ne trouve l'ordonnancement humain des groupes de pixels. L'ensemble des méthodes sauf le Minimum trouve quand même quatre classes sur les cinq retenues par l'expert. Dans le tableau 2, le triple Π et la moyenne triple Π, tout comme le maximum, se détachent des autres méthodes en ordonnant les groupes comme l'expert. Dans le tableau 3, on donne des résultats quantitatifs et qualitatifs pour les dix images testées. L'estimation quantitative consiste à évaluer pour chaque opérateur la présence de groupe de pixels corrects (donnés par l'expert) quelque soit leur position dans les cinq premiers groupes évalués. L'estimation qualitative consiste à évaluer le nombre de groupes de pixels correctement positionnés pour chaque opérateur.

Ces résultats montrent que quatre méthodes sur les cinq étudiées se démarquent. Le triple Π, la moyenne triple Π, le maximum et la moyenne arithmétique affichent plus de 70% de réussite quantitative. Au niveau qualitatif, le triple Π, la moyenne arithmétique et la moyenne triple Π se démarquent avec plus de 30 % de réussite avec un léger avantage pour les méthodes triple Π. Le renforcement effectué par les critères en triple Π est visible sur les valeurs obtenues avec un avantage pour le triple Π, la moyenne triple Π et la moyenne arithmétique pour finir.

5 Conclusion

Dans cet article, nous avons présenté une méthode de segmentation pour effectuer la reconnaissance de végétaux dans des images. Nous avons proposé d'utiliser des opérateurs de fusions pour améliorer la structuration de données au niveau des agrégats de pixels fournis par la segmentation. La structuration de données dans une image permet de focaliser les traitements sur des éléments plus pertinents que d'autres. On peut prendre l'exemple pour une fleur où il est plus intéressant d'observer son bulbe et ses pétales avant de regarder la nature de sa tige ou de ses feuilles. Ce processus peut accélérer la reconnaissance ou la guider, le but étant d'éviter de noyer le système de reconnaissance dans une multitude de données hétérogènes. On utilise aussi l'ordonnancement des classes afin d'ajouter de la sémantique en combinant des critères appartenant à des classes différentes. Cette opération se fait à partir de caractéristiques spatiales qu'il est nécessaire alors de fusionner. Parmi l'ensemble des opérateurs de fusion étudiés, le triple Π et la moyenne triple Π se démarquent en proposant un classement proche de celui de l'expert. Des tests supplémentaires permettront de mieux comprendre l'influence de ces opérateurs sur la qualité de la structuration.

Références

Aitkenhead, M., I. Dalgetty, C. Mullins, A. McDonald, et N. Strachan (2003). Weed and crop discrimination using image analysis and artificial intelligence methods. *Computers and Electronics in Agriculture 39*, 157–171.

BINET, J. et J. P. BRUNEL (1967). *Physiologie végétale*.

Bloch, I. et A. H. (Eds) (2001). Fusion : General concepts and characteristics. *International Journal of Intelligent Systems 16*, 1107–1134.

Calvo, T., B. D. Baets, et J. Fodor (2001). The functional equations of franck and alsina for uninorms and nullnorms. *Fuzzy Sets and Systems 120*, 385–394.

Camargoa, A. et J. Smith (2009). Image pattern classification for the identification of disease causing agents in plants. *Computers and Electronics in Agriculture 66*, 121–125.

D. Comaniciu, P. M. (May, 2002). Mean shift : A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24 no. 5*, 603–619.

Doncescu, A., S. Régis, K. Inoue, et R. Emilion (2007). Analysis of new aggregation operators : Mean 3pi. *Journal of Advanced Computational Intelligence and Intelligent Informatics 11*(6).

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

Emilion, R., S. Régis, A. Méril, et A. Doncescu (2004). Un nouvel opérateur de type moyenne pour la fusion de données. In *Logique Floue et Application (LFA)*, Nantes.

Nagau, J. et J. L. Henry (2010). An optimal global methode for classification of color pixels. In *FCSIS, Pologne (accepted)*.

Nelsen, R. (1999). *An introduction to Copulas*, Volume 139 of *Lecture Notes in Statistics*. New York : Springer.

Philipp, I. et T. Rath (2002). Improving plant discrimination in image processing by use of different colour space transformations. *Computers and Electronics in Agriculture 35*, 1–15.

Piera-Carreté, N., J. Aguilar-Martin, et M. Sanchez (1988). Mixed connectives between min and max. In *8th Inter. Symp. on Multiple Valued Logic*.

Silvert, W. (1979). Symmetric summation : A class of operations on fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics 9*(10), 657–659.

Wang, X.-F., D.-S. Huang, J.-X. Dua, H. Xu, et L. Heutte (2008). Classification of plant leaf images with complicated background. *Applied Mathematics and Computation 205*, 916–926.

Wertheimer, M. (1958). Principles of perceptual organization. *readings in Perception*, 115–135.

Yager, R. (1988). On ordered weighted averaging operators in multi-criteria decision making. *IEEE Transaction on Systems, Man, and Cybernetics*, 183–190.

Yager, R. (2004). Owa aggregation over a continuous interval argument with applications to decision making. *IEEE Transactions on Systems, Man and Cybernetics-Part B : Cybernetics 34*(5), 1952–1963.

Yager, R. et A. Rybalov (1996). Uninorm aggregation operators. *Fuzzy Sets and Systems 80*, 111–120.

Yager, R. et A. Rybalov (1998). Full reinforcement operators in aggregation techniques. *IEEE Transactions on Systems, Man, Cybernetics-Part B : Cybernetics 28*(6).

Zarghami, M., M. Szidarovszky, et R. Ardakanian (2008). Owa aggregation over a continuous interval argument with applications to decision making. *IEEE Transactions on Systems, Man and Cybernetics-Part B : Cybernetics 38*(2), 547–552.

Zimmerman, H. et P. Zynso (1980). Latent connectives in human decision making. *Fuzzy Sets and Systems 4*, 37–51.

Summary

In this paper, we use a segmentation method applied to the recognition of plants in images. Segmentation provides groups of pixels, and we propose to use aggregation operators to direct these groups to improve the performance of the recognition system. In particular, we use an operator called totally reinforced triple Π. The first tests were made and show that the use of triple Π improves outcomes for the structuring of groups of pixels.

Visualisation de données spatiotemporelles imprécises : application en archéologie

Cyril de Runz*, Frédéric Blanchard*
Philippe Vautrot*, Eric Desjardin*

* CReSTIC
IUT de Reims Châlons Charleville
Rue des Crayères, BP 1035, 51687 Reims Cedex 2
cyril.de-runz@univ-reims.fr

Résumé. Dans cet article, nous proposons d'exploiter une technique spécifique d'exploration visuelle d'un ensemble d'objets archéologiques dont les composantes spatiales et temporelles sont représentées par des ensembles flous convexes et normalisés. Pour cela, en nous basant sur la définition de vecteurs multidimensionnels issus de défuzzifications ou de comparaison entre deux nombres flous, nous construisons une image couleur dans laquelle chaque pixel représente un objet. L'image couleur donne un rendu synthétique de l'information permettant à l'utilisateur de l'observer et de l'analyser.

1 Introduction

L'étude intuitive et visuelle de l'ensemble des données associées aux objets d'une base de données archéologiques est complexe dans les systèmes d'information géographique (SIG). En effet, bien que l'on puisse actuellement avoir une légende combinant un certain nombre d'attributs, ce nombre est limité. L'exploration visuelle nécessite alors d'utiliser une technique spécifique de visualisation. Pour cela, il faut considérer ces composantes de l'information archéologique comme une collection de données multidimensionnelles (Guptill, 2005).

L'approche générale de l'exploration visuelle de grands volumes de données multicomposantes consiste à présenter un résumé en image de ces informations à l'instar de la démarche proposée par Keim (2000). Afin de visualiser la plus grande quantité d'information possible, nous utilisons une technique de visualisation qui construit une image couleur à partir de ces informations et qui fut introduite dans Blanchard et al. (2005).

Cette technique orientée-pixel consiste à représenter une collection par une image où chaque pixel correspond à une et une seule donnée. Les couleurs des pixels sont déterminées « objectivement [1] ». La couleur et la spatialisation fournissent alors une image qui constitue un résumé des données en permettant de voir de manière intuitive les principales structures. Ce travail a montré son efficacité sur des bases de données classiques (Blanchard et al., 2005).

Pour cela, les données sont préalablement réduites par une Analyse en Composantes Principales à des données tridimensionnelles regroupant les trois composantes principales. En uti-

1. La couleur est calculée à partir des données sans utilisation d'une échelle de couleurs.

lisant ces données réduites, on associe un pixel de l'image couleur à chaque donnée ; la couleur est affectée objectivement et calculée par la transformée inverse de celle proposée dans (Ohta et al., 1980). Afin de regrouper au maximum, dans l'image de visualisation, les données proches selon leurs trois premières composantes principales, les pixels représentant les données sont organisés spatialement à l'aide d'une courbe de remplissage dite de Peano-Hilbert (Moon et al., 2001).

Dans cet article, deux processus d'exploration visuelle sont étudiés. Le premier a pour but de visualiser les composantes temporelles de l'information archéologique. Ces informations sont difficiles à visualiser dans le cas de grands volumes de données et rendent presque impossible l'exploration intuitive et directe de ces composantes. Dans le second processus, l'objectif est de visualiser les dissimilarités à un objet sélectionné dans la base de données.

Dans le cas du projet SIGRem (Desjardin et de Runz, 2009), les objets de *BDFRues* représentent des tronçons de rues romaines trouvées à Reims. Les composantes temporelles, spatiales et orientationnelles des données sont modélisées en tenant compte de leurs imprécisions par des ensembles flous (de Runz et al., 2008). Il faut donc pré-traiter l'information afin d'en dégager des évaluations quantitatives qui seront dès lors considérées comme des vecteurs multidimensionnels. Ainsi, la technique proposée est appliquée aux vecteurs multidimensionnels pour visualiser les objets archéologiques.

Nous proposons, dans cet article, de visualiser les composantes temporelles des objets de *BDFRues*. Pour cela, comme énoncé précédemment, il est nécessaire de quantifier les données avant même de lancer le processus de visualisation. Dans ce but, nous construisons un vecteur d'évaluation pour chaque nombre flou représentant la période d'activité d'un objet archéologique. Les différentes valeurs de ces vecteurs seront déterminées par différents estimateurs de quantités floues.

Pour visualiser les dissimilarités entre objets archéologiques, les vecteurs multidimensionnels nécessaires sont issus de trois indices de dissimilarité entre objets archéologiques présentés dans de Runz et al. (2008). Ces indices de dissimilarités permettent d'évaluer les dissimilarités temporelles, d'orientation et de localisation entre objets archéologiques. Pour un objet sélectionné, l'image couleur de visualisation regroupe alors spatialement, autour de sa représentation pixel, les pixels associés aux objets qui lui sont le moins dissimilaires d'un point de vue spatial, directionnel et temporel.

Nous présenterons dans la section 2 le contexte applicatif. La section 3 sera dédiée à la description de la technique de visualisation choisie. La section 4 exposera le processus de visualisation des composantes temporelles associées aux tronçons de rues trouvés à Reims datant de l'époque romaine. Dans la section 5, nous proposerons de visualiser la dissimilarité des objets de la base vis-à-vis d'un objet présélectionné.

2 Projet SIGRem

Dans la problématique de la valorisation et de la gestion du patrimoine archéologique, la démarche développée par l'Université de Reims Champagne Ardenne, l'Institut National de Recherches Archéologiques Préventives et Ministère de la Culture et de la Communication dans le Centre Interinstitutionnel de Recherches Archéologiques de Reims peut être considérée comme novatrice par l'intégration de la géomatique au cœur de l'analyse urbaine et régionale.

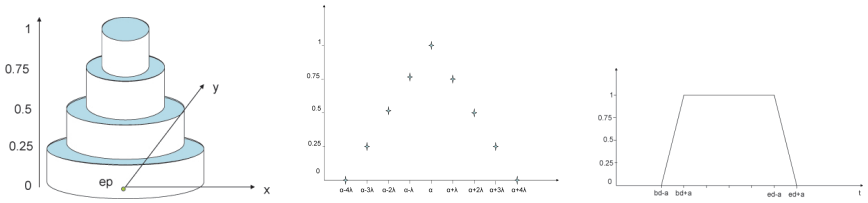


FIG. 1 – *Modèles flous pour la localisation, l'orientation et les périodes d'activité des tronçons de rues romaines*

Au-delà de l'élaboration de la cartographie archéologique de la cité des Rèmes [2], le projet SIGRem, soutenu par la région Champagne Ardenne, l'État et la ville de Reims, et cadre applicatif de ce travail, porte sur la mise en place d'un SIG pluridisciplinaire intégrant les données archéologiques recueillies depuis les 30 dernières années. Dans cet article, nous proposons d'appliquer le processus exploratoire proposé sur la base de données *BDFRues*, partie intégrante du projet SIGRem. Cette base est dédiée aux éléments de rues romaines à Reims. Elle est constituée de 33 objets à l'heure actuelle. Son enrichissement est en cours. Les tronçons de rues sont caractérisés par des points ayant une orientation et une période d'activité.

La datation de la période d'activité des objets est généralement issue d'interprétations ou d'estimations dépendantes de l'environnement de la découverte (lieux de fouilles, stratigraphie, comparaison aux objets se situant dans la même pièce…). Elle est donc largement imprécise. Le géoréférencement est lui aussi sujet à de l'imprécision liée à différents facteurs : positionnement du point de fouilles, position par rapport à la route, référentiel utilisé, mouvement de terrain. L'orientation de la route est aussi à redéfinir dans ce cadre. En effet, l'orientation est notamment dépendante de la technique d'estimation utilisée à l'époque de la fouille.

Nous représentons les orientations, les périodes d'activité, et les localisations par des ensembles flous convexes et normalisés à savoir respectivement par des nombres flous, des intervalles flous et des ensembles flous spatiaux (2D). On peut ainsi prendre en compte cette incertitude (voir Figure 1).

Afin d'obtenir un rendu synthétique visuel et pertinent de ces données, nous proposons d'exploiter une technique de visualisation orientée-pixel présentée dans la section suivante.

3 Visualisation de données par une image couleur

Afin de fournir une image couleur des objets, nous nous intéressons plus particulièrement à la visualisation statique et plane de données multidimensionnelles quantitatives. L'utilisation des outils de visualisation se heurte alors à deux difficultés principales : la dimension et l'effectif des échantillons de données.

La dimension de l'espace dans lequel se situent les données peut être importante (la dimensionnalité peut être supérieure à 100 dans certains cas). Ceci conduit à un ensemble de phénomènes dissimulant l'information pertinente que l'on recherche. Ces phénomènes sont

2. Cité des Rèmes : Reims et ses environs à l'époque romaine

connus sous le nom de « malédiction de la dimensionnalité » (Donoho, 2000). Par ailleurs, l'effectif de l'échantillon peut être considérable : il peut dépasser le million d'individus. Les techniques de visualisation ont alors tendance à masquer l'information pertinente du fait de cet effectif.

Dans ce cadre, la méthode de Blanchard et al. (2005) est une approche orientée-pixel qui résume les données, et en fournit un résumé sous forme d'une image couleur. Cette approche permet de s'affranchir de la première difficulté par la réduction de la dimensionnalité et de la seconde par l'association d'un pixel à chaque donnée permettant ainsi de visualiser autant de données qu'il y a de pixels affichables.

3.1 Réduction de la dimensionnalité

L'analyse de données multidimensionnelles nécessite une réduction de dimensionnalité pour des raisons pratiques et théoriques (représentations des données, la malédiction de la dimensionnalité). Dans l'approche de visualisation présentée ici, les données sont dans un espace initial de dimension supérieure à trois.

Une approche classique, simple et généralement efficace de la réduction de dimensionnalité est utilisée : on conserve les trois premières composantes générées par une Analyse en Composantes Principales. Une revue des techniques d'ACP est proposée dans Hyvärinen (1999).

Le principe est de projeter les données dans un sous-espace de dimension trois, les axes de projection étant orthogonaux et décorrélés. L'avantage de l'ACP est de déterminer les composantes, itérativement, par ordre décroissant de l'information portée. Ainsi, la première composante contient plus d'information que la seconde, qui en contient plus que la troisième, et ainsi de suite. Ainsi, en réduisant les données de dimension $n > 3$ à des données de dimension 3 par la sélection des trois premières composantes (C_1, C_2, C_3) de l'ACP, on maximise l'information contenue dans ces trois composantes.

Les données réduites guident ensuite le processus de visualisation. À chaque donnée de dimension trois est affecté un pixel que l'on place spatialement dans l'image à l'aide d'une courbe de Peano-Hilbert.

3.2 Remplissage de l'image de visualisation

Pour construire une image d'un échantillon de données, chaque donnée est associée à un pixel de l'image. Cette approche de la visualisation orientée-pixel permet de représenter des échantillons de grande taille (Keim, 2000). La construction de l'image consiste à déterminer les coordonnées des pixels (i.e. des représentations des données) dans l'espace image.

Si les pixels sont placés arbitrairement ou dispersés dans l'image, il devient difficile d'effectuer des rapprochements entre les données. Pour que l'image soit un outil de visualisation efficace, lisible au premier coup d'oeil de manière très intuitive, il faut que les proximités entre données soient faciles à déterminer. Pour cela, il faut que, dans l'image résultat, des données similaires soient spatialement très proches. La construction de l'image s'effectue en deux étapes : les pixels (i.e. les représentations des données) seront d'abord triés de manière à former une suite de pixels successifs ; ensuite cette « ligne » sera utilisée pour remplir l'image.

Ainsi la première étape du remplissage de l'image de visualisation consiste à trier les pixels associés aux données afin de produire une liste de pixels. Le tri des pixels est effectué en utilisant les résultats de la réduction de dimension des données. Les trois composantes obtenues

(C_1, C_2, C_3) donnent trois clefs pour effectuer un tri sur l'ensemble des données de l'échantillon. En effet, les composantes issues de l'ACP sont classées selon la quantité d'information qu'elles fournissent. Ainsi, le tri se fera en majeur sur la première composante car elle contient le plus d'information, puis sur la seconde composante et enfin en mineur sur la troisième composante. La récupération des trois premières composantes de l'ACP et leur tri sont d'ordre polynomial.

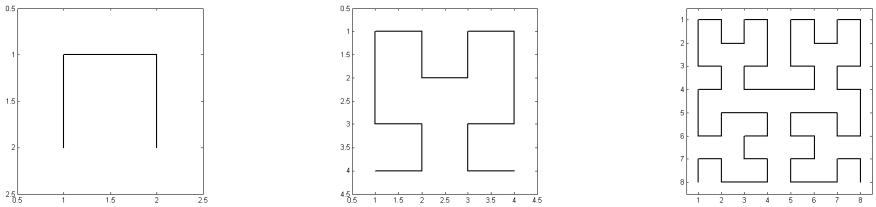


FIG. 2 – *Étapes de la construction d'une courbe de Peano-Hilbert*

L'étape suivante consiste à remplir l'image avec cette liste de pixels successifs. La courbe de Peano-Hilbert constitue le moyen le plus classique pour effectuer cette construction (Moon et al., 2001) (voir sur la Figure 2 la description de la procédure récursive de construction d'une telle courbe). Le principal avantage de cette courbe est de préserver au mieux les regroupements des classes de données.

Avec ces deux étapes de tri des données puis le remplissage de l'image par une courbe de Peano-Hilbert, on évite de disperser les pixels dans l'image construite. Cette approche tend à préserver la cohérence spatiale des données permettant ainsi une visualisation très intuitive des échantillons de données. Il faut maintenant déterminer la couleur de chaque pixel en fonction des valeurs contenues dans la donnée réduite associée au pixel.

3.3 À propos de la couleur

En imagerie, la couleur est souvent définie par un triplet (R,V,B) de trois valeurs Rouge, Vert et Bleu, codées sur 8 bits (entre 0 et 255). Après avoir réduit la dimension de l'échantillon, chaque donnée est représentée par un triplet (C_1, C_2, C_3). Cependant, les trois composantes principales ne peuvent former directement le triplet RVB.

La technique de visualisation proposée se base pour l'affectation de la couleur sur l'étude statistique de la couleur de Ohta et al. (1980). Ces derniers proposent d'approximer l'ACP d'une image couleur par une transformation linéaire. A partir des données (R, V, B) des pixels couleur, ils calculent les triplets (C_1, C_2, C_3) qui approximent les trois composantes de l'ACP (Ohta et al., 1980). La technique de visualisation exposée ici cherche à associer une couleur à chaque triplet (C_1, C_2, C_3). Par la transformation inverse de celle de Ohta *et al.*, elle associe à chaque donnée une couleur (R, V, B). Ainsi les composantes couleurs R, V et B sont liées aux composantes C_1, C_2 et C_3 par la relation suivante :

$$\left\{ \begin{array}{lcl} R & = & (6 \times C_1 + 3 \times C_2 - 2 \times C_3)/6 \\ V & = & (3 \times C_1 + 2 \times C_3)/3 \\ B & = & (6 \times C_1 - 3 \times C_2 - 2 \times C_3)/6 \end{array} \right.$$

Ce type d'approche présente l'avantage d'être objectif et non supervisé contrairement aux méthodes traditionnelles de détermination de palettes ou d'échelles de couleurs. Cette approche de la couleur dépend de l'échantillon de données. Si l'échantillon change, les couleurs changent. Elle propose un résumé coloré associé à un échantillon. Cette technique de visualisation a été appliquée avec succès à des bases de données classiques, des données simulées et à des images de fluorescence X.

Notre idée est d'utiliser cette technique pour visualiser les objets selon l'information des composantes temporelles, et les dissimilarités des objets archéologiques en tenant compte de leurs imperfections.

4 Visualisation des objets selon leurs périodes d'activité par une image couleur

Les données archéologiques sont spatiotemporelles et imparfaites. Afin de donner une lecture intuitive des données, il est nécessaire de les visualiser. Lorsqu'elles sont stockées dans une base de données associée à un SIG, une visualisation classique consiste à la production d'une ou plusieurs cartes thématiques. Cependant ces cartes ne permettent pas de rapprocher spatialement les objets aux localisations éloignées, et les couleurs dépendent d'une échelle fixée (d'une légende). Afin de rapprocher les données archéologiques selon le temps en prenant en considération l'imperfection, il est nécessaire d'utiliser une autre approche pour la visualisation.

La visualisation d'un grand nombre de données spatiotemporelles imprécises est difficile. En effet, dans le cadre de données représentées par des quantités floues, représenter l'ensemble des fonctions d'appartenance sur un même repère complique la lecture des quantités floues à considérer. Cependant, de nombreuses techniques de visualisation de données multi-composantes, à l'instar de celle présentée précédemment, utilisent les informations quantitatives contenues dans ces données afin de les visualiser. Une solution à la visualisation de quantités floues consiste à visualiser les données par l'intermédiaire d'évaluations des différentes quantités floues. Ces évaluations forment des données quantitatives multi-composantes décrivant l'information à visualiser.

L'analyse de données floues nécessite généralement une défuzzification des données. La défuzzification est le processus qui amène à produire un résultat quantifiable à partir de données floues. Ainsi, par exemple, les méthodes de comparaison de quantités floues rangent le plus souvent celles-ci par le biais d'évaluations (Wang et Kerre, 2001).

Le principe de la visualisation consiste en premier lieu à décrire une quantité floue par plusieurs évaluations quantitatives obtenues avec différentes méthodes de défuzzification. Une quantité floue est alors représentée par un vecteur d'évaluations. Les données sont ensuite visualisées en utilisant les vecteurs d'évaluations.

Dans *BDFRues*, les objets archéologiques sont temporellement modélisés par des nombres flous. Le but de la visualisation est alors d'associer à chaque objet archéologique un pixel couleur de l'image résultat en fonction du nombre flou représentant sa période d'activité.

L'objectif ici est d'abord d'évaluer chaque nombre flou séparément puis de le positionner par rapport aux autres par la technique de visualisation précédente via ses évaluations. Les

évaluations des nombres flous doivent donc ne prendre en compte que le nombre flou devant être visualisé.

4.1 Méthodes de défuzzification des nombres flous

Les méthodes de défuzzification présentées ici ne considèrent que le nombre flou à évaluer. VanLeekwijck et Kerre (1999) les séparent en trois classes. Bien que chaque méthode ait ses particularités, les classes proposées suggèrent à des utilisations différentes. Le choix de l'utilisation de l'une de ces méthodes dépend donc fortement de l'analyse voulue.

Les méthodes de type maxima et les méthodes dérivées forment la première classe. Elles sélectionnent un élément du cœur de la quantité à évaluer comme valeur de défuzzification. Selon Van Leekwijck et Kerre, l'utilisation première de ces méthodes se situe dans le cadre des systèmes de connaissances floues. De plus, ces méthodes sont efficaces d'un point de vue calculatoire.

Dans la seconde classe, les opérateurs de défuzzification convertissent d'abord les fonctions d'appartenance en distribution de probabilités afin de calculer la valeur espérée. Au regard du manque de fondement théorique de ces conversions, la principale raison de leur utilisation est que ces méthodes vérifient l'hypothèse de continuité, essentielle pour les contrôleurs flous.

Dans la troisième classe, les méthodes utilisent les aires sous les fonctions d'appartenance afin d'évaluer les quantités floues. Comme pour les méthodes de la seconde classe, elles sont principalement dédiées au contrôle flou.

Afin d'explorer visuellement les objets archéologiques selon les représentations de leurs périodes d'activité, nous souhaitons définir pour chacun des nombres flous, modélisant la période d'activité de ces objets, un vecteur d'évaluation le représentant dans le processus de visualisation.

4.2 Construction du vecteur multidimensionnel d'évaluation de la représentation d'une période d'activité

Afin de construire simplement un vecteur avec des méthodes de chaque classe, nous proposons de n'utiliser que des méthodes de défuzzification sélectionnées parmi celles présentées dans VanLeekwijck et Kerre (1999) afin qu'elles ne prennent pas en considération de paramètre autre que l'ensemble à considérer, l'objectif final étant de proposer une visualisation temporelle non supervisée.

Pour la première classe, nous choisissons les méthodes suivantes : le "first of maximum" (FOM) qui retourne le plus petit élément du cœur d'un nombre flou ; le "last of maximum" (LOM) qui renvoie le plus grand élément du cœur d'un nombre flou ; le "middle of maximum" (MOM) qui permet de récupérer l'élément médian du cœur d'un nombre flou.

En ce qui concerne la seconde classe, nous sélectionnons les méthodes suivantes : le "center of gravity" (COG) qui donne en sortie le centre de gravité de la fonction d'appartenance d'un nombre flou ; le "mean of maxima' (MeOM) qui calcule la moyenne du cœur d'un nombre flou ; le "mean of support" (MeOS) par lequel on obtient la moyenne du support d'un nombre flou.

Enfin, pour la dernière classe, nous prenons le "center of area" (COA) car celui-ci permet d'obtenir l'élément du support minimisant la différence des aires de la fonction d'appartenance avant et après ce dernier.

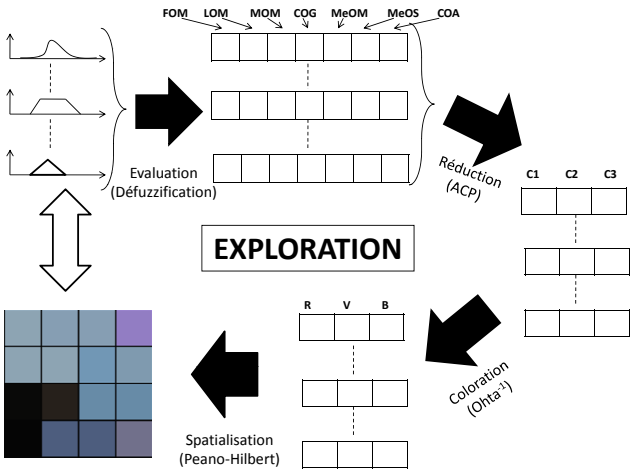


FIG. 3 – *Visualisation des objets selon les représentations floues des périodes d'activité — schéma récapitulatif*

Nous associons donc à chaque période d'activité un vecteur d'évaluation de dimension 7. C'est par le prisme de ce vecteur que nous explorons les données. Pour cela, nous utilisons l'ensemble des vecteurs en entrée du processus de visualisation de Blanchard et Herbin. Le processus général de l'exploration est présenté dans la figure 3.

4.3 Visualisation des objets selon les représentations de leurs périodes d'activité

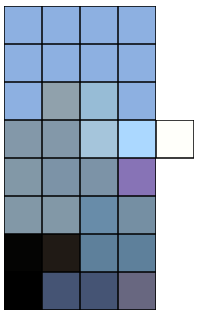


FIG. 4 – *Visualisation des objets de BDFRues par une image couleur selon les représentations de leurs périodes d'activité*

L'image résultat est présentée sur la figure 4. Elle contient 33 pixels en couleurs. Chaque pixel représente un objet selon l'ensemble flou associé à la période d'activité de l'objet. L'organisation spatiale et l'information couleur des pixels permettent d'observer de façon immédiate des informations de structuration de cet ensemble de périodes. Cette image suggère des regroupements des données par couleurs semblables.

Les regroupements observés correspondent à des objets ayant des représentations de leurs périodes d'activité de profils proches. En effet, les objets, dont les représentations des périodes d'activité ont les plus larges supports, sont visualisés par des pixels de couleur bleue claire (haut de l'image de visualisation), tandis que ceux dont les cœurs des représentations des périodes d'activité sont de cardinalité moyenne, sont coloriés dans les gris (milieu de l'image).

Par ailleurs, les composantes principales calculées sur l'ensemble des vecteurs décrivant les représentations des périodes d'activité des objets issus de *BDFRues* permettent d'expliquer plus de 99% de la variance totale. Cette opération de projection conserve donc la quasi totalité de l'information apportée par les différentes évaluations. La visualisation porte donc sur l'essentiel de l'information temporelle contenue dans *BDFRues*.

