

Extraction d'indices spatiaux et temporels dans des séquences vidéo couleur

Sébastien Lefèvre*, Nicole Vincent*

* LSIIT – Université Louis Pasteur (Strasbourg I)
Parc d'Innovation, Bd Brant, BP 10413, 67412 Illkirch Cedex
lefevre@lsiit.u-strasbg.fr

** CRIP5 – Université René Descartes (Paris V)
45 rue des Saints Pères, 75270 Paris Cedex 06
nicole.vincent@math-info.univ-paris5.fr

Résumé. Dans cet article, nous considérons les séquences vidéo couleur comme des données complexes. Notre contribution porte sur deux méthodes adaptées à ce type de données et permettant d'extraire des indices spatiaux et temporels. Nous pensons que ces méthodes peuvent être intégrées avec succès dans un processus plus complexe de fouille de données multimédia, aspect qui ne sera pas abordé ici. Les méthodes présentées sont basées sur l'espace Teinte Saturation Luminance. L'extraction d'indices spatiaux est assimilée au problème de la séparation du fond et des objets, résolu par une approche multirésolution ne nécessitant qu'une seule image. L'extraction d'indices temporels correspond à la détection des changements de plans dans une séquence d'images, obtenue par l'utilisation de mesures de distances indépendantes du contexte. Les caractéristiques communes de nos deux méthodes sont l'utilisation de l'espace TSL, l'efficacité calculatoire, et la robustesse aux artefacts. Nous illustrons ces approches par des résultats obtenus sur des séquences vidéo sportives.

1 Introduction

A l'ère de la société de l'information et de la communication, les données numériques occupent une place de plus en plus importante et il devient nécessaire de disposer d'outils adaptés pour les traiter, les synthétiser, les fouiller. En particulier, les séquences vidéo issues des canaux télévisuels fournissent des volumes de données dont la taille ne permet plus aujourd'hui un parcours linéaire. L'accès aux éléments pertinents requiert la description des données par des indices.

Nous nous intéressons ici au problème de l'extraction d'indices dans les séquences vidéo. Puisque celles-ci sont le plus souvent composées d'images couleur, nous proposons d'utiliser l'espace couleur Teinte Saturation Luminance qui fournit des caractéristiques intéressantes. En se basant sur cet espace, nous cherchons tout d'abord à extraire des indices spatiaux, que nous assimilons aux différents éléments contenus dans les images : les objets et l'arrière-plan de la scène. Puis nous nous focalisons sur l'extraction d'indices temporels représentant les limites des différents plans d'une séquence. Notre article sera donc organisé de la manière suivante : après avoir présenté

l'espace TSL, nous décrirons nos méthodes d'extraction d'indices spatiaux et temporels, et enfin nous commenterons les résultats obtenus sur des séquences vidéo sportives.

2 L'espace Teinte Saturation Luminance

Le codage de la couleur dans des images numériques peut être effectué en utilisant différents espaces de représentation, appelés traditionnellement espaces couleur. Pour plus d'informations, le lecteur pourra se référer au récent ouvrage de Trémeau *et al.* (Trémeau et al. 2004) ou au panorama de Chang *et al.* (Chang et al. 2001) sur la segmentation couleur. Dans cet article, nous nous focalisons sur l'espace Teinte Saturation Luminance (TSL) que nous présentons ici.

L'espace TSL est représenté par trois composantes : la teinte, la saturation, et la luminance. Tandis que la saturation et la luminance sont codées "de manière classique" sous forme scalaire, la teinte est pour sa part une valeur angulaire. Ces composantes peuvent être interprétées de la manière suivante : la teinte représente la couleur perçue (rouge, jaune, vert, *etc.*) ; la saturation mesure la pureté de la couleur (par exemple pour une teinte rouge, le rose se caractérise par une saturation plus faible que le rouge, tandis que le noir, le blanc, et le gris sont caractérisés par une saturation nulle) ; la luminance représente le niveau de gris, de "sombre" pour une valeur faible à "clair" pour une valeur élevée. L'espace TSL fournit, au travers de ses 3 composantes, des informations complémentaires. Il a l'avantage de permettre l'élaboration de méthodes robustes aux changements d'illumination. En effet, ces artefacts affectent principalement la composante de luminance. En ne tenant compte que des composantes de chrominance (teinte et saturation), il est donc possible de diminuer la sensibilité aux changements d'illumination. Nous avons également observé que la teinte était une composante plus robuste que la saturation ou la luminance dans un cadre multirésolution, où les données peuvent être analysées à une échelle plus ou moins fine. En effet, la teinte est moins sensible aux artefacts dus à des moyennages successifs des valeurs des pixels, nécessaires dans le cas d'une représentation multirésolution pyramidale. Nous avons ainsi observé une indépendance vis-à-vis de la résolution de certaines mesures calculées sur les teintes (comme la plage de valeur ou l'écart-type).

La teinte est donc une composante intéressante, invariante aux changements d'illumination et aux cadres multirésolutions. Cependant elle doit être analysée avec précaution. En effet, sa fiabilité dépend du niveau de saturation et la teinte n'est significative que si la saturation est élevée. Les méthodes d'analyse basées sur la teinte des pixels doivent donc vérifier que ceux-ci ne sont pas achromatiques. Une autre contrainte de la teinte provient de sa nature mathématique (mesure angulaire) qui nécessite l'utilisation de mesures statistiques spécifiques. Ainsi, lorsqu'on cherche à mesurer la similarité entre deux valeurs, on est logiquement amené à calculer la différence absolue entre ces deux valeurs. Dans le cas de valeurs angulaires, la transposition est relativement triviale :

$$|\theta_i - \theta_j|_{\angle} = \min(|\theta_i - \theta_j|, 2\pi - |\theta_i - \theta_j|) \quad (1)$$

en notant $|\theta_i - \theta_j|_{\angle}$ la différence absolue de deux valeurs angulaires θ_i et θ_j . En cas de fusion ou de combinaison d'informations, la moyenne est fréquemment utilisée en

analyse d'images. Nous avons choisi d'utiliser la définition donnée dans (Mardia et Jupp 2000) pour le calcul de la moyenne $\bar{\theta}$ d'un ensemble de mesures d'angles $\{\theta_i\}_{i \in [1, \Theta]}$:

$$\bar{\theta} = \begin{cases} \omega & \text{si } \sum_{i=1}^{\Theta} \cos(\theta_i) \geq 0, \\ \omega + \pi & \text{sinon} \end{cases} \quad \text{avec } \omega = \text{Arctan} \left(\frac{\sum_{i=1}^{\Theta} \sin(\theta_i)}{\sum_{i=1}^{\Theta} \cos(\theta_i)} \right) \quad (2)$$

Dans (Mardia et Jupp 2000) est donné également un algorithme permettant le calcul de l'amplitude de variation de l'ensemble $\{\theta_i\}_{i \in [1, \Theta]}$. Cette méthode nécessite un tri croissant préalable de tous les angles considérés, et est donc caractérisée par une complexité algorithmique relativement élevée. Nous proposons ici une méthode applicable dans le cas où l'angle moyen $\bar{\theta}$ a déjà été calculé. Cette méthode nécessite, en plus de la moyenne, la connaissance des angles minimum et maximum dans l'intervalle de longueur 2π choisi, respectivement notés θ_{\min} et θ_{\max} . Si la moyenne appartient à l'intervalle limité par les deux angles, c'est-à-dire si $\theta_{\min} \leq \bar{\theta} \leq \theta_{\max}$, alors l'amplitude de variation est égale à $\theta_{\max} - \theta_{\min}$. Dans le cas contraire, l'angle complémentaire doit être considéré et l'amplitude de variation est égale à $2\pi - (\theta_{\max} - \theta_{\min})$.

Nous avons décrit ici l'espace TSL et introduit des mesures relatives aux valeurs angulaires. En se basant sur cet espace, nous allons maintenant décrire deux méthodes d'extraction d'indices spatiaux et temporels.

3 Extraction d'indices spatiaux

Au sein des trames d'une séquence vidéo, il est possible d'extraire différents types d'indices spatiaux au cours d'un processus de fouille de données. Nous avons choisi de considérer comme indices les régions (position, taille, *etc*) appartenant soit à des objets soit à l'arrière-plan de la scène. Ce choix nous semble opportun dans la mesure où la séparation fond/objets est cruciale dans de nombreuses applications, telles que le suivi d'objet, l'interprétation du contenu des images et des séquences d'images, ou encore la compression. En effet, la norme de compression MPEG-4 décrit une scène par les différents objets qui la composent et par son arrière-plan (Haskell et al. 1998).

Afin d'extraire des indices relatifs aux objets ou à l'arrière-plan d'une scène, il est nécessaire de disposer de plusieurs images et de les comparer : ainsi, les objets sont délimités par les zones de l'image dont le contenu a évolué au cours du temps. Si une image de référence (sans objet) est disponible, les objets correspondent aux zones de l'image analysée qui diffèrent de l'image de référence. Ce principe n'est malheureusement valable que si la caméra est statique. Dans le cas d'une caméra mobile, une étape coûteuse d'estimation/compensation de mouvement est nécessaire : le temps de calcul du processus de fouille croît alors considérablement. Nous présentons ici une méthode d'extraction d'indices spatiaux relatifs aux éléments d'une scène (objets et arrière-plan) qui ne requiert pas d'information de mouvement, pouvant être appliquée sur chaque image de manière indépendante. Notre méthode est adaptée aux scènes où l'arrière-plan occupe une partie du champ de vision plus importante que les objets (statiques ou en mouvement). Elle est basée sur le constat suivant : au fur et à mesure que la résolution d'une image diminue, les détails disparaissent et le contenu de l'image tend à représenter exclusivement l'arrière-plan au détriment des objets présents dans la scène.

Cette réflexion nous amène à proposer une approche multirésolution : en considérant une image originale I_0 , il est possible de diminuer fortement sa résolution pour obtenir une image à très faible résolution $I_{r_{\max}}$. Aucun objet n'est plus perceptible dans cette image qui n'est composée que de l'arrière-plan. Un modèle du fond peut donc être calculé à partir de cette image. En augmentant la résolution r de manière itérative, il est alors possible de comparer les différentes régions de l'image I_r avec le modèle de l'arrière-plan suivant un critère donné. Cette comparaison permet de déterminer si chaque région correspond ou non à l'arrière-plan. De plus, le principe de multirésolution permet d'analyser des données à un degré de précision plus ou moins fin selon le résultat attendu. Une analyse multirésolution a aussi généralement l'avantage de limiter le nombre de calculs nécessaires par rapport à son homologue monorésolution. L'inconvénient principal est lié à la difficulté du choix des résolutions initiale et finale de l'analyse. Le traitement décrit précédemment s'effectue donc en quatre étapes successives : création de la représentation multirésolution, estimation du modèle de l'arrière-plan, segmentation itérative aux différentes résolutions, et enfin segmentation finale et obtention des indices relatifs aux objets et à l'arrière-plan. La première étape consiste en la création de la représentation multirésolution de l'image, par décomposition pyramidale, où la valeur d'un pixel $P(x, y)$ à la résolution $r + 1$ est calculée comme la moyenne d'un ensemble de p^2 pixels à la résolution r . La taille de l'image dépend alors de la résolution. La moyenne a été préférée à d'autres mesures nécessitant des calculs plus importants, comme par exemple la valeur médiane. Le calcul est effectué itérativement, à partir de la résolution originale $r = 0$, et jusqu'à obtenir la résolution voulue $r = r_{\max}$. Il est alors possible de modéliser l'arrière-plan à l'aide de l'image $I_{r_{\max}}$, sommet de la pyramide. Cette considération n'est bien sûr valable que si l'arrière-plan occupe une partie importante de l'image, s'il diffère suffisamment des objets présents dans la scène, et s'il est assez homogène. Afin de garantir une faible complexité tout en assurant une robustesse aux changements d'illumination, nous avons décidé d'utiliser comme modèle la moyenne des teintes d'une région, puisque la teinte nous semblait une composante relativement robuste aux moyennages successifs nécessaires pour construire la représentation multirésolution de l'image. Le modèle de l'image I est noté $\varphi(I)$ et l'arrière-plan sera donc caractérisé par $\varphi(I_{r_{\max}})$. Ce choix requiert la validité de deux hypothèses : il est nécessaire que les pixels ne soient pas achromatiques (saturation non nulle) pour que la valeur de leur teinte soit fiable, et l'arrière-plan doit être de couleur (ou plutôt de teinte) homogène. La caractérisation de chaque région (et donc la génération des indices correspondants) peut ensuite être effectuée et améliorée de manière itérative (à la manière d'un *quadtree* incomplet) depuis la résolution $r_{\max} - 1$ jusqu'à la résolution initiale $r = 0$, base de la pyramide. A une résolution donnée r' avec $r_{\max} > r' > 0$, l'image $I_{r'}$ doit être analysée. Cette image est comparable à l'image initiale I_0 qui aurait été découpée en $K = (K_0)^{r_{\max} - r'}$ régions, avec K_0 une constante dont la valeur pourrait être en toute logique égale à p^2 . Chacune de ces régions est alors comparée avec le modèle de l'arrière-plan. Cet appariement entre deux régions I^m et I^n s'effectue en calculant sur les moyennes respectives des teintes la mesure de similarité $\delta(\varphi(I^m), \varphi(I^n)) = |\varphi(I^m) - \varphi(I^n)|_{\angle}$. Une fois cette mesure δ calculée entre le modèle d'une région donnée et le modèle de l'arrière-plan, elle est comparée à un seuil S_1 afin d'effectuer ou non la reconnaissance. Une valeur inférieure au seuil signifie