Ainsi, l'image couleur résultant de la visualisation permet une lecture intuitive — par proximité spatiale et de couleur — d'un grand nombre d'objets archéologiques selon la proximité entre les représentations de leurs périodes d'activité.

La section suivante porte sur une visualisation analogue (par le biais d'un vecteur d'évaluation) des dissimilarités entre les objets archéologiques de la base et un objet sélectionné.

5 Visualisation des dissimilarités à un objet sélectionné

L'objectif de ce processus exploratoire est de visualiser par une image couleur les objets selon leur dissimilarité à un objet sélectionné. On utilise pour cela les mêmes indices de dissimilarité que dans de Runz et al. (2008) en termes d'orientation, de localisation, de période d'activité, c'est à dire D_{date}, D_{orien} et D_{loc}.

5.1 Construction du vecteur multidimensionnel d'évaluation des dissimilarités à un objet sélectionné

Nous proposons d'utiliser une distance classique (Grzegorzewski, 1998) entre nombres et/ou intervalles flous comme mesure de dissimilarité. Soit F et G deux nombres et/ou intervalles flous, soit $F_{\alpha-}$ (resp. $G_{\alpha-}$) et $F_{\alpha+}$ (resp. $G_{\alpha+}$) les bornes inférieure et supérieure de l'α-coupe F_α de F (resp G_α de G), alors la distance entre F et G est obtenue par :

$$D(F,G) = \int_0^1 |F_{\alpha-} - G_{\alpha-}| + |F_{\alpha+} - G_{\alpha+}| d\alpha.$$

Nous utiliserons cette mesure pour le calcul de la dissimilarité d'orientations (D_{orien}) et de périodes d'activité entre éléments (D_{date}).

Pour le calcul de la dissimilarité de localisation, en raison du caractère cylindrique de la fonction d'appartenance des ensembles flous spatiaux associés aux données, nous calculons la mesure de dissimilarité D_{loc} à partir de leurs projections floues sur le plan passant par les centres des localisations (voir Figure 5).

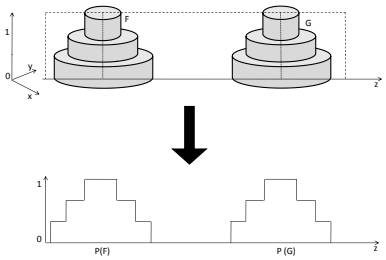


FIG. 5 – *Projections pour le calcul de dissimilarité des localisations*

Dans *BDFRues*, une fois l'objet A_j sélectionné, le vecteur d'évaluation $v_{A_j}(A_i)$ de chaque objet A_i à évaluer est déterminé par les dissimilarités de cet objet avec A_i. Ainsi, $v_{A_j}(A_i)$ est défini de la manière suivante :

$$v_{A_j}(A_i) = (D_{date}(A_j, A_i), D_{orien}(A_j, A_i), D_{loc}(A_j, A_i)).$$

Ce vecteur contient trois informations *a priori* décorrélées — D_{date}, D_{orien} et D_{loc}.

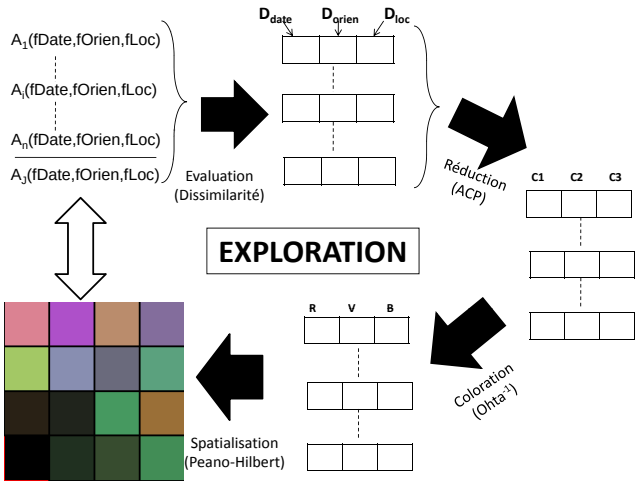


FIG. 6 – *Visualisation des objets selon leurs dissimilarités à un objet sélectionné — schéma récapitulatif*

Nous proposons de visualiser la dissimilarité des objets à un objet sélectionné dans la base, en donnant en entrée du processus de visualisation ces vecteurs de dimension 3. Le processus exploratoire est présenté dans la figure 6.

5.2 Visualisation des dissimilarités

Le processus de visualisation des dissimilarités des objets à un objet sélectionné donne la Figure 7.

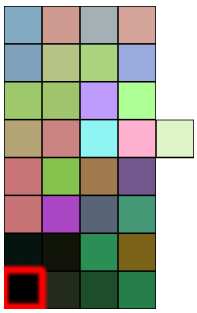


FIG. 7 – *Visualisation de la dissimilarité des objets à un objet sélectionné (contour rouge) par une image couleur*

Comme cette application prend à la fois en compte les dissimilarités d'orientation, de localisation et de datation, celle-ci ne fait plus apparaître les classes observées dans la visualisation strictement temporelle.

Les objets les plus similaires à l'objet sélectionné sont dans l'image les plus proches spatialement de celui sélectionné (de contour rouge dans la figure). Il y a donc *a priori* trois objets très similaires, en couleur sombre, à celui sélectionné. En effet, de par le fait que leurs dissimilarités vis-à-vis de l'objet sélectionné est faible, cela se traduit par une faible disparité et donc une faible variance. Les composantes couleurs des pixels associés aux dits objets auront donc des valeurs faibles ce qui donne une teinte proche du noir.

6 Conclusion

Nous avons présenté dans cet article une méthode originale d'exploration visuelle intuitive d'un ensemble d'objets archéologiques dont les composantes spatiales et temporelles sont représentées par des ensembles flous convexes et normalisés. Cette méthode s'est basée sur la construction de vecteurs dont les valeurs furent obtenues soit par plusieurs défuzzifications des représentations de la composante observée, soit par calcul des indices de dissimilarité des objets à un objet en entrée. L'étape de visualisation a consisté à affecter à chaque objet archéologique une couleur pour obtenir des pixels que l'on organise spatialement dans une image. Dans ce but, nous avons réduit les vecteurs d'évaluations par une ACP à des vecteurs de dimension 3. Par la transformée inverse de celle d'Ohta et al. (1980), nous avons déterminé les couleurs des pixels représentant les objets. L'image fut alors construite en utilisant une courbe de Peano-Hilbert.

Cette visualisation est strictement exploratoire. Elle permet de faire des rapprochements entre données et de les regrouper pour aider à les interpréter. C'est un outil qui présente d'autant plus d'intérêt que le nombre de données augmente (il offre la possibilité de visualiser plusieurs

millions de données). L'image résultante fournit une carte synthétique de la base archéologique étudiée en fonction de l'objectif du processus exploratoire. Cette image peut être considérée comme une légende organisée de l'information multidimensionnelle visualisée qui associe à chaque objet une couleur de manière objective.

Références

Blanchard, F., M. Herbin, et L. Lucas (2005). A New Pixel-Oriented Visualization Technique Through Color Image. *Information Visualization 4*(4), 257–265.

de Runz, C., F. Blanchard, E. Desjardin, et M. Herbin (2008). Fouilles archéologiques : à la recherche d'éléments représentatifs. In *Atelier Fouilles de Données Complexes - Conférence Extraction et Gestion des Connaissances - EGC'08*, Sophia Antipolis, France, pp. 95–103.

Desjardin, E. et C. de Runz (2009). Gissar : de la saisie de fouilles à l'analyse spatiotemporelle en archéologie. In *Spatial Analysis and GEOmatics*, Paris, France.

Donoho, D. L. (2000). High-Dimensional Data Analysis : The Curses and Blessings of Dimensionality. In *AMS Conference Mathematical Challenges of the 21st Century*.

Grzegorzewski, P. (1998). Metrics and orders in space of fuzzy numbers. *Fuzzy Sets and Systems 97*, 83–94.

Guptill, S. C. (2005). Metadata and data catalogues. In P. A. Longley, M. F. Goodchild, D. J. Maguire, et D. W. Rhind (Eds.), *Geographical Information Systems. Principles, Techniques, Management and Applications*, Volume 2, Chapter 49, pp. 677–692. Wiley. Seconde Edition.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys 2*, 94–128.

Keim, D. A. (2000). Designing Pixel-oriented Visualization Techniques : Theory and Applications. *IEEE Transaction on Visualization and Computer Graphics (TVCG) 6*(1), 59–78.

Moon, B., H. V. Jagadish, C. Faloutsos, et J. H. Saltz (2001). Analysis of the Clustering Properties of the Hilbert Space-Filling Curve. *IEEE Transactions on Knowledge and Data Engineering 13*(1), 124–141.

Ohta, Y., T. Kanade, et T. Sakai (1980). Color Information for Region Segmentation. *Computer Graphics and Image Processing 13*, 222–241.

VanLeekwijck, W. et E. E. Kerre (1999). Defuzzification : criteria and classification. *Fuzzy Sets and Systems 108*, 159–178.

Wang, X. et E. E. Kerre (2001). Reasonable properties for the ordering of fuzzy quantities (I). *Fuzzy Sets anf Systems 118*, 375–385.

Summary

In this article, we use a specific technique for the visualization of an archaeological dataset in which object components are modeled using normalized convex fuzzy sets. In order to build a color image of data, we use definition of multidimensional vector for each object. The color image gives us a resume of information allowing users to observe and analyze it graphically.

Étude de données multisources par simulation de capteurs et clustering collaboratif

Germain Forestier*, Cédric Wemmert*, Pierre Gançarski*

*Université de Strasbourg - LSIIT - CNRS - UMR 7005
Pôle API, Bd Sébastien Brant - 67412 Illkirch, France
{forestier,wemmert,gancarski}@unistra.fr

Résumé. Depuis quelques années les données issues de capteur satellitaire deviennent de plus en plus accessibles. Différents systèmes satellitaires sont maintenant disponibles et produisent une masse de données importante utilisée pour l'observation de la Terre. Pour mieux comprendre la complexité de la surface terrestre, il devient courant d'utiliser plusieurs données provenant de capteurs différents. Cependant, il est souvent difficile de prédire le gain potentiel d'un système multisource avant de réellement acquérir les données. Dans cet article nous présentons une approche par simulation qui permet de créer différentes vues de données satellitaire à partir des caractéristiques des capteurs. Ces différentes vues sont ensuite étudiées à travers des algorithmes de classification supervisée, d'outils statistiques et de clustering collaboratif pour évaluer l'intérêt d'utiliser ces données conjointement. Les premières expériences permettent de mettre en avant des couples de capteurs pouvant tirer partie de cette utilisation multisource.

1 Introduction

Un nombre important de capteurs satellitaires sont disponibles pour capturer des images de télédétection de la surface terrestre. Ces données de télédétection sont utilisées intensivement pour étudier la Terre et les systèmes satellitaires sont utilisés pour collecter des données dans les domaines de l'agriculture, la production de nourriture, la géologie, la prospection pétrolière, l'exploration minière, la géographie ou l'étude des milieux urbains.

Chaque satellite a ses propres caractéristiques, l'une des plus importantes étant sa résolution spectrale. La résolution spectrale d'un capteur peut être caractérisée par le nombre de bandes spectrales, leurs largeurs, et leurs positions sur le spectre (Herold et al., 2003).

Plusieurs études ont montrées (Herold et al., 2003; Meyer et Chander, 2007; Heidena et al., 2007) que la résolution spectrale des capteurs est un point critique, particulièrement pour discriminer différents types d'occupation du sol dans des environnements complexes tels que le milieu urbain. Malgré l'augmentation de la disponibilité des données hyperspectrales, les capteurs multispectraux embarqués à bord de plusieurs satellites acquierent chaque jour une très grande masse de données avec une résolution spectrale relativement pauvre. La plupart des systèmes satellitaires possède 4 à 7 bandes spectrales allant du visible à l'infrarouge sur le

spectre électromagnétique (Govender et al., 2007). Il existe cependant quelques capteurs qui utilisent également des bandes thermiques. L'un des avantages de ces capteurs multispectraux, comparés aux capteurs hyperspectraux, est la couverture spatiale plus importante, qui permet de cartographier plus rapidement de grandes zones.

Comme le nombre de systèmes satellitaires existant augmente avec la complexité des données, un challenge important est d'évaluer la complémentarité de ces capteurs pour une application donnée. En effet, les informations fournies par les différents capteurs peuvent être complémentaires, et un problème clé est de proposer des systèmes capables d'utiliser ces sources d'informations hétérogènes dans un processus unique. Cependant, acquérir des images satellites est toujours très coûteux, c'est pourquoi il serait intéressant d'évaluer a priori la complémentarité des capteurs. Pour résoudre ce problème, nous utilisons dans cet article une approche par simulation. La simulation de capteurs, également appelée simulation de bande, consiste à générer des données multispecrales à partir de données acquises par un autre capteur existant ayant une meilleur résolution spatiale. Cette simulation consiste à combiner des bandes fines en bandes plus larges. Ce type d'approche a déjà été utilisé, particulièrement pour la calibration de capteur. L'étape de simulation utilise les Relative Spectral Response (RSR) des capteurs multispectraux, qui décrivent la réponse spectrale de chaque bande simulée.

Les spectres utilisés pour la simulation proviennent de librairies spectrales qui sont des dépôts de spectres de différents matériaux (par exemple des minéraux, de la végétation etc...) généralement capturés sur le terrain en utilisant des spectromètres. Nous avons utilisé ces librairies ainsi que les caractéristiques techniques de plusieurs capteurs pour simuler différentes vues des spectres de ces librairies. A l'issue de cette étape de simulation nous disposons donc d'un ensemble d'objets tels qu'ils auraient été perçus par les différents capteurs. Chaque objet est donc décrit par un certain nombre d'attributs correspondant au nombre de bandes spectrales du capteur. Ce nombre ainsi que la nature des bandes sont différents pour chaque capteur. On dispose bien alors de vues différents de la même donnée originelle (la libraire spectrale). La seule différence entre les données étant les caractéristiques du capteur utilisé lors de la simulation.

Pour évaluer l'intérêt d'utiliser plusieurs vues conjointement, nous nous sommes intéressés à évaluer l'utilisation de couples de capteurs. L'objectif est d'étudier si l'utilisation de couples de vues provenant de satellites différents améliore la qualité des résultats. Dans la Section 2, nous présentons la problématique de la simulation de capteur, et comment celle-ci est utilisée pour créer des données multisources. Dans la Section 3, nous étudions les approches misent en oeuvre pour traiter ces données et évaluer leur intérêt en utilisation conjointe. Finalement, la Section 4 conclue cet article.

2 Simulation de capteurs

2.1 Définition et applications

Comme indiqué précédemment, le principe de la simulation de capteurs est de générer un spectre multispectral simulé à partir de données acquises à partir de capteurs ayant une meilleure résolution spectrale (hyperspectrale). La simulation consiste à combiner plusieurs bandes hyperspectrales voisines pour former la bande plus large correspondante dans la simulation multispectrale. Elle est réalisée par l'utilisation des fonctions de réponse spectrale rel-

ative ou *Relative Spectral Response* (RSR) du capteur multispectral à simuler. Ces fonctions décrivent la réponse spectrale de chacune des bandes spectrales du capteur. La FIG. 1 présente la fonction de RSR de trois systèmes satellites bien connus : Quickbird, Spot 5 et Landsat.

La simulation de capteurs a été utilisée dans plusieurs types d'applications. Par exemple, Salvatore et al. (1999) ont simulé la réponse d'un nouveau capteur à partir de données AVIRIS hyperspectrales. Cela a permis aux scientifiques d'évaluer le potentiel de leur nouveau capteur multispectral et de paramétrer au mieux les RSR afin d'obtenir de meilleurs résultats en fonction de leurs objectifs. Herold et al. (2003) ont étudié différentes résolutions spectrales pour l'analyse de tissu urbain. Pour cela, ils ont utilisé des données hyperspectrales AVIRIS et une librairie de spectres mesurés sur des matériaux (bitume, tuiles, végétation, etc.) afin de trouver quelles bandes spectrales permettaient de séparer au mieux les classes urbaines d'occupation du sol (bâtiments, routes, végétation, etc.). Les données AVIRIS ont aussi été utilisées pour simuler des données Landsat et Ikonos. Les résultats ont montré que Ikonos et Landsat manquaient de finesse spectrale pour séparer certaines classes urbaines.

2.2 Méthodes de simulation

Afin de simuler des données multispectrales à partir de données hyperspectrales, il faut fusionner les bandes hyperspectrales voisines afin de simuler les bandes multispectrales. Cependant, les réflectances des bandes proches à fusionner ne peuvent pas être simplement additionnées. En fait, elles doivent être pondérées en fonction du RSR des bandes multispectrales. Cette sensibilité est décrite pour chaque capteur par sa fonction de RSR.

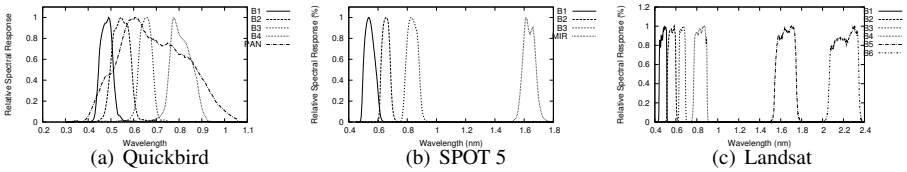
Comme évoqué par Clark et al. (2002), plusieurs stratégies différentes ont été proposées pour calculer les pondérations à appliquer à chaque bande hyperspectrale. Pour la simulation effectuée dans cet article, chaque longueur d'onde hypespectrale a été liée avec la moyenne de la RSR (dans l'intervale de la largeur à mi-hauteur de chaque bande hyperspectral simulée). Cette approche est similaire à celle proposée par Franke et al. (2006) et est décrite en détail dans Forestier et al. (2009).

Il est important de préciser que certains paramètres externes ne sont pas utilisés dans cette étude. Par exemple, d'autres approches (Kavzoglu, 2004) prennent en compte d'autres variables comme les paramètres atmosphériques ou les différences géométriques entre les différents capteurs. Dans nos travaux, nous nous intéressons à la capacité des différents capteurs à séparer les différentes classes en fonction de leur RSR, c'est pourquoi nous nous concentrons sur les différences spectrales entre les capteurs uniquement. Cependant, d'autres aspects comme la résolution spatiale devraient aussi être étudiés afin de mieux apréhender les différences entre les capteurs. Les six capteurs étudiés ici sont : Spot 5, Quickbird, Pleiades, Landsat TM, Ikonos and Formosat (voir TAB. 1).

2.3 Librairies spectrales

Une librairie spectrale est une base de données de spectres de plusieurs types de matériaux (minéraux, objets artificiels, végétation, etc.) mesurés *in situ* à partir de spectromètres.

Plusieurs librairies spectrales libres de droit existent. Dans ces travaux, nous avons utilisé la librairie ASTER (Baldridge et al., 2008) qui se compose de spectres provenant du *Jet Propulsion Laboratory* (JPL), de la *John Hopkins University* (JHU) et du *United States Geological Survey* (USGS). Cette librairie comporte des spectres de différentes roches, minéraux,



(a) Quickbird (b) SPOT 5 (c) Landsat

FIG. 1 – *Exemple de trois réponses spectrales relatives (RSR).*

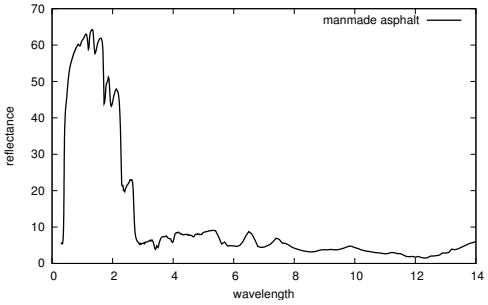
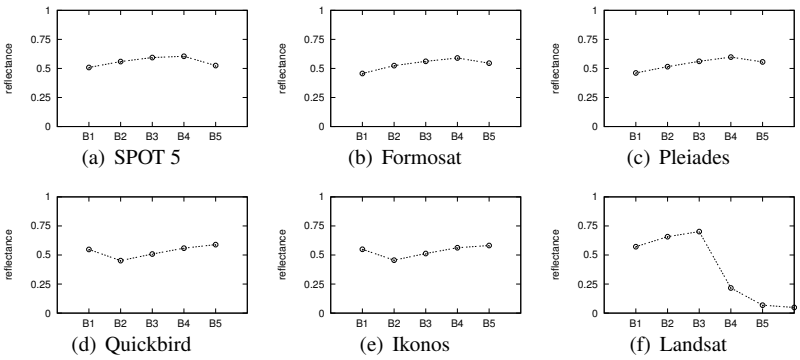


FIG. 2 – *Exemple du spectre complet de l'asphalte extrait de la librairie ASTER.*



(a) SPOT 5 (b) Formosat (c) Pleiades

(d) Quickbird (e) Ikonos (f) Landsat

FIG. 3 – *Exemple de spectre simulé extrait de la librairie ASTER.*

Nom	# Bandes spectrales	Propriétaire
Spot 5	5	CNES
Quickbird	5	Digital Globe
Pleiades	5	CNES
Landsat TM	6	NASA
Ikonos	5	Satellite Imaging Corporation
Formosat	5	Taiwan

TAB. 1 – *Liste des capteurs étudiés dans cet article.*

sols lunaires, sols terrestres, matériaux artificiels, météorites, végétation, neige et glace, qui couvrent les longueurs d'onde du visible à l'infrarouge thermique (0.4-15.4 μm). La première version date de juillet 1998 et la seconde est disponible depuis 2007 sur simple demande via le site web d'ASTER[1]. La librairie ASTER est, à notre avis, la plus simple et complète disponible librement.

Les spectres de cette librairie ont été convolués avec le RSR de chaque capteur afin de créer plusieurs points de vue de la librairie. En fait, le processus de simulation permet de construire la *vue* qu'aurait chaque capteur des données disponibles dans la librairie. Ceci est illustré sur la FIG. 2 qui représente le spectre complet de l'asphalte extrait de la librairie ASTER, et la FIG. 3 qui présente le spectre simulé obtenu pour chacun des capteurs étudiés. Le but est d'utiliser ces différentes vues de la même donnée afin d'évaluer le gain apporté par l'utilisation conjointe de plusieurs satellites pour identifier les différentes occupations du sol.

3 Évaluation des données multisources

L'objectif de cette section est d'étudier l'intérêt d'utiliser conjointement plusieurs données simulées par la méthode proposée dans la section précédente. Trois approches simples ont été mises en oeuvre pour évaluer le gain potentiel d'utiliser ces données multisources conjointement. Nous nous sommes intéressé ici à étudier la collaboration entre couple de capteurs. L'objectif est d'étudier si le fait d'utiliser des données provenant de deux capteurs permet d'améliorer les performances par rapport à l'utilisation monosource de chacune des données individuellement.

3.1 Clustering collaboratif

Un nombre important de nouveaux algorithmes de clustering ont été développés ces dernières années, et des méthodes existantes ont également été modifiées et améliorées. Cette abondance de méthodes peut être expliquée par la difficulté de proposer des méthodes génériques s'adaptant à toutes les types de données disponibles. En effet, chaque méthode comporte un biais induit par l'objectif choisi pour créer les clusters. Par conséquent, deux méthodes différentes peuvent proposer des résultats de clustering très différents à partir des mêmes données. De plus, le même algorithme peut fournir des résultats différents en fonction de son initialisation ou de ses paramètres. Pour résoudre ce problème, certaines méthodes proposent

[1] `http://speclib.jpl.nasa.gov`

d'utiliser plusieurs résultats de clustering différents pour mieux refléter la diversité potentielle des résultats. Ces approches tirent parti des informations fournies par les différents résultats de manière sensiblement différente. Certaines cherchent à créer un résultat unique à partir de plusieurs résultats (Strehl et Ghosh, 2002) en s'intéressant uniquement à la fusion de plusieurs partitions produites par les différentes méthodes. D'autres utilisent des parties de chaque résultat pour construire le résultat final (Law et al., 2004). Une approche faisant également appel à plusieurs résultats est appelée le clustering collaboratif (Wemmert et al., 2000). Celle-ci consiste à utiliser plusieurs méthodes de clustering qui vont collaborer pour proposer un clustering commun d'un même jeu de données. La collaboration peut être définie comme un processus, où deux acteurs ou plus travaillent ensemble pour arriver à un but commun en partageant des connaissances. La première étape du clustering collaboratif consiste à effectuer plusieurs clusterings différents des données. Puis, ces différents résultats sont modifiés pendant une étape de raffinement. Lors de cette étape, chaque résultat est remis en cause à partir des informations proposées par les autres résultats. Dans nos travaux précédent, nous nous sommes majoritairement intéressés à l'utilisation monosource du clustering collaboratif, c'est à dire que chaque méthode de clustering travaille sur la même donnée (voir FIG. 4 (a)). Dans cet article nous avons adopté une approche multisource où chaque méthode va travailler sur une vue différente des données, ici la vue de la librairie spectrale par un des capteurs (voir FIG. 4 (b)). Ce type d'approche a déjà été étudiée récemment dans le but d'utiliser plusieurs représentations des données pour produire un résultat final unique par (**??**). Pour vérifier l'intérêt d'utiliser des données multisources provenant de plusieurs capteurs différents, plusieurs configurations ont été évaluées pour chaque couple de capteurs. Soit D_1 les données issues du premier capteur et D_2 les données issues du second capteur. Les configurations évaluées sont les suivantes :

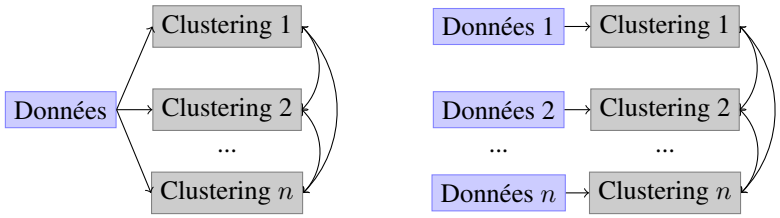
- D_1 : seulement la première vue
- D_2 : seulement la seconde vue
- $D_1 + D_2$: fusion de D_1 et de D_2
- $D_1 \circlearrowleft D_1$: collaboration utilisant deux fois D_1
- $D_2 \circlearrowleft D_2$: collaboration utilisant deux fois D_2
- $D_1 \circlearrowleft D_2$: collaboration utlisant les deux vues D_1 et D_2

La FIG. 5 illustre ces différentes configurations. Dans chacune des expériences, nous avons utilisé l'algorithme KMeans comme méthode de base. Pour chaque configuration ne contenant qu'une seule vue (les trois premières), l'algorithme des KMeans a été initialisé pour trouver un nombre de clusters égal au nombre de classes. Pour les configurations en mode collaboratif (les trois autres), chaque méthode a été initialisée aléatoirement avec un nombre de clusters choisi dans $\{5 \dots 10\}$. Ce choix a été fait pour assurer une certaine diversité des deux résultats impliqués dans la collaboration, ce qui est nécessaire pour obtenir une collaboration intéressante entre les deux résultats. Les résultats de clustering obtenus avec et sans collaboration ont été évalués à l'aide d'indices d'évaluation de partition (ex : Rand, Jaccard). Le tableau TAB. 2 illustre les résultats sous forme d'une matrice pour chaque couple de capteurs. Le symbole • indique que la configuration $D_1 \circlearrowleft D_2$ a fourni de meilleurs résultats que toute les autres configurations, ◦ indique que au moins une autre configuration a donné un meilleur résultat que $D_1 \circlearrowleft D_2$.

Dans ce tableau, il apparait que les capteurs ayant des fonctions RSR proches ne tire pas parti de la collaboration. Par exemple, la collaboration des capteurs Quickbird et Pleiades dont les RSR sont très proches n'est pas bénéfique. A contrario, la collaboration de Spot 5 avec

D_1 \ D_2	formosat	pleiades	quickbird	spot5	landsat	ikonos
formosat	-	○	○	●	●	○
pleiades		-	○	●	●	●
quickbird			-	●	●	○
spot5				-	○	●
landsat					-	●
ikonos						-

TAB. 2 – *Evaluation de la collaboration de couples de capteurs*



(a) Monosource (b) Multisource

FIG. 4 – *Cas monosource (a) et multisource (b)*

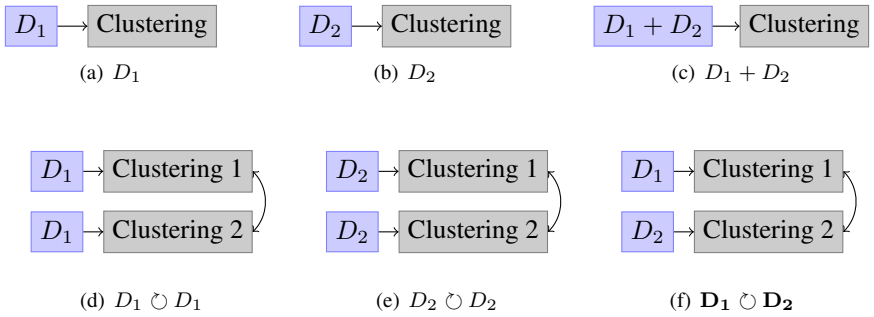
Quickbird ou Pleiades semble bénéfique, Spot 5 possèdant une bande dans le moyen infrarouge que ces deux autres capteurs ne possèdent pas. La conclusion que l'on peut tirer de ces premiers résulats est qu'il semble que la collaboration est bénéfique quand les capteurs ne portent pas exactement la même information.

3.2 Évaluation par classifieurs supervisés

Les résultats obtenus dans les expériences menées dans la section précédente, montrent qu'il semble être intéressant de faire collaborer plusieurs vues si les données ne sont pas similaires. En effet, d'après les résultats obtenus, plus les capteurs sont différents, et plus ceux-ci apportent des informations différentes et potentiellement complémentaires. Pour valider cette hypothèse, nous avons appris un modèle prédictif à l'aide d'une méthode de classification supervisée (Bayésien Naïf) pour chacune des vues. Nous avons ensuite comparé les prédictions des différents modèles appris. Soit p_1 et p_2 deux modèles prédictifs et $p_1(o_i)$ et $p_2(o_i)$ les prédiction pour ces deux modèles pour l'objet o_i (même objet mais représenté de manière différente par les deux vues). Nous avons calculé un coefficient d'accord qui correspond au pourcentage d'accord sur l'ensemble des prédictions des N objets :

$$pr = \sum_{i=0}^{N} \frac{(p_1(o_i) = p_2(o_i))}{N} \quad (1)$$

Le tableau TAB. 3 présente les résultats pour chaque couple de capteurs. Si on met ces résultats en rapport avec ceux obtenus en clustering collaboratif on peut y observer un lien fort



(a) D_1 (b) D_2 (c) $D_1 + D_2$

(d) $D_1 \circlearrowleft D_1$ (e) $D_2 \circlearrowleft D_2$ (f) $\mathbf{D_1 \circlearrowleft D_2}$

FIG. 5 – *Les différentes configurations évaluées.*

D_1 \ D_2	formosat	pleiades	quickbird	spot5	landsat	ikonos
formosat	100	91, 85	97, 78	**85, 93**	**80, 74**	99, 26
pleiades		100	94, 07	**92, 59**	**83, 7**	**92, 59**
quickbird			100	**88, 15**	**82, 96**	98, 52
spot5				100	91, 11	**86, 67**
landsat5					100	**81, 48**
ikonos						100

TAB. 3 – *Pourcentage du nombre de fois où les deux classifieurs sont en accord.*

(voir valeurs en gras). On observe que pour que l'utilisation de différentes vues soit bénéfique il est nécessaire que celles-ci portent des informations différentes. En effet, on peut imaginer qu'utiliser des données fortement similaires risque de ne pas améliorer le processus. Il cependant nécessaire que ces données partagent une certaine cohérence commune, et ne diffèrent que sur certains objets spécifiques. Cela pose le problème de pouvoir évaluer ces différences entre les vues et savoir *à quel point* ces données doivent diverger, et également de connaître l'impact de données *trop* dissimilaires.

3.3 Évaluation par critère statistique

Enfin, dans cette section, nous avons évalué les similarités entre les différentes vues en utilisant un coefficient de corrélation dans le but de vérifier les résultats obtenus dans les deux sections précédentes. Ce coefficient de corrélation a été calculé comme la moyenne entre les corrélations des différents attributs des différentes vues. Comme chaque satellite possède un nombre de bandes différent et que celles-ci ne sont pas similaires ni ordonnées, il est nécessaire de calculer ce coefficient pour tous les couples de bandes pour un couple de capteurs donné. Soit $D_1 = (a_1, \ldots, a_{n_1}), D_2 = (a_1, \ldots, a_{n_2})$ les attributs (ou bandes) pour deux capteurs. La corrélation moyenne est évaluée telle que :

$$\mu_{corr} = \sum_{i}^{n_1} \sum_{j}^{n_2} \frac{r(a_i, a_j)}{(n_1 * n_2)} \tag{2}$$

D_1 \ D_2	formosat	pleiades	quickbird	spot5	landsat	ikonos
formosat	-	$0,630$	$0,642$	$\mathbf{0,592}$	$\mathbf{0,632}$	$0,643$
pleiades		-	$0,640$	$\mathbf{0,590}$	$\mathbf{0,631}$	$\mathbf{0,641}$
quickbird			-	$\mathbf{0,602}$	$\mathbf{0,628}$	$0,632$
spot5				-	$0,676$	$\mathbf{0,654}$
landsat					-	$\mathbf{0,653}$
ikonos						-

TAB. 4 – *Corrélation moyenne entre les différentes vues.*

avec $r(a_i, a_j)$ la coefficient de corrélation linéaire de Bravais-Pearson entre les valeurs des bandes considérées comme deux variables aléatoires :

$$r(a_i, a_j) = \frac{\rho_{a_i a_j}}{\rho_{a_i} \rho_{a_j}} \tag{3}$$

avec $\rho_{a_i a_j}$ la covariance de (a_i, a_j) et ρ_{a_i} et ρ_{a_j}, respectivement l'écart type de a_i et a_j.

4 Conclusion

Dans cet article nous avons présenté des travaux sur l'utilisation de données multisources provenant de la simulation de capteurs. Une étape de simulation utilisant une librairie spectrale permet de générer des vues de cette librairie à la résolution de différents capteurs. Ces données ont ensuite été utilisées dans un processus de clustering collaboratif pour étudier l'intérêt de faire collaborer des données issues de couples de capteurs. Une étude utilisant des modèles prédictifs supervisés ainsi que le calcul de coefficients de corrélation entre différentes vues ont permis d'identifier qu'il est intéressant de faire collaborer les vues si celle-ci sont légèrement dissimilaire. La question maintenant posée est de pouvoir mieux évaluer ces dissimilarités, les quantifier, et évaluer quel niveau de dissimilarité est nécessaire pour obtenir une amélioration significative des résultats.

Références

Baldridge, A. M., S. J. Hook, C. I. Grove, et R. g. (2008). The aster spectral library version 2.0. *Remote Sensing of Environment*.

Clark, R., G. A. Swayze, K. Livo, R. F. Kokaly, T. V. V. King, J. B. Dalton, J. S. Vance, B. W. Rockwell, T. Hoefen, et R. R. McDougal (2002). Synthesis of multispectral bands from hyperspectral data : Validation based on images acquired by aviris, hyperion, ali, and etm+.

Forestier, G., J. Inglada, C. Wemmert, et P. Gancarski (2009). Mining spectral libraries to study sensors' discrimination ability. In *SPIE Europe Remote Sensing*.

Franke, J., V. Heinzel, et G. Menz (2006). Assessment of ndvi- differences caused by sensor specific relative spectral response functions. *IEEE International Geoscience and Remote Sensing Symposium*, 1138–1141.

Govender, M., K. Chetty, et H. Bulcock (2007). A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water SA 33*(2), 145–152.

Greene, D. et P. Cunningham (2009). A matrix factorization approach for integrating multiple data views.

Guillaume Cleuziou, Matthieu Exbrayat, L. M. et J.-H. Sublemontier (2009). Cofkm : A centralized method for multiview clustering.

Heidena, U., K. Segl, S. Roessner, et H. Kaufmann (2007). Determination of robust spectral features for identification of urban surface materials in hyperspectral remote sensing data. *Remote Sensing of Environment 111*, 537–552.

Herold, M., M. Gardner, et D. Roberts (2003). Spectral resolution requirements for mapping urban areas. *Geoscience and Remote Sensing, IEEE Transactions on 41*(9), 1907–1919.

Kavzoglu, T. (2004). Simulating landsat etm+ imagery using dais 7915 hyperspectral scanner data. *International journal of remote sensing 25*(22), 5049–5067.

Law, M., A. Topchy, et A. Jain (2004). Multiobjective data clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, Volume 2, pp. 424–430.

Meyer, D. et G. Chander (2007). The effect of variations in relative spectral response on the retrieval of land surface parameters from multiple sources of remotely sensed imagery. *IEEE International Geoscience and Remote Sensing Symposium*, 5150–5153.

Salvatore, E., C. Esposito, T. Krug, et R. Green (1999). Simulation of the spectral bands of the ccd and wfi cameras of the cbers satellite using aviris data.

Strehl, A. et J. Ghosh (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research 3*, 583–617.

Wemmert, C., P. Gançarski, et J. Korczak (2000). A collaborative approach to combine multiple learning methods. *International Journal on Artificial Intelligence Tools (World Scientific) 9*(1), 59–78.

Summary

In recent years, satellite sensor data have become easier to acquire. Several different satellite systems are now available and produce a large amount of data used for Earth observation. To better grasp the complexity of the Earth surface, it became usual to use different images from different satellites. However, it is generally difficult to predict the potential gain of using multisource satellite sensor data before actually acquiring the data. In this paper, we present a simulation approach to create different views of remote sensing sensor data according to different satellite characteristics. These different views are then used in a collaborative clustering approach to assess the interest of using these multisource data together. Experiments provide some insights on couple of satellite systems able to leverage the complementary of the sources.

Graphes multidimensionels : Approche Coopératives

Lydia Boudjeloud-Assala*, Hanane Azzag**

*LITA, EA 3097
Laboratoire d'Informatique Théorique et Appliquée
Université Paul Verlaine Metz
Ile du Saulcy, F-57045 METZ CEDEX 1
lydia.boudjeloud@univ-metz.fr
http://www.lita.univ-metz.fr/
**LIPN, UMR CNRS 7030
Laboratoire d'Informatique de Paris-Nord Institut Galilée
99 Av. J.B. Clément, F-93430 VILLETANEUSE
hanane.azzag@lipn.univ-paris13.fr
http://www-lipn.univ-paris13.fr/

Résumé. Nous proposons dans cet article une alternative originale pour résoudre le problème de la recherche d'espaces de visualisation pour découvrir la structure complexe des données, tout en respectant leur topologie. Notre proposition, fournie une visualisation multidimensionnelle à partir des données et de leur matrice de similarité obtenue dans des sous espaces. Une première partie concerne la sélection des sous espaces à partir de l'espace global des données. Cette sélection se fait par un algorithme génétique réduisant l'espace de description des données en déterminant uniquement les descripteurs les plus pertinents. Chaque sous ensemble de descripteurs pertinents engendre un sous espace des données. L'évaluation des descripteurs se base sur une mesure de répartition des données, une fois un sous espace sélectionné le graphe de voisinages est construit par un algorithme utilisant des fourmis (agents) artificielles. Dans ce travail, nous proposons plusieurs visualisations et nous les comparons avec celles obtenues sur l'ensemble total (lorsqu'il est possible de l'obtenir). Nous constatons que dans certains cas on se rapproche de la solution globale, et dans d'autres cas nous obtenons des sous espaces qui mettent en évidence des caractéristiques de données qui ne peuvent être vues dans l'espace total des données (des groupes plus homogènes, des données atypiques ...). Des expérimentations sur des ensembles de données ayant un très grand nombre de dimensions montrent l'efficacité et la rapidité de la coopération d'approches biomimétiques.

1 Introduction

Le but de l'extraction des connaissances à partir des données (ECD) est de pouvoir extraire des informations pertinentes contenues dans les grands ensembles de données pour une application connue a priori. L'intérêt des connaissances extraites est validé en fonction du but

de l'application. Ainsi seul l'utilisateur peut expliquer et déterminer la pertinence des résultats obtenus en fonction de ses objectifs. Dans la plupart des méthodes utilisées en ECD, le rôle de l'utilisateur est réduit au choix de l'algorithme utilisé et au réglage des paramètres d'entrée, puis il lance l'exécution et récupère ses résultats sous une forme plus ou moins compréhensible. La première idée traitée dans cet article est d'utiliser des méthodes de visualisation graphique pour permettre à l'utilisateur de mieux comprendre ces données et d'être plus impliqué dans le processus de fouille de données.

De plus, face à la quantité sans cesse grandissante de données stockées, les algorithmes de fouille et de visualisation de données doivent pouvoir être capable de traiter de grandes quantités de données. Une des solutions est d'effectuer un prétraitement des données permettant la réduction de la dimension des données sans perte significative d'informations. La seconde idée traitée est d'utiliser des approches biomimétiques pour prétraiter et visualiser sous forme d'un graphe des ensembles de données de grandes dimensions. Un graphe est une représentation structurée de l'information contenue au sein d'un ensemble de données. Un noeud du graphe représente un individu de l'ensemble de données tandis qu'un lien entre deux noeuds représente une distance. Selon le niveau de détail d'information que l'on désire représenter, il est possible de produire un graphe particulier. Cette structuration sert le plus souvent à disposer d'un outil de modélisation efficace que l'on peut facilement interroger et en extraire des connaissances. Pour l'utilisateur, une bonne représentation visuelle d'un graphe est souvent explicite et parle d'elle même. Elle renseigne globalement sur la structuration de l'information relationnelle et localement sur les relations entre les noeuds. La visualisation d'un graphe ne doit pas modifier sa structure. Elle doit, au contraire, aider à la compréhension de la structure interne du graphe (connexions entre les noeuds et mise en valeur de l'information qu'il contient).

La section suivante présente un bref état de l'art sur la visualisation et la malédiction de la dimentionalité, la troisième section présente l'approche biomimétique coopérative, nous présentons ensuite les différents résultats obtenus, nous terminerons par une conclusion et perspectives.

2 Etat de l'art

2.1 Visualisation multidimensionnelle de graphes

Le cas des grands graphes représente un problème à part entière dans le domaine de la visualisation. Un grand graphe peut être perçu de deux façons, un graphe représentant l'information issue d'un ensemble de données ayant un très grand nombre d'individus (on parlera dans ce cas de grands graphes) ou bien issue d'un ensemble de données ayant un très grand nombre de descripteurs (on parlera dans ce cas de graphes multidimensionels). La visualisation de grands graphes est assimilé à d'importantes quantités d'informations que l'on cherche à visualiser et explorer. Les différentes familles de méthodes de dessin de graphes, présentées dans Lavergne (2008), traitent de ce problème et admettent en conclusion que la visualisation de grands graphes devient complexe et contraignante sur plusieurs plans. Nous pouvons citer à titre d'exemple les contraintes les plus couramment rencontrées dans le domaine :

- importante complexité du problème.
- temps de calcul de la visualisation.
- temps d'afichage (successions et grand nombre d'opérations de dessin, 2D et 3D).

– quantité de mémoire utilisée (i.e. mémoire vive et disque).
– difficulté à maintenir le respect des contraintes esthétiques (l'information perçue dans le graphe devient moins explicite).

Il existe différentes solutions afin de dessiner et ainsi visualiser un grand graphe. Cependant, il n'existe pas de méthode parfaite et chacune dépend des contraintes imposées par la nature des données du graphe et de sa structuration. Certaines solutions, tel que le framework de visualisation Tulip de Auber (2002), réalisent une visualisation de grands graphes (plusieurs milliers de noeuds à la fois). Ceci dans le but de visualiser un maximum de données. Cependant, ces solutions ne permettent pas de mettre l'accent sur une partie précise du graphe ni d'interagir sur son ensemble. Elles produisent le plus souvent une vue statique du dessin de graphe. Des techniques de filtrage existent pour cacher tout ou partie des propriétés structurelles du graphe lors de la visualisation. Un graphe avec une structure complexe est difficile à rendre visuellement. Il est également délicat à interpréter de manière intelligible. D'autres méthodes consiste à considérer voire décomposer le graphe de départ en une succession de sous-graphes. Un algorithme de dessin est ensuite appliqué à chacun des sous-graphes pour les visualiser. Nous nous interessons dans cet article à la visualisation de graphes multidimensionels produit à partir de données de grandes dimensions, très peu de travaux ont été réalisés dans ce domaine. Cependant on retrouve les mêmes contraintes que la visualisation de grands graphes citées précedement. Nous allons dans ce qui suit présenter brièvement le problème de la dimensionnalité que peut rencontrer les méthodes de visualisation de graphes multidimensionels.

2.2 La malédiction de la dimensionnalité

Comme cela se passe aussi pour les méthodes de fouille de données, la plupart de ces méthodes de visualisation de grands graphes, et plus particulièrement les graphes multidimensionels, considèrent des concepts de voisinage en se basant sur les relations entre les données. Dans les ensembles à grandes dimensions, les données sont rares et les notions de distance ou de voisinage perdent leur sens. L'éparpillement, la rareté dans les espaces de grandes dimensions implique que toute relation devient complexe et très coûteuse en temps d'exécution.

Prenons par exemple la notion de distance euclidienne. Considérons la distance euclidienne dans R^n où la dimension est égale à n. Soient trois points de R^n : l'origine A qui a pour coordonnées $(0, 0, \ldots, 0)$, le point B qui a pour coordonnées $(1, 0, \ldots, 0)$ et un point A' de coordonnées $(\epsilon, \epsilon, \ldots, \epsilon)$ où ϵ est un nombre positif très petit. Comparons les distances AA' et AB. Quand la dimension n est inférieure à trois, on a AA'$<<$ AB (AA' est plus petite ou égale à $\epsilon\sqrt{3}$ et AB est égale à 1). Quand la dimension n croit, la distance AB reste égale à 1 mais la distance AA' peut devenir plus grande que AB. Cet exemple montre que la perception de distance dans les espaces 2D ou 3D ne peut pas être facilement extrapolée aux espaces de dimensions supérieures. Ceci illustre un aspect de "la malédiction de la dimensionnalité" ou "curse of dimensionality".

Les techniques de sélection de dimensions (descripteurs) permettent de s'affranchir de ces difficultés et consistent donc à réduire l'ensemble des descripteurs considérés. L'objectif de la sélection de descripteurs, en fouille de données, est de réduire la complexité, augmenter la précision de la prédiction et/ou réduire le temps de traitement des données, en sélectionnant le sous ensemble de descripteurs de taille minimale.

L'approche la plus intuitive pour appréhender le problème des grandes dimensions est d'énumérer tous les sous ensembles de descripteurs possibles et de rechercher le sous ensemble qui satisfait la problématique traitée. Cependant, le fait d'énumérer (rechercher) toutes les combinaisons possibles est un problème NP-difficile, (Narendra et Fukunaga, 1977).

Pour $|D|$ dimensions (ou de descripteurs), la recherche exhaustive consiste à explorer $2^{|D|} - 1$ sous ensembles possibles. La recherche d'un sous ensemble de s dimensions parmi D consiste à appliquer le critère d'évaluation $C^s_{|D|}$ fois, soit $\frac{|D|!}{s!(|D|-s)!}$ fois. Si l'on trouve S' ensembles, on aura donc une complexité de :

$$\sum_{s=0}^{S'} C^s_{|D|} = O(|D|^{S'}) \tag{1}$$

Les étapes de traitement et de visualisation faites avec les dimensions sélectionnées nécessiteront donc moins de temps pour les calculs et moins d'espace mémoire. Il existe d'autres arguments en faveur de la sélection de dimensions :

- en présence d'un ensemble de données de taille N fixe et limité, en réduisant la dimension D on peut éviter le phénomène de "curse of dimensionality" et obtenir des résultats avec de bonnes performances,
- la complexité des algorithmes augmente lorsque la dimension D croit,
- certaines dimensions (variables, attributs) sont séparément pertinentes, mais le gain est faible lorsqu'elles sont combinées. Certaines dimensions peuvent aussi être corrélées ou redondantes.

Pour remédier à celà, le recours à des heuristiques est une solution possible.

3 Approche coopérative

Comme pour les algorithme de fouille de données, beaucoup de méthodes de visualisation perdent de leur efficacité lorsque le nombre de dimensions devient trop important. Pour pouvoir traiter efficacement de tels ensembles de données il est alors nécessaire soit de concevoir de nouvelles méthodes, soit de ramener l'exécution d'une méthode existante à un cas plus favorable. C'est cette deuxième solution que nous allons suivre. L'idée est donc de réduire l'ensemble de descripteurs avant de faire appel à la méthode de visualisation sous forme d'un graphe. La solution proposée dans cet article est de faire coopérer deux méthodes biomimétiques pour la visualisation de graphes multidimensionnels. La première approche est un algorithme génétique pour la sélection de descripteurs (dimensions) pertinents développé par Boudjeloud-Assala (2005). L'algorithme génétique réduit l'espace de description des données en déterminant uniquement les descripteurs les plus pertinents. Chaque sous ensemble de descripteurs pertinents engendre un sous espace des données. L'évaluation des descripteurs se base sur une mesure de répartition des données que nous allons introduir dans la suite, une fois un sous espace sélectionné le graphe de voisinages est construit par une autre approche biomimétique : algorithme utilisant des fourmis (agents) artificielles *AntGraph* (Lavergne, 2008).

Notre but à moyen terme est de proposer une méthode complète de fusion des résultats (graphes) obtenues. Nous souhaitons ainsi faire coopérer les différents graphes afin de trouver la meuilleure solution, en d'autre termes : le graphe qui définie le mieux l'ensemble total des données.

3.1 Evaluation d'un sous ensemble de descripteurs

Nous présentons donc, dans cette partie une nouvelle mesure de pertinence d'un sous ensemble de descripteurs utilisée comme fonction d'évaluation d'un algorithme génétique pour la sélection de descripteurs (Boudjeloud-Assala, 2005). Notre objectif est de trouver un sous ensemble de descripteurs qui représente au mieux la configuration de l'ensemble de départ c'est-à-dire que l'on puisse retrouver la même configuration des résultats du clustering (taille, nombre, contenu, ..., pour chaque cluster). Nous évaluons les sous espaces de descripteurs à l'aide d'un nouvel indice qu'on va appeler *SE* pour *Subset Evaluation* ou mesure de pertinence d'un sous espace de descripteurs et qui représente une F-mesure combinant deux critères : critère de qualité (validité) d'un clustering et une mesure de distribution (répartition) des données dans chaque cluster (groupe d'individus). Cette nouvelle mesure va nous permettre de retrouver des sous espaces de dimensions fidèles aux données originales. Enfin nous comparons les visualisations obtenus dans le sous ensemble de dimensions avec les visualisations de l'ensemble complet des descripteurs (lorsqu'il est possible de l'obtenir). Nous constatons que dans certains cas on se rapproche de la solution globale, et dans d'autres cas nous obtenons des sous espaces qui mettent en évidence des caractéristiques de données qui ne peuvent être vues dans l'espace total des données (des groupes plus homogènes, des données atypiques ...).

3.1.1 Mesure de qualité

Nous utilisons l'indice de validité de Calinski et Harabasz classé premier par (Milligan et Cooper, 1985) pour retrouver le nombre de clusters optimal et valider la qualité du résultat de clustering dans les différents sous ensembles de descripteurs car il décrit parfaitement l'homogénéité d'un cluster et la séparabilité entre différents clusters.

$$CH = (SSB/(k-1))/(SSW/(n-k)) \tag{2}$$

avec :

k : nombre de clusters,

n : nombre de points de l'ensemble des données,

$|C_k|$: cardinalité du cluster k,

m_k : centre du cluster k,

m : centre de l'ensemble de données,

SSW : inertie intra clusters,

SSB : inertie inter clusters.

Lorsque nous maximisons la valeur de CH, nous obtenons l'homogénéité et la séparabilité optimales des différents clusters de l'ensemble de données. L'indice CH de validité d'un résultat de clustering peut être remplacé par un autre indice de validité d'un résultat de clustering pour trouver par exemple des clusters de formes variées ou des chevauchement de clusters. Cependant, en utilisant cet indice individuellement pour évaluer les différents sous espaces de descripteurs, il est essentiel de vérifier la répartition (ou distribution des éléments) des différents clusters pour vérifier l'adéquation avec le résultat de clustering de l'ensemble total des données, d'où la nécessité d'introduire une nouvelle mesure de répartition.

3.1.2 Mesure de répartition

Nous introduisons une mesure de répartition pour évaluer l'adéquation du sous espace de descripteurs avec l'ensemble total des données en terme de répartition des éléments dans les différents clusters. Pour définir cette mesure, nous adoptons le formalisme suivant : soit S un sous ensemble de descripteurs des D descripteurs originales de l'ensemble de données T contenant N éléments ($T : (N * D), |S| < |D|$). Soient R_S et R_T deux mesures représentant l'inverse de la moyenne harmonique de la distribution des points de l'ensemble de données dans les différents clusters ($i/i = 1 \dots k$), aussi bien dans l'ensemble total des données T (N_{i_T}) que dans les sous espaces de descripteurs S (N_{i_S}).

$$R_S = \sum_{i=1}^{k} \frac{N}{N_{i_S}} \tag{3}$$

$$R_T = \sum_{i=1}^{k} \frac{N}{N_{i_T}} \tag{4}$$

R_T représente la répartition obtenue sur l'ensemble total des données (avec le k optimal retrouvé) et nous recherchons le sous espace S qui obtient le meilleur R_S. Cette moyenne harmonique permet de vérifier les répartitions des points dans les différents clusters, et aussi d'avoir un taux de répartition des données dans les clusters qui nous permettra de comparer et de vérifier l'adéquation de la répartition obtenue dans les sous espaces de descripteurs avec celle obtenue dans l'ensemble total des données en ayant tous les descripteurs.

3.1.3 Mesure de pertinence

Pour évaluer la qualité du résultat de clustering global dans les sous espaces de données, nous proposons un nouveau critère combinant une mesure de qualité du résultat de clustering (nous prenons CH qui peut être remplacé par un autre indice de validité) et une mesure de répartition (d'organisation ou de distribution) des données, R_S/R_T que nous venons d'introduire. Pour combiner ces deux mesures, nous proposons d'utiliser une F-mesure, définie par VanRijsbergen (1979) dans le contexte de la recherche d'information pour combiner les mesures de rappel et de précision. Nous l'appliquons à CH et R_S/R_T qui doivent être toutes deux maximisées, ce qui conduit au critère que l'on va noter SE pour *Subset Evaluation*, et que l'on doit maximiser aussi.

$$SE = \frac{(\beta^2 + 1).CH.\frac{R_S}{R_T}}{(\beta^2.CH) + \frac{R_S}{R_T}} \tag{5}$$

Une F-mesure est une moyenne harmonique entre deux mesures, paramétrée par β qui détermine l'importance relative accordée à chacun des deux objectifs (mesures) lors de l'évaluation. Ainsi, si $\beta > 1$, SE favorise davantage une forte répartition qu'un bon résultat de clustering.

3.2 Algorithme *AntGraph*

Nous allons présenté brièvement dans cette section le principe de la méthode de visualisation interactive en 2D/3D de graphe de voisinage construit avec la méthode *AntGraph* de Lavergne (2008). Nous considérons un ensemble de N données triées de manière aléatoire. Une fourmi représente une donnée et devient à terme un noeud du graphe que nous cherchons à construire. Initialement, nous choisissons la première fourmi du tri de données (notée a_1) que nous considérons comme le point d'entrée dans l'étape de construction du graphe. Il s'agit du support fixe et du premier noeud du graphe. Puis les $N-1$ fourmis restantes sont insérées dans le graphe de la manière suivante : chaque fourmi (notée a_i) entre dans le graphe par le noeud a_1, se déplace de noeud en noeud, et ce jusqu'à ce qu'elle se connecte dans le graphe. On peut alors passer à la fourmi suivante. Lorsque a_i est en déplacement, on note a_{pos} la fourmi sur laquelle elle se trouve. Ensuite, le voisinage perçu par a_i correspond à a_{pos} ainsi qu'aux fourmis connectées à a_{pos} (les fourmis voisines). Intuitivement, lorsque la fourmi a_i entre dans le graphe, elle suit le chemin de "similarité maximum" qui se dessine selon le voisinage qu'elle perçoît, puis elle se connecte en établissant un ou plusieurs liens. Ces derniers constituent les arêtes du graphe. Les fourmis vont progressivement se fixer sur ce point initial, puis successivement sur les fourmis fixées à ce point, et ainsi de suite jusqu'à ce que toutes les fourmis soient rattachées à la structure. Les déplacements et les connexions d'une fourmi a_i dépendent donc de la valeur retournée par la fonction basée sur la mesure de similarité $sim(i,j)$ entre les données, et du voisinage local de la fourmi en déplacement. Finalement, la structure construite par des fourmis artificielles, en adaptant le principe d'auto-assemblage des fourmis réelles, est un graphe où les noeuds sont des données et les arêtes sont les relations de voisinage entre ces données. Nous généralisons ces principes dans le but de construire un graphe de fourmis avec comme connaissance de départ un ensemble de données et la mesure de similarité entre ces données (matrice de similarité). Dans une logique d'interaction avec l'utilisateur, l'algorithme *AntGraph* produit un dessin du graphe à chaque itération de l'algorithme. En pratique, c'est une mise à jour l'affichage de ce graphe. Cette dernière étape intervient après le calcul des nouvelles positions de l'ensemble des noeuds du graphe, nous utilisons et commentons dans cet article la visualisation finale du graphe construit par les fourmis. Plusieurs travaux de recherche ont validés la qualité de l'approche biomimétique en terme de visualisation et de construction de graphes de voisinage.

Ens de données	Nbr individus	Nbr descripteurs
CNS	60	7129
Ovarian	253	15154
Prostate	102	12600
MLL	57	12582

TAB. 1 – *Ensembles de données tests.*

4 Expérimentation

Nous allons dans ce qui suit montrer l'efficacité de la coopération de deux approches biomimétique en terme d'extraction d'information et de rapidité. La première partie consiste à sélectionner un sous ensemble de descripteurs pertinents à l'aide de l'algrithme génétique ayant comme fonction objectif la mesure de pertinence présentée précedement. Nous construisons et visualisons ensuite un graphe avec la méthode *AntGraph* à partir de l'ensemble de données sur l'espace réduit par l'algrithme génétique. Nous avons réalisé plusieurs tests comparatives sur des ensembles de données ayant un trés grand nombre de dimensions issue du Kent Ridge Biomedical Dataset Repository (Jinyan et Huiqing, 2002) présentés dans le tableau 1.

4.1 Tests

Le tableau 2 présente le temps d'execution de la première approche sélection de dimmensions sur les différents ensembles de données testsprésentés dans le tableau 1.