que la région est considérée appartenir à l'arrière-plan. Cependant, un second test est nécessaire pour vérifier la cohérence de la région étudiée et éviter les artefacts liés à l'utilisation de la moyenne. En effet, si une région est composée de deux pixels ayant des teintes opposées, la moyenne ne reflétera pas correctement le contenu de la région. Nous analysons donc la cohérence de chaque région appariée avec l'arrière-plan. Nous avons préféré l'amplitude de variation à d'autres mesures de dispersion telles que la variance pour évaluer ici la cohérence d'une région : plus la plage de valeurs d'une région est faible, plus celle-ci est homogène. Pour calculer cette plage de valeurs angulaires, nous n'utilisons pas la méthode donnée dans (Mardia et Jupp 2000) mais l'approche originale présentée précédemment. Une fois la plage de valeurs calculée pour une région I_r^k , nous comparons cette mesure de dispersion avec un second seuil S_2 . Une plage inférieure au seuil assure l'homogénéité de la région concernée. Celle-ci est alors étiquetée en fond ou arrière-plan. Dans le cas contraire, l'hétérogénéité de la région candidate à l'étiquetage implique son rejet. Si une région fournit une réponse positive à ces deux tests successifs, elle est affectée à l'arrière-plan. Dans ce cas la région n'est plus analysée à de meilleures résolutions. A l'opposé, une région sans étiquette sera analysée plus en détail à la résolution $r' - 1$. Cette segmentation est effectuée si nécessaire jusqu'à la résolution initiale $r = 0$. Dans le cas d'applications nécessitant une segmentation et des indices très précis, les régions correspondant aux objets peuvent être analysées par la suite afin d'affiner les contours des objets. Au contraire, si la précision des contours des objets n'est pas nécessaire, le processus peut être arrêté à une résolution r_{final} avec $r_{\text{max}} > r_{\text{final}} \geq 0$. Dans ce cas, nous considérons que les régions sans étiquette représentent les objets. Afin d'améliorer la qualité du modèle de l'arrière-plan, il est également possible de recalculer celui-ci au cours du processus de segmentation. Dans ce cas, le modèle obtenu à la résolution r_{max} ne représente que l'état initial de l'arrière-plan. A mesure que la résolution devient plus fine, les résultats sont plus précis, et il est possible d'obtenir un modèle de l'arrière-plan plus fiable en ne considérant que les parties de l'image déjà étiquetées comme appartenant au fond.

L'espace TSL, utilisé ici dans un cadre multirésolution pour extraire des indices spatiaux, peut également être employé pour fournir des indices temporels.

4 Extraction d'indices temporels

Après avoir montré comment les séquences vidéo couleur pouvaient être analysées dans l'espace TSL pour fournir des indices de nature spatiale, nous allons étudier maintenant l'extraction d'indices temporels. Plus précisément, les indices que nous cherchons à extraire sont les frontières des différents plans de la séquence (ou changements de plans). Nous rappellerons brièvement les différents types de transitions et présenterons ensuite notre approche en détaillant le prétraitement des données, la définition de la mesure de distance utilisée, et enfin la solution globale proposée.

Un plan est défini comme une suite d'images issues d'une acquisition continue d'une caméra donnée. Ainsi, toutes les images d'un plan ont été acquises avec la même caméra. Le plan est souvent l'unité temporelle la plus petite pour une séquence vidéo si l'on ne prend pas en compte l'image pour laquelle la notion de temps a disparu. Chaque plan est séparé du précédent et du suivant par une transition, qui peut être brusque ou

progressive. Lors d'une transition brusque (appelée *cut*), la dernière image du premier plan est directement suivie par la première image du second plan. Dans le cas où les deux plans sont connectés en utilisant un effet particulier, on parle de transition progressive : fondu, volet, *etc.* On distingue le fondu du noir vers un plan, d'un plan vers le noir, ou d'un plan vers un autre plan. Au cours d'un fondu, le niveau de chaque pixel des images intermédiaires (appartenant à la transition progressive) est calculé en fonction des niveaux des pixels de la dernière image du premier plan et de la première image du second plan. La proportion varie au cours de la transition de 0 à 1 pour la première image du second plan et de 1 à 0 pour la dernière image du premier plan. Lors d'un volet, chaque pixel des images intermédiaires a un niveau égal à celui du pixel de mêmes coordonnées spatiales soit dans la dernière image du premier plan soit dans la première image du second plan. Les images appartenant à un volet vont donc contenir de plus en plus de pixels extraits de la première image du second plan et de moins en moins de pixels extraits de la dernière image du premier plan. La plupart des méthodes proposées pour résoudre le problème étudié ici fonctionnent en deux étapes : le calcul d'une mesure de dissimilarité entre deux trames successives d'une séquence vidéo, puis la comparaison de la valeur obtenue avec un seuil, afin de déterminer ou non la présence d'un changement de plans. Suivant ce principe, la détection d'un changement de plans est effective si la condition $d(I_t, I_{t-1}) > S$ est respectée. Dans cette section, nous adoptons les notations I_t pour représenter l'image de la séquence vidéo obtenue à l'instant t , d une distance, et S un seuil. On trouvera dans (Lefèvre et al. 2003) un panoramas approfondi des méthodes adaptées aux données non-compressées.

Afin de garantir un temps de calcul relativement faible et d'assurer une certaine robustesse au bruit, nous proposons d'introduire un prétraitement des données. Celui-ci consistera à diminuer la résolution spatiale et sera obtenu par l'approche multirésolution décrite dans la section précédente, en considérant des blocs de 8×8 pixels (choix qui nous permet de traiter également des données compressées JPEG ou MPEG à l'aide des coefficients DC). De manière à accroître la robustesse aux changements d'illumination et à réduire les temps de calcul, nous avons choisi de limiter la représentation des pixels à un espace de dimension 2 composé de la teinte et de la saturation. Pour toutes les scènes d'extérieur qui sont fréquentes dans les séquences vidéo, on peut noter une amélioration par rapport à l'usage de l'espace RVB. Comme (Carron 1995), nous mesurons la différence entre deux images dans le sous-espace TS. Cette différence est obtenue par la distance d définie comme :

$$d_k(I_{t_1}, I_{t_2}) = \sum_{x=1}^X \sum_{y=1}^Y I_{t_1}(x, y) \ominus I_{t_2}(x, y) \quad (3)$$

avec \ominus un opérateur algébrique utilisé pour comparer deux pixels et défini par :

$$I_{t_1}(x, y) \ominus I_{t_2}(x, y) = \alpha_{T,S}(I_{t_1}(x, y, T) - I_{t_2}(x, y, T)) \pmod{2\pi} + (1 - \alpha_{T,S})|I_{t_1}(x, y, S) - I_{t_2}(x, y, S)| \quad (4)$$

où $\alpha_{T,S}$ est un coefficient permettant de donner plus ou moins d'influence aux composantes T et S. En effet, dans le cas de pixels achromatiques, il est important de ne pas donner d'importance à la composante T qui n'est alors pas fiable. La mesure de distance d , quoique relativement simple, permet d'estimer correctement la

différence entre deux images en assurant une invariance à l'illumination. L'utilisation d'une mesure de distance plus complexe pourrait apporter une quantité d'information supplémentaire mais engendrerait également un surcoût en terme de temps de calcul. Cependant, l'utilisation directe de cette mesure de dissimilarité entre deux images successives nécessiterait la comparaison à un seuil S . Le seuil utilisé doit souvent être fixé de manière empirique, et dépend du domaine vidéo étudié (sport, bulletin d'informations, *etc.*) ou du type de plans présents dans la séquence. Ainsi, un plan éloigné, où les objets en mouvement sont petits, sera caractérisé par une valeur d relativement faible tandis qu'un plan proche ou serré, où les objets en mouvement occupent une portion importante de l'image, sera caractérisé par une valeur d plus élevée. Le seuil S devra donc être ajusté en conséquence afin d'éviter les fausses détections ou l'absence de détection. Comme certaines séquences vidéo, notamment les retransmissions télévisées d'événements sportifs, contiennent des plans éloignés et des plans proches, il est nécessaire d'utiliser une méthode plus générale qui puisse s'adapter à ces différents types de plans. Nous proposons donc d'introduire un seuil adaptatif, noté S_d , qui dépend du temps. La valeur du seuil est mise à jour pour chaque nouvelle image de la séquence, soit $S_d(t) = \alpha_{S_d} S_d(t-1) + (1 - \alpha_{S_d}) d(I_t, I_{t-1})$ où $S_d(t)$ représente la valeur du seuil S_d à l'instant t . De cette façon, il s'adapte automatiquement avec une certaine inertie (représentée par le coefficient α_{S_d}) au contenu de la vidéo étudiée, sa valeur étant modifiée en fonction des valeurs de $d(I_t, I_{t-1})$. Les mesures de précision et de rappel dépendront évidemment de α_{S_d} . L'utilisation directe d'une mesure d entre deux images successives est très sensible au bruit et au mouvement présent dans la séquence étudiée. L'introduction d'un seuil adaptatif permet de limiter cette sensibilité dans une certaine mesure, mais évidemment pas dans sa totalité. Nous proposons donc de considérer une mesure relative et non une mesure absolue. Cette mesure relative, notée d' , permet d'accroître la robustesse au bruit et aux mouvements importants présents dans la séquence, et est définie par $d'(I_t) = |d(I_t, I_{t-1}) - d(I_{t-1}, I_{t-2})|$. Contrairement à la mesure d , la mesure d' est définie de manière relative et son ordre de grandeur dépend donc moins du type de plan ou de vidéo étudié. Afin de détecter un changement de plans, cette mesure peut donc être comparée à un seuil $S_{d'}$ fixé empiriquement au début de la séquence. La valeur de $S_{d'}$ pourra évoluer en fonction du type de vidéo ou de plan analysé.