Ens de données	32 descripteurs	9 descripteurs	5 descripteurs
CNS	38	25	22
Ovarian	165	131	120
Prostate	84	49	45
MLL	54	46	43

TAB. 2 – *Temps d'éxecution (en secondes) de l'AG par taille du sous ensemble de données.*

Nous nous intéressons principalement dans ce travail à la forme des graphes et non pas au résultat de la classification puisque celle ci dépend des résultats obtenues sur l'ensemble total des données. Nous comparons, donc, les visualisations obtenues d'une part sur l'ensemble total des descripteurs et d'autre part sur les différents sous ensembles de descripteurs obtenues par l'algorithme génétique. Nous remarquerons que dans certains cas (selon le sous espace sélectionné et les paramètres de la mesure d'évaluation) on se rapproche de la solution globale, et dans d'autres cas nous obtenons des sous espaces qui mettent en évidence des caractéristiques de données qui ne peuvent être vues dans l'espace total des données (des groupes plus homogènes, des données atypiques ...). Notons que le temps d'execution (donné en secondes) de l'algorithme génétique est très résonable et est proportionel à la $taille\ des\ individus * taille\ des\ dimensions$.

4.2 Visualisations

Une fois la procédure de sélection de descripteurs réalisée, nous présentons à l'algorithme *AntGraph* une matrice de similarité calculée sur les différents descripteurs sélectionnés pour créer et visualiser le graphe de voisinage résultant. Pour tous les ensembles de données, nous avons tout d'abord essayer de construire et de visualiser le graphe de voisinage sur l'ensemble total de description (figures 1)et 3, puis nous recherchons le sous ensemble de descripteurs à l'aide de l'algorithme génétique. Nous recherchons le sous ensemble fidèle aux données et d'autres sous ensembles d'espace de description qui permettent de mettre en évidence d'autres

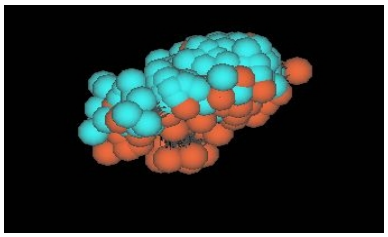
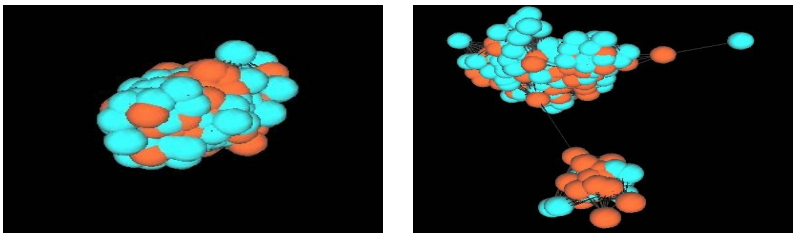


FIG. 1 – *Visualisation du graphe obtenue sur l'ensemble total de données Ovarian.*



(a) Ovarian 9 dimensions. (b) Ovarian 5 dimensions.

FIG. 2 – *Visualisation des graphes pour l'ensemble de données Ovarian, obtenues sur les différents sous ensembles de dimensions.*

caractéristiques des données (des groupes plus homogènes, des données atypiques ...) (figures 2-(a et b)).

La figure 1 présente la visualisation du graphe de voisinage obtenue avec *AntGraph* sur l'ensemble total des descripteurs de l'ensemble de données Ovarian (253 individus et 15154 descripteurs, variables, dimensions) du Kent Ridge Biomedical Dataset Repository (Jinyan et Huiqing, 2002). La visualisation montre bien que l'ensemble de données est complètement homogène, on ne voit aucune séparation des données. Nous observons la même chose dans la visualisation du graphe de voisinage de l'ensemble de données Ovarian sur un sous ensemble de 9 descripteurs (figure 2-(a)). Cependant, lorsque la mesure d'évaluation est differement paramétrée nous arrivons à avoir (et voir) d'autres caractéristiques des données, comme sur la figure 2-(b) où l'on visualise le graphe de voisinage de l'ensemble de données Ovarian sur un sous ensemble de 5 descripteurs, on arrive à distinguer clairement 2 groupes de données et des individus qui se détachent du groupe (pouvant être caractérisés comme individus atypiques).

Nous pouvons également voir sur la figure 3 le graphe de voisinage obtenue sur l'ensemble total des descripteurs de l'ensemble de données MLL (57 individus et 12582 descripteurs). Les données sont imbriquées entre elles et ce n'est pas évident de voir les groupes. Cependant les données se distingent mieux que pour l'ensemble de données Ovarian.

Lorsque la mesure d'évaluation est differement paramétrée nous arrivons à avoir des sous ensembles de descripteurs qui mettent en évidences d'autres caractéristiques des données. Dans le cas de l'ensemble de données MLL (57 individus et 12582 descripteurs), avec un sous ensemble de 4 descripteurs (figure 3-(a)) nous arrivons à distinguer 2 à 3 groupes d'individus

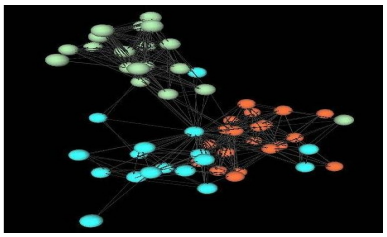


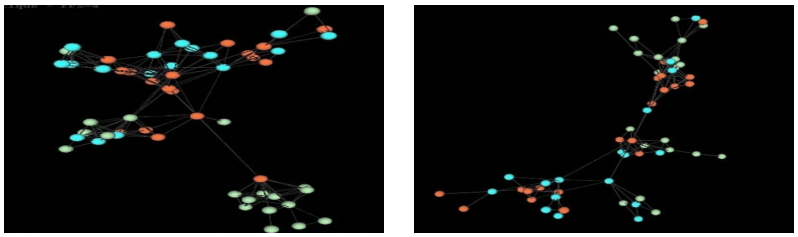
FIG. 3 – *Visualisation du graphe obtenue sur l'ensemble total de données MLL.*

idem avec un sous ensemble de 5 descripteurs (figure 3-(b)). Les données ainsi que les groupes se détachent clairement dans le sous ensemble de 9 descripteurs (figure 3-(d)), contrairement à la figure 3-(c) qui est pratiquement fidèle à l'espace de représentation initial des données (12582 descripteurs) sauf que la visualisation est sur 9 descriteurs sélectionnés. Il en est de même pour les autres ensembles de données CNS (60 individus et 7129 descripteurs) (figures 5-(a et b)) ainsi que l'ensemble de données Prostate (102 individus et 12600 descripteurs) figures 5-(c et d)), certaines caractéristiques apparaissent mieux sur un sous ensemble réduit figures 5-(a et c)) que sur l'ensemble total des descripteurs figures 5-(b et d)).

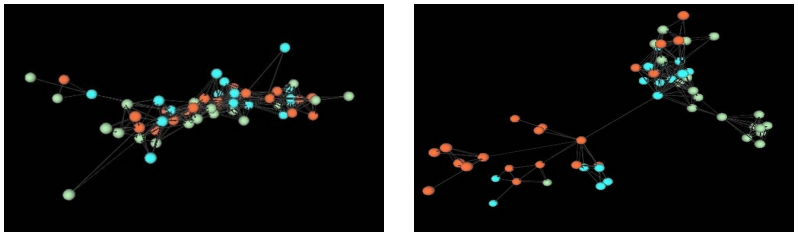
5 Conclusion et travaux futurs

Nous avons présenté dans cet article une façon de résoudre le problème de la recherche d'espaces de visualisation pour découvrir la structure complexe des données, tout en respectant leur topologie. Notre proposition, fournit une visualisation multidimensionnelle à partir des données et de leur matrice de similarité obtenue dans des sous espaces de descripteurs. Ces descripteurs sont sélectionné à partir de l'espace global des données. Cette sélection se fait par un algorithme génétique réduisant l'espace de description en déterminant uniquement les descripteurs les plus pertinents. L'évaluation des descripteurs se base sur une mesure de répartition des données, une fois un sous espace sélectionné le graphe de voisinages est construit par un algorithme utilisant des fourmis (agents) artificielles *AntGraph*. Nous avons proposer plusieurs visualisations et nous les avons comparer avec celles obtenues sur l'ensemble total. Les tests réalisés sur les ensembles de données ayant un très grand nombre de dimensions montrent l'efficacité et la rapidité de la coopération d'approches biomimétiques. Dans certains cas on se rapproche de la solution globale, et dans d'autres cas nous obtenons des sous espaces qui mettent en évidence des caractéristiques de données qui ne peuvent être vues dans l'espace total des données (des groupes plus homogènes, des données atypiques ...). L'intérêt de notre méthode de réduire l'espace de description en déterminant uniquement les descripteurs les plus pertinents par rapport à l'information recherchée.

Ces expérimentations préliminaires sont encourageants et plusieurs perspectives peuvent se découler de ce travail. Nous pensons par la suite proposer à l'utilisateur un algorithme biomimétique intéractif qui lui proposerait plusieurs visualisations obtenues directement par l'approche coopérative, et qu'il évalue de façon intéractive selon l'information qu'il souhaite obtenir de ces données. Nous pourrions également étendre l'intéractivité au problème de classi-

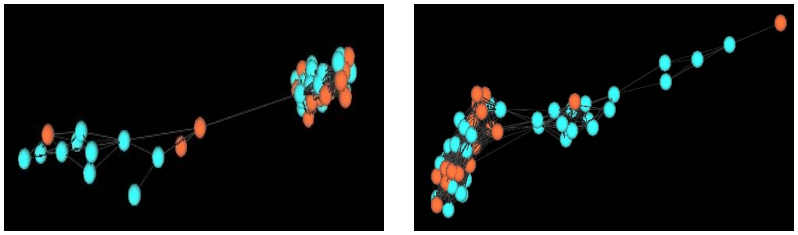


(a) MLL 4 dimensions. (b) MLL 5 dimensions.

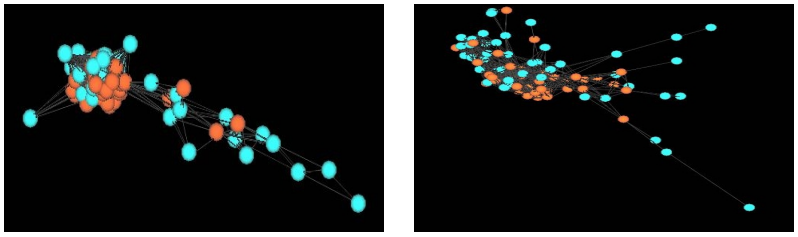


(c) MLL 9 dimensions. (d) MLL 9 dimensions.

FIG. 4 – *Visualisation des graphes pour l'ensemble de données MLL, obtenues sur les différents sous ensembles de dimensions.*



(a) CNS 9 dimensions. (b) CNS ensemble total.



(c) Prostate 9 dimensions. (d) Prostate ensemble total.

FIG. 5 – *Visualisation des graphes pour les ensembles de données Prostate et CNS.*

fictation, en déterminant à partir des visualisations l'identification des groupes et des individus atypiques.

A Moyen terme, notre objectif est de proposer une méthode complète de fusion des résultats (graphes) obtenues. Nous souhaitons ainsi faire coopérer les différents graphes afin de trouver la meuilleure solution, en d'autre termes : le graphe qui définie le mieux l'ensemble total des données (le graphe le plus pertinent).

Références

Auber, D. (2002). *Outils de visualisation de larges structures de données*. Thèse de Doctorat de l'Ecole Polytechnique de l'Université Bordeaux I.

Boudjeloud-Assala, L. (2005). *Visualisation et Algorithmes Génétiques pour la Fouille de Grands Ensembles de Données*. Thèse de Doctorat de l'Ecole Polytechnique de l'Université de Nantes.

Jinyan, L. et L. Huiqing (2002). Kent ridge bio-medical data set repository. http ://sdmc.-lit.org.sg/GEDatasets.

Lavergne, J. (2008). *Algorithmes de fourmis artifficielles pour la construction incrémentale et la visualisation interactive de grands graphes de voisinage*. Thèse de Doctorat de l'Université François Rabelais.

Milligan, G. W. et M. C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*(1), 159–179.

Narendra, P. et K. Fukunaga (1977). A branch and bound algorithm for feature subset selection. In *IEEE Transactions in Computers*, Volume 26, pp. 914–922.

VanRijsbergen, C. (1979). *Information retrivial*. Butterworth, London.

Summary

We propose in this paper an original alternative to solve the problem of the search space visualization to explore the complex structure of data, withe respecting their topology. Our proposal, provided a multi-dimensional visualization from the data and their similarity matrix obtained in different sub-spaces. The first part concerns the subspaces selection from whole data set. We use a genetic algorithm with subset evaluation tio select a pertinent features. Once a sub-space selected neighborhood graph is constructed by an algorithm using artificial ants (agents). In this work, we propose several visualizations and we compare them with those obtained on the whole data set (where we can get it). We note that in some cases they are same, and in other cases we obtain subspaces where some data characteristics appears that can't be seen in the whole data set (clusters, outliers, ...). Experiments on different data sets with very high dimensions seen effective cooperation of biomimetic approaches.

Un langage et un générateur pour représenter les résumés visuels de bases de données géographiques

Ibtissem Cherni*,**, Karla Lopez*
Robert Laurini*, Sami Faiz**

*LIRIS – INSA de Lyon
69621 – Villeurbane Cedex - France
{Ibtissem.Cherni, Karla.Lopez, Robert.Laurini}@insa-lyon.fr
** Faculté des Sciences Juridique, Economique et de Gestion de Jendouba
Avenue de l'UMA - 8189 Jendouba – Tunisie
Sami.faiz@insat.rnu.tn

Résumé. Les chorèmes sont des représentations schématisées du territoire. Ils permettent de représenter visuellement des connaissances géographiques alors que jusqu'à présent les chorèmes avaient été conçus manuellement par les géographes en utilisant leurs propres connaissances mentales du territoire. C'est dans ce contexte qu'est né un projet international afin de découvrir automatiquement les connaissances géographiques à partir d'une fouille d'une base de données géographiques, et de les visualiser en s'appuyant sur la théorie des chorèmes. Cet article présente ce projet en mettant l'accent sur ChorML, langage de description des résumés visuels intermédiaire entre les résultats de la fouille de données et la visualisation, un générateur permettant de traduire les requêtes SQL de fouille de données en ChorML..

1 Introduction

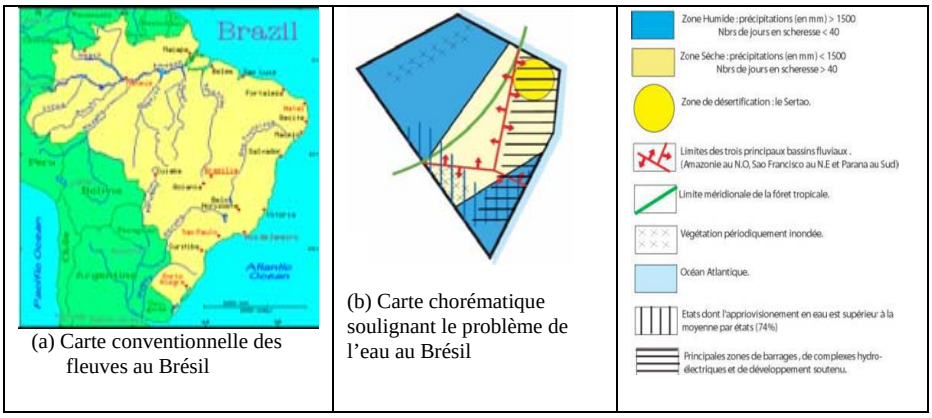
Alors que les données géographiques sont des informations portant sur des objets et événements situés sur le globe terrestre, les chorèmes (Brunet, 1986) sont des représentations schématisées des territoires et représentent la structure et l'organisation de ces territoires. Pour ces raisons, ils semblent constituer une solution intéressante pour générer des résumés de bases de données géographiques afin de faciliter l'aide à la décision spatiale. Dans ce cadre, un projet international, intitulé ChEVIS (Chorem and Visualisation System) entre la France, le Mexique, l'Italie et la Tunisie vise à définir des solutions cartographiques capables de représenter les informations issues de fouille des données géographiques en s'appuyant sur les chorèmes.

Cet article a pour objectif de présenter ce projet en mettant l'accent sur le langage de description des chorèmes nommé ChorML (Chorem Markup Language), ainsi que son système de génération. En premier lieu, nous définissons les chorèmes grâce à un exemple, puis nous présentons le projet ChEVIS. Ensuite, nous décrivons le langage ChorML et en dernier lieu le système de génération du code ChorML.

2 Les chorèmes

Un chorème (Brunet, 1986) est défini comme une représentation schématisée d'un espace géographique : ils servent à visualiser les caractéristiques importantes d'un territoire, et ainsi facilitent sa compréhension dans une optique de prise de décision.

Un motif géographique (Del Fatto, 2009) ou pattern est une régularité intéressante d'un certain phénomène découvert dans une base des données géographiques. Les motifs peuvent être utilisés comme point de départ pour la localisation des phénomènes spatio-temporels et des relations entre eux (cf. figure 1).Chaque paragraphe est en Times 10 points, en utilisant le style RNTI Paragraphe de texte. Le texte de chaque page y compris les entêtes et pieds de pages doit appartenir à une zone de 19.3 cm par 13.2 cm. Merci de ne pas changer les marges du document.



(a) Carte conventionnelle des fleuves au Brésil

(b) Carte chorématique soulignant le problème de l'eau au Brésil

FIG. 1 – *Fleuves au Brésil ; (a) carte conventionnelle ; (b) carte chorématique soulignant le problème de l'eau dans ce pays ; (c) légende (Lafon et al., 2005)..*

Si la figure 1a donne une carte classique des fleuves au Brésil, elle ne dit rien sur les grands problèmes que doit affronter ce pays dans le domaine de l'eau, alors que la Figure 1b (ou carte chorématique) met en évidence, par un formalisme associé la nature et la localisation des problèmes. Cette carte chorématique a été conçue par un géographe à partir de ses connaissances mentales sur ce pays. L'objectif de notre projet de recherche est, en partant d'une base de données géographiques, d'extraire les connaissances importantes et de les visualiser de manière à donner un résumé visuel du contenu de la base de données.

3 Présentation générale du projet ChEVIS

Le projet de recherche ChEVIS developpé en commun par plusieurs laboratoires de recherches s'est donné pour objectif d'extraire les chorèmes à partir de la fouille des données géographiques. En d'autres termes, les connaissances géographiques extraites ne seront pas rédigées avec la logique descriptive, mais par des représentations visuelles. La figure 2 montre l'architecture du système ChEVIS (Chorem Extraction and Visualisation System) qui est composé de deux sous-systèmes, le système d'extraction des chorèmes (Chorem Extraction

System) et le système de visualisation des chorèmes (Chorem Visualization System) reliés par les divers niveaux du langage ChorML.

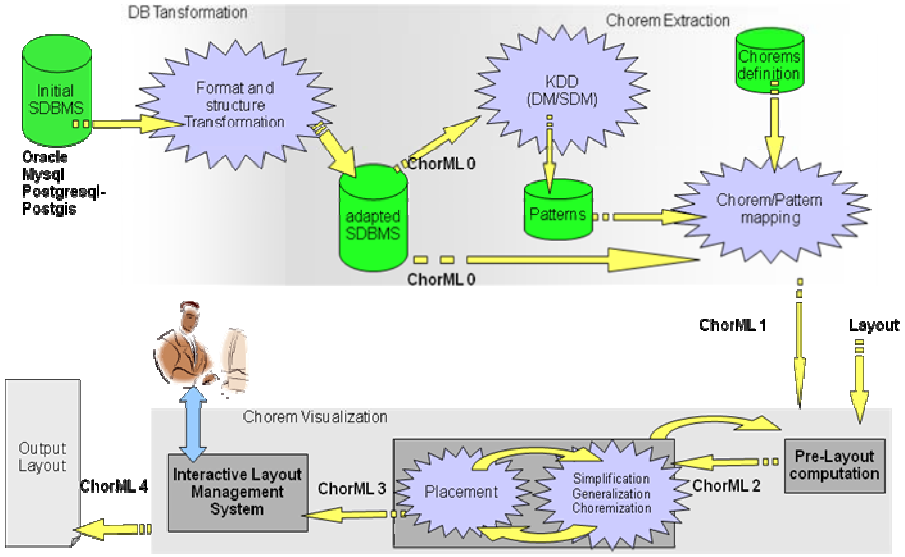


FIG. 2 – *Architecture du système (Laurini, 2008).*

3.1 Le système d'extraction des chorèmes

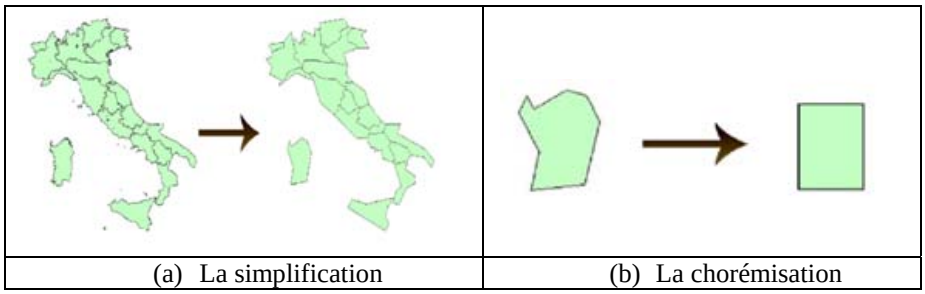
Comme dit précédemment, le point de départ est une base de données géographique dont il s'agit d'extraire les motifs les plus importants. Dès lors, le système d'extraction des chorèmes (Del Fatto, 2009) permet de transformer une liste de motifs en chorèmes. Dans une première phase, un dictionnaire des chorèmes est importé dans la base de données pour être employé dans la phase suivante du processus d'extraction des chorèmes. Par la suite, le système exécute un processus d'association entre les motifs et les chorèmes, ainsi qu'un processus de réduction, afin d'obtenir un ensemble restreint de chorèmes. En effet, du processus d'extraction, une grande quantité de motifs peuvent être obtenus, en risquant ainsi de surcharger d'un nombre excessif de chorèmes de la carte chorématique en cours de construction. En termes quantitatifs, l'objectif est d'obtenir entre cinq et dix chorèmes. Les deux phases permettent la participation des utilisateurs afin de sélectionner les motifs et les chorèmes les plus significatifs. Finalement, le système exécute le calcul des relations spatiales existantes parmi les éléments du chorème. Ainsi, des fonctions d'Oracle Spatial sont utilisées par le système et les résultats sont codés en ChorML, qui sera ensuite envoyé au système de visualisation (Laurini et al., 2006).

3.2 Le système de visualisation des chorèmes (carte chorématique)

Le système de visualisation des chorèmes (Del Fatto, 2009) permet de transformer une liste de chorèmes dans une représentation visuelle; deux phases différentes sont développées par ce système, la création des chorèmes et leur modification.

La création des chorèmes a été développée à travers un système multi-agent qui exécute trois phases : (1) la préparation des dessins, (2) le calcul des coordonnées (depuis longitude/latitude en pixels) puis (3) le placement optimal, qui par des déplacements et simplifications successives détermine le meilleur agencement possible. Cette dernière est effectuée à travers trois opérations (cf. Figure 3) :

- L'opération de simplification qui détermine une version simplifiée des données géométriques, en réduisant le nombre de sommets des polygones de la forme originale. Cette technique est basée sur des fonctions spatiales, qui incorporent l'algorithme de Ramer-Douglas-Peucker (RDP) et ses variantes Douglas et Peucker, 1973).
- L'opération de chorémisation qui conditionne la simplification en associant à la composante spatiale du chorème déjà simplifié, la forme d'un polygone régulier avec un faible nombre de côtés.
- L'opération de satisfaction des contraintes topologiques entre les chorèmes afin de respecter les règles de positionnement relatif, notamment après la phase de simplification ; par exemple pour vérifier que les ports sont bien sur la partie terrestre et non pas au milieu de la mer.



(a) La simplification | (b) La chorémisation

FIG. 3 – *Les deux phases principales de la visualisation des chorèmes (Del Fatto 2009).*

FIG. 4 – *Exemple de carte chorématique produite par le système CheVIS représentant les flux de migration dominants en Italie.*

Un exemple de carte chorématique est donné dans la Figure 4. Finalement, la phase de modification des chorèmes fournit aux utilisateurs la possibilité d'éditer celle-ci (Chorem Editor).

4 ChorML : Chorem Markup Language

ChorML [COI 08] est un langage de type XML, multi-niveaux, dont l'objectif est de mémoriser les informations relatives aux chorèmes et de permettre la communication de telles informations entre les différents modules du système. Ces niveaux sont les suivants :

- *Le niveau 0* est composé de XML et GML (GML 1998). Son objectif est de stocker la BDG en y incluant des informations complémentaires comme les métadonnées (ISO 2003) et des données extérieures au territoire. Par exemple : La France ; généralement les noms des pays limitrophes ne sont pas stockés dans la BD initiale alors qu'il est courant de les voir mentionner dans les cartes chorématiques.
- *Le niveau 1* de ChorML est également une combinaison de XML et de GML. Il spécifie les procédures et les résultats des algorithmes de fouille des données spatiales. Les éléments du langage sont en particulier :
  - Les informations de caractère général, contenant l'identificateur de la carte, le nom de la carte, le nom de l'auteur, la date de création, le système de référence, le nom de la base des données originale, la dernière date de mise à jour, etc.
  - La liste des motifs et leur origine (traçabilité) dans laquelle les données géographiques sont codées en GML.
  - Une pré-légende, qui contient une description en format texte de chaque chorème.
  - La liste des relations topologiques et non-topologiques parmi les chorèmes.
- *Le niveau 2 de ChorML est une combinaison de XML et de SVG (SVG 1999). Les éléments complémentaires de ce niveau sont en particulier :*
  - La liste des chorèmes simplifiés résultant de la phrase de Chorem Création.
  - La liste des chorèmes résultant de l'opération de Chorem Editing.
  - Une légende qui contient une description en format texte et visuelle de chaque chorème.
  - La liste des relations topologiques et non-topologiques parmi les chorèmes.

Si le niveau 0 est essentiellement une extension de GML, et le niveau 2 de SVG, le niveau 1 est le plus original et sa structure est donnée Figure 5 (Coimbra, 2008).

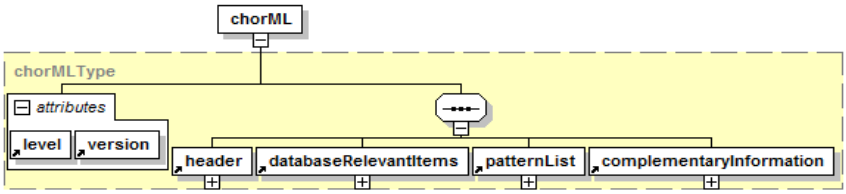


FIG. 5 – *La structure de ChorML (Coimbra, 2008).*

Le tableau expose un exemple d'évolution d'un élément géographique ponctuel selon les niveaux de ChorML.

Niveau	**Représentation**
Niveau 0	Point de longitude / latitude, représentant une ville avec sa population.
Niveau 1	Point de longitude / latitude et son importance s'il est choisi comme motif ; autrement il n'apparaît plus.
Niveau 2	Point avec coordonnées en pixel, représenté par un cercle de rayon

	déterminé avec une couleur associée.

TAB. 1 – *Présentation d'un point à travers les différents niveaux de ChorML.*

Suite à notre analyse des chorèmes, il a été convenu de ne garder que les motifs suivants faits, flux, clusters et relations de co-localisation, et à titre d'illustration, la Figure 6 donne la structure de représentation des flux. De même, il a été décidé de coder les contraintes topologiques en se basant sur les relations topologiques au sens d'Egenhofer (Egenhofer, 1989) en considération de manière à les respecter suite aux opérations de schématisation. A titre d'exemple, dans la Figure 7, les contraintes topologiques à conserver seront le fait que les ports restent sur le continent (point-surface), et que les fleuves se jettent dans la mer (ligne-surface) encodées par des contraintes de type touch-inside.

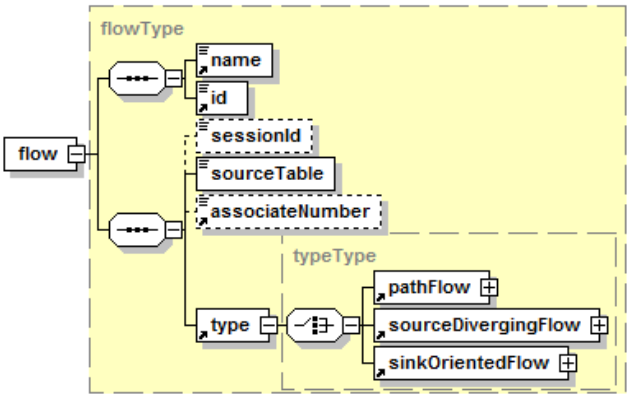


FIG. 6 – *Représentation des flux (Coimbra 2008).*

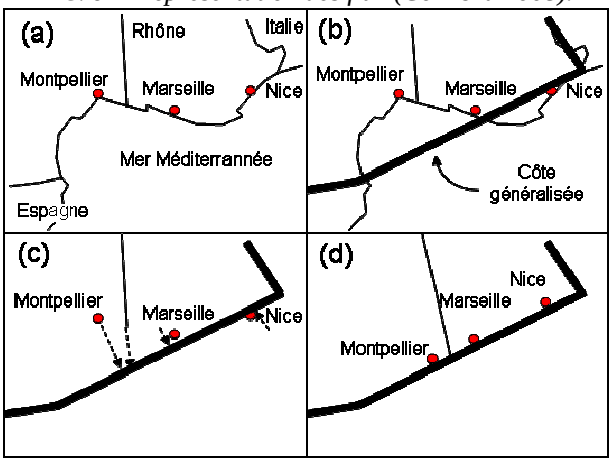


FIG. 7 – *Exemples de contraintes topologiques à respecter de type point-surface et ligne-surface : (a) carte originale ; (b) opération de généralisation ; (c) éléments à déplacer ; (d) résultat final.*

5 Génération automatique de ChorML

Comme dit précédemment, l'objectif de cet article est, une fois exposé le contexte de présenter le système de génération de ChorML en supposant partir, non pas d'une base de données géographiques en GML, mais décrite avec ORACLE Spatial. Comme ce logiciel dispose d'un arsenal important de procédures de fouille de données et de fonctions géographiques, ce générateur se présentera de fait comme un traducteur de SQL ou PL/SQL en ChorML ; pour simplifier l'exposé par la suite nous ne parlerons que de requêtes SQL. Afin de suivre la traçabilité des motifs, il faut mémoriser le texte initial de la requête et ses résultats.

Dès lors, ce système de génération des documents ChorML est composé de quatre modèles comme le montre la Figure 8.

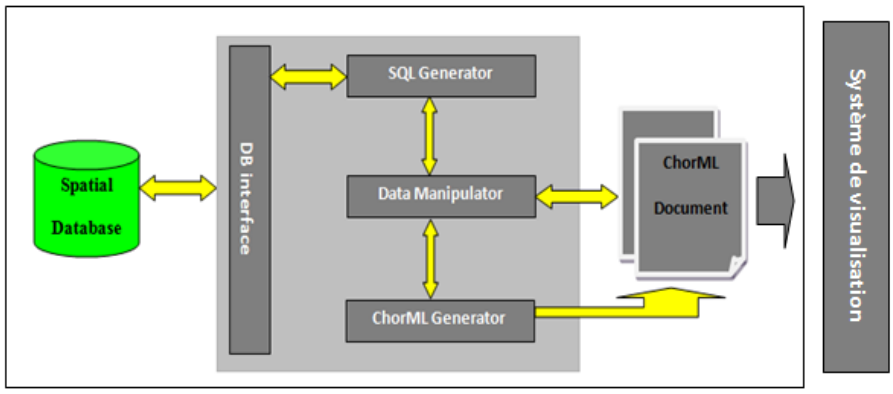


FIG. 8 – *Architecture du générateur ChorML à partir d'Oracle Spatial.*

Dans ce qui suit nous allons présenter rapidement les quatre parties de notre système.

a) Connexion : ce module a pour finalité d'assurer la communication entre le générateur et le SGBD utilisé (Oracle dans notre cas), et ce du moment où on lance la requête jusqu'au moment où l'on récupère le résultat. Ce module assure également le bon déroulement de l'exécution de notre connexion.

b) Génération : c'est à travers ce module que les requêtes seront lancées et exécutées. En fait SQL Generator assure (1) la préparation et l'exécution de la requête SQL appropriée et (2) la récupération du résultat et stockage des données dans une structure bien déterminée.

c) Manipulation : la tâche principale de ce module est de récupérer le résultat du SQL Generator et d'injecter les données dans le module ChorML Generator ; il permet aussi de faire des modifications sur le ChorML déjà généré. Ce dernier est composé de trois méthodes : (1) Une méthode qui permet de récupérer un résultat d'une requête d'un type donné (sql/plsql) afin d'appeler les méthodes nécessaires au module ChorMLGenerator, elle prend comme paramètres une requête, un nom du fichier pour stocker le résultat de génération et un type de requête. (2) Une méthode qui permet de

récupérer le résultat d'une requête cluster afin d'appeler les méthodes de génération de résultats correspondantes au type de requête. (3) Une méthode de génération de contraintes qui seront intégrées dans le fichier final ChorML.

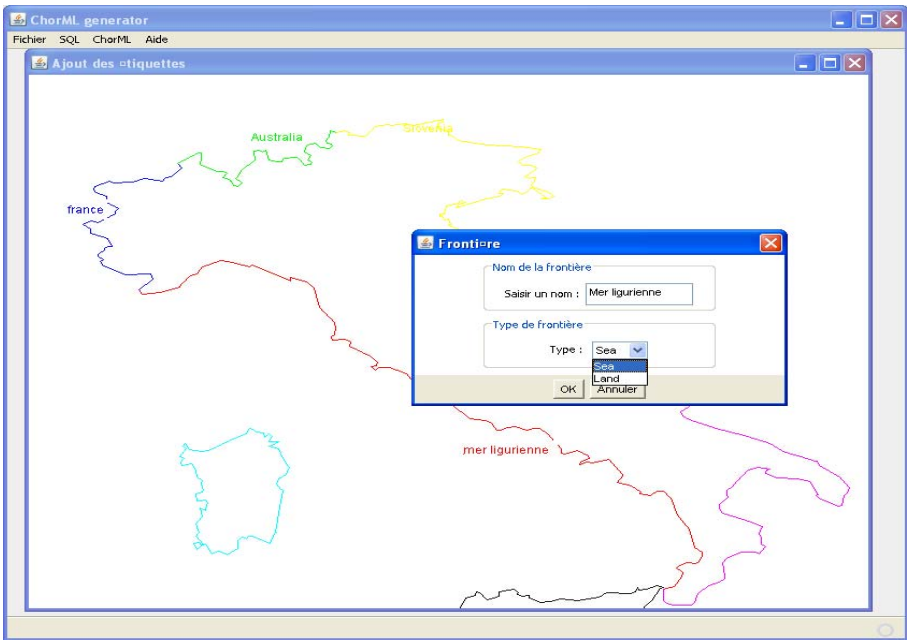


FIG. 9 – *ChorML generator : génération des frontières.*

d) ChorML Generator : ce module a pour objectif la construction d'un document XML selon la spécification ChorML. Ce module est un composant essentiel dans l'application développée car il permet la génération de document XML selon la spécification ChorML, il contient (1) Un module qui permet de générer un document XML selon la spécification ChorML d'un élément simple ou d'un cluster ou région ; (2) Un module qui soit le plus important, c'est celui qui fait la transformation d'un élément de type SDO_GEOMETRY en ChorML en fonction de type de ce champ récupéré de la base de données. (3) Un module qui permet de générer un document XML qui représente les contraintes ChorML. (4) Un module qui permet de stocker des requêtes SQL sous forme d'un document XML. (5) Une méthode qui permet la génération complète d'un document ChorML, elle permet de fusionner les documents générés « l'entête » et « les informations complémentaires ».

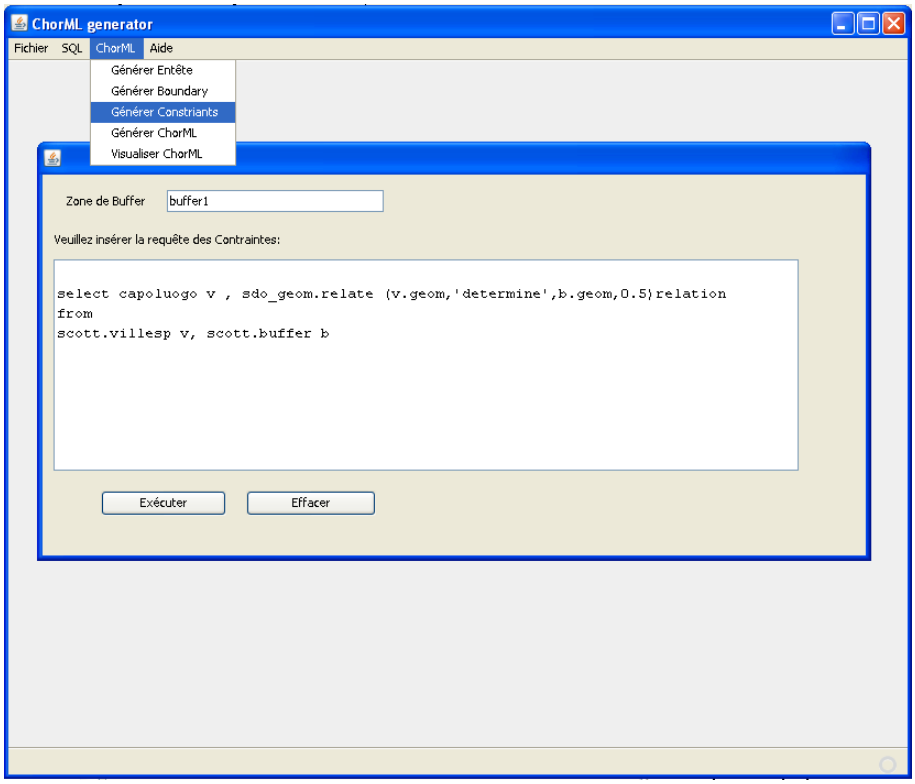


FIG. 10 – *ChorML generator : génération des relations topologiques.*

5.1 Etat de l'implémentation

Dans cette section, nous proposons quelques étapes de la génération d'un document ChorML par notre système ChorML Generator : La génération des frontières, la génération des relations topologiques et la génération du document ChorML final.

En effet, rares sont les bases de données géographiques qui décrivent rapidement l'extérieur de leur territoire de juridiction. Afin d'assurer une plus grande intelligibilité des chorèmes, il a été décidé d'ajouter des informations extérieures comme les noms des mers, des pays voisins, etc.

La figure 10 montre l'interface de génération des frontières d'un pays à partir d'une interface qui permet de valider le nom de frontière et son type (mer, terre) aussi l'utilisateur doit avoir le choix d'ajouter des frontières régionales de la même manière que pour les frontières. Les frontières régionales sont les frontières entre un pays et une région d'un autre pays.

La génération des relations topologiques se fait par l'exécution de deux requêtes spatiales : SDO-GEOM_RELATE et SDO-BUFFER, ces deux requêtes vont donner les relations topologiques entre les différents types géométriques. L'interface de génération des relations est montrée par la Figure 10.

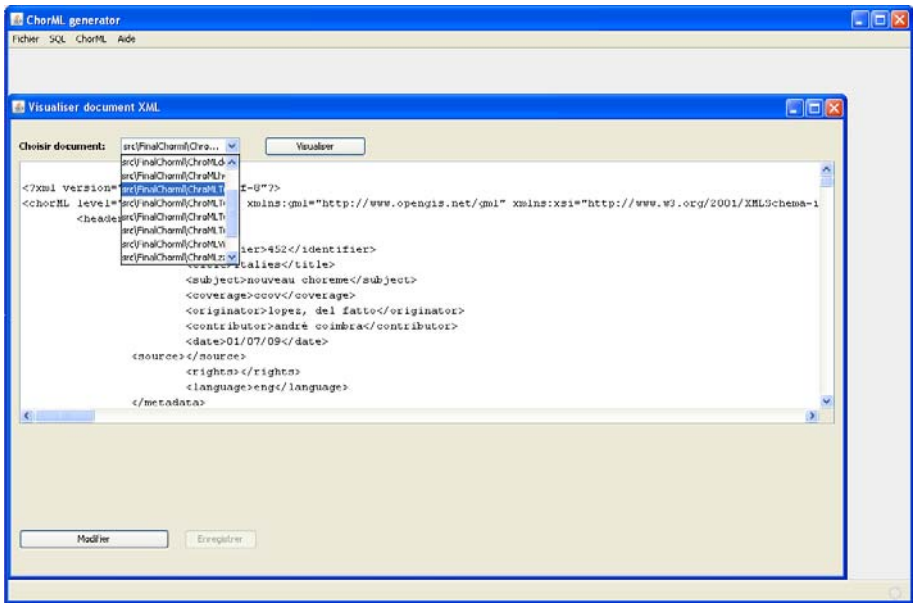


FIG. 11 – *ChorML generator : génération et visualisation des documents ChorML.*

La Figure 11 montre la génération d'un document ChorML qui est le résultat de fusion de trois documents XML : l'entête, la requête et les informations complémentaires.

5.2 Validation des résultats

Après avoir généré les documents ChorML, il faudrait disposer d'un outil pour valider les résultats obtenus par notre système; c'est pourquoi nous avons décidé de transformer ces résultats en KML (Keyhole Markup Language) [KML 07] qui est un langage basé sur le formalisme XML et destiné à la gestion de l'affichage de données géographiques dans les logiciels Google Earth, Google Maps, Google Mobile et World Wind.

Dans la Figure 12, nous présentons la visualisation d'une requête de motif et de frontière ayant été transformée en KML et puis visualisée avec Google Earth.

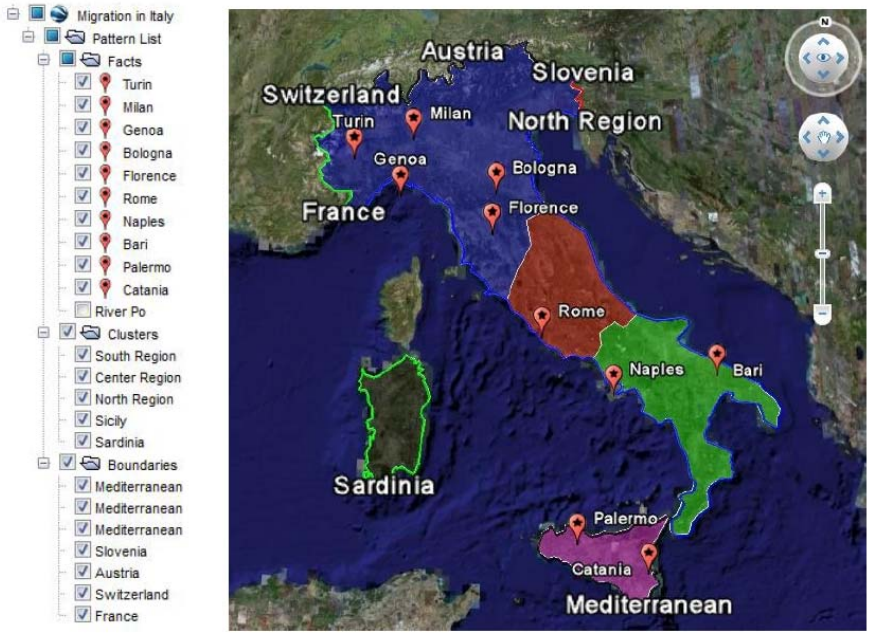


Figure 12. Visualisation avec Google Earth (cluster et frontières).

6 Conclusion

Cet article présente notre projet ChEVIS qui permet de définir des solutions cartographiques capables de représenter convenablement les informations extraites d'une base des données géographique. La solution proposée se fonde sur le concept de chorèmes et sur sa capacité de synthétiser des scènes qui comprennent des objets géographiques et des phénomènes spatio-temporels en leur associant des notations visuelles schématisées.

Nous tenons à remercier Pierre Lauret pour la programmation des modules relatifs aux frontières et aux contraintes topologiques.

Références

Brunet, R (1986). *La carte-modèle et les chorèmes*, Mappemonde 86/4 pp. 4-6.

André Rocha Coimbra (2008). *ChorML: XML Extension For Modeling Visual Summaries of Geographic Databases Based on Chorems*, Master project, INSA Lyon.

VinCenzo Del Fatto **(**2009)**.** *Visual Summaries of Geographic Databases by Chorems*,Ph.D. Thesis, INSA Lyon,University of Salerno.

Egenhofer M.J (1989). *A Formal Definition of Binary Topological Relationships.* Foundations of Data Organization and Algorithms, 3rd International Conference, FODO 1989,

Paris, France, June 21-23, 1989, Proceedings. Lecture Notes in Computer Science 367 Springer 1989, ISBN 3-540-51295-0, pp. 457-472.

GML (1998). *The Open Geospatial Consortium: Geography Markup Language (GML),* http://schemas.opengis.net/gml

ISO 19115 (2003). *Geographic Information - Metadata. International Organization for Standardization (ISO),* Geneva, 2003.

KML (2007). *Keyhole Markup Language (KML),* (http://code.google.com/apis/kml/documentation/)

Lafon B., Codemard C. et Lafon F (2005**)**. *Essai de chorème sur la thématique de l'eau au Brésil*, :(http://webetab.ac-bordeaux.fr/Pedagogie/Histgeo/espaceeleve/bresil/eau/eau.htm).

Laurini, R., Milleret-Raffort, F. et Lopez, K (2006). *A Primer of Geographic Databases Based on Chorems*, In proceedings of the SebGIS Conference, Montpellier, Published by Springer Verlag LNCS 4278, pp. 1693-1702.

Laurini, R. (2008). *Visual Summaries of Geographic Databases,* 14th Int'l conf. On Distibuted Mutimedia System, Invited speaker.

Douglas, D., Peucker, T (1973). *Algorithms for the reduction of the number of points required to represent a digitized line or its caricature*", The Canadian Cartographer 10(2), 112-122.

SVG (1999) Scalable Vector Graphics (SVG), (http://www.w3.org/Graphics/SVG/).

Summary

Chorems are schematic representations of territory and they can represent visually geographic knowledge. Until now chorems were made manually by geographers using their own mental knowledge of the territory. So, a new international project has been launched in order to automatically discover spatial patterns from a geographic database and layout the results, based on chorem theory. This article presents this project and stresses ChorML as a language for representing geographic knowledge to be displayed in a subsequent subsystem, together with a generator translation data mining SQL queries into ChorML.

Liaisons complexes entre variables : les repérer, les valider Application à l'économie du mariage.

Martine Cadot*, Dhouha El Haj Ali**

* Université Henri Poincaré / LORIA, Nancy, France
martine.cadot@loria.fr
http://www.loria.fr/~cadot
** Université Manar I, Faculté des sciences économiques et de gestion de Tunis, Tunisie
elhajali.dhouha@yahoo.fr

Résumé. Nous nous intéressons ici à un type particulier de complexité qui est celle des liaisons entre variables qualitatives. Il existe des modèles statistiques qui ont été construits pour traiter certains aspects de cette complexité. Ainsi le modèle linéaire général (Azaïs et Bardet 2005) parvient à rendre compte d'aspects intéressants de la complexité comme les interactions d'ordre quelconque, les liaisons négatives au même titre que les positives, et les contrastes. Mais ces méthodes sont mal adaptées au cas d'un grand nombre de variables et elles exigent une explicitation a priori des liaisons en jeu. Nous présentons une méthode qui extrait directement à partir des données le même type de liaisons que le modèle linéaire général, sans nécessiter d'hypothèses contraignantes, tout en étant compatible avec un grand nombre de variables. Nous l'illustrons en l'appliquant à des données issues de l'enquête PAPFEM, réalisée en 2001 par l'Office National de la Famille et de la Population en Tunisie, et nous mettons au jour le lien particulièrement complexe entre la pauvreté du ménage et la situation socio-économique des deux époux.

1 Introduction

Notre but est d'extraire des données ce que nous appelons les *liaisons complexes* entre variables, pour les opposer aux liaisons simples qui se limiteraient, par exemple, à la liste des valeurs pour un indice de liaison (corrélation linéaire, etc.) des variables prises 2 à 2. Certains modèles statistiques permettent une représentation des liaisons complexes entre variables issues de données de tableaux individusXvariables, c'est-à-dire contenant pour chaque individu sa valeur à chaque variable. Nous les décrivons dans la section suivante en nous intéressant plus particulièrement au modèle statistique le plus utilisé par les chercheurs en sciences humaines, le modèle linéaire général[1], car il se base sur une décomposition des liaisons complexes en effets simples, interactions, contrastes. Nous décrivons également les

[1] Le modèle linéaire général (Azaïs et Bardet 2005), ou plus simplement le modèle linéaire, exprime la variable à expliquer comme combinaison linéaire des paramètres du modèle, et non des variables explicatives. Par exemple l'équation de régression polynomiale $Y=aX^2+bX+c+\varepsilon$ est linéaire en les paramètres inconnus a, b et c.

conditions d'application de ces modèles statistiques qui les rendent inopérants pour ce que nous souhaitons faire : extraire automatiquement de grands tableaux individusXvariables les liaisons complexes entre variables sans faire aucune hypothèse sur les données. Nous terminons cette deuxième section par un rapide tour d'horizon des méthodes d'EGC (Extraction et Gestion de Connaissances) permettant de brasser plus de variables. La dernière de ces méthodes est celle sur laquelle nous nous appuyons pour extraire les liaisons complexes entre variables.

Dans la section 3, nous détaillons les principes de cette méthode, puis nous montrons comment nous la modifions afin qu'elle nous permette de repérer et valider les liaisons complexes entre variables. Dans la section suivante, nous l'appliquons à des données réelles issues d'une enquête portant sur les conditions de vie des ménages tunisiens. La dernière section est consacrée au bilan et aux perspectives.

2 Les liaisons complexes dans le traitement des données

2.1 Les liaisons complexes en statistique « classique »

La complexité des relations entre variables n'est pas un fait nouveau en statistique. Dans la figure 1, nous avons représenté quatre modèles statistiques permettant d'exprimer divers aspects de la complexité des liens entre un petit nombre de variables. Les variables y sont représentées par des lettres de A à J, les liens entre elles par des lignes les joignant. Un lien peut comporter selon les cas une flèche, un signe et/ou un libellé.

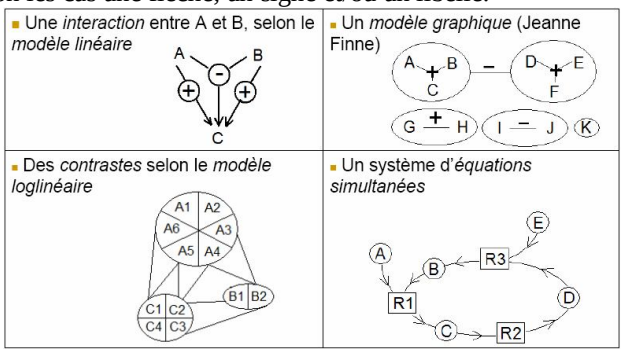


FIG. 1 – *4 modèles statistiques de liaisons complexes entre variables*

Le modèle statistique le plus utilisé pour décrire cette complexité est le modèle linéaire (schéma en haut à gauche de la figure 1), (Fisher 1936) : il permet de décomposer l'influence de quelques variables, ici A et B, sur une autre variable, ici C, en plusieurs constituants indépendants. Dans ce modèle, dont le type causal[2] est indiqué par des flèches, les variables ont deux statuts différents : 1) la variable C est appelée variable *dépendante*, ou *à expliquer*, et doit être *quantitative*, c'est-à-dire mesurée sur une échelle numérique, comme la température par exemple, 2) les variables A et B sont appelées variables *indépendantes* ou

[2] Le modèle linéaire n'a pas pour rôle d'établir des liens de cause à effet, mais de les accepter ou de les rejeter, sur des critères numériques et non sémantiques. Une fois ces liens retenus, c'est l'interprétation qui décide de leur attribuer un sens causal ou non.

explicatives et peuvent être quantitatives ou *qualitatives*, c'est-à-dire pouvant prendre plusieurs valeurs différentes qu'on ne peut pas unanimement ordonner, appelées *modalités*, comme les couleurs par exemple. Dans le schéma, on a représenté un cas où A et B n'ont pas le même effet sur C selon qu'ils agissent séparément ou conjointement. Par exemple l'équation C=20+3A+5B-10AB+e signifie, dans le cas de variables A et B quantitatives, qu'une augmentation d'une unité de A, indépendamment de B, produit une augmentation de 3 unités de C, indiquée par un signe + sur le lien allant de A vers C, qu'une augmentation d'une unité de B, indépendamment de A, produit une augmentation de C de 5 unités, mais qu'une augmentation conjointe d'une unité de A et de B produit une augmentation de 3+5 unités de C ainsi qu'une diminution de 10 unités de C, ce qu'indique le – sur le lien de (A,B) vers C. Cette décomposition linéaire de la valeur de C en composants, ici la moyenne de C, l'effet de A, de B, de l'interaction AB et l'effet individuel e appelé également *résidu*, ne peut être utilisable que si les nombres présents dans l'équation sont des estimations fiables des coefficients réels. L'ANOVA (ANalysis Of VAriance) et la régression linéaire permettent de les estimer au mieux quand diverses conditions concernant la distribution de probabilité des résidus sont vérifiées (Winer 1991). Un modèle linéaire est généralement d'autant plus performant qu'il contient peu d'interactions, et que celles-ci mettent en jeu peu de variables.

Le modèle log-linéaire (Morineau 1996) est le modèle statistique de liaisons complexes entre variables le plus répandu après le modèle linéaire. Il est utilisé pour un petit nombre de variables toutes qualitatives et de même statut (schéma en bas à gauche de la figure 1), et dans sa version la plus élémentaire, il permet de contrôler l'indépendance de deux variables. Dans le schéma, la variable A dispose de six modalités (A1 à A6), B en ayant deux et C quatre, qui interagissent entre elles, chaque effet pouvant se décomposer en *contrastes*, qui sont des oppositions entre groupes de modalités[3]. Pour pouvoir se raccrocher au formalisme du modèle linéaire, on utilise comme variable quantitative dépendante le logarithme de l'effectif correspondant à chaque modalité. Aux contraintes du modèle précédent s'ajoutent des exigences d'effectifs suffisants à chaque croisement de modalités. Cette exigence le rend en pratique inutilisable dès que le nombre de variables augmente.

Les deux schémas à droite de la figure 1, montrent des modèles plus spécifiques de liaisons complexes entre variables quantitatives. Celui du haut est un *modèle graphique* dans lequel les liens entre variables sont les corrélations significatives, la complexité venant des regroupements faits entre variables, et des liens entre groupes de variables. Cette complexité est toutefois simplifiée, car elle s'appuie sur les coefficients de corrélation de Bravais-Pearson, également appelée corrélation linéaire. Celui du bas est un *système d'équations simultanées* (Joreskog 1984), formé de trois équations de régression, R1, R2 et R3, qui s'enchaînent en un système dans lequel des variables prennent un statut intermédiaire entre variables expliquées et explicatives. L'estimation de ces modèles ne peut se faire que sous certaines conditions excluant notamment la possibilité de « cercles vicieux ».

Les quatre modèles que nous venons de décrire présentent un point de vue statistique de la complexité des liens entre variables, et produisent des estimations fiables des composants des liens quand certaines conditions sont réalisées : quelques variables bien choisies, les autres n'étant pas censées intervenir, une grande partie des liens et de leurs composants jugés négligeables pour se concentrer sur les seuls liens et composants intéressants, et une

[3] Les contrastes ne sont pas une spécificité du modèle log-linéaire : ils interviennent essentiellement dans le modèle linéaire quand les variables explicatives sont qualitatives et se déclinent en contrastes a priori, a posteriori (Howel 1998).

spécification plus ou moins précise des lois de probabilités suivies par les données. Et les résultats, s'ils ont été obtenus à partir d'un échantillon tiré d'une population selon certaines règles, seront valides pour cette « population ».

Pour réaliser notre but, qui est d'extraire et de valider les liaisons complexes entre les variables des données, nous devons sortir de ce cadre statistique. D'abord, nous ne souhaitons pas privilégier certaines variables, certains liens ou composants, car notre but est de les trouver tous, même (Suzuki et al. 1998) ceux qui auraient pu échapper aux experts des sciences humaines qui sont à l'origine des données. Quant aux lois de probabilités suivies par les données, elles s'éloignent souvent des lois préconisées dans les modèles statistiques, ce qui est le cas des lois zipfiennes (Breslau et al. 1999) par exemple. Si on choisit de rester dans ce cadre statistique, il faut alors corriger, recoder les données pour se rapprocher des distributions théoriques, et cela a pour conséquence d'entraîner la perte d'une partie de l'information, ce que nous souhaitons également éviter. Nous examinons maintenant d'autres modèles contenant plus de variables, et moins de contraintes.

2.2 Les liaisons complexes en EGC

L'analyse des données traite différemment la complexité entre variables. La variabilité intrinsèque des individus figurant dans le tableau individusXvariables est d'abord éliminée en le remplaçant par un tableau variablesXvariables contenant la valeur des liens entre toutes les variables prises 2 à 2. Les analyses factorielles et les classifications font partie des méthodes les plus anciennes et encore les plus utilisées actuellement. Dans l'analyse factorielle, si les données sont quantitatives c'est le plus souvent le tableau des corrélations qui est factorisé pour créer p nouvelles variables de niveau supérieur (Analyse en Composantes Principales), et si les données sont qualitatives, on factorise le tableau des fréquences de chaque couple de modalités des variables (Analyse Factorielle des Correspondances Multiples). Dans la classification hiérarchique ascendante (Lerman I.C., 1995), on part d'un tableau de dissimilarités ou de distances entre variables, et on agrège en q étapes les variables de départ en de nouvelles variables. L'utilisation de ces méthodes requiert moins de conditions sur les données que les méthodes statistiques décrites précédemment et sont adaptées à un grand nombre de variables. Notons toutefois qu'elles ne sont pas très stables, car elles dépendent d'un paramètre (nombre de facteurs, niveau d'agrégation ou nombre de classes), et les résultats différent de façon parfois importante selon les valeurs de ce paramètre. De plus, la seule complexité qui est mise au jour est celle figurant à l'intérieur du tableau de liens 2 à 2. S'il y avait des interactions entre variables de niveau supérieur, elles ne sont plus décelables, et même, elles peuvent créer des perturbations dans l'analyse qui devient moins fiable (Benzecri 1970).

Depuis quelques décennies, l'accroissement constant des capacités de calcul et de stockage des ordinateurs, ainsi que le libre accès sur le Web à des bases de données de plus en plus nombreuses ont eu pour conséquence l'apparition de nouvelles méthodes de traitement des données. Parmi les méthodes actuelles, les trois plus aptes, à notre connaissance, à prendre en compte la complexité des liaisons entre variables sont 1) les méthodes à noyaux, comme les « Support Vector Machines » (Vapnik 1995), qui permettent de prédire la valeur d'une variable à expliquer par des fonctions bien choisies (les noyaux) des variables explicatives, 2) les réseaux bayésiens (Leray 2004) qui permettent de tenir un raisonnement de type cause à effet en s'appuyant sur une estimation des lois de probabilité de petits groupes de variables supposés sans liens avec les autres et 3) les méthodes

d'extraction des motifs qui extraient toutes les associations de k (k>=1) variables dépassant un seuil d'indice de qualité donné. Les SVM font partie des méthodes d'apprentissage supervisé et fonctionnent comme une boîte noire, fournissant un taux global de reconnaissance de la variable à expliquer très fiable, mais sans pouvoir préciser les fonctions de quelles variables ont le plus contribué à ce taux. La construction d'un réseau bayésien à partir des données nécessite de poser certaines hypothèses sur ce réseau (par exemple pas plus de 4 liens qui se suivent), et malgré cela le réseau peut comporter des contresens du point de vue de la causalité que l'utilisateur doit corriger avant de pouvoir utiliser le réseau (Cadot 2009). Les méthodes à base d'extraction de motifs ont l'avantage d'être entièrement automatiques, une fois certains paramètres fixés, de donner explicitement la liste des liaisons complexes retenues et de ne pas exiger que les données suivent des lois de probabilités spécifiques. Elles ont été construites initialement pour traiter des données binaires comme les tickets de caisse des supermarchés (Han 2001). C'est la méthode qui nous a paru la plus facilement adaptable à notre recherche de liaisons complexes entre variables.