Comme il a été précisé précédemment, un changement de plans peut être brusque ou progressif. En considérant *une transition progressive comme une transition brusque dont les effets sont étalés sur plusieurs images*, nous proposons ici une approche permettant de détecter les transitions brusques ou progressives de manière relativement similaire. Pour détecter un changement brusque, nous comparons directement la valeur d' avec un seuil $S_{d'}$. En effet, si la valeur de d' est élevée, c'est-à-dire si la différence absolue entre $d(I_t, I_{t-1})$ et $d(I_{t-1}, I_{t-2})$ est significative, alors l'évolution du contenu de la séquence entre les images I_{t-2} et I_{t-1} n'est pas cohérente avec celle observée entre les images I_{t-1} et I_t . Un changement brusque existe donc à l'instant $t-1$. Si aucun changement brusque n'a été détecté, il est encore possible de se trouver en présence d'une transition progressive. La valeur d' ne peut être utilisée directement dans le cas d'une transition progressive puisqu'elle ne reflète l'évolution de la mesure de distance d qu'à un instant donné. Il est donc nécessaire de cumuler les valeurs d' obtenues pour toutes

les trames composant une transition progressive afin d'obtenir une mesure qui sera du même ordre de grandeur que la valeur du seuil considéré dans le cas d'une transition brusque. La détection des transitions progressives s'effectue en deux étapes successives. La première étape consiste en la détection des trames susceptibles d'être les frontières d'une transition progressive. Pour détecter celles-ci, nous analysons l'évolution de la mesure de distance d et nous comparons à chaque instant t la valeur $d(I_t, I_{t-1})$ avec le seuil adaptatif $S_d(t)$ défini précédemment. L'utilisation de ce seuil adaptatif nous permet de gérer tout type de situation (plan proche ou éloigné, mouvement important ou pas, *etc*). Si la condition $d(I_{t_1}, I_{t_1-1}) > S_d(t_1)$ est vérifiée, alors il est possible qu'une transition soit présente dans la séquence à partir de l'instant t_1 . L'instant t_2 de fin de cette transition correspondrait à la première trame vérifiant la condition $d(I_{t_2}, I_{t_2-1}) < S_d(t_2)$ avec $t_2 > t_1$. Une fois les frontières t_1 et t_2 d'une possible transition déterminées, il est nécessaire d'analyser les trames t de cet intervalle de temps. Pour cela, nous calculons la valeur cumulée de d' sur l'ensemble des trames $[t_1, t_2]$ notée $d'_{\text{cumul}}(t_1, t_2)$, soit $d'_{\text{cumul}}(t_1, t_2) = \sum_{t=t_1}^{t_2} d'(I_t)$. La comparaison de $d'_{\text{cumul}}(t_1, t_2)$ avec le seuil $S_{d'}$ permet alors de valider ou non la présence d'un changement entre les trames I_{t_1} et I_{t_2} .

Les deux méthodes présentées ici ont été testées dans le contexte de l'analyse de séquences vidéo de football. Les résultats obtenus vont maintenant être présentés.

5 Résultats

Nous commentons ici les résultats obtenus avec les deux méthodes proposées, et validant l'intérêt de l'espace TSL et les capacités de nos méthodes à extraire des indices pertinents. Le contexte considéré est celui de l'analyse temps-réel de séquences vidéo de match de football, représentant des scènes d'extérieur d'illumination variable. Les temps de calcul ont été mesurés à l'aide d'un PC Pentium 4 1.7 GHz.

La méthode d'extraction des indices spatiaux consiste à séparer les objets et le fond. Dans le contexte proposé, l'objectif est d'identifier les joueurs par rapport au terrain de football. Une étude des résultats obtenus sur différentes images (où les tailles des objets varient considérablement) nous a permis d'observer que les objets sont correctement détectés, indépendamment de l'aire qu'ils occupent dans l'image. Les temps de calcul observés dépendent de la taille de l'image (30 millisecondes sont nécessaires pour traiter une image de taille 192×128 pixels). Le processus de segmentation est itératif, comme l'illustre la figure 1. Le choix de la résolution finale r_{final} influe directement sur la précision du résultat. Cependant, même en considérant une résolution finale similaire à la résolution originale (*i.e.* $r_{\text{final}} = 0$), les contours des objets détectés resteront grossiers et parallèles aux côtés de l'image. Cependant, ce résultat peut être suffisant dans de nombreux cas. Nous avons également observé que l'utilisation de la teinte fournit un résultat de meilleure qualité que l'espace RVB. Aucune des composantes R, V ou B ne contient l'information suffisante pour atteindre des résultats aussi précis qu'avec la teinte (l'information pertinente étant dispersée dans les trois canaux de base). De plus, nous avons évalué la robustesse de la composante couleur utilisée au réglage des paramètres. Là encore, la teinte fournit les résultats les plus intéressants et les plus robustes. Les principales limites de la méthode proposée ont été identifiées *a priori* :

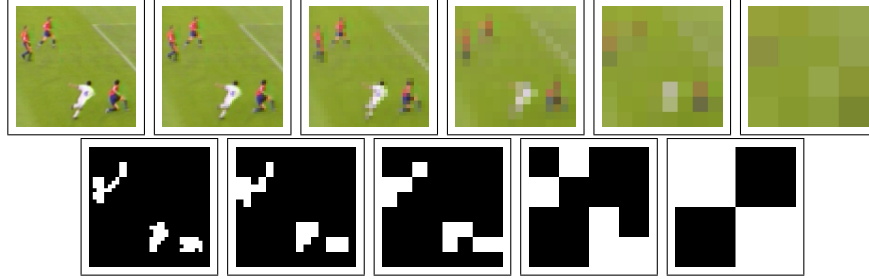


FIG. 1 – Image et résultat obtenu à différentes résolutions.

les pixels ne doivent pas être achromatiques (puisque seule la teinte est utilisée dans le processus de segmentation), et l'arrière-plan doit être relativement uniforme (puisque'il est modélisé ensuite par une moyenne).

La méthode d'extraction des indices temporels débute par une étape de réduction qui fournit des images de 20×15 pixels. Le temps de calcul alors nécessaire est égal à 4 millisecondes par image. Nous avons étudié l'évolution des mesures d , d' , et d'_{cumul} pour des séquences vidéo contenant différents types de transition et avons constaté que le contraste des valeurs au niveau des extremums locaux est beaucoup plus marqué dans l'espace TSL que dans l'espace RVB. Ce choix d'espace assure donc à la méthode proposée une plus grande robustesse et confirme notre hypothèse théorique de départ. La qualité d'une méthode de détection de changements de plans peut être évaluée grâce aux mesures de rappel et de précision, tenant compte respectivement des détections manquées et des fausses détections. Des tests effectués sur une vingtaine de séquences, chacune composée de 500 images et contenant entre un et trois changements de plans (brusques ou progressifs), ont permis d'obtenir des mesures de rappel et de précision respectivement égales à 80 % (100 % dans le cas de cuts) et 100 %. Un ensemble de tests nous a permis de vérifier expérimentalement que les résultats obtenus avec cette méthode étaient meilleurs que ceux obtenus avec d'autres méthodes de la littérature. La principale limite de l'approche proposée ici est sa sensibilité au mouvement présent dans la séquence. Ce mouvement apparent peut être provoqué par une accélération brusque de la caméra ou par le mouvement d'un objet occupant quasiment toute l'image dans un plan rapproché.

6 Conclusion

De nos jours, les données complexes telles que les séquences vidéo couleur sont la plupart du temps représentées dans l'espace Rouge Vert Bleu prévu pour l'affichage des images à l'écran. Nous avons montré ici comment un autre espace de représentation de la couleur, l'espace Teinte Saturation Luminance, pouvait apporter une amélioration en analyse d'images dans l'optique d'un processus de fouille. Pour cela nous avons cherché à extraire deux types d'indices, spatiaux et temporels, à l'aide de l'espace TSL. Tandis que l'extraction des indices spatiaux ne nécessite qu'une seule image grâce à un cadre d'analyse multirésolution, l'obtention des indices temporels utilise une méthode

s'adaptant au contenu, notamment par l'usage d'une mesure de distance intertrames différentielle. Notre contribution a donc porté sur ces trois aspects : caractéristiques et calculs dans l'espace TSL, extraction d'indices spatiaux par séparation des objets et du fond, extraction d'indices temporels par détection des changements de plans.

Outre la validation de nos approches par leur intégration à un processus de fouille, les perspectives des travaux présentés ici sont d'une part l'approfondissement des résultats présentés, en continuant d'identifier les caractéristiques de l'espace TSL et de proposer des modes de calcul appropriés à cet espace, et d'autre part l'atténuation des limites des méthodes proposées (meilleure prise en compte des données achromatiques, considération de scènes au contenu plus complexe).

Références

- Carron T. (1995), Segmentations d'images couleur dans la base Teinte-Luminance-Saturation : approche numérique et symbolique, Thèse de Doctorat, Université de Savoie.
- Chang H., Jiang X., Sun Y., et Wang J. (2001), Color image segmentation : Advances and prospects, *Pattern Recognition*, Vol. 34, pp. 2259-2281.
- Haskell B. et al. (1998), Image and video coding : emerging standards and beyond, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 8, pp. 814-837.
- Lefèvre S., Holler. J, et Vincent N. (2003), A study of real-time segmentation of uncompressed video sequences for content-based search and retrieval, *Real-Time Imaging*, Vol. 9, pp. 73-98.
- Mardia K. et Jupp P. (2000), *Directional Statistics*, Wiley & Sons Ltd.
- Trémeau A., Fernandez-Maloigne C., et Bonton P. (2004), *Imagerie Numérique Couleur*, Dunod.

Summary

In this paper, we consider colour video sequences as complex data. Our contribution consists in two methods which are adapted to this kind of data and able to extract some spatial and temporal descriptors. We think these methods can be successfully involved in a more complex process of multimedia data mining, the description of which is out of focus of this paper. The methods we are presented here are based on the Hue Saturation Luminance colour space. The extraction of spatial features is related to the problem of separation between background and foreground parts, solved using a multiresolution approach based on a single frame. The extraction of temporal features is related to the detection of shot changes in video sequences, obtained with the use of context-independent distances measures. Both methods share some common features : use of the HSL colour space, computational efficiency, and robustness to artefacts. We illustrate our two methods with some results obtained on sport video sequences.

Fusion de classifieurs pour la classification d'images sonar

Arnaud Martin

ENSIETA / E³I², EA3876
2, rue François Verny, 29806 Brest cedex 9
Arnaud.Martin@ensieta.fr
<http://www.ensieta.fr/e3i2>

Résumé. Nous présentons dans ce papier des approches de fusion d'informations haut niveau applicables pour des données numériques ou des données symboliques. Nous étudions l'intérêt des telles approches particulièrement pour la fusion de classifieurs. Une étude comparative est présentée dans le cadre de la caractérisation des fonds marins à partir d'images sonar. Reconnaître le type de sédiments sur des images sonar est un problème difficile en soi en partie à cause de la complexité des données. Nous comparons les approches de fusion d'informations haut niveau et montrons le gain obtenu.

1 Introduction

La fusion d'informations est apparue afin de gérer des quantités très importantes de données multisources dans le domaine militaire. Depuis quelques années des méthodes de fusion ont été adaptées et développées pour des applications en traitement du signal. Plusieurs sens sont donnés à la fusion d'informations, nous reprenons ici la définition proposée par (Bloch 2003) : La fusion d'informations consiste à combiner des informations issues de plusieurs sources afin d'aider à la prise de décision.

Nous ne cherchons pas ici à réduire les redondances contenues dans les informations issues de plusieurs sources, mais au contraire à en tenir compte afin d'améliorer la prise de décision. De même nous cherchons à modéliser au mieux les différentes imperfections des données (imprécisions, incertitudes, conflit, ambiguïté, incomplétude, fiabilité des sources, ...) non pas pour les supprimer, mais encore pour l'aide à la décision.

Différents niveaux de fusion ont été proposés dans la littérature. Ce qui est communément retenu, est une division en trois niveaux (Dasarathy 1997), celui des données (ou bas niveau), celui des caractéristiques (*i.e.* des paramètres extraits) (ou fusion de niveau intermédiaire) et celui des décisions (ou fusion de haut niveau).

Le choix du niveau de fusion doit se faire en fonction des données disponibles et de l'architecture de la fusion retenue (centralisée, distribuée, ...) qui sont liées à l'application recherchée. Ainsi, nous pouvons chercher à fusionner des informations issues de différents capteurs tels que des radars de fréquences différentes afin d'estimer au mieux la réflexion d'une cible. Dans ce cas une approche de fusion bas niveau sera préférable.