3 Liaisons complexes et k-motifs de variables booléennes

3.1 Les motifs de variables

Supposons que l'on dispose d'un ensemble S de N individus, d'un ensemble V de p variables booléennes (c.à.d. à deux valeurs 1 : Vrai, 0 : Faux), et d'une relation R entre S et V mise sous la forme d'un tableau booléen individusXvariables. Un *k-motif* est une association de k variables de V. Il se nomme généralement par la liste de ses variables. Il y a autant de motifs que de sous-ensembles de variables dans V, soit 2^p si l'on compte le motif vide. L'explosion combinatoire est contenue en ne gardant que les motifs les plus « intéressants » selon divers critères. Par exemple dans l'algorithme Apriori (Agrawal et al. 1996), c'est le choix d'un seuil de *support* du motif, nombre d'individus pour lesquels les variables du motif sont simultanément vraies, qui va enrayer cette explosion. De plus, comme le support d'un k-motif est une fonction non croissante de k, leur choix d'un algorithme d'extraction de motifs fonctionnant par niveau, c'est-à-dire construisant tous les k-motifs avant de construire les (k+1) motifs, permet de rendre cette recherche d'autant plus rapide que le seuil de support choisi est élevé.

A partir des couples de motifs, ont été définies des *règles d'association* (notées AR dans la suite de cette article) permettant de déduire le second motif du premier. Leur qualité est mesurée par deux indices, le support, nombre d'exemples vérifiant simultanément les deux motifs, et la *confiance*, proportion d'exemples vérifiant le second parmi ceux vérifiant le premier (Agrawal et al. 1996). Par exemple, la règle AB→C, où A, B et C représentent respectivement l'achat d'écharpe, de bonnet et de paire de gants, se traduit par « les personnes qui ont acheté écharpe et bonnet ont aussi acheté des gants ». Si on suppose que parmi les 50 personnes ayant acheté écharpe et bonnet, 40 ont acheté des gants, le support de la règle est égal à 40 (c'est le support du motif ABC, c.à.d. le nombre de personnes ayant acheté simultanément les trois articles), et sa confiance, rapport entre le support de ABC et celui de AB, est de 40/50 soit 80%. Ces premiers travaux en informatique dans le domaine de la grande distribution ont produit des règles très proches de celles issues de travaux précédents en statistique dans le domaine de la didactique (règles *d'implication statistique*, Gras 1979) et en algèbre dans le domaine des sciences humaines (règles *d'implication*

informative, Guigues et al 1986). Cet héritage multiple a eu des retombées depuis une dizaine d'années non seulement sur les RA, mais également sur les motifs. Par exemple, la qualité des RA peut se mesurer actuellement par plus de cinquante indices (Guillet 2004) pris séparément ou combinés (Lenca et al. 2003), qui représentent autant de nuances dans l'interprétation de l'association, selon les statistiques, probabilités, ou la sémantique du domaine d'application. Et il existe actuellement de nombreux algorithmes d'extraction de motifs qui diffèrent par les notions algébriques qu'ils utilisent ou par leur façon de parcourir les données (Cadot 2006) et de stocker l'information lue au fur et à mesure.

Nous avons défini au début de cette section un motif comme une association de variables booléennes, puis nous avons décrit son mode d'utilisation, au travers des règles d'association. La règle AB→C liant les achats de 3 articles, que nous avons donnée en exemple, pourrait s'interpréter en terme d'interaction positive (ou négative) du modèle linéaire décrit précédemment : les achats simultanés d'écharpe et de bonnet produisent des achats de gants plus (ou moins) importants que ceux produits par des achats séparés d'écharpe et de bonnet. C'est un des buts de notre méthode MIDOVA (Multidimensional Interaction differential Of VAriance, Cadot 2006) que d'extraire les motifs de variables correspondant à des interactions, positives ou négatives. Extraire des motifs ou RA produisant des liaisons positives ou négatives n'est pas nouveau (Suzuki et al. 1998), mais MIDOVA a pour objectif d'extraire les motifs produisant des différentiels de liaison, mesurés par un indice appelé *reste* et noté Mr : par exemple si la règle AC→E est exacte (100% de confiance), cela correspond à une liaison « totale » pour ACE, de différentiel Mr=0, et aucun sur-motif du motif ACE, comme ACDE par exemple, ne sera créé, pas plus que la règle exacte ACD→E ne sera construite. MIDOVA ne construit que les motifs auxquels on ne pouvait pas s'attendre, sachant leurs sous-motifs. La qualité des motifs extraits par MIDOVA est proportionnelle à *l'étonnement* qu'ils créent, notion empruntée aux règles d'implication statistique de Régis Gras, selon une construction algébrique inspirée des règles d'implication informative de Guigues et Duquenne.

3.2 Extraction des k-motifs avec MIDOVA

MIDOVA est une méthode d'extraction par niveau des k-motifs qui a été construite pour réaliser deux objectifs :

- extraire des k-motifs qui représentent non seulement les liaisons positives entre variables mais également des liaisons négatives.
- créer un k-motif à partir de ses (k-1)-motifs sous-motifs seulement s'il apporte une information supplémentaire à celle apportée par ses sous-motifs.

Ces deux objectifs sont inspirés du modèle linéaire, les motifs s'interprétant en termes d'interaction : les interactions de niveau k représentent les différentiels d'information, c'est-à-dire les informations qui s'ajoutent aux informations de niveau k-1, et s'amenuisent généralement au fur et à mesure que k augmente. Et leur opérationnalisation s'inspire du modèle log-linéaire, version du modèle linéaire adaptée aux variables qualitatives.

Donnons quelques précisions sur le fonctionnement de l'algorithme MIDOVA : c'est un algorithme par niveau, qui fonctionne comme Apriori, à la seule différence que le passage en dessous du seuil de support s n'est plus un critère d'arrêt de la construction des sur-motifs d'un motif. Il est remplacé en cela par le reste dont la formule est $Mr = 2^{k-1}|b-s|$ où b est la borne de l'intervalle de variation de s la plus proche de s. L'intervalle de variation du support d'un motif est déterminé par ses sous-motifs et son amplitude représente la part de variabilité

qu'ils lui ont laissé. Plus il en prend, moins il en laisse à ses éventuels sur-motifs. Nous décrivons dans le paragraphe suivant les différences essentielles entre notre méthode et l'algorithme de référence Apriori, en renvoyant le lecteur intéressé par un exposé détaillé du fonctionnement de cette méthode et de l'algorithme à Cadot (2006).

En pratique MIDOVA produit moins de niveaux qu'Apriori : la valeur de Mr ne peut pas croître, et devient nulle pour les motifs de longueur k>=L avec L tel que $2^{L-1}<=N<2^{L}$. Par exemple si $N<4096=2^{12}$, comme dans l'application relatée dans la section suivante, MIDOVA n'extrait que des k-motifs de longueur k<=12, limite qu'Apriori peut dépasser. Et la longueur maximale des motifs diminue d'autant plus que les liaisons dans les données sont importantes. Si nous nous plaçons dans le cas extrême où toutes les variables sont vraies pour tous les individus, Apriori générera les 2^{p} motifs possibles, alors que MIDOVA ne générera que des 1-motifs ayant tous leur reste à 0. Et en cas de variables toutes identiques sans être toujours vraies, Apriori générera à nouveau tous les motifs possibles, alors que MIDOVA ne construira que les 1 et 2-motifs, ces derniers étant de reste nul. Si un seuil de reste est choisi, plus il est grand, plus le nombre de niveaux a tendance à diminuer. Nous avons fixé un seuil de reste à 10 dans l'application de la section suivante, et la longueur maximale des motifs que nous avons obtenus est de 8.

4 Application à des données socio-économiques

4.1 Description des données

Les données proviennent de l'Enquête Nationale sur la santé de la femme Tunisienne réalisée en 2001 par l'office national de la famille et de la population de la Tunisie (ONFP) et financée par le programme arabe de la santé de la famille. Elle couvre 6691 ménages qui représentent 4346 couples où les femmes sont âgées de 15 à 54. Après une étude préliminaire des données, correction des variables manquantes et suppression des variables aberrantes nous avons retenu 4087 couples.

L'enquête est riche en données portant sur les deux membres du couple, avec des informations approfondies et détaillées sur la femme car l'enquête a été réalisée principalement pour étudier des phénomènes liés à la femme. Le questionnaire comporte un nombre important de rubriques dont celle relative au ménage qui apporte des informations sur sa composition, les conditions de vie (type de logement, nombre des chambres, toilettes, source d'eau potable, etc), les caractéristiques socio-économiques de chaque membre du ménage, la profession, la stabilité du travail, l'âge des conjoints, niveau d'instruction de chacun d'eux, l'état matrimonial, les résidences et des informations sur les parents et les enfants , etc.

Le mariage est récemment considéré comme un moyen de lutte contre la pauvreté et un moyen de redistribution de richesse (Sigle Rushton et al. 2002). Mais peu d'études économiques et statistiques ont analysé et modélisé la liaison complexe entre la décision du mariage et la pauvreté. Nous utilisons la méthode MIDOVA pour déterminer ce lien dans la société tunisienne.

4.2 Préparation des données pour le traitement

Nous avons repris les 104 variables booléennes telles qu'elles ont été construites (El Haj Ali 2007). Elles sont essentiellement de deux types : 1) la réponse de la personne interrogée à une question de type oui/non a donné une variable codée en 1/0 et 2) la réponse à une question par choix dans une liste a fourni autant de variables que de réponses possibles, la seule variable mise à 1 correspondant au choix de la personne interrogée (i.e. recodage par dichotomisation). Quant à la variable « pauvreté », c'est une variable également booléenne construite à partir du résultat d'une analyse factorielle en composantes principales (ACP) de 23 variables décrivant les conditions de vie des ménages telles que : type du sol, possession de voiture, vélo, climatiseur, source d'eau potable…. Nous avons gardé certaines variables « en double », parfois recodées en sens inverse, ce qui nous a permis de contrôler que les relations significatives entre elles apparaissaient bien comme attendu, avec leur signe et avec le niveau de significativité le plus élevé. Après nettoyage des données (correction par recodage et élimination des valeurs manquantes), nous avons obtenu les valeurs de 4072 individus sur 72 variables, dont nous avons tiré un échantillon de 3300 individus (les 8/10es environ, pris au hasard, les 2/10es restants devant être consacrés à des vérifications ultérieures du pouvoir de généralisation des liaisons trouvées), et c'est le tableau booléen complet de 3300 individus et 72 variables que nous avons traité avec MIDOVA.

4.3 Résultats obtenus et interprétations

Nous avons recopié dans le tableau 1 le nombre de k-motifs extraits par MIDOVA de ces données. Les motifs obtenus ont une longueur k allant de 1 à 8. Le seuil de notre paramètre de reste, Mr a été fixé à 10, mais le support n'a pas été fixé, ce qui fait que les supports de ces motifs peuvent prendre toutes les valeurs comprises entre 0 et 3300. Toutefois dès qu'un k-motif atteint de telles valeurs, son reste est nul et il ne permet pas de construire les (k+1)-motifs (c'est le cas de 4 variables, ou 1-motifs, parmi les 72 de la ligne 1 du tableau, 3 sont de support inférieur à 10, et une de support 3300). C'est ainsi que les motifs construits par MIDOVA restent de longueur et en nombre raisonnables.

Longueur k des motifs	Nombre de Variables		Nombre de k-motifs avec		
	Avant	Après	Mr>=0 (tous)	« pauvreté »	Mr>10
1	72	68	72	1	68
2	68	68	2278	67	1627
3	68	67	22946	830	14945
4	65	65	85801	4355	51357
5	65	64	128160	4615	98263
6	64	59	96075	2189	22155
7	53	42	5715	8	841
8	30	16	39	0	0

TAB. 1 – *Les k-motifs extraits par Midova des données*

On a trouvé 67 motifs de longueur 2 contenant la variable pauvreté. On a procédé pour chaque 2-motif à un test du Chi2 d'indépendance entre la variable pauvreté et l'autre variable du 2-motif. On a indiqué par 2* et -2* le fait que les variables sont liées de façon très

significative ($p<0.01$) selon ce test, le signe indiquant le sens du lien, par 1* et -1* quand elles sont liées de façon significative ($p<0.05$, 1* et -1*), et « ns » quand elles ne sont pas significativement liées. Les 67 variables sont en annexe avec leurs intitulés et leurs numéros d'ordre selon leurs niveaux de significativité. Nous donnons ci-dessous un exemple d'interprétation du premier lien de significativité -2* de l'annexe, qui est donc un lien négatif très significatif entre la pauvreté et le « milieu urbain », d'autres interprétations se trouvant en annexe à la suite de la liste des 67 variables .

- On observa que la pauvreté est moins importante dans le milieu urbain que dans le « milieu rural ». Ceci s'explique par la multiplicité des possibilités de trouver un travail dans tous les secteurs où les femmes sont mieux rémunérées, alors que celles qui vivent dans les zones rurales travaillent principalement dans l'agriculture où la rémunération des femmes est faible.

Parmi les 830 motifs de longueur 3 contenant la variable « pauvreté », seuls 69 indiquent des interactions de niveau 3 significatives selon un test du Chi2 d'adéquation. Pour réaliser ce test, les 8 effectifs théoriques du tableau de contingence de dimension 3 ont été calculés de telle façon que les effectifs de leurs marges soient conservés et que *l'odd-ratio* (Morineau 1996) du tableau de dimension 3 soit 1. L'interprétation est plus délicate que pour les 2-motifs, car elle porte sur le différentiel entre les deux effets conditionnels et l'effet conjoint. Voici un exemple d'interprétation d'interaction positive très significative (2*) trouvée entre la pauvreté et deux autres variables : « homme ouvrier spécialisé » et « père ouvrier spécialisé », qui sont chacune liée positivement de façon très significative à la pauvreté, leurs numéros respectifs dans l'annexe 1 étant 10 et 11.

- le fait que l'homme soit ouvrier spécialisé augmente la pauvreté de son ménage, et le fait que le père de la femme soit ouvrier spécialisé augmente la pauvreté du ménage de la femme. Si ces deux personnes sont mariées ensembles, l'augmentation de la pauvreté de leur ménage est encore plus grande que les deux augmentations réunies. Ce résultat confirme que le mariage des pauvres avec des pauvres aggrave la pauvreté chez cette catégorie.

5 Bilan et perspectives

Nous avons présenté dans cet article des modèles statistiques permettant d'exprimer des relations complexes entre variables. Ces modèles fonctionnent très bien quand toutes leurs conditions d'application sont réunies, mais ce n'est pas souvent le cas pour les données que nous voulons traiter, avec un nombre parfois important de variables. Nous avons exposé une méthode appelée MIDOVA permettant d'extraire les relations complexes entre variables qui est une variante de la méthode d'extraction de motifs utilisant l'algorithme Apriori. Nous lui avons transposé certains éléments des modèles linéaire et log-linéaire. La méthode MIDOVA permet, par construction, d'extraire des motifs moins longs, moins nombreux et plus pertinents. L'application faite de la méthode MIDOVA sur des données réelles a permis d'apprécier la qualité des motifs obtenus. L'interprétation des motifs, qui est plus délicate quand leur longueur est plus importante, pourrait être facilitée par des techniques de post-mining. La validation des liaisons extraites des $8/10^{es}$ des données pourrait être renforcée par un contrôle sur les $2/10^{es}$ restants.

Références

Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A.I. Press (1996). Fast discovery of association rules. In Fayyad, U.M. et al., eds., *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California : AAAI Press , MIT Press. pp. 307-328.

Azaïs, J.-M., Bardet, J.-M. (2005). *Le modèle linéaire par l'exemple. Régression, analyse de la variance et plans d'expériences illustrés avec R, SAS et Splus*. Paris. Dunod

Benzécri J.-P. et coll., (1970) *Pratique de l'analyse des données*, Tome 5, Paris, Dunod,

Breslau L., Cao P., Fan L., Phillips G., Shenker S., Web Caching and Zipf-like Distributions : Evidence and Implications, *Proceedings of IEEE Infocom*, NewYork, 1999, p. 126-134.

Cadot, M. (2009). Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayésiens et les treillis de Galois, dans *Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités*. Revue RNTI, n° E-16, p. 223-250.

Cadot M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Thèse, université de Franche-Comté

El Haj Ali Dhouha et Zaiem M Hedi (2007). Un test de la théorie du mariage de Becker sur des données tunisiennes. *75ème congrés de l'ACFAS*, Mai 2007, Montreal, Canada.

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7: 179-188

Gras R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse, univ. Rennes I.

Guigues J.L. et Duquenne V. (1986) Familles minimales d'implications informatives résultat d'un tableau de données binaires, *Math. Sci. Hum.* n°95, pp. 5-18

Guillet, F. (2004). Mesure de qualité des connaissances en ECD. *EGC 2004*, France.

Han J. and Kamber M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco.

Howell, D.C.,(1998), *Méthode statistique en sciences humaines*, De Boeck

Joreskog K.G et Sorbom D. (1984) *Lisrel VI, user's guide*, 3ème édition, Mooresville, IN : Scientific Software. 1984

Lenca P., Meyer P., Picouet P., Vaillant B., Lallich S. (2003). Critères d'évaluation des mesures de qualité en ECD, *JS 2003, Proceedings* , Lyon, pp. 647-650.

Leray P., Francois O., (2004) Réseaux bayésiens pour la classification, Méthodologie et illustration dans le cadre du diagnostic médical, *Revue d'intelligence artificielle, RSTI série RIA*, Vol 18, no 2/2004, Lavoisier, Paris, , p. 168-193

Lerman I.C.,(1995) Rôle de l'inférence statistique dans une approche de l'analyse classificatoire des données, *Méthodes d'analyses statistiques multidimensionnelles en didactique des mathématiques*. IUFM de Caen,

Morineau, A., Nakache, J.-pp, Krzyzanowski, C. (1996), *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris.

Sigle Rushton. W., McLanahan, S. (2002) Richer or poorer ? Marriage as an antipoverty strategy in the United States. *Population*, 57(3); p. 519-538;

Suzuki E., Kodratoff Y., (1998) Discovery of Surprising Exception Rules Based on Intensity of Implication. *Second European Symposium on Principles of Data Mining and Knowledge Discovery*. .Springer-Verlag, London, LN In Computer Science. , p. 10-18.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag

Winer, B.J., Brown D.R., and Michels, K.M. (1991*). Statistical principles in experimental design* 3rd ed. New York: McGraw-Hill.

Annexe : 2-motifs (pauvreté et une autre variable V2)

Liste des variables V2 numérotées de 1 à 67 ordonnées par niveau de significativité

2* : 1) Dépenses du mariage Dm : autre a dépensé pour mariage, 2) Dm de la famille du mari, 3) Enfants décédés, 4) Derni. naiss pd 5DA, 5) A connu son mari dans famille, 6) Femme n'a aucun niveau d'instruction, 7) Homme non instruit, 8) Femme ne travaille pas, 9) Homme ne travaille pas, 10) Homme ouvrier spécialisé, 11) Père ouvrier spécialisé

-2* : 12) Milieu urbain, 13) A travaillé avant, 14) Dm : Pere/mere/titulaire/frere, 15) Dm : elle-même, 16) Dm : mari sans crédit, 17) Dm : mari avec crédit, 18) Connaît Sida, 19) Niveau d'instruction de la femme (Nif) : primaire, 20) Nif 2eme cycle primaire, 21) Nif technique 22) Nif secondaire, 23) Nif universitaire, 24) Niveau d'instruction de l'homme (Nih) : technique, 25) Nih secondaire, 26) Nih universitaire, 27) Femme cadre supérieur, 28) Homme cadre moyen, 29) Homme cadre moyen, 30) Père cadre moyen

1* : 31) Dm d'une autre source

-1* : 32) Nombre de mariages : une seule fois, 33) Grossesses perdues, 34) nombre de mariage de la femme est 1, 35) Homme de niveau 2eme cycle de base, 36) Femme ouvrière spécialisée, 37) Homme ouvrier, 38) Homme de profession libérale/ grand agriculteur,

ns : 39) A participe au programme d'illettré, 40) Dm de sa famille à elle, 41) Ne sait pas, 42) Époux a été marie auparavant, 43) Naissances vivantes, 44) Enfants vivant a la maison, 45) Enfants vivant ailleurs, 46) Au moins une naissance, 47) Entendu parler d'autre MST, 48) Nombre de mariages de l'époux est égal à 1, 49) Ses parents habitent chez elle, 50) Ses beaux-parents habitent avec elle ?: Elle habite seule, 51) Elle habite seule, 52) Homme de niveau d'instruction primaire, 53) Femme ouvrière, 54) Femme artisan/petit commerçant, 55) Femme cadre moyen, 56) Femme profession libérale/grand agriculteur, 57) Homme artisan /petit commerçant, 58) Père ne travaille pas, 59) Père ouvrier, 60) Père artisan/petit commerçant, 61) Père cadre supérieur, 62) Mère ne travaille pas, 63) Mère ouvrière, 64) Mère ouvrière spécialisée, 65) Mère artisan/petit commerçant, 66) Mère cadre moyen, 67) Mère cadre supérieur

Voici les interprétations qui peuvent en être tirées en termes d'économie de la famille :

5 : Plus le lien de parenté est important plus le taux de pauvreté est élevé, ceci s'explique par l'appartenance des deux membres du couple à la même catégorie sociale. En effet, dans la société tunisienne, les petites familles et « les familles mères » découlant de même racine ont pratiquement le même niveau de vie, et comme cette variable est reliée positivement et fortement à la pauvreté, cela confirme que le mariage des pauvres avec les pauvres aggrave encore la pauvreté des ménages.

6 à 10 : On constate que les ménages composés par des couples non instruits ou chômeurs ou de basse qualification professionnelle (ouvriers) sont pauvres. Un tel résultat est attendu puisque l'éducation est un critère primordial pour occuper une profession et la profession représente la source principale du revenu ou de la richesse de l'individu ou du ménage.

11 : Nous remarquons que le capital initial de la femme (capital de sa famille) a un impact sur ses conditions de vie après le mariage. Plus le capital initial est important, plus la vie de la femme est confortable et inversement. Ce résultat s'explique par deux points : 1- la famille intervient souvent pour aider leur fille (après le mariage) financièrement et socialement, ce résultat est confirmé par la variation de la variable w416A. 2- la fille de capital initial important est plus demandée sur le marché du mariage tunisien (El Hadj Ali 2007), cette compétition lui permet de choisir l'homme avec qui elle maximise son utilité et qui lui garantit une vie confortable.

19 à 29 : Plus la femme et son mari sont éduqués, plus ces couples vivent au dessus de la pauvreté, ceci s'explique par la forte relation positive entre le niveau d'éducation et la profession occupée. Les couples composés des femmes participant au marché de l'emploi et des maris non instruits sont pauvres. Ceci s'explique par la composition du marché du mariage tunisien (El Hadj Ali 2007) où les riches s'apparient avec les riches et les pauvres s'apparient avec les pauvres (la richesse est mesurée en terme de niveau d'éducation). En effet, si le mari est non instruit, sa femme est très probablement non instruite ou de niveau primaire et suite à son niveau d'études elle occupe une profession de faible revenu. Par conséquent le ménage composé par de tels membres est forcément pauvre.

62 à 67 : la situation professionnelle de la mère de la femme n'a pas d'impact sur la pauvreté ou non de la femme.

Summary

We deal in this paper with a particular type of complexity, i.e. the one underlying the links between qualitative variables. Some statistical models have been designed for handling a few aspects of this complexity: the general linear model (GLM) gets to account for interesting types of complexity such as interactions of various orders, negative links, and contrasts. But these methods badly fit to the case of a large number of variables, and they constrain the user to explicit beforehand the links at work. We present a method for mining the same type of relations as GLM directly from the data, without any constraining hypotheses, while being compatible with a large number of variables. We illustrate this method applying it to data extracted from the PAPFEM survey achieved in 2001 by the Tunisian Office National de la Famille et de la Population; we have brought to light the specially complex link between the household poverty and the socio-economic position of the husband and wife.

Les multi-sources dans un contexte d'Intelligence Economique

Anass EL HADDADI*,**, Bernard DOUSSET*, Ilham BERRADA**, Iloïse LOUBIER*

*IRIT, SIG, Univesité Paul Sabatier, route de Narbonne, 31062 Toulouse cedex 09
anass.el-haddadi@irit.fr; dousset@irit.fr; loubier@irit.fr
**ENSIAS, équipe AL BIRONI, Université Mohamed V – Souissi, B.P. 713 AGDAL, Rabat
iberrada@ensias.ma

Résumé. Dans cet article nous présentons une approche originale du traitement des multi-sources dans un contexte d'Intelligence Economique. Cette originalité repose sur des descripteurs spécifiques adaptés à chaque source (métadonnées de premier niveau) et sur un descripteur générique (métadonnées de second niveau) qui permet de piloter les outils d'extraction et d'analyse. Nous pouvons ainsi traiter simultanément des corpus d'information hétérogènes dans leur format natif et les combiner à l'infini.

1 Introduction

Les entreprises font face aujourd'hui à une concurrence accrue sur des marchés extrêmement dynamiques et imprévisibles : nouveaux entrants, fusions et acquisitions, baisses tarifaires brutales, évolution rapide des modes de consommation et des valeurs, fragilité des marques et de leur réputation, … Les facteurs de risques externes n'ont jamais été aussi nombreux, sans s'étendre sur la crise financière actuelle.

Il existe une discipline pour mieux anticiper les risques et détecter les opportunités : l'Intelligence Economique. Cette dernière reste encore, quinze ans après la définition canonique proposée par Martre (Martre, 1994), une notion aux frontières peu stables. Les dernières années ont vu une multiplication des définitions de l'IE : passant de définitions centrées sur la description des processus et techniques d'IE, à des définitions incluant les objectifs stratégique de l'IE, et d'autres incluant les notions de gestion de connaissance, d'apprentissage collectif ou de coopération (Salles, 2000).

Dans le contexte de nos travaux, nous retenons la notion de l'IE telle qu'elle a été définie par Henri Martre c.à.d. en tant qu'ensemble des actions coordonnées de recherche, de traitement et distribution de l'information utile aux acteurs pour permettre l'action et la prise de décision. Ceci dépasse les actions partielles désignées sous le nom de documentation, de veille (scientifique et technologique, concurrentielle, financière, juridique et réglementaire) et invite de surcroît à "passer d'un traitement individuel de l'information à la gestion de l'information et à un processus d'actions collectives" (Martre, 1994).

La réussite d'un tel processus passe par la maîtrise de l'information. Cependant, il existe plusieurs types de sources à surveiller : les bases de données scientifiques, les bases de brevets, les media, la presse, les blogs, les flux RSS, l'Internet, l'Intranet, les forums, ….

Dans cet article, nous allons aborder les différents aspects de notre démarche de traitement des multi-sources dans un contexte d'IE qui repose sur des descripteurs spécifiques adaptés à chaque source (métadonnées de premier niveau) et sur un descripteur générique

(métadonnées de second niveau) qui permet de piloter de façon générique les outils d'extraction, d'analyse et de restitution..

2 Source d'information pour la veille

La première étape du processus de veille est la sélection d'information, qui consiste à élaborer un corpus ciblé, en fonction de l'objectif visé, qui par la suite sera analysé via des méthodes de fouille de texte. On emploie souvent le terme corpus pour désigner de vastes ensembles de données textuelles semi ou totalement structurées et disponibles sous forme électronique.

Cette étape permet de se focaliser, suivant des critères prédéfinis, sur des données supposées à la fois « interprétable » et à fort potentiel informatif. De plus, la préparation des données consiste, dans un premier temps, à les sélectionner en accord avec les objectifs que l'on s'impose, en ayant recours aux techniques de recherche d'informations, (Maron et Kuhns, 1960), (Saltan, 1970), (Rocchio, 1971). Ce processus (Saltan et McGill, 1984) cherche à mettre en correspondance une collection de documents et le besoin de l'utilisateur (Maniez et Grolier, 1991), traduit sous la forme d'une requête (Kleinberg, 1999) à travers un système d'information. Ce dernier est composé d'un module d'indexation automatique ou semi-automatique, d'un module d'appariement document-requête et éventuellement d'un module de reformulation de la requête.

Différents modèles sont utilisés dans les moteurs de recherches, pour l'appariement entre la requête et le document, tels que le modèle probabilistes (Maron et Khuns, 1960), booléen (Salton, 1971), (Roberston, 1977), flou (Paice, 1984), connexionniste (Mothe, 1994), (Boughanem et al. 2000), flexible (Sauvagnat, 2005), …

Les dispositifs de veille se basent en fait sur deux types de source d'information.

2.1 Sources formelles

L'information est dite formelle dès qu'elle est publiée sur support papiers, informatique, microfilms, ... Elle peut être structurée ou non, mais il s'agit dans tous les cas d'une information directement accessible et exploitable. Ce type de sources correspond à l'information blanche.

Les sources formelles sont composées principalement de la presse, la télévision, la radio, les livres, les banques de donnée et CD-ROM, les brevets, les informations légales, les études réalisée par des prestataires publics ou privés, Internet. Ces sources ont l'avantage d'être sûres et assez exhaustives, facile d'accès.

Dans un contexte de veille stratégique, les bases de données les plus consultées sont à dominante scientifique, technologique, réglementaire et se trouvent sur des bases bibliographiques. Parmi les bases de données les plus intéressantes, nous pouvons citer dans le domaine économique Factiva, physique avec Inspec, orienté entreprise avec Kompass Europe, multidisciplinaire avec Pascal, médical avec PubMed, …

2.2 Sources informelles

L'information informelle est constitué de toute les informations non formalisées et non disponibles directement. Il est donc nécessaire d'entreprendre les démarches directes auprès

des détenteurs supposés de cette information. Ce type de sources correspond à l'information grise. Ces sources peuvent être les expositions et les salons, les fournisseurs, les colloques, les congrès, les clubs : on y échange des informations, on y communique. L'information qui circule alors peut être d'une grande valeur stratégique, les concurrents (portes ouverts, communication commerciale et financière, publication de journal interne,…), les sources internes de l'entreprise : 80% des informations que recherche un décideur se trouvent dans son entreprise, des sites personnels, des études de recherche menées par un groupe d'étudiants ou de thésard, etc.… les réseaux personnels : le cousin, l'ami commercial de chez X, le représentant de Y, le voisin qui travaille chez Z, … dans la limite de la légalité et de la déontologie.

Les données récoltées son ensuite analysées par rapport aux besoins émis au début de projet. A ce stade du projet, on arrive à la tâche principale de veille qu'est la diffusion des analyses auprès des collaborateurs pour interpréter les résultats. Il faut donc que l'information sélectionnée et mise en avant remonte vers les acteurs cibles. Les résultats de traitements de données représentent une base de travail pour les autres services : recherche et développement, commercial,…. Une information est fondamentalement une action en devenir pour qui sait la mettre en perspective, elle procure la capacité à mettre en œuvre des actions en vue d'influer sur l'environnement. Néanmoins, l'information est périssable, sa valeur démunie avec le temps et globalement plus la source est formalisée, plus l'information est obsolète.

Dans notre approche les données cibles sont sélectionnées (Dkaki et al., 1997) en fonction de l'objectif d'exploitation. Comme le décrit (Rousseau-Hans, 1998), l'utilisation d'outils infométriques dans une démarche de veille technologique permet une approche globale de l'information contenue dans un corpus. Ces outils découpent d'abord les données en unités (mots, dates ou chaînes de caractères), puis appliquent des calculs mathématiques et statistiques afin d'obtenir sous forme de graphiques ou des cartes une représentation des unités en fonction de relations ou proximités calculées.

L'utilisateur va effectuer une recherche d'information, en interrogeant des sources identifiées comme pertinentes issues de déchargement de CD/Rom, télédéchargement de base de données en ligne, aspirateur d'URL (Wisigot, MémoWeb, Teleport pro), aspirateurs de site tels que MémoWeb ou Teleport pro permettant de récupérer l'intégralité ou une partie d'un site.

Les corpus utilisés sont composés de notices, c'est-à-dire des documents structurés en champs (Dkaki et al., 2000). Le mot, unité sémantiquement trop pauvre, a été supplanté par la notion de termes que l'on peut associer à un concept dans une ontologie (Chrisment et al., 2006), (Chrisment et al., 2008).

Un champ, l'unité de base, est le contenu informationnel identifié par une balise et une donnée, par exemple auteur, date, adresse, organisme. Un item est le contenant du champ, un terme, c'est-à-dire la donnée. Il peut être (Mothe, 2000), (Dousset, 2003) :

- mono-valué ne pouvant avoir qu'une seule valeur possible telle que la date ou encore la langue. Par exemple PUBLICATION YEAR=2007 ;
- multi-valué en ayant plusieurs valeurs, comme par exemple plusieurs noms d'auteurs pour un article coécrit, délimités par des séparateurs ;
- diversifié, si le champ contient plusieurs valeurs représentant des concepts différents. Par exemple SOURCE= Lancet, 2007-11, -32p., ce champ peut se décomposer en une revue, une date de publication : 2007-11, qui se divise elle-même en année et en mois, une référence : 32p. indiquant le numéro de la page.

3 Différentes étapes de la veille

La chaîne de traitement débute par la détection et le recueil d'information par les documentalistes depuis des sources souvent hétérogènes mais complémentaires pour aboutir à un corpus multi formats représentatif du domaine étudié. Chaque format est ensuite décrit par des méta données de premier niveau (Md1) qui analysent finement sa structure, décomposent son contenu en unités sémantiques et définissent les modes d'extraction les mieux appropriés à chaque cas. Des méta données de second niveau (Md2) permettent de mettre en correspondance les contenus de même type (mais pas de même format) issus de chaque source et servent d'interface unique avec les techniques, que nous allons développer, d'identification, d'extraction, de dénombrement et de croisement des informations utiles.

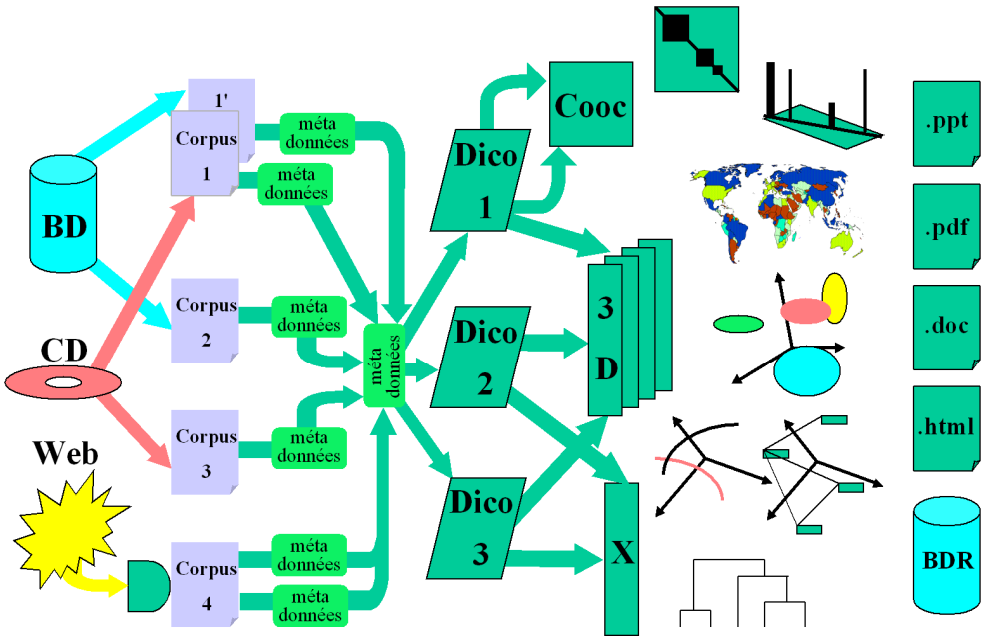


FIG. 1 - *Principe général du processus de veille sur des sources hétérogènes.*

Ensuite les informations sont croisées, soit au sein d'un même type (associations) soit entre deux types différents afin de réaliser une première étude statique de leurs influences mutuelles. Il est aussi possible d'éclater les tableaux obtenus en fonction du temps (par périodes homogènes) afin de réaliser des analyses évolutives et prospectives qui sont les seules à faire ressortir la dimension stratégique d'un domaine.

Cette étape franchie, un ensemble de méthodes d'analyse est déployé pour extraire de ces structures l'information endogène (cachée dans les données, non explicitée par les auteurs, mais particulièrement utile au plan stratégique). De nombreuses sorties graphiques sont alors proposées : visualisations en 2D ou 3D, cartes factorielles, arbres de classification, graphes

de liens, cartes géographiques. Enfin, vient le moment de la restitution : feuille de synthèse, rapport, présentation vidéo, document hypertexte, portail, …

Nous donnons, figure 1, une illustration générale de la séquence des opérations de ce processus de veille dont nous allons préciser, dans la suite de cet article, l'ensemble des méthodes et des enjeux :

4 Extraction de l'information explicite

4.1 Traitement de structures de données

Notre principal objectif étant la réactivité, nous avons constaté qu'une majorité des logiciels du commerce utilisait un format de données propriétaire et laissait le soin aux utilisateurs d'amener leurs informations dans ce format, d'où une perte de temps considérable et souvent l'impossibilité d'y arriver faute de compétence suffisante en informatique. Nous avons donc décidé de traiter, dans la mesure du possible, les informations dans leur format natif. Il y a plusieurs avantages à cela : meilleure réactivité, mise-à-jour du corpus facilitée, conservation de l'ensemble de l'information. Mais pour s'adapter à quasiment toutes les structures, il était nécessaire d'utiliser des outils de description des formats : les méta-données. Leur principe est le suivant :

- **Trouver une technique pour différencier les documents les uns des autres (ou les unités textuelles).**
- **Déterminer les balises des champs sémantiques présentes dans la base, leur donner un nom et un sigle standard.**
- **Définir leur utilité et leur priorité.**
- **Déterminer d'astucieuses techniques de découpage pour extraire au mieux chaque type d'information.**

Plus de 90% des cas rencontrés peuvent ainsi être traités sans aucun reformatage. De plus, comme nous allons le voir, il est possible de travailler simultanément sur plusieurs formats et donc sur plusieurs sources en donnant des règles de correspondances entre les champs utiles à l'aide de méta-données de second niveau. Celles-ci permettent à la fois d'orchestrer la synchronisation de tous les formats et de les interfacer de façon unique avec les outils d'extraction sémantique. Chaque source a son format, chaque format a son descripteur spécifique (méta-données de premier niveau), une collection de formats est gérée par un descripteur générique (méta-données de second niveau : le chef d'orchestre).

4.2 Objets définis dans les métadonnées

- **Bannière de synchronisation :** elle sert à détecter le changement de document à l'intérieur d'un fichier séquentiel contenant un ensemble de documents au même format. Si une ligne vide sépare toujours deux documents eux mêmes sans saut de ligne, cette séparation peut éventuellement servir de bannière de synchronisation « Vide ».

- **Nom standard d'un champ :** c'est le nom donné, dans la langue de l'utilisateur, à un type d'information balisée, il doit être le même dans tous les descripteurs compatibles avec les analyses multi-bases (exemple : Titre, Résumé, Auteurs, Dates, Pays, …).
- **Sigle standard d'un champ :** il permet de nommer de façon standard (si possible avec deux lettres) répertoires, dictionnaires et matrices (PA.ind, PA.Filtre, PA.Syn, MC-PA, AU-AU-DT,…), c'est très utile pour normaliser la structure des analyses et leur appliquer des traitements automatiques (diffusion au format html, compilation dans une base de donnée pour alimenter un portail, …).
- **Bannière du champ :** c'est le nom pris par un champ de la base au moment de sa collecte (long ou court) ou à l'issue d'un éventuel reformatage (html) : TI:, Title:, TI-, Titre :, …
- **Drapeau d'utilisation :** il sert à masquer certains champs inutiles à l'analyse ou à masquer temporairement des champs peu utiles. Les interfaces utilisateur des fonctions de la plate-forme n'affichent donc que les champs requis.
- **Liste des séparateurs :** elle définit l'ensemble des chaînes de caractères qui servent à séparer les unités sémantiques à extraire de chaque champ. Certains jokers sont prévus : le saut de ligne « \n », l'espace « b », … Un opérateur permet aussi de ne conserver, à l'extraction, que l'élément de rang i : « ORDi », « ORD0 » pour le dernier.

4.3 Descripteurs de formats spécifiques

Pour chaque base structurée ou semi-structurée, il convient donc de définir son descripteur de format spécifique qui permet de l'interfacer définitivement avec notre plate-forme de traitement de l'information. Voici quelques exemples partiels de descripteurs de formats :

- **Current Contents sur CD**

```
Enregistrement
# descripteurs des champs de la base Current_Contents
# nom           abrev   champ visible separateurs #
Multitermes     MT      MTM:  True    b"
Numero          NO      AN:   True    "
RT              RT      RT:   False   "
Type_doc        PT      PT:   False   "
Titre           TI      TI:   True    "
Titrec          Ti      TI:   True    s,"s."s:"s;"s?"sb"s)"s]","."":";"?"b"(")"["]"
RP              RP      RP:   True    "
Adresse         AD      IN:   True    ;"
```

Bannière de synchronisation

Nom standardisé du champ

Organisme	OR	IN:	True	,";"
Email	EM	EM:	False	"
Journal	JN	SO:	True	."ORD1"
Date	DP	PY:		
IS	IS	IS:		

Seul le premier item rencontré est retenu

Descripteur	DE	KA:	True	;"
Classification	CL	KP:	True	;"
CC	CC	CC:	False	"
References	RF	RF:	False	"
GA	GA	GA:	False	"
UD	UD	UD:	False	"
Auteur_l	AL	AU:	True	;":"
Auteur_c	AC	AU:	True	;":"
Ville	VI	IN:	True	,"^"|";"0"1"2"3"4"5"6"7"8"9"
Pays	PA	IN:	True	,"^"|";"0"1"2"3"4"5"6"7"8"9"
Resume	AB	AB:	True	,"s."s:"s;"s?"sb"s)"s]""",".":";"?"b")"["]"
FTXT	FX	FTXT:	False	"
FIN	FIN	FIN	FIN	"

- **SCI (Web of science)**

PT
descripteurs des champs de la base Web of Science

# nom	abrev	champ	visible	separateurs #
Numero	NO	UT	True	:000"
Type_pub	PT	PT	False	"
Auteur_l	AL	AU	True	\n"
Auteur_c	AC	AU	True	\n"
Titre	TI	TI	True	"
Journal	JN	SO	True	\n"LA"
Type_doc	DT	DT	False	"
SN	SN	SN	False	"
Adresse	AD	C1	True	"
Organisme	OR	C1	True	\n"
Ville	VI	C1	True	,"^"|";"\n"0"1"2"3"4"5"6"7"8"9"
Pays	PA	C1	True	,"^"|";"\n"0"1"2"3"4"5"6"7"8"9"
Descripteur	DE	DE	True	;"
Index	ID	ID	True	;"
Resume	AB	AB	True	s,"s."s:"s;"s?"sb"s)"s]",".":";"?"b")"["]"|"-"
Multitermes	MT	MTM:	False	b"
TC	TC	TC	False	"
EP	EP	EP	False	"
PG	PG	PG	False	"
JI	JI	JI	False	"
Date	DP	PY	True	b"

Saut de ligne

Espace

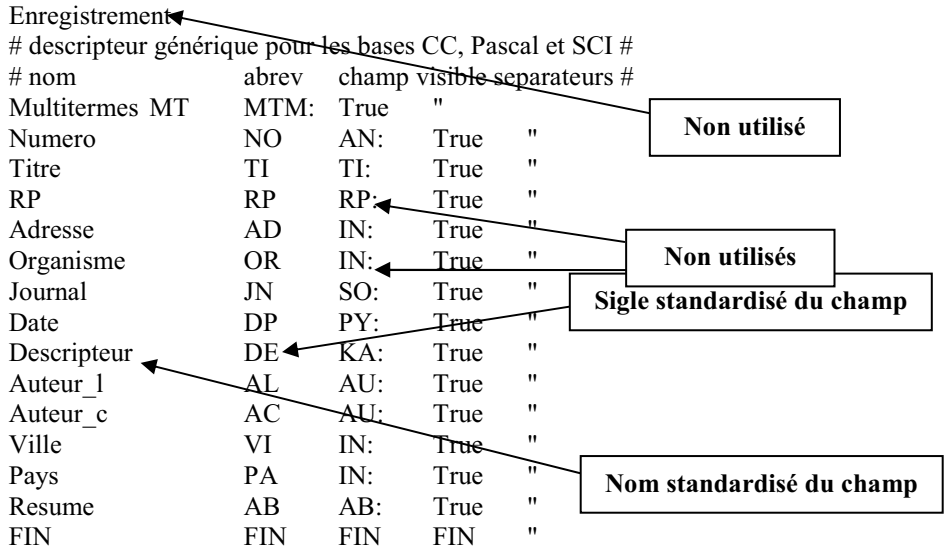
Fin des méta-données

SU		SU	SU	False "

FIN FIN FIN FIN "

4.4 Descripteur générique

Si nous désirons réaliser une étude simultanée sur les 3 bases décrites plus haut, nous devons définir un descripteur générique en donnant pour chaque champ utile un nom standard, un sigle standard et un drapeau d'utilisation (Dkaki, 1996). Les autres éléments (bannière de synchronisation, nom effectif du champ dans chaque base et liste des séparateurs) sont tour à tour empruntés aux méta données de chaque base. Voici un descripteur générique compatible simultanément avec les trois bases (Current-contents, Pascal et Web of science) :



- La bannière de synchronisation est, ici, inutile puisqu'elle est spécifique à chaque base, elle est tout de même conservée par soucis d'unicité du format des descripteurs.
- La première colonne permet de désigner explicitement chaque champ utile, son nom standard permet de faire coïncider les champs de même nature dans les différentes bases simultanément analysées.
- La seconde permet de nommer répertoires et matrices de croisement de façon standard (habituellement deux lettres en majuscule).
- La troisième est inutilisée car remplacée par celle qui est spécifique à chaque base (liste des noms exacts des bannières des champs de cette base).

- La quatrième sert à rendre actifs les champs utiles pour toutes les bases.
- La cinquième est remplacée par celle venant du descripteur de chaque base, puisqu'il s'agit de définir les séparateurs spécifiques à chaque champ.

L'extraction de la terminologie venant des trois bases est alors pilotée par le descripteur générique, par contre, les spécificités de chaque base (séparation des documents, noms des champs et techniques de découpage) sont tour à tour prises en charge par le descripteur associé à chaque base. Nous avons ainsi pu fusionner, sans aucun reformatage, les contenus de tous les champs utiles de 12 bases bibliographiques issues du Web invisible.

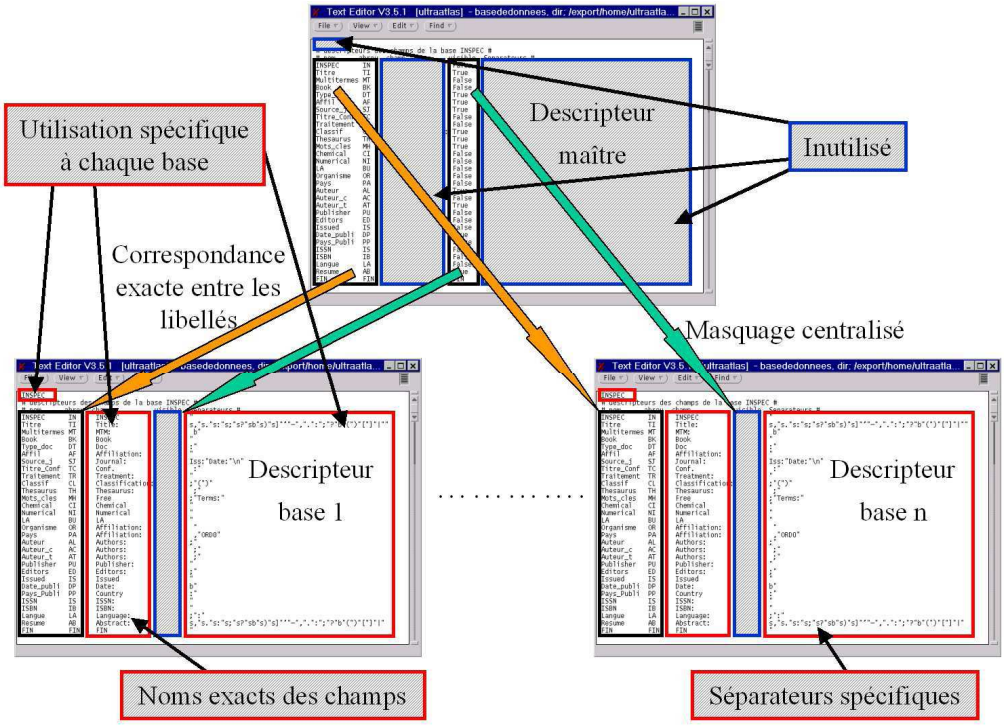


FIG. 2 - *Principe des méta-données de premier et de second niveau.*

5 Multi-termes et synonymie

L'extraction d'information permet de passer d'une représentation brute du texte (succession de mots) à une succession d'unités terminologiques. Les différents items sont alors identifiées (titre, auteur, etc.) par association de chaque modalité et de sa fréquence dans le corpus. Il est important de souligner qu'une unité terminologique peut être sous la forme de

radicaux auxquels il est possible de ramener certains uni-termes par des algorithmes de radicalisation (Lovins, 1986), (Porter, 1980), de (Paice, 1996) et de Carry (Paternostre et al., 2002), c'est-à-dire par suppression de tous les affixes d'un mot. Par affixes on entend suffixe (cafe-**tière**), préfixe (**sur**-population) et infixe (rêv-**ass**-er).

Enfin, il peut être un mot composé qu'on nommera alors « multi-terme » (Dousset et kanoun, 1998). Si nous prenons par l'exemple du multi-terme « date-mining », la plupart des méthodes traditionnelles disoocient le mot « data » du mot « mininig ». L'objectif de la méthode des multi-termes est d'associer « data » à « mininig ». Les traitements statistiques (Diday, 2005) (Diday, 2008b) qui seront appliqués par la suite porteront alors sur « data mining » et non pas sur les deux termes distincts.

La détection des multi-termes s'effectue de la manière suivante. L'utilisateur choisit dans un premier temps les traitements de détection des multi termes qui lui semblent les plus adéquats. Pour chaque champ sémantique (mot-clés, titres, résumés…), un dictionnaire est crée, contenant toutes les valeurs du champ rencontrées. Puis, ces dictionnaires sont fusionnés et les doublons sont supprimés. Tous les mots composés, séparés par des tirets dans les titres, résumés, texte intégral, sont conservés sans leurs acronymes, générant alors un dictionnaire unique de la spécialité. L'utilisateur peut ajouter manuellement d'autres multi-termes à ce dictionnaire.

Un dictionnaire de synonymie est généré pour prendre en compte les acronymes, ainsi que les variations morphologiques (inversion, terminaisons, pluriels,…). Grâce à ce dernier, le système considère le multi-terme et sa variante comme une même et seule entité. Ici aussi, l'utilisateur peut compléter manuellement ce dictionnaire en rajoutant à la main d'autre synonymies de mutli-termes.

6 Conclusion

Dans cet article nous avons présenté une approche originale de traitements de multi-sources dans un contexte d'Intelligence économique. Dans laquelle nous traitons les informations dans leurs formats natifs, et ceci dans le but d'avoir une meilleure réactivité et une facilité de mis-à-jour du corpus. Cette démarche permet d'optimiser le processus de veille et de surveiller n'importe quel type d'informations, puisque plus de 90% des cas rencontrés peuvent ainsi être traités sans aucun reformatage.

Les sources ne sont bien évidemment pas équivalentes, aussi est-il intéressant de les combiner pour avoir une information plus exhaustive et moins biaisée. Deux bases qui indexent le même journal ne choisissent pas les mêmes articles. L'une va proposer le résumé et de bons mots-clés, l'autre les citations et toutes les adresses. Faut-il dédoublonner ou doit-on conserver toutes les informations complémentaires ? La question reste posée. Nous laissons toujours le choix aux utilisateurs, mais nous préconisons de garder toutes les versions disponibles, car une information redondante est moins gênante qu'une information partielle ou manquante.

Références

BOUGHANEM M., CHRISMENT C., MOTHE J., SOULE-DUPUY C., TAMINE L. (2000). *Connectionist and genetic approaches to perform IR. F. Crestani and G.* Pasi edi

tors. Soft computing in information retrieval: techniques and applications. Physica Verlag, pages 173-198, Heidelberg.

CHRISMENT C., HERNANDEZ N., GENOVA F., MOTHE J. (2006). *D'un thesaurus vers une ontologie de domaine pour l'exploitation d'un corpus.* AMETIST, INIST, Vol.0, pages 59-92.

CHRISMENT C., HAEMMERLE O., HERNANDEZ N., MOTHE J. (2008). *Méthodologie de transformation d'un thesaurus en une ontologie de domaine.* Revue d'Intelligence Artificielle (RIA) 22(1) :7-37.

DKAKI T., MOTHE J., DOUSSET B., CHRISMENT C. (1996*). Extraction et synthèse de connaissances à partir de bases de données hétérogènes.* INFORSID'96, pp 287-308, (Bordeaux, France).

DKAKI T., DOUSSET B., MOTHE J. (1997). *Recherche de l'information stratégique dans les bases de données : veille scientifique et technique.* 15ème congrès INFORSID, INFORSID'97, pp 673-690.

DKAKI T., DOUSSET B., EGRET D., MOTHE J. (2000). *Information discovery from semi-structured sources – Application to astronomical literature.* Computer Physics Communication.

DIDAY E., (2005). *De la statistique des données à la statistique des connaissances: avancées récentes en Analyse des Données Symboliques.* EGC 2005, page 703.

DIDAY E., (2008). *Principe d'Analyse des données symboliques et application à la détection d'anomalies sur des ouvrages publics.* EGC 2008, pages 211-212.

DOUSSET B. (2003). *Intégration de méthodes interactives de découverte de connaissances pour la veille stratégique.* Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse.

DOUSSET B., KANOUN S., *Optimisation du choix de la terminologie pour la reformulation de requêtes : cas des mutli-termes.* VSST'98, pp107-119.

KLEINBERG J.M. (1999). *Authoritative sources in a hyperlinked environment.* Journal of the ACM, pages 604-632.

LOVINS J. B., (1986). *Development of a Stemming Algorithm.* Machanical Translation and computational Linguistics, pages 22.31.

MANIEZ J., DE GROLIER E. (1991). *A decade of research in classification*.

MARON M., KUHNS J. (1960). *On relevance, probabilistic indexing and information retrienval.* Journal of the Association for Computing Machinery, pages 216-244.

MARTRE H. (1994), *IE et stratégie des entreprises,* Œuvre Collective du Commissariat au Plan, Paris, la documentation Française.

MOTHE J. (1994). *Modèle connexionniste pour la recherche d'informations – Expansion dirigée de requêtes et apprentissage*. Thèse de doctorat, Université Paul Sabatier, Toulouse.

MOTHE J. (2000). *Recherche et exploitation d'information – Découverte de connaissance pour l'accès à l'information.* Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse.

PAICE C. (1984). *Soft evaluation of Boolean search queries in information retrieval systems.* Information Technology: Research and Development, pages 33-42.

PAICE C. (1996). *Method for evaluation of stemming algorithms based on error counting.* Journal on the American Society for Information Science, pages 632-649.

PATERNOSTRE M., FRANCQ P., SAERENS M., LAMORAL J., WARTEL D. (2002). *Carry, un algorithme de désuffixation pour le français.* URL : http//www.galilei.ulb.ac.be

PAZIENZA M-T. (1997). *Information Extraction: A multidisciplinary approach to an emerging information technology*. International summer school, SCIE, ISBN 3-540-63438-X.

PORTER M. F. (1980), *An algorithm fur suffux stripping.* Program, pages 130-137.

ROBERSTON S.E. (1977). *The probability ranking principle in IR.* Journal of Documentation. Pages 294-304.

ROCCHIO J. (1971). *Relevance feedback in information retrieval*. Prentice Hall Inc., Englewwod Cliffs, NJ.

ROUSSOU-HANS F. (1998). *L'analyse de corpus d'information comme support de la veille stratégique.* Document numérique. Volume 2, page 189.

SALLES M., CLERMONT Ph., DOUSSET B., (2000). *Une méthode de conception de système d'IE*. Communication au colloque IDMME'2000, Montréal.

SALTON G. (1971). *A comparison between manual and automatic indexing methods.* Journal of American Documentation. Pages 61-71.

SALTON G., MCGILL M. (1984). *Introduction to modern information retrevial*. McGraw-Hill Int. Book Co.

SAUVAGNAT K. (2005). *Modèle flexible pour la recherche d'information dans le corpus de documents semi-structurés.* Thèse de doctorat, Université Paul Sabatier , Toulouse, 2005.

Event Annotation based on Machine Learning

Aymen Elkhlifi*, Rim Faiz **
*LaLIC,Université Paris-Sorbonne, 28 rue Serpente, 75006, Paris, France
Aymen.Elkhlifi@paris4.sorbonne.fr
**LARODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie
Rim.Faiz@ihec.rnu.tn

Abstract. After the beginning of the extension of current Web towards the semantics, the annotation system starts to take a significant role. In fact, it participates in giving the semantic aspect to the different types of documents. Daily, several news agencies publish thousands of articles concerning several events of all types (political, economic, cultural, etc.). The decision makers find themselves in front of a great number of events, a few of which concern them. The automatic processing of such events becomes increasingly necessary. Thus, we propose a machine learning-based approach that allows annotating news articles to generate an automatic summary of the events. We propose a new similarity measurement between events and we validate our approach by the development of the "AnnotEv" system.

1 Introduction

Acquiring knowledge from texts is a need which has increased during the last years. With the considerable rise of the documents volumes available in electronic format, it is necessary to extract and filter relevant information from the contents of those documents. As an example, the stock market events are numerous and diversified. The stock market experts must analyze these events in a relatively reasonable time in order to make important decisions. It is a question, therefore, of annotating the documents presenting the events to be able to extract those which are relevant. In this respect, our work aims at the development of an approach that annotates the news articles.

Furthermore, with the proliferation of news articles from different Web sources, summing up such information is becoming increasingly indispensable. Due to the large number of news sources (such as: BBC, Reuters, CNN, Aljazeera, etc.), everyday, thousands of articles are produced in the entire world concerning a given event. That is why we should think of automatizing the annotation process of such articles.

The documents indexing and events extracting are becoming more and more tiresome. Thus, we are urged to generate an easily consultable semantic annotation that takes into consideration the increase in size of the document and enriches its indexing. By looking a given event via a sequential analysis of the article, we noticed that there are sentences which do not refer to any event.. We also remarked that several other sentences refer to the same event. That is why we intend to eliminate the non-event sentences and to group the others in cluster form.