Dans ce papier, nous considérons une application dans le cadre de la classification. Plusieurs classifieurs peuvent fournir une information sur la classe de l'objet observé. Ainsi, nous retenons des approches de fusion haut niveau pour résoudre un tel problème. Les données exprimant une décision peuvent être de type numérique (tel que les sorties des classifieurs) ou symbolique (tel que les classes décidées par les classifieurs exprimées sous forme de sym-

boles). Nous présentons ici une étude comparative des méthodes de fusion haut niveau dans le cadre de la classification d'images sonar.

Les images sonar sont caractérisées par un grand nombre d'imperfections telles que l'incertitude du milieu, les imprécisions des mesures et de reconstruction de l'image. C'est pourquoi, nous cherchons ici à extraire des paramètres de textures sur ces images, en considérant que la physique du problème a été au mieux prise en compte lors de l'étape de reconstruction. Nous retenons quatre jeux de paramètres extraits selon quatre méthodes différentes de traitement d'images. Les images sonar sont ensuite classées par quatre perceptrons multicouche, chacun considérant un des quatre jeux de paramètres.

Nous présentons tout d'abord trois grandes classes de méthodes de fusion haut niveau, les approches par vote, celles issues de la théorie des possibilités et celles issues de la théorie des croyances. Nous exposons ensuite la complexité des images sonar pour leur classification automatique, ainsi que les quatre méthodes retenues pour l'extraction de paramètres de texture. Enfin nous présentons une évaluation comparative des méthodes de fusion d'informations haut niveau selon la configuration retenue pour la classification d'images sonar.

2 Méthodes de fusion d'informations

Nous présentons ici trois cadres théoriques de fusion d'informations haut niveau, le principe du vote, la théorie des possibilités et la théorie des croyances. Nous considérons le problème de la fusion de m sources S_j afin de déterminer une des n classes C_i possibles.

2.1 Principe du vote

Le principe du vote est la méthode de fusion d'informations la plus simple à mettre en œuvre. Plus qu'une approche de fusion, le principe du vote est une méthode de combinaison particulièrement adaptée aux décisions de type symbolique. Notons $S_j(x) = i$ le fait que la source S_j attribue la classe C_i à l'observation x . Nous supposons ici que les classes C_i sont exclusives. A chaque source nous associons la fonction indicatrice :

$$M_i^j(x) = \begin{cases} 1 & \text{si } S_j(x) = i, \\ 0 & \text{sinon.} \end{cases} \quad (1)$$

La combinaison des sources s'écrit par :

$$M_k^E(x) = \sum_{j=1}^m M_k^j(x), \quad (2)$$

pour tout k . L'opérateur de combinaison est donc associatif et commutatif. La règle du vote majoritaire consiste à choisir la décision prise par le maximum de sources, c'est-à-dire le maximum de M_k^E . Cependant cette règle simple n'admet pas toujours de solutions dans l'ensemble des classes $D = \{C_1, \dots, C_n\}$. En effet, par exemple si le nombre de sources m est paire et que $m/2$ sources décident C_{i_1} et $m/2$ autres sources disent C_{i_2} , ou encore dans le cas où chaque source affecte à x une classe différente. Nous sommes donc obligé d'ajouter une classe C_{n+1} qui représente l'incertitude totale liée au conflit des sources sous l'hypothèse de l'exhaustivité

des classes $C_{n+1} = \{C_1, \dots, C_n\}$. La décision finale de l'expert prise par cette règle s'écrit donc par :

$$E(x) = \begin{cases} k & \text{si } \max_k M_k^E(x), \\ n+1 & \text{sinon.} \end{cases} \quad (3)$$

Cette règle est cependant peu satisfaisante dans les cas où deux sources donnent le maximum pour des classes différentes. La règle la plus employée est la règle du vote majoritaire absolu qui s'écrit :

$$E(x) = \begin{cases} k & \text{si } \max_k M_k^E(x) > \frac{m}{2}, \\ n+1 & \text{sinon.} \end{cases} \quad (4)$$

A partir de cette règle il a été démontré (Lam et Suen 1997) plusieurs résultats prouvant que la méthode du vote permet d'obtenir de meilleures performances que toutes les sources prises séparément, sous des hypothèses d'indépendance statistique des sources et de même probabilité, et ceci est d'autant plus vrai que m est impaire.

Il est possible de généraliser le principe du vote majoritaire afin de supprimer le conflit. Au lieu de combiner les réponses des sources par une somme simple comme dans l'équation (2), l'idée est d'employer une somme pondérée (Xu et al. 1992) :

$$M_k^E(x) = \sum_{j=1}^m \alpha_{jk} M_k^j(x), \quad (5)$$

où $\sum_{j=1}^m \sum_{k=1}^n \alpha_{jk} = 1$. Les poids α_{jk} représentent la fiabilité d'une source pour une décision

donnée, et l'estimation de ces poids peut se faire à partir des taux normalisés de réussite pour chaque classe et chaque classifieur. Notons qu'alors nous introduisons une connaissance *a priori* non nécessaire précédemment. Les différentes règles de décision possibles peuvent se résumer par la formule suivante :

$$E(x) = \begin{cases} k & \text{si } M_k^E(x) = \max_i M_i^E(x) \geq c m + b(x), \\ n+1 & \text{sinon,} \end{cases} \quad (6)$$

où c est une constante de $[0,1]$ et $b(x)$ est une fonction de $M_k^E(x)$.

2.2 Théorie des possibilités

La théorie des possibilités proposée par (Zadeh 1978, Dubois et Prade 1988) permet de tenir compte de l'imprécision des données ainsi que de l'incertitude à partir de deux fonctions de possibilité et de nécessité. Ces deux fonctions sont obtenues à partir des distributions de possibilités définies sur $D = \{C_1, \dots, C_n\}$ par :

$$\pi : D \rightarrow [0,1], \sup_{x \in D} \pi(x) = 1. \quad (7)$$

Ces distributions donnent le degré d'appartenance au domaine D , qui n'est autre qu'un opérateur flou. Afin d'extraire l'imprécision et l'incertitude des données, deux fonctions spécifiques sont

définies à partir de ces distributions. La fonction de possibilité est définie pour tout $A \in 2^D$ par :

$$\Pi(A) = \sup_{x \in A} \pi(x). \quad (8)$$

La fonction de nécessité est donnée pour tout $A \in 2^D$ par :

$$N(A) = 1 - \Pi(A^c), \quad (9)$$

où A^c représente l'évènement contraire de A .

Un des avantages de la théorie des possibilités est le nombre d'opérateurs de combinaison disponibles. Il est ainsi possible de combiner l'information issue des distributions de possibilité, à partir d'opérateurs de type t -norme, t -conorme, moyenne, sommes symétriques, etc... Le choix du type de combinaison est un problème délicat *a priori* dans la théorie des possibilités, et doit être fait selon l'application et l'objectif recherché. Ce choix peut se faire selon le comportement général de l'opérateur (conjonctif, disjonctif, ou des compromis), selon les propriétés désirées, selon sa capacité à discriminer les classes, ou encore selon son comportement dans des situations de conflit. En pratique de nombreux opérateurs sont employés et testés dans les applications, tels que max (opérateur de type t -norme), min (opérateur de type t -conorme), ou la moyenne, la médiane et les intégrales floues (opérateurs de type moyenne).

La dernière étape de la fusion d'informations est l'étape de décision. Dans le cadre de la théorie des possibilités, elle est généralement faite selon la règle suivante : la classe C_k est décidée pour l'observation x si :

$$C_k = \operatorname{argmax}_{1 \leq i \leq n} \mu_i(x), \quad (10)$$

où $\mu_i(x)$ représente le coefficient d'appartenance de x à la classe C_i , qui sera ici donné par les sorties du classifieur.

Par construction des opérateurs de combinaison et de la règle de décision, la théorie des possibilités est davantage adaptée à la fusion d'informations de type numérique. Ainsi les coefficients d'appartenance peuvent être facilement obtenus dans le cadre de la classification par les sorties numériques des classifieurs. Nous emploierons donc ici cette théorie pour la fusion d'informations haut niveau sur des données de type numérique.

2.3 Théorie des croyances

La théorie des croyances (ou théorie de Dempster-Shafer) permet également de représenter à la fois l'imprécision et l'incertitude au travers deux fonctions : la fonction de croyance et la fonction de plausibilité (Bloch 2003, Appriou 2002). Ces deux fonctions sont dérivées des fonctions de masses. Le principe de la théorie des croyances repose sur la manipulation de ces fonctions de masse définies sur des sous-ensembles et non sur des singletons comme dans la théorie des probabilités. En effet, elles sont définies sur chaque sous-espace de l'ensemble des disjonctions du cadre de discernement $D = \{C_1, \dots, C_n\}$ et à valeurs dans $[0,1]$. Généralement, il est ajouté une condition donnée par :

$$\sum_{A \in 2^D} m_j(A) = 1, \quad (11)$$

où $m(\cdot)$ représente la fonction de masse. Dans cette théorie la première difficulté est le choix de la fonction de masse. Plusieurs approches ont été proposées, nous en retenons ici deux : l'une fondée sur un modèle probabiliste (Appriou 2002) et l'autre sur une transformation en distance (Denœux 1995).

(Appriou 2002) propose deux modèles répondant à trois axiomes qui impliquent la considération de $n * m$ fonctions de masse aux seuls éléments focaux possibles $\{C_i\}$, $\{C_i^c\}$ et D . Un axiome garantit de plus l'équivalence avec l'approche bayésienne dans le cas où la réalité est parfaitement connue (méthode optimale dans ce cas). Ces deux modèles sont sensiblement équivalents sur nos données, nous utilisons dans cet article le modèle donné par :

$$\begin{cases} m_{ij}(C_i)(x) = \frac{\alpha_{ij} R_j p(S_j/C_i)}{1 + R_j p(S_j/C_i)} \\ m_{ij}(C_i^c)(x) = \frac{\alpha_{ij} R_j}{1 + R_j p(S_j/C_i)} \\ m_{ij}(D)(x) = 1 - \alpha_{ij} \end{cases} \quad (12)$$

où p est une probabilité, $R_j = (\max_{i,j} p(S_j/C_i))^{-1}$ est un facteur de normalisation, et $\alpha_{ij} \in [0,1]$ est un coefficient d'affaiblissement permettant de tenir compte de la fiabilité d'une source S_j pour une classe C_i , que nous choisissons ici égale à 1.

La difficulté de ce modèle est alors l'estimation des probabilités $p(S_j/C_i)$. Dans le cas où la donnée de la source S_j est la réponse d'un classifieur exprimée sous la forme de la classe (donnée symbolique), l'estimation de ces probabilités peut être faite par les matrices de confusion sur une base d'apprentissage. Si la réponse du classifieur est une donnée numérique, l'estimation de telles probabilités peut se faire soit par une approche fondée sur les fréquences, soit sous l'hypothèse de la distribution suivie par ces probabilités. Dans ce dernier cas l'estimation est généralement plus délicate, nous retiendrons donc ce modèle pour la fusion d'informations haut niveau des données symboliques.

En revanche l'approche fondée sur une transformation en distance proposée par (Denœux 1995) est plus adaptée à la fusion d'informations haut niveau des données numériques. En effet, les fonctions de masse sont définies par :

$$\begin{cases} m_{ij}(C_i/x^{(t)})(x) = \alpha_{ij} \varphi_i(d(x, x^{(t)})) \\ m_{ij}(D/x^{(t)})(x) = 1 - \alpha_{ij} \varphi_i(d(x, x^{(t)})) \end{cases} \quad (13)$$

où $(x^{(t)})$ est un vecteur d'apprentissage des réponses des sources, $\alpha_{ij} \in [0,1]$ est un coefficient d'affaiblissement, d est une distance à déterminer entre x et $x^{(t)}$, C_i est la classe associée à $x^{(t)}$, et φ_i est une fonction vérifiant :

$$\begin{cases} \varphi_i(0) = 1, \\ \lim_{d \rightarrow +\infty} \varphi_i(d) = 0. \end{cases} \quad (14)$$

Il existe un grand nombre de fonctions φ_i vérifiant ces égalités, sans qu'il y ait une méthode pour le choix de ces fonctions. Dans le cas d'une distance euclidienne, Denœux propose la fonction :

$$\varphi_i(d) = \exp(-\gamma_i d^2), \quad (15)$$

où $\gamma_i > 0$ est un paramètre associé à la classe C_i . γ_i peut être initialisé comme l'inverse de la distance moyenne entre les vecteurs d'apprentissage vérifiant C_i . La distance $d(x, x^{(t)})$ peut

être considérée uniquement pour les k -plus proches voisins de x afin de réduire le temps de calcul.

La différence de fond avec les modèles d'Appriou est qu'ici il faut estimer une distance au lieu d'une probabilité. La distance, généralement euclidienne est plus adaptée aux données numériques, tandis que l'estimation des probabilités est ici plus aisée pour des données symboliques. Dans le cas d'une distance euclidienne, notons que nous obtenons une fonction de masse proche de celle obtenue par un modèle d'Appriou sous l'hypothèse d'une distribution gaussienne. Nous comparons donc ces deux approches pour la fusion d'informations avec des données de type symétrique pour le modèle probabiliste et numérique pour le modèle des distances.

La combinaison des fonctions de masse est fondée sur la règle orthogonale de Dempster-Shafer non normalisée proposée par (Smets 1990) définie pour deux fonctions de masse m_1 et m_2 et pour tout $A \in 2^D$ par :

$$m(A) = (m_1 \oplus m_2)(A) = \sum_{B \cap C = A} m_1(B)m_2(C). \quad (16)$$

Cette opérateur est associatif et commutatif. La masse affectée sur l'ensemble vide s'interprète comme une mesure de conflit. Pour les modèles d'Appriou et de Denœux, nous obtenons donc une unique fonction de masse en combinant les fonctions m_{ij} .

La dernière étape de la fusion est la décision sur la classe la plus vraisemblable. La théorie des croyances offrent plusieurs règles de décision fondées sur la maximisation d'un critère. En particulier, nous pouvons employer le maximum des fonctions de croyance ou des fonctions de plausibilité. Si les premières peuvent être trop pessimistes, les secondes peuvent être trop optimistes. Un compromis est le maximum des probabilités pignistiques proposées par (Smets 1990) qui est le critère retenu dans cet article.

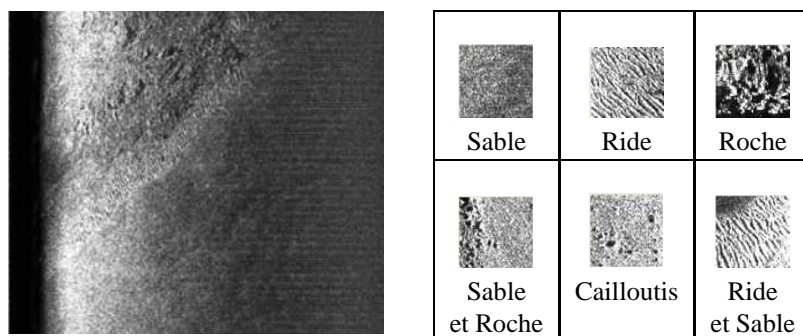
3 Classification d'images sonar

La classification des images sonar est un problème difficile en soi. Les méthodes de caractérisation automatique consistent en des méthodes d'analyse de texture, les images de fonds marins présentant en effet des zones de sédiments homogènes ou non qui peuvent s'apparenter à des textures. La littérature concernant les méthodes d'analyse de la texture est abondante et le choix de l'une ou de plusieurs d'entre elles dépend très souvent de l'image et de l'application. Ces méthodes fournissent généralement un ensemble assez restreint de paramètres pertinents qui permettent de classer l'image en un ou plusieurs type de sédiments. Nous exposons ici toute la complexité des images sonar due aux nombreuses imperfections, puis nous présentons les classifieurs à base de méthodes d'analyse de texture.

Les images sonar sont obtenues à partir des mesures temporelles faites en traînant à l'arrière d'un bateau un sonar qui peut être latéral, frontal, ou multifaisceaux. Chaque signal émis est réfléchi sur le fond puis reçu sur l'antenne du sonar avec un décalage et une intensité variable. Pour la reconstruction sous forme d'images un grand nombre de données physiques (géométrie du dispositif, coordonnées du bateau, mouvements du sonar, ...) est pris en compte, mais elles sont entachées des bruits de mesures dues à l'instrumentation. A ceci viennent s'ajouter des interférences dues à des trajets multiples du signal (sur le fond ou la surface), à des bruits de

chatoiement, ou encore à la faune et à la flore. Les images sont donc entachées d'un grand nombre d'imperfections telles que l'imprécision et l'incertitude.

26 images fournies par le GESMA (Groupe d'Etudes Sous-Marine de l'Atlantique) ont été obtenues à partir d'un sonar Klein 5400 permettant une bonne résolution. Ces images ont été segmentées en imagerie de taille 64×64 pixels (voir Tab. 1) étiquetées selon le type de sédiment. Nous avons ainsi distingué le sable (54.52%), la roche (21.35%), les rides de sable



TAB. 1 – Exemple d'image sonar (fournit par le GESMA) et d'images extraites et étiquetées.

(8.80%), la vase (5.50%), les cailloutis (0.77%) et l'ombre (2.40%) qui représente l'absence d'information sur le type de sédiment. De plus, nous avons indiqué lorsque ces images comprennent plus d'un sédiment (homogènes ou non), ce qui représente 39.70% des images. Le type de sédiment de ces images est le plus présent. Notons également que ces bases de données sont très délicates à réaliser, car elles sont entachées des erreurs éventuelles de l'expert.

Les classifieurs sont composés chacun d'une méthode d'extraction de paramètres de texture et d'un perceptron multicouche. L'approche retenue pour l'architecture de fusion est celle présentée sur la Fig. 1. La fusion d'informations haut niveau se fait donc soit au niveau des sorties numériques des perceptrons soit au niveau des sorties symboliques représentant les classes affectées.

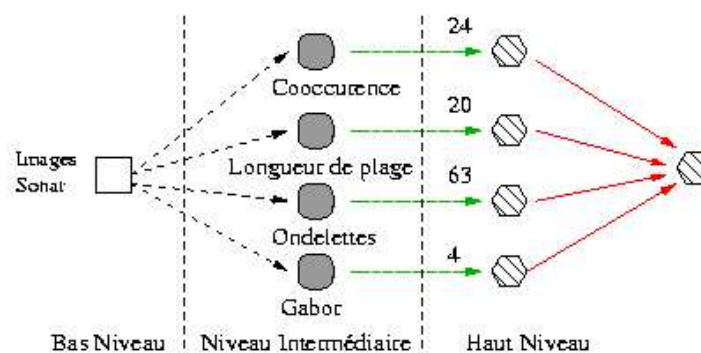


FIG. 1 – Architecture de fusion de classifieurs retenue.

Les méthodes d'extraction de paramètres de texture sont celles déjà présentées dans (Martin et al. 2004). Chaque méthode permet de calculer des paramètres différents, parfois redondants entre eux, mais avec des caractéristiques propres à la méthode.

Les matrices de co-occurrence sont calculées en comptant les occurrences identiques de niveaux de gris entre deux pixels contigus. Quatre directions sont considérées : 0, 45, 90 et 135 degrés. Dans ces quatre directions six paramètres d'Haralick sont calculés : l'homogénéité, le contraste, l'entropie, la corrélation et l'uniformité. Un des problèmes principaux de cette approche est la non invariance en translation. Ainsi les imagerie de rides auront des paramètres différents selon la direction de celles-ci.

Les matrices de longueurs de plages sont obtenues en comptabilisant les pixels consécutifs possédant le même niveau de gris dans les quatre directions précédemment considérées. Dans chacune des directions cinq paramètres sont calculés : la proportion de petite longueur de plage, la dispersion des plages entre les niveaux de gris et entre les longueurs et le pourcentage des longueurs de plage. Cette approche est bien adaptée aux images optiques faiblement bruitées. Dans le cas des images sonar, où un bruit de chatoiement est fortement présent, il faudrait soit supprimer ce bruit soit adapter le calcul des paramètres. Nous conservons cependant cette approche afin d'évaluer la robustesse de la fusion aux mauvais paramètres de texture.

La troisième approche retenue est une transformée en ondelettes. Les deux approches précédentes ne permettent pas de tenir compte de l'invariance dans les directions. La transformée en ondelettes discrète invariante en translation est fondée sur le choix de la transformation optimale pour chaque niveau de décomposition. Chaque niveau de décomposition fournit quatre images, sur lesquelles nous calculons trois paramètres : l'énergie, l'entropie, et une moyenne. Nous retenons un niveau de décomposition de trois ce qui fournit 63 paramètres au classifieur.

Enfin, une approche fondée sur les filtres de Gabor permet de résoudre le problème des rides. En effet, nous considérons cinq fréquences différentes pour six directions ce qui donne trente filtres. Sur ces filtres, nous calculons quatre paramètres statistiques : le maximum de l'écart-type d'un sédiment considéré, la moyenne de tous les filtres, la moyenne dans la direction horizontale (celle des pings du sonar) et l'écart-type avant filtrage.

Ces quatre jeux de paramètres sont ensuite considérés indépendamment à l'entrée de quatre perceptrons multicouche ayant ainsi des couches d'entrée de 24, 63, 20 et 4 neurones et une couche de sortie de 6 neurones correspondant aux six classes de sédiments considérés. L'apprentissage est réalisé pour une fonction sigmoïde de sortie donnant ainsi pour chacun des neurones k de la couche de sortie une valeur réelle $o_k \in [0,1]$. Ces valeurs o_k constituent les données numériques sur la décision des classifieurs. Les décisions symboliques sont obtenues en considérant pour chaque classifieur le maximum des o_k , indiquant ainsi la classe C_k préférée par chaque perceptron.

4 Résultats

La base de données a été divisée aléatoirement en trois parties égales. La première sert à l'apprentissage des perceptrons multicouche, la deuxième à l'apprentissage de la fusion et la troisième pour les tests. Afin d'accroître la qualité de l'estimation des taux de classification, nous avons répété le tirage aléatoire 10 fois et moyenné les résultats. Dans l'approche par vote majoritaire nous avons obtenu un conflit de 18.59%, afin de supprimer ce conflit nous avons

Coocc.	longueur de plages	ondel.	Gabor	PMC	vote	poss.	croyances	
							proba.	distance
70.0	50.3	68.9	66.4	50.0	62.0	69.9	68.8	79.5

TAB. 2 – Taux de classification avant et après fusion d'informations (%).

roche	87.3	sable	84.9
ride	61.3	vase	4.9
cailloutis	0.9	ombre	71.5
homogènes	91.3	non homogènes	63.1

TAB. 3 – taux de classification par type de sédiment pour le modèle de distance (%).

considéré l'approche avec les pondérations α_{jk} estimées par les matrices de confusion. Dans le cadre de la théorie des possibilités de nombreux opérateurs de combinaison ont été testés ; nous présentons ici celui donnant les résultats les plus probants donnés par l'opérateur max (t -conorme).

Le Tab. 2 présente les taux de bonne classification définis par le rapport du nombre d'images bien classées sur le nombre total d'images de la base de test. Nous constatons que les quatre méthodes de fusion présentées sont plus robustes aux données erronées fournies par les longueurs de plages que le PMC (perceptron global prenant en entrée l'ensemble des paramètres extraits). Cependant la fusion par vote reste moins bonne que les trois classifieurs issus des matrices de co-occurrence, ondelettes et Gabor, les hypothèses de (Lam et Suen 1997) ne sont pas vérifiées. Les deux approches de fusion d'informations haut niveau à partir des données numériques sont plus performantes que les approches à partir des données symboliques. Notons de plus que la théorie des croyances avec le modèle de distance donne significativement les meilleurs résultats que nous détaillons dans le Tab. 3. Les meilleurs taux sont atteints pour les sédiments sable et roche, ceci est dû à l'apprentissage du perceptron qui est meilleur pour les sédiments les plus représentés numériquement. Les cailloutis et la vase offrent de mauvais résultats car leur effectif est faible dans la base. Notons encore que les taux pour les images homogènes sont bien meilleurs, mais est-il raisonnable de chercher à affecter un type de sédiment à une image qui en contient plusieurs. Ceci doit entraîner une remise en cause de la constitution même de la base de données.

5 Conclusion

Nous avons étudié les différentes approches de fusion d'informations haut niveau, en faisant ressortir leurs avantages et inconvénients, et notamment la facilité pour chacune d'entre elles à être employées pour des données numériques et symboliques. Ces approches ont été comparées dans le cadre d'une application particulièrement délicate : la classification d'images sonar. En effet, nous avons vu la complexité pour l'expert à interpréter ces images, et la difficulté de les classer automatiquement. La fusion d'informations apporte une solution intéressante pour la résolution de tels problèmes particulièrement grâce à sa facilité de mise en œuvre

pour des applications de classification. Nous avons ici fait ressortir de meilleures performances pour la fusion d'informations haut niveau à partir de données numériques et plus particulièrement dans le cadre de la théorie des croyances. Cependant, nous devons bien nous garder de généraliser de tels résultats à tout type de données.

Pour cette application nous avons fait ressortir l'influence du sur-apprentissage du perceptron employé qui provient de la différence d'effectifs des sédiments dans la base de données. La gestion des événements rares peut être réalisée par la fusion, mais dans ce cas avant le classifieur. Une fusion d'informations bas niveau doit alors être envisagée. Une autre difficulté est issue de la constitution même de la base. Le fait d'avoir des imagerie possédant plusieurs sédiments augmente l'incertitude, qui est dans ce cas dure à mesurer. Nous travaillons sur la réalisation d'une base de zones homogènes où l'incertitude sera mesurable plus finement.

Références

- Appriou, A. (2002), Discrimination multisignal par la théorie de l'évidence, In *Décision et Reconnaissance des formes en signal*, Hermes Science Publication, pp 219-258, 2002.
- Bloch, I. (2003), *Fusion d'informations en traitement du signal et des images*, Lavoisier (eds), Hermes Science Publication, 2003.
- Dasarathy, B.V. (1997), Sensor Fusion Potential Exploitation - Innovative Architectures and Illustrative Applications, *Proceeding of the IEEE* 1997, 85(1), pp 24-38.
- Dencœux, T. (1995), A k -Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory, *IEEE Transactions on Systems, Man Cybernetics* 1995, 25(5), pp 804-813.
- Dubois, D. et Prade, H. (1988), *Possibility Theory*, Plenum Press, New York, 1988.
- Lam, L. et Suen, C.Y. (1997), Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance, *IEEE Transactions on Systems, Man Cybernetics* 1997, 27(5), pp 553-568.
- Martin, A., Sévellec, G. et Leblond, I. (2004), Characteristics vs decision fusion for sea-bottom characterization, *Caractérisation in-situ des fonds marins*, Brest, France, 2004.
- Smets, Ph. (1990), The Combination of Evidence in the Transferable Belief Model, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1990, 12(5), pp 447-458.
- Xu, B.V., Krzyzak, A. et Suen, C.Y. (1992), Methods of Combining Multiple Classifiers and Their Application to Handwriting Recognition, *IEEE Transactions on Systems, Man Cybernetics* 1992, 22(3), pp 418-435.
- Zadeh, L.A. (1978), Fuzzy Sets as a Basis For a Theory of Possibility, *Fuzzy Sets and Systems* 1978, 1, pp 3-28.

Summary

In this paper, we present some high level information fusion approaches for numeric and symbolic data. We study the interest of such method particularly for classifier fusion. A comparative study is made in a context of sea bed characterization from sonar images. The classification of kind of sediment is a difficult problem because of the data complexity. We compare high level information fusion and give the obtained performance.

Fouille de collections de documents en vue d'une caractérisation thématique de connaissances textuelles

Abdenour Mokrane, Gérard Dray, Pascal Poncelet

Groupe Connaissance et Systèmes Complexes
LGI2P – Site EERIE – EMA
Parc scientifique Georges Besse, 30035 Nîmes cedex 1 - France
Tél : +33 (0)4 66 38 70 94 Fax : +33 (0)4 66 38 70 74
{abdenour.mokrane, gerard.dray, pascal.poncelet}@ema.fr

Résumé. De nos jours, les entreprises, organismes ou individus se trouvent submergés par la quantité d'information et de documents disponibles. Les utilisateurs ne sont plus capables d'analyser ou d'appréhender ces informations dans leur globalité. Dans ce contexte, il devient indispensable de proposer de nouvelles méthodes pour extraire et caractériser de manière automatique les informations contenues dans les bases documentaires. Nous proposons dans cet article l'approche *IC-Doc* de caractérisation automatique et thématique du contenu de collections de documents textuels. *IC-Doc* est basée sur une méthode originale d'extraction et de classification de connaissances textuelles prenant en considération les co-occurrences contextuelles et le partage de contextes entre les différents termes représentatifs du contenu. *IC-Doc* permet ainsi une extraction automatique de *KDMs* (*Knowledge Dynamic Maps*) sur les contenus des bases documentaires. Ces *KDMs* permettent de guider et d'aider les utilisateurs dans leurs tâches de consultations documentaires. Ce papier présente également une expérimentation de notre approche sur des collections de documents textuels.

Mots-Clefs. Caractérisation thématique, Similarité textuelle, Partage de contextes, Knowledge Dynamic Map.

1 Introduction

La fouille de données textuelles vise essentiellement à résoudre les problèmes de surabondance d'informations et faciliter l'extraction des connaissances enfouies dans les documents disponibles sur les bases de données ou sur le Web. Chaque jour, en particulier en raison de l'essor des communications électroniques, le nombre de documents disponibles croît de manière exponentielle et l'utilisateur (entreprise, organisme ou individu) se trouve submergé par la quantité d'informations disponibles. Ces utilisateurs ne sont donc plus capables d'analyser ou d'appréhender ces informations dans leur globalité.

De nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent aux traitements de bases de documents textuels (Baldi et Di meglio 2004,

Chung et al. 2003, Hongyuan et al. 2001, Mokrane et al. 2004b, Poibeu 2003, Ihadjadene 2004). Ces travaux ont donné naissance à des systèmes de catégorisation et de cartographie de documents comme *Kartoo* (Chung et al. 2003) ou *Mapstan* (Spinat 2002). Ces outils retrouvent des liens entre les différents documents ou sites Web et représentent ces liens sous forme de cartes de navigation. Cependant les modèles d'informations proposés sont peu représentatifs du contenu global ou de chacun des documents par rapport aux différentes thématiques des bases documentaires. Ces modèles s'inspirent des outils de recherches documentaires qui demandent à l'utilisateur de décrire l'information qu'il n'a pas (Ihadjadene 2004). En outre, ces systèmes sont peu adaptés à une caractérisation thématique en vue d'une navigation par le contenu dans une collection de documents. Il devient donc indispensable de proposer de nouvelles méthodes et systèmes pour extraire et caractériser de manière automatique les informations contenues dans les bases de documents textuels.

Dans cet article, nous proposons l'approche *IC-Doc* de caractérisation automatique et thématique du contenu de collections de documents textuels. *IC-Doc* est basée sur une méthode originale d'extraction et de classification de connaissances textuelles prenant en considération les co-occurrences contextuelles et le partage de contextes entre les différents termes représentatifs du contenu. *IC-Doc* permet ainsi une extraction automatique de *KDMs* (*Knowledge Dynamic Maps*) à partir du contenu d'une base documentaire. Ces *KDMs* permettent de guider les utilisateurs dans leurs tâches de consultations documentaires.

L'article est organisé de la manière suivante. La section 2 présente les étapes générales de caractérisation de collections de documents, les différents pré-traitements linguistiques ainsi que l'analyse statistique pour l'extraction des termes représentatifs. La section 3 détaille la méthodologie d'extraction de connaissances textuelles en vue de la construction de *KDMs*. La section 4 expose les résultats de nos expérimentations sur des collections de documents. La section 5 synthétise brièvement les travaux de fouille de données textuelles liés à notre problématique. Enfin, la section 6 résume notre approche et présente les perspectives de recherche associées.

2 Approche IC-Doc

L'approche *IC-Doc* proposée dans cet article fait suite à nos travaux sur la fouille de données textuelles (Mokrane et al. 2004a, Mokrane et al. 2004b). Cette approche permet une caractérisation automatique et thématique du contenu de collections de documents textuels. Les différentes étapes de cette approche sont les suivantes :

I. Pré-traitements linguistiques et analyse statistique des documents

- (a) Lemmatisation et étiquetage morpho-syntaxique.
- (b) Elimination des mots vides et détection des contextes.
- (c) Analyse statistique en vue de l'extraction des Termes Représentatifs (*TR*).

II. Extraction des connaissances textuelles

- (a) Représentation des termes.
- (b) Mesures de similarités.
- (c) Clustering et caractérisation thématique.

Dans la suite de cette section, nous présentons succinctement la phase I concernant les pré-traitements linguistiques et l'analyse statistique des documents. La phase II de l'approche *IC-Doc*, objet de cet article, sera décrite dans la section 3.

2.1 Pré-traitements linguistiques

La première étape des pré-traitements linguistiques consiste en la lemmatisation et l'étiquetage morphosyntaxique des documents. L'étape suivante concerne l'élimination des mots vides (articles, pronoms, prépositions, etc.) et la détection des différents contextes. À l'aide des étiquettes, nous conservons les noms, les verbes et les adjectifs. Dans le cadre de notre modèle un contexte correspond à une phrase. De manière générale, dans les différentes approches existantes, un contexte peut être une phrase, un paragraphe ou même l'ensemble du document. Dans le cadre de notre modèle, un contexte correspond à une phrase et ainsi la détection des contextes va correspondre à l'annotation des différentes phrases de la base documentaire.

2.2 Analyse statistique en vue de l'extraction des termes représentatifs

Avant de préciser comment extraire les termes représentatifs, nous donnons quelques définitions et notations utilisées par la suite.

2.2.1 Définitions et notations

Soit $\{T_1, T_2, \dots, T_n\}$ l'ensemble des termes de la base de données textuelles obtenus à l'étape des pré-traitements linguistiques.

Co-occurrence contextuelle (CO) : Deux termes T_i et T_j appartenant, en même temps au même contexte, forment une co-occurrence appelée *CO* et notée $\{CO : T_i - T_j\}$. L'ensemble des co-occurrences contextuelles d'une base documentaire *BDoc* est notée *COD* de *BDoc*.

Fréquences d'un terme (FTC et FTD) : La fréquence *FTC* d'un terme T dans une base de documents textuels correspond au nombre d'occurrences du terme T dans la base. La fréquence *FTD* d'un terme T dans une base de documents textuels correspond au nombre de documents contenant T . Les fréquences *FTC* et *FTD* d'un terme T_i sont notées respectivement FTC_i et FTD_i .

Fréquences d'une co-occurrence (FCC et FCD) : La fréquence *FCC* d'une co-occurrence *CO* dans une base de documents textuels correspond au nombre d'occurrences de *CO* dans la base. La fréquence *FCD* d'une co-occurrence *CO* dans un document D correspond au nombre d'occurrences de *CO* dans D .

Matrice de co-occurrences brute (MATCO) : Soit N le nombre de termes d'un corpus documentaire et E l'ensemble de ces termes. La matrice de co-occurrence brute d'une base documentaire notée *MATCO* de E correspond à une matrice de N lignes et N colonnes. La ligne i de la matrice correspond à un terme T_i de la base et la colonne j de la matrice correspond à un terme T_j de la base ($i = 1..N, j = 1..N$).

$$\text{Si } (i \neq j) \text{ } MATCO(i, j) = FCC \text{ de } \{CO : T_i - T_j\} \text{ sinon } MATCO(i, j) = FTC_i \quad (1)$$

Matrice de co-occurrences réduite (RMATCO) : À partir de la matrice de co-occurrences brute d'une base documentaire, nous pouvons construire une matrice de co-occurrences réduite définie comme suit : soit E l'ensemble des termes d'un corpus documentaire et

considérant les deux ensembles E_1 et E_2 avec $E_1 \subset E$ et $E_2 \subset E$, contenant respectivement M et K termes. La matrice de co-occurrences réduite notée $RMATCO$ de E_1 sur E_2 correspond à une matrice de M lignes et K colonnes. La ligne i de la matrice correspond à un terme T_i de l'ensemble E_1 et la colonne j de la matrice correspond à un terme T_j de l'ensemble E_2 . ($i = 1..M, j = 1..K$).

$$Si (T_i \neq T_j) \quad RMATCO(i,j) = FCC \text{ de } \{CO : T_i - T_j\} \quad \text{sinon} \quad RMATCO(i,j) = FTC_i \quad (2)$$

L'analyse statistique de la base documentaire consiste à calculer tout d'abord les FTC , FTD , et FCC de l'ensemble des termes E de la base documentaire $BDoc$. Elle consiste ensuite à construire la matrice de co-occurrences brute $MATCO$ de E pour l'extraction de l'ensemble des termes représentatifs.

2.2.2 Extraction des termes représentatifs

Dans le cadre de notre modèle, nous sélectionnons l'ensemble des termes représentatifs à l'aide de l'Algorithme 1. Plus de détails sont disponibles dans (Mokrane et al. 2004a, Mokrane et al. 2004b).

Algorithme 1 : Termes Représentatifs TR

Input: E ensemble des termes d'une base documentaire $BDoc$; $MATCO$ de E ;
 Vecteur $Vdist = \langle (T_1, FTD_1) \dots (T_i, FTD_i) \dots (T_n, FTD_n) \rangle$; $n = |E|$; $i = 1..n$
Output: TR (ensemble des Termes Représentatifs)
Begin
 1. $TR \leftarrow \emptyset$;
 2. **foreach** $T_i \in E$ **do**
 if $\frac{FTC_i}{FTD_i} > \alpha$ **then** $TR = TR \cup \{T_i\}$;
 3. **foreach** $\{CO : T_i - T_j\} \in COD \text{ de } BDoc$ **do**
 if $FCC \text{ de } \{CO : T_i - T_j\} > \beta$ **then** $TR = TR \cup \{T_i\} \cup \{T_j\}$;
End
 Les paramètres α et β sont les seuils de sélection des termes (Mokrane et al. 2004b).

A partir de l'ensemble des termes représentatifs (TR) et de la matrice de co-occurrences brute ($MATCO$) de la base documentaire, nous construisons la matrice de co-occurrences réduite $RMATCO$ de TR sur TR , cette matrice est utilisée dans la phase II de l'approche $IC-Doc$, décrite dans la section suivante.

3 Extraction des connaissances textuelles

Dans cette section, nous présentons la méthodologie de représentation de l'ensemble TR en se basant sur les relations textuelles entre les termes. L'objectif visé est de classer les

termes représentatifs par thématiques et de construire des *KDMs*. Avant de détailler les différentes étapes de notre méthode, nous définissons la notion de *KDM*.

Une *KDM* (*Knowledge Dynamic Map*) est un graphe $G = \langle X, U \rangle$ où X est un ensemble de N sommets modélisant N termes représentatifs et U un ensemble d'arêtes représentant les relations textuelles. Les sommets du graphe sont des hyperliens (liens dynamiques) à deux fonctionnalités. La première fonctionnalité permet d'organiser d'une manière automatique le graphe autour d'un thème central. La seconde permet d'atteindre une nouvelle *KDM*. La dimension d'une *KDM* correspond au nombre de ses termes représentatifs.

3.1 Représentation des termes

Pour représenter l'ensemble des termes TR qui seront utilisés lors du calcul des mesures de similarités, nous définissons les deux relations suivantes : Soient T_i et T_j deux termes représentatifs, nous notons $(T_i \wedge T_j)$ l'ensemble des termes représentatifs appartenant à des contextes d'apparition de T_i et de T_j . Cet ensemble est défini comme suit :

$$(T_i \wedge T_j) = \{T_k \in E / \{CO : T_i \text{---} T_k\} \wedge \{CO : T_j \text{---} T_k\}\} \quad (3)$$

Nous notons $(T_i \wedge \neg T_j)$ l'ensemble des termes représentatifs appartenant aux contextes de A et non pas aux contextes de B . Cet ensemble est défini comme suit :

$$(T_i \wedge \neg T_j) = \{T_k \in E / \{CO : T_i \text{---} T_k\} \wedge \neg \{CO : T_j \text{---} T_k\}\} \quad (4)$$

où $\neg \{CO : T_j \text{---} T_k\}$ signifie que le couples de termes $\langle T_j, T_k \rangle$ ne forme pas une co-occurrence contextuelle.

Nous représentons les termes de l'ensemble TR par deux matrices notées respectivement $MatR1$ et $MatR2$. La première matrice ($MatR1$) prend en considération la relation de co-occurrences contextuelles et la deuxième matrice ($MatR2$) prend en considération la notion de partage de contextes entre les termes représentatifs. Les deux matrices $MatR1$ et $MatR2$ sont calculées suivant l'*Algorithme 2*.

Algorithme 2 : Représentation de l'ensemble TR

Input: $TR = \{T_1, \dots, T_m\}$; $RMATCO$ de TR sur TR ; $m = |TR|$;

Output: Matrices $MatR1$ et $MatR2$

Begin

```

    for ( $i = 1$ ;  $i \leq m$ ;  $i++$ ) do
        for ( $j = 1$ ;  $j \leq m$ ;  $j++$ ) do
             $MatR1(i, j) = (FCC \text{ de } \{CO : T_i \text{---} T_j\}) / FTC_i$  ;
             $A = |(T_i \wedge T_j)|$  ;
             $B = |(T_i \wedge \neg T_j)|$  ;
             $MatR2(i, j) = \frac{A}{A + B}$  ;

```

End

3.2 Mesures de similarités, Clustering et KDMs

Après la représentation de l'ensemble des TR , nous calculons, à partir des deux matrices $MatR1$ et $MatR2$, les mesures de similarités entre les différents termes représentatifs TR de

la base documentaire. Nous notons $KDMAT$, la matrice de mesures de similarités entre les TR , cette matrice est calculée de la manière suivante :

Soit $T_i \in TR$, $T_j \in TR$ et $m = |TR|$; la similarité textuelle entre T_i et T_j est donnée par la formule (5).

$$KDMAT(i, j) = \alpha * Dist1(i, j) + (1 - \alpha)Dist2(i, j) \quad (5)$$

où $Dist1(i, j)$ et $Dist2(i, j)$ sont des distances euclidiennes calculés à partir des matrices $MatR1$ et $MatR2$ suivant les formules (6) et (7) :

$$Dist1(i, j) = \sqrt{\sum_{k=1}^m [MatR1(i, k) - MatR1(j, k)]^2} \quad (6)$$

$$Dist2(i, j) = \sqrt{\sum_{k=1}^m [MatR2(i, k) - MatR2(j, k)]^2} \quad (7)$$

Les expérimentations ont permis de fixer le paramètre α à 0.3. (i.e. le critère de co-occurrences contextuelles contribue à 30 % à la pertinence des résultats tandis que le critère de partage de contextes contribue à 70% à la pertinence des résultats.

Dans le but de classer les termes représentatifs (TR) par thématiques et de construire des $KDMs$ cohérentes, nous appliquons un algorithme de clustering aux données de la matrice $KDMAT$ adapté aux données de cette matrice. Nous avons choisi d'utiliser l'algorithme k-means, simple et robuste (Jain et al. 1999), qui nous a permis de mettre en œuvre notre approche.

Nous appliquons cette méthode de la manière suivante : soit NB le nombre de thématiques de la base documentaire et DK la dimension d'une KDM définie a priori de façon à ne pas surcharger l'utilisateur. Dans une première étape, nous appliquons k-means (NB) aux données de $KDMAT$. A l'issue de cette étape nous obtenons NB clusters, chaque cluster correspond aux termes représentatifs d'une thématique (sous ensemble des TR). De la même manière, l'étape suivante du processus de clustering consiste en l'application du k-means(DK) aux sous ensembles des termes représentatifs obtenus à l'étape 1. A la fin de cette seconde étape, chaque cluster de termes représente un ensemble de sommets d'une KDM . Le processus de clustering est relancé sur chacun des ensembles de termes d'une KDM si la dimension de cette dernière dépasse DK . Le processus de clustering de cette seconde étape est itératif. L'objectif du processus de clustering itératif est de permettre d'écarter un cluster en plusieurs autres clusters, i.e. une KDM en plusieurs autres $KDMs$, permettant ainsi une visualisation des résultats du clustering et une navigation par le contenu dans une base documentaire.

4 Expérimentation

Etant donné que nous ne nous intéressons pas dans cet article au traitement automatique du langage naturel ($TALN$), nous avons utilisé pour l'analyse linguistique des documents, cordial analyseur (Web 1 – Cf. Références) qui intègre un étiqueteur morphosyntaxique et un lemmatiseur fonctionnant pour les documents textuels en Français. Nous avons développé

une collection d'outils permettant de mettre en oeuvre l'approche *IC-Doc* ainsi qu'un prototype de visualisation des résultats en *KDMs* (FIG 1 illustre l'interface de ce prototype).

Différentes expérimentations ont été réalisées, dans l'objectif de montrer la pertinence et la capacité de notre approche pour une caractérisation thématique indépendamment des poids donnés aux thématiques dans les collections de documents.

4.1 Données

Nous expérimentons notre approche sur des collections de documents composées de trois thématiques qui sont : économie, informatique et cinéma. Les compositions des différentes collections de documents sont illustrées dans le tableau 1.

Documents analysés par étape	Economie <i>Nb_Doc</i>	Informatique <i>Nb_Doc</i>	Cinema <i>Nb_Doc</i>
C1	10	10	10
C2	40	40	40
C3	100	100	100
C4	10	40	100
C5	40	100	10
C6	100	10	40
C7	40	10	100

TAB 1 – Composition des collections de documents

4.2 Méthode

Après l'extraction des différents termes représentatifs *TR* à partir de chacune des collections de documents, nous appliquons le clustering suivant l'approche *IC-Doc* sur les *TR*, nous évaluons les résultats obtenus par les mesures de *Précision* et de *Rappel* sur les termes représentatifs extraits pour chacune des thématiques dans chaque collection de documents. La précision et le rappel dans le cadre de notre expérimentation sont définis comme suit : soit *S* l'ensemble des *TR* d'une thématique extraits par le système dans une collection de documents ; soit *V* l'ensemble des *TR* de la thématique dans la collection de documents, la précision et le rappel sont calculés comme suit :

$$\textbf{Précision} = |S \cap V| / |S| \quad \textbf{Rappel} = |S \cap V| / |V|$$

La précision détermine la quantité d'informations extraite appartenant à chacune des thématiques ; le rappel détermine la quantité d'information extraite par rapport aux thématiques.

4.3 Résultats

Les résultats obtenus sont illustrés sur le tableau 2. L'objectif de l'expérimentation sur la collection C1 (10 documents pour chacune des thématiques) est de montrer que les résultats pour une thématique ne sont pas significatifs dans le cas d'une quantité très réduite de

Fouille de collections de documents pour une caractérisation thématique

documents, en raison de données pauvres sur la thématique dans la collection de documents, ce qui se traduit par des chutes de précisions ou de rappels.

Dans tous les autres cas la précision dépasse les 75% et le rappel dépasse les 50% pour chacune des thématiques. Comme illustre TAB 2, la précision ou le rappel ne peuvent chuter pour une thématique que dans le cas de thématiques pauvres dans une collection (par exemple 10 documents) comme dans la collection C6 et C7 pour informatique, C5 pour cinéma ou C4 pour économie.

Résultats	Nombre des TR	Economie		Informatique		Cinéma	
		Précision	Rappel	Précision	Rappel	Précision	Rappel
C1	533	0.996	0.852	1.000	0.223	0.490	0.387
C2	1526	0.915	0.671	0.985	0.683	0.821	0.533
C3	2383	0.943	0.675	0.997	0.616	0.902	0.557
C4	1631	0.660	0.559	0.994	0.664	0.958	0.565
C5	1510	0.868	0.649	0.996	0.611	0.316	0.348
C6	1391	0.992	0.905	0.920	0.080	0.751	0.513
C7	1394	0.982	0.721	0.575	0.381	0.982	0.622

TAB 2 – Résultats par collection de documents

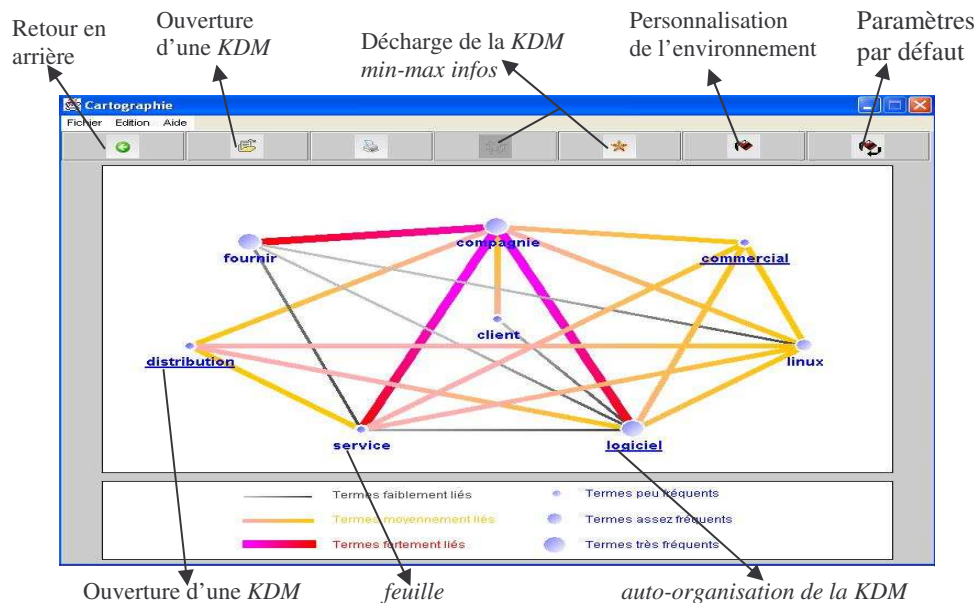


FIG 1 – Interface du prototype de visualisation des résultats en KDMs

5 Travaux connexes

Les outils et les méthodes de fouille de textes permettent l'acquisition, le classement, l'analyse, l'interprétation, l'exploitation et la visualisation d'informations contenues dans des documents textuels (Poibeau 2003). Actuellement, de nombreux travaux de recherche,

notamment issus du Web Mining (Kosala et Blockeel 2000) et du Text Mining, s'intéressent à la fouille de corpus documentaires (Baldi et Di meglio 2004, Besançon R 2001, Chen et al. 2001, Hongyuan et al. 2001, Han et Kamber 2000, Turenne 2000, Ihadjadene 2004). L'objectif de ces travaux est généralement d'analyser le contenu des documents pour en extraire des termes significatifs ainsi que les liaisons qui peuvent exister entre ces différents termes. Dans ce cadre, les modèles de similarités textuelles et la notion de co-occurrences sont les plus utilisées pour l'analyse du contenu (Poibeau 2003). Dans un contexte proche, celui de la recherche documentaire, la recherche de co-occurrences a également été largement étudiée, elle consiste à rechercher les associations de termes les plus fréquentes dans les documents afin de retrouver rapidement les documents pertinents qui peuvent répondre aux requêtes de l'utilisateur. Dans (Pereira et al. 1993) cette co-occurrence est utilisée pour la classification des termes selon la distribution de leurs contextes syntaxiques. TANAKA et IWASAKI (Tanaka et Iwasaki 1996) utilisent la matrice de co-occurrences pour la désambiguïsation des termes. Dans (Besançon R 2002) un modèle de filtrages syntaxiques de co-occurrences est proposé pour la représentation vectorielle de documents et la recherche documentaire. Tous ces travaux ne prennent pas en considération, à la fois, la notion de co-occurrences avec la notion de partage de contextes pour l'extraction des connaissances textuelles ou le choix des termes représentatifs d'une base de documents textuels. Ce qui implique une pénalisation d'une partie importante des relations textuelles. L'application de ces approches pour la caractérisation de bases documentaires est donc limitée dans la mesure où elle ne permet pas une extraction pertinente et représentative des informations sur les différents contenus textuels.

6 Conclusion

Dans le cadre de cet article, nous avons présenté l'approche *IC-Doc* de caractérisation thématique de connaissances textuelles à partir de collections de documents. *IC-Doc* est basée sur une méthode originale d'extraction et de classification de connaissances textuelles prenant en considération les co-occurrences contextuelles et le partage de contextes entre les différents termes représentatifs du contenu. Les résultats de nos expérimentations montrent la pertinence de notre approche et sa capacité pour une caractérisation thématique des collections de documents indépendamment des poids donnés aux thématiques dans les collections de documents. Lorsque les thématiques se chevauchent, les résultats du clustering pourraient être améliorés par des techniques de clustering flou. Ces techniques font l'objet de nos travaux en cours sur l'extraction et la caractérisation automatique de connaissances textuelles à partir de diverses collections de documents, facilitant les consultations documentaires et favorisant les échanges d'expériences entre utilisateurs.

Références

- Baldi S. et Di meglio E. (2004), A text mining strategy based on local contexts of words, Proceedings of JADT'04, Le poids des mots, Presses Universitaires de Louvain, Vol. 2, pp 79-87.
- Besançon R. (2002), Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents, Actes de TALN'02, pp 135-144.
- Besançon R. (2001), Intégration de connaissances syntaxiques et sémantiques dans les

- représentations vectorielles des textes, PhD thesis, Ecole polytechnique Fédérale de Lausanne, 2001.
- Chung W., Chen H. et Nunamaker J. (2003), Business intelligence explorer: A knowledge map framework for discovering business intelligence on the Web, Proceedings of HICSS'03, 10 p.
- Jain A.K., Murty M.N. et Flynn P.J. (1999). Data clustering: A review. ACM Computing Surveys, Vol. 31, Issue 3, pp 264-323.
- Chen H., Fan H., Chau M. et Zeng D. (2001), MetaSpider: Meta-searching and categorization on the Web, Journal of the American Society for Information Science and Technology, Vol. 52, pp 1134 –1147.
- Kosala R. et Blockeel H. (2000), Web Mining research: A survey. SIGKDD Explorations, 2(1), pp 1-15.
- Hongyuan Z., Xiaofeng H., Chris D., Ming G., et Horst S. (2001), Automatic topic identification using webpage clustering, Proceedings of ICDM'01, pp 25-31.
- Han. J. et Kamber. M. (2000), Data mining: concepts and techniques, Morgan Kaufmann Publishers, ISBN 1-55860-489-8, 2000.
- Mokrane A, Poncelet. P et Dray. G. (2004), Visualisation automatique du contenu d'une base de documents textuels via les hyper-cartes d'information, Actes des VSST'04, pp 239-250.
- Mokrane. A, Arezki. R, Dray. G et Poncelet. P. (2004): Cartographie automatique du contenu d'un corpus de documents textuels. Actes des JADT'04, Le poids des mots, Presses Universitaires de Louvain, Vol. 2, pp 816-823.
- Poibeau T. (2003), Extraction automatique d'information, du text mining au Web sémantique. Hermès sciences publications, ISBN : 2-7462-0610-2, 2003.
- Pereira, F., Tishby, N. et Lee, L. (1993), Distributional clustering of English words. Proceedings of the 31th Meeting of the Association for Computational Linguistics, pp183-190.
- Spinat E. (2002): Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ?, Colloque Cartographie de l'information : De la visualisation à la prise de décision dans la veille et le management de la connaissance.
- Tanaka K. et Iwasaki H. (1996), Extraction of lexical translations from non-aligned corpora, Proceedings of the 16th International Conference on Computational Linguistics, pp 580-585.
- Turenne N. (2000), Apprentissage statistique pour l'extraction de concepts à partir de textes - Application au filtrage d'informations textuelles, Thèse de doctorat, Université Louis Pasteur de Strasbourg, 2000.
- Ihadjadene M. (2004), Méthodes avancées pour les systèmes de recherche d'informations, Ouvrage collectif sous la direction de M. Ihadjadene, Hermès sciences publications, ISBN : 2-7462-0846-6, 2004.
- Web 1. Cordial analyseur, Site Web,
http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm

Recherche d'information multimédia : Apport de la fouille de données et des ontologies

Marie-Aude Aufaure *, Marinette Bouet **

* Supélec, Plateau du Moulon, Département Informatique,
F-91192 Gif-sur Yvette Cedex, France
Marie-Aude.Aufaure@supelec.fr
www.supelec.fr/ecole/si/pages_perso/aufaure.html

** LIMOS, UMR 6158 CNRS – Université Blaise Pascal (Clermont-Ferrand II)
Campus des Cézeaux – 24, Avenue des Landais – 63173 AUBIERE Cedex – France
Marinette.Bouet@cust.univ-bpclermont.fr

Résumé. A ce jour, le média image est omniprésent dans de nombreuses applications. Un volume de données considérable est produit ce qui conduit à la nécessité de développer des outils permettant de retrouver efficacement de l'information pertinente. Les systèmes de recherche actuels montrent aujourd'hui leurs limites en raison de l'absence de sémantique. Une voie qui semble intéressante à explorer afin de combler le fossé existant entre les propriétés extraites et le contenu sémantique, est la fouille de données. C'est un domaine de recherche encore immature mais très prometteur. Cet article présente des travaux préliminaires sur la manière de définir de nouveaux descripteurs intégrant la sémantique. Le clustering et la caractérisation des classes obtenues sont utilisés pour réduire l'espace de recherche et produire une vue résumée de la base. La navigation basée sur une ontologie visuelle est un moyen puissant et convivial pour retrouver de l'information pertinente.

1 Introduction

Durant la dernière décennie, un volume considérable de données multimédia a été produit. Ces données sont par essence complexes, non structurées et volumineuses et les applications ayant besoin de rechercher des images pertinentes de manière efficace, de plus en plus nombreuses. Du fait qu'une image ne contient pas directement d'information interprétable de manière automatique, les méta-données vont jouer un rôle très important. L'étape de pré-traitement permet d'extraire un ensemble de méta-données comme : (1) les méta-données relatives au type de donnée multimédia, (2) les méta-données descriptives : nom de l'auteur, date, etc., (3) les méta-données relatives au contenu (sémantique, visuel, relations spatiales) : le contenu visuel est décrit en termes de couleur, forme et texture, et le contenu sémantique est une interprétation de l'image.

Le but est de pouvoir traiter les données du pixel à la connaissance puisque par le vocable « image », on entend image numérique c'est-à-dire une image qui se présente sous la forme d'une matrice de pixels. Il est à noter aussi que le média image ne concerne que les images fixes; les images animées étant dénotées par l'expression « média animation ». Au niveau pixel, des descripteurs visuels sont extraits et les requêtes sont basées sur le contenu. De nombreux travaux existent dans le domaine de la vision par ordinateur sur la partie descripteurs visuels. Dans ce cas, la recherche d'information consiste en une recherche par similarité (**C**ontent **B**ased **I**mage **R**etrieval) basée sur une distance entre les descripteurs visuels extraits des images (Venters et Cooper 2000). Le