Accordingly, our work focuses on the annotation of documents: First, we prepare the text in a preprocessing stage. Second, we omit the non-event sentences. Then, we group the sentences indicating the same or similar events. Finally, we generate a summary article.

* We thank the France embassy in Tunis IFC for their financing

The rest of the document is organized as follows: Section (2) introduces the related work on annotation methods. Then, the particular methods of temporal information annotation are exhibited. In section (3), we present our approach for automatic events annotation. In order to validate our survey, we describe the different steps we followed to carry out the AnnotEv system. The experimentation is described in section (4). In section (5) we evaluate the system in order to demonstrate its ability. In section (6) we conclude with few notes and some perspectives.

2 Related Works on Methods of Annotation

It is worth noting that the annotation definition varies according to the application domain: Linguistics, E-learning, Biology, Web application etc. But it can be said that the annotation is all graphic or textual information attached to a document. It refers to various entities including a set of documents, a document, a passage, a sentence, a term, a word or an image.

Several methods and techniques are used for the current annotation systems such as contextual exploration (Desclés, 1997), conceptual graphs (Roussey et al, 2005), meta-thesauri (Khelif et al., 2007), linguistic indicators (Dau et al., 2004) and machine learning (Elkhlifi et al, 2007). We describe, thereafter, the principal existing systems of annotation for instance:

SyDoM (Roussey et al., 2002) is a semantic annotation system of Web pages. It allows the enrichment of these pages in order to find them without taking account of their writing language. It is devoted to the management of textual documents stored with XML formats. We remark that SyDoM has two main advantages: firstly, multilingual research and secondly, the improvement of the Web pages representation. But, we notice that SyDoM can carry out research only on Web pages that have been already annotated, namely it is unable to interrogate Web pages when the annotations were created using different semantic thesauri.

EXCOM (Desclés et al., 1997) is an annotation engine that uses a set of linguistic tools which aim at annotating a document by a bloc of internal/external knowledge. This engine is under development and, at the present time, it allows the production of a temporal organization of the stories and an automatic reformulation of the questions. Yet, we observe that an important part of this system is still not implemented, i.e. the semantic indexing of the documents by taking account of annotated information.

Annotea (Kahan, 2001) is a collaborative Customer Server system for the annotation of documents. The annotations are stored on a specialized server. They are divided in such a way that anyone which has an access to the annotation server will be able to consult all annotations related to a given document and to add his/her own annotations. These annotations can be typographical comments, corrections, assumptions or estimates. This system was developed using the W3C standards. Nevertheless, the only possible form of annotation is the text; we cannot annotate by images or icons.

The system developed by the ACACIA team (Khelif et al, 2007) allows the annotation of genes. It helps the biologists carry out the experiments on the biochips to validate and interpret the obtained results. Its need emanates from the fact that biologists spend too much time to scan and analyze the key words corresponding to genes and the biological phenomena studied in documents or genetic databases.

All previous works are interested in general documents annotation like scientific articles, Web documents and multimedia documents. Only few of them focus on the events annotation. Among these works we can mention: The annotation of temporal information in texts

(Muller et al., 2004): this work focused more specifically on relations between events introduced by verbs in finite clause. It proposes a procedure that achieves the task of annotation and a way of measuring the results. The authors of this work tested the feasibility of this procedure on newswire articles with promising results. Then, they developed two measures of evaluation of the annotation: Fineness and Consistency.

The annotation scheme for annotating features and relations in texts (Setzer et al., 2000): it enables to determine the relative order and, if possible, the absolute time of the events reported in them. Such a scheme could be used to construct an annotated corpus which would yield the benefits normally associated with the construction of such resources. It can be also used to better understand the phenomena. Moreover, it represents a resource for training and evaluation of adaptive algorithms to automatically identify features and relations of interest. However, we noted that this work is based only on the temporal markers to determine the relations between events. This technique is not completely correct since there exist implicit inter-events relations which are expressed without using temporal markers.

The annotation of time (Mani et al., 2001) with a canonized representation of the times expressions: a method was described for extracting such time expressions in multiple languages. The annotation process is divided into two steps: first, flagging a temporal expression in a document (based on the presence of specific lexical trigger words) and, second, identifying the time value that the expression designates or the speaker intends for it to designate.

We note that the temporal information annotation are generally concerned with the detection of dates and temporal markers (Setzer, 2000), event descriptions and finding the events date (Faiz et Elkhlifi, 2008, 2009) and the temporal relations between events in a text (Muller, 2004).

However, in our study we are interested rather in the annotation of the events in the form of metadata on the document.

3 The Proposed Approach of Event Annotation

We note that the aforementioned approaches for the annotation of temporal information were mainly linguistic. As well, they are based on the temporal indices. We are however interested in event annotation and exploitation based on machine learning.. Our approach is not restricted to the events detection, but it allows also gathering the similar events in order to facilitate an ulterior treatment: indexing, storage in a database, summarization, etc.

The automatic process of documents annotation which we present is carried out in four stages (see Figure 1):

1. **Preprocessing**: it consists, on the one hand, in the segmentation of text and, on the other hand, in the identification of entities.
2. **Events annotation**: it uses a classifier playing the role of a filter for the non-event sentences.
3. **Clustering**: it consists in gathering the sentences referring to the same or similar events. We propose in this stage a new similarity measurement between the events.

4. **Document annotation:** it takes various forms such as: sentences, form, concept, according to the field of application of our approach (Elkhlifi et al., 2009)

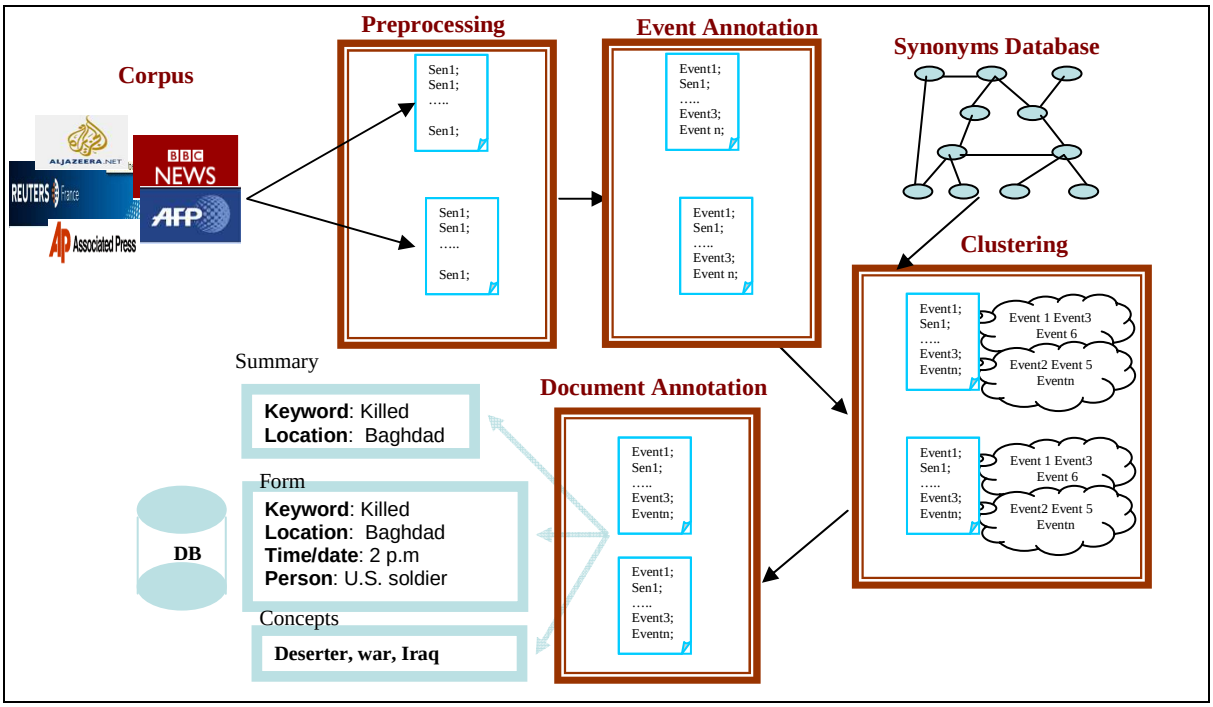


FIG. 1 – *Proposed approach for automatic annotation of events*

3.1 The Preprocessing

In the case study, the preprocessing consists in the application of some Natural Language Processing (NLP) tools to the rough text in order to segment and annotate entities.

We remarked that the segmentation is often neglected by the annotation systems in spite of its importance compared to the annotation quality. In addition, the entity identification is often used in other contexts like the question-answer systems. We present, thereafter, the segmentation and the named entity recognition.

The segmentation: is the determination of the sentences borders. It is a hardly-realisable task. Given that a point followed by a capital letter is not enough to detect the end or the beginning of a segment, it is necessary to take into account all typographical markers. Moreover, other linguistic bases are engaged like the syntactic structure of a sentence and the significance of each typographical marker in a well defined context. The existing tools segment the well structured texts into paragraphs.

There exist some works related to the monolingual segmentation, in French language, English and Arabic (Belguith et al., 2005). Other more recent works considered the multilin-

gual aspect, like the work of Mourad (2001) which proposed an approach that consists in defining a textual segment starting from a systematic study of the punctuation marks.

We developed our own segmentor while basing ourselves on punctuation marks. Due to the great number of the linguistic rules to program, we have to integrate in our knowledge base all the rules developed in the system Segatex (Mourad, 2001). The result of the segmentation of a text is to detail below.

```
<?xml version="1.0" encoding="UTF-8"?>
   <article lang="Ang">
      <para id="1">
         <title id="1">
            <phrase id="1">
               Iraqi leader denies civil war as 50 people die
            </phrase>
         </title>
      </para>
      <para id="1">
         <phrase id="2">
            BAGHDAD, Iraq (CNN) -- On a day in which at least 50 people were killed, Iraqi
            Prime Minister Nuri al-  Maliki said he did not foresee a civil war in Iraq and that
            violence in his country was abating.
         </phrase>
      </para>
      <para id="3">
         <phrase id="3">
            <quote id="1"> « In Iraq, we'll never be in civil war, » </quote>
              al-Maliki told CNN's  « Late Edition » on Sunday. </phrase>
      </para>
         <para id="4">
         <phrase id="4">
            Attacks on American troops around the Iraqi capital Sunday left six soldiers dead, the U.S
            command in Baghdad reported.
         </phrase>  <!-- ..... –
   </article >
```

FIG. 2 – *Extract of the article segmentation " Iraqi leader denies civil war as 50 people die ". BBC 2008*

Named Entities are types of particular lexemes which refer to an entity of the concrete world in certain fields, namely human, social, political, economic or geographical and which has a name (typically a proper name or an acronym). The entities are identified in the documents by a tag whose type corresponds to the types of answers. The types selected are recognized by rules thanks to the joint exploitation of two information sources (Elkhlifi et al. 2008):

- General lexicons allowing finding syntactic and semantic features associated with the simple words in complement of the lexical features.
- Dictionaries of named entities.

It may be noted that the named entity recognition is in the middle of text information extraction. The majority of the current systems are able to annotate the dates and the places. It is possible to meet within one document several mentions which refer to only one entity.

Figure 3 presents the same text of Figure 2 which named entities were extracted. Initially the position of each term is fixed. Then, the annotations concerning each entity are mentioned by specifying their attributes (Figure 3).

```
<?xml version="1.0" encoding="windows-1252" ?>
  <!-- The document content area with serialized nodes –
     <TextWithNodes>
          <Node id="4"/>Iraqi<Node id="9"/>  <Node id="10"/>leader<Node id="16"/>
          <Node id="17"/>denies<Node id="23"/>  <Node id="24"/>civil<Node id="29"/>
          <Node id="30"/>war<Node id="33"/>  <Node id="34"/>as<Node id="36"/>
          <Node id="37"/>50<Node id="39"/>  <Node id="40"/>people<Node id="46"/>
          <Node id="47"/>die<Node id="50"/> <Node id="56"/>BAGHDAD <Node id="63"/>
       </TextWithNodes>
  <AnnotationSet>
    <Annotation Id="485" Type="Location" StartNode="56 " EndNode ="63">
          <Feature>
               <Name className="java.lang.String">rule2</Name>
               <Value className="java.lang.String">LocFinal</Value>
          </Feature>
          <Feature>   <Name className="java.lang.String">rule1</Name>
                    <Value className="java.lang.String">Location1</Value> </Feature>
          <Feature> <Name className="java.lang.String">locType</Name>
               <Value className="java.lang.String">city</Value> </Feature>
    </Annotation>          <!-- …… –
  </AnnotationSet>
```

FIG. 3 – *Extract of the article figure 1 after the named entity recognition*

The result of the first stage of our approach is the set of segmented sentences the entities of which are annotated.

3.2 Events Annotation

An event is a specific object which occurs at one specific moment and in a well defined place. Our objective is to identify all events present in a document. We mark each detected event by a tag. Accordingly, a model of classification is built automatically from the training set which permits to predict whether a sentence contains an event or not. We initially used the following attributes: Length of the sentence; Numbers of capital letter; Numbers of stop words; Number of city/town and Number of numerical marks.

Within the framework of our study, and through the analysis of the news articles, we noticed that the addition of other attributes to the preceding list is possible like, for example, the temporal markers (after, before, simultaneously, etc.) and the calendar terms (Sunday, 9/12/2004, Mars). The problem of choosing significant attributes can be solved by using feature selection algorithms which leads to select a subset of relevant attributes in order to find a predictive model. There is a variety of feature selection algorithms (chi-public garden, Relief and Principal component analysis). After having carried out experimentation, we added the attribute "number of calendar terms".

Several machine learning techniques can be used for classification problem such as the neural network, the decision tree, the Bayesian network, etc. We chose the decision tree for numerous reasons; it is easily interpretable by people. Moreover, the decision tree construction is less skeletal compared to the other techniques in such a way to allow the reduction of the system complexity.

The training set is annotated by experts. For each news article the events are annotated as follows: the annotator is brought to assign labels for each sentence representing an event. If a sentence refers to an event, they assign the label "yes", if not "no". We applied to this same training set various algorithms of decision trees construction. Then, we chose the model which has the biggest PCC (Percentage Correctly Classified).

The result of this stage is the set of sentences referring to events. Moreover, the classification of sentences as an event or not, represents a kind of filtering; on the basis of a segmented text, we filter the non-event sentences.

In this stage we gather the sentences referring to the same or similar events by the application of the algorithm 'Hierarchical Agglomerative Clustering (HAC)'. This algorithm initially assigns each object with a cluster, then collects on several occasions the clusters until one of the stop criteria is satisfied.

Our contribution consists in putting forward a new similarity measurement between the events. Given the importance of similarity measurements in clustering, we noted that there are several of such measurements between documents including: Salton's cosinus, Khi-Deux distance, Cosine in distributional space (Salton et al., 1983). Other measurements, which are more interesting for us, are linked to the similarity between sentences, the most recent of which is Naughton's measurement (Naughton et al., 2006).

We put forward a new similarity measurement between events inspired by tf-idf "weight term frequency–inverse document frequency". This measurement also takes into account the clusters position in the article.

In order to gather sentences expressing the same or similar event by two different lexicons, we use a synonyms database for the replacement of the instances by their classes.

For example, let us have the two following event- sentences, initially considered as two clusters C_1 and C_2.

C_1: In Baquba, two separate shooting incidents left six dead and 15 wounded Sunday afternoon.

C_2: In other attacks reported by security and hospital officials, two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the southern city of Basra killed five and injured 15.

We notice that the words (bombardments and bombings), (wounded and injured) imply the same meanings. Hence, there is a need to replace these words by their classes from the synonyms database in order to increase the similarity between both clusters. In general, the similarity between two classes expressing the same or similar event by means of two different lexes. We define SIM between two clusters C_1 and C_2, then, as follows:

$$SIM\ (C_1, C_2) = \frac{\sum_{j=1}^{t} Ct_{1j} Ct_{2j}}{\sqrt{\sum_{j=1}^{t} Ct_{1j}^2 + \sum_{j=1}^{t} Ct_{2j}^2}}$$

with Ct_{ij} as the weight of each term in a cluster after the replacement of instances by their classes from synonyms database. It is calculated as follows:

$$Ct_{ij} = tf(t_i, c) \times \log\left(N / df(t_i)\right) \text{ with:}$$

- *tf(t_i , c)* the frequency of the term t_i in a cluster *c*.
- *N* the number of clusters.
- *df (t_i)* the number of clusters containing the term t_i.

Based on what has been said in so far, and by taking into consideration the sentence position in the article, we propose the new similarity measurement FSIM which combines the similarity between sentences and the distance between them:

With D (Ct_1, Ct_2) the distance between both clusters in the article and $\alpha \in [0, 1]$ fixed during the experimentation. Therefore, for N clusters, we have n×(n-1)/2 possible combinations.

It is important to group the sentences indicating the same or similar events since they will be gathered even if they use various words. Figure 3, for example, presents the application of HAC algorithm by using FSIM on a press article:

C_1: Iraqi leader denies civil war as 50 people die.

C_2: On a day in which at least 50 people were killed, Iraqi Prime Minister Nuri al-Maliki said he did not foresee a civil war in Iraq and that violence in his country was abating.

In Iraq, we'll never be in civil war," al-Maliki told CNN's "Late Edition" on Sunday.

C_3: One U.S. soldier was killed by gunfire in eastern Baghdad about 2 p.m.

C_4: U.S. commander wounded since 1 p.m

……

Levin, the ranking Democrat on the Senate Armed Services Committee, called for the United States to set a date to begin withdrawing its forces.

C_5: In Baquba, two separate shooting incidents Sunday afternoon left six dead and 15 wounded, officials said.

C_6 : Two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the southern city of Basra killed five and injured 15

FSIM (C_1, C_2) = **0.91** FSIM (C_1, C_3) = **0.12** FSIM (C_1, C_4) = **0.1** FSIM (C_1, C_5) = **0.05** FSIM (C_1, C_6) = **0.02**

FSIM (C_2, C_3) = **0.08** FSIM (C_2, C_4) = **0.1** FSIM (C_2, C_5) = **0.32** FSIM (C_2, C_6) = **0.36** FSM (C_3, C_4) = **0.84**

FSIM (C_3, C_5) = **0.28** FSIM (C_3, C_6) = **0.23** FSIM (C_4, C_5) = **0.19** FSIM (C_4, C_6) = **0.15** FSIM (C_5, C_6) = **0.79**

FIG. 4 – *first step of HAC*

The sentences in bold indicate an event. First of all, we calculate the FSIM between these sentences where each sentence is a cluster. Then, we obtain values at the bottom of figure 4. Besides, we Group together C_1 and C_2 into only one cluster CA (because they have the biggest FSIM). Finally, we reapply HAC on the new clusters.

After 3 iterations, we obtain two new clusters C_B and D_C (Figure 5):

Iraqi leader denies civil war as 50 people die.

On a day in which at least 50 people were killed, Iraqi Prime Minister Nuri al-Maliki said he did not foresee a civil war in Iraq and that violence in his country was abating.

In Iraq, we'll never be in civil war," al-Maliki told CNN's "Late Edition" on Sunday.

One U.S. soldier was killed by gunfire in eastern Baghdad about 2 p.m.

U.S. commander wounded since 1 p.m ……

Levin, the ranking Democrat on the Senate Armed Services Committee, called for the United States to set a date to begin withdrawing its forces.

In Baquba, two separate shooting incidents Sunday afternoon left six dead

and 15 wounded, officials said.

Two car bombings in the northern city of Kirkuk killed 10 and wounded 32, and a blast in the southern city of Basra killed five and injured 15

FSIM (C_A, C_B) = **0.14** FSIM (C_A, C_C) = **0.07** FSIM (C_B, C_C) = **0.09**

FIG. 5 last step of HAC

In this HAC iteration we stop the clustering since we have a value of FSIM inferior to the threshold of similarity, fixed initially at 0,6.

Using the clusters and their positions in the article we generate a description which combines the events and presents the annotation of the article under three types:

- To extract sentences which sum up the article.
- To structure the annotation in a standard form to store events in databases.
- To extract concepts (future work).

Thus, we continue the enrichment of the document by other metadata which will be very useful for all ulterior treatment (information retrieval, automatic summarization, question-answer systems, storage of the events in a database, indexation, etc).

A possible form of metadata is the forms filling, i.e. stoking the events in database while answering has well determined questions (where, when, who) for example:

Location: Baghdad; Time/date: 2 p.m.; Person: U.S. soldier; Keyword: Killed.

For each event sentence, we have this information since the preprocessing. After having stored this information in a relational database, we can find events by date, person, or time by a simple request on the selected fields. Another form of metadata is the automatic summary, which consists in marking the sentences which form the summary of a document. In general, the goal of a summary system is to produce a condensed representation of the contents where the important information of the original text is preserved. It is also necessary to consider the user needs and the specified task.

In this context, we propose an informative summary containing the essential information of the article. This summary is also selective since it neglects the general aspects of the article. In addition, it can be said that it is targeted since it is correlated to the events.

For each cluster generated by the third stage we annotate the article by the principal events it contains. We use the following heuristics: the sentence having the maximum value attributes in the classification stage is the best to annotate the cluster. Let us take again the preceding example, the summary of which is presented as follows (Figure 6):

Iraqi leader denies civil war as 50 people die.
One U.S. soldier was killed by gunfire in eastern Baghdad about 2 p.m.
In Baquba, two separate shooting incidents Sunday afternoon left six dead and 15 wounded, officials said.

FIG.6. Summarization of article

4 Experimentation

To validate our approach, we develop the AnnotEv system with Java language; Visual Studio platform, particularly Visual J++.

AnnotEv includes the four following modules:

- Module 1: The segmentation and the recognition of the named entities
- Module 2: Events annotation.
- Module 3: Clustering.
- Module 4: Automatic summary.

We prepared a corpus containing 82 articles which are related to the Iraq War published in 2006 starting from 5 news agencies: (CNN: 13 articles), (Reutres: 17 articles), (BBC: 14 articles), (Associated Press: 22 articles) and (AFP: 16 articles).

The average length of a sentence is of 12.23 words, with an average of 6.05 events per article, for a total of approximately 91.500 words, 7479 sentences and 496 events.

We also use the Weka libraries edition 3.5 and Gate 4.0 (Bontcheva, 2005) respectively for decision tree and entity identification.

After removing the images and the legends of the article we segment them into sentences and, then, we annotate entities by calling upon the Gazettee-Annie method from Gate class. We use WordNet (Fellbaum et al., 1998) as a synonyms database. Besides, we annotate the events and group them according to their similarities. We develop several interfaces to en-

sure the management of corpus and training set annotation. We also develop other interfaces for the events search and summarization. We have to integrate a speaker agent who is able to read the texts in two different versions. For the system user, it is enough to select a new article in order to listen to its summary.

The training set is part of the group of obtained sentences after the preprocessing stage. It is annotated by two experts. For each sentence the default value of the attribute “Event” is 'No' (sentence not indicating an event), the commentator has to put

'Yes' if the sentence refers to an event. An ARFF file (input format of Weka) is generated automatically for each article. It will be useful like a data source for the algorithms of classification. We adopted J48, ADTREE and Random Tree with the cases of the events.

To evaluate the method of clustering we employ the precision and recall measurements. We assign each pair of sentences to one of the four following categories:

- a: grouped together (and annotated like referring to the same event).
- b: not grouped together (but annotated as referring to the same event).
- c: grouped inaccurately together.
- d: correctly not grouped together.

The Precision, the Recall and F1 prove that to be calculated as:

$$P = \frac{a}{a + c}, R = \frac{a}{a + b} \text{ and } F1 = \frac{2 \times P \times R}{(P + R)}$$

1. Results

The evaluation is done at several levels. We start with the classification evaluation by using the PCC, then, the clustering by measuring the Precision and the Recall. We exploit the following algorithms:

J48: implementation of C4.5 algorithm (Quinlan, 1993) which selects for each level the tree node as the attribute which differentiate better the data. Then, it divides the training set into sub-groups in order to reflect the values of the attribute of the selected node. We repeat the same treatment for under group until we obtain under homogeneous groups (all the instances or the majority have the same attribute of decision).

ADTree: construction of the decision trees extended to the cases of multiclasses and multi-labels.

Random Tree: begin with tree random and chosen by the majority best vote.

We obtained the following results:

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.706	0.313	0.706	0.706	0.706	yes
0.688	0.294	0.688	0.688	0.688	no

TAB. 1 – *Result with Random Tree.*

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.824	0.25	0.778	0.824	0.8	yes
0.75	0.176	0.8	0.75	0.774	no

TAB. 2 – *Result with J48.*

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.765	0.313	0.722	0.765	0.743	yes
0.688	0.235	0.733	0.688	0.71	no

TAB. 3 – Result with ADTree.

For the clustering stage, we obtained an improvement of Recall (R) and Precision (P) and the function F1: R = 85%, P=87% and F1=73.333%.

This improvement is made of the similarity measurement that we proposed. Indeed, it detects the similarity between the sentences even if it contains different terms.

2. Conclusion and Future Work

In this article we have proposed four stages to annotate press articles starting, in a first stage, by the preprocessing that consists in applying NLP tools to prepare the data. In a second stage, the filtering of the non event-driven sentences has been done by dint of a classifier. In the third stage we have gathered the sentences in clusters according to their degree of similarity (FSIM). Finally, we have generated an automatic summary of the principal events constituting the article. Our approach was evaluated on news articles corpus concerning the Iraq War published in 2008.

This approach comes within the framework of the Information Extraction from texts, particularly the extraction and the exploitation of the events. Actually, it constitutes a considerable target in many application domains like the national security, the economy or the industry. In such fields the concepts of technological/ economic survey became essential, in particular for the help in decision making: definition of strategies, placement towards competition, etc. In short term, one of the first future works which we propose is to adopt AnnotEv for news articles in Arabic. In long term, we look forward to fuse the events. In effect, we have the idea of adopting, to the case of the events, the MCT model for the fusion of information in general.

Références

Belguith Hadrich L., L. Baccour, G. Mourad (2005). *Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules*. In Actes de la 12ème conférence TALN'2005, Dourdan France, Vol. 1, p. 451–456.

Bontcheva K., V. Tablan et D. Maynard, H. Cunningham (2004). *Evolving GATE to Meet New Challenges.* in Language Engineering. Natural Language Engineering. pp. 349-373.

Dau F., M-L Mugnier, G. Stumme (2005). *Conceptual Structures: Common Semantics for Sharing Knowledge*, 13th Inter Conf on Conceptual Structures 2005.

Desclés J.P. (1997*). Systèmes d'exploration contextuelle, Co-texte et calcul du sens, Claude Guimier*, Presses de l'université de Caen, pp. 215-232.

Elkhlifi A., R, Faiz (2009). Automatic Annotation Approach of Events, News Articles. *International Journal of Computing & Information Sciences (IJCIS)*, December 2009.

Elkhlifi A., R, Faiz (2007) .*Machine Learning Approach for the Automatic Annotation of Events*. Proceedings of the 20th Intern FLAIRS AAAI Press, California, ,pp 362-367

Elkhlifi A., R, Faiz (2008). *AnnotEv : Système d'annotation des événements*. Actes des huitièmes journées scientifiques en Génie Electronique et Informatique (GEI 2008) session spéciale Web Sémantique, Sousse, Tunisie, pp 173-182.

Faïz R., A, Elkhlifi (2009). *Annotation sémantique des événements*, in Annotations automatiques et recherche d'informations ", Eds. dir. Desclés Jean-Pierre, Le Priol Florence, Hermes - Traite IC2 -- Serie Cognition et Traitement de l'information.

Faïz R., A, Elkhlifi (2008). *Approche d'annotation automatique des événements*, Revue des Nouvelles Technologies de l'Information (RNTI), D. A. Zighed et G. Venturini (Editeurs), Vol.1, pp 37-42

Fellbaum C., J. Grabowski, S. Landes (1998). *Performance and confidence in a semantic annotation task*, In WordNet: an electronic lexical database, Language, Speech and Communication, chapter 9, pp. 216-237. Cambridge, Massachusetts: The MIT Press.

Frank E., I. Witten (2005). *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann.525 pages.

Kahan J., M-R. Koivunen (2001). *Annotea: an open RDF infrastructure for shared Web annotations*. Proceedings of the 10th international conference on WWW.

Khelif K., R. Dieng-Kuntz, P. Barbry (2007). *An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain*, in J. UCS 13(12).

Mani I., L. Ferro, B. Sundheim, G. Wilson (2001). *Guidelines for Annotating Temporal Information*. In Human Language Technology Conf..

Mourad Gh. (2001) *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique de citations*. Thèse de doctorat Univ Paris-Sorbonne.

Muller P., X. Tannier (2004). *Annotating and measuring temporal relations in texts*. In Proceedings of Coling, volume I. Genève, ACL.

Naughton M., N. Kushmerick, and J. Carthy (2006). *Event extraction from heterogeneous news sources.* Proc Event Extraction and Synthesis, American Nat. Conf. A I.

Quinlan J. R. (1993). C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers.

Roussey C., S. Calabretto (2005). *An experiment using Conceptual Graph Structure for a Multilingual Information System*, in the 13th Inter Conference on Conceptual Structures.

Salton G., M.J. Mac Gill (1983). *Introduction to Modern Information Retrieval.* In International Student Edition.

Setzer A., R. Gaizauskas (2000). *TimeML: Robust specification of event and temporal expressions in text*. In The second inter conf on language resources and evaluation.

Résumé

Après les tendances vers le Web Sémantique, l'annotation commence à prendre un rôle significatif, puisque elle donne un aspect sémantique à l'information. Dans cet article, nous proposons une approche, qui se base sur l'apprentissage automatique, et qui permet d'annoter les événements pour générer un résumé automatique. Nous avons validé notre approche par le développement du système "AnnotEv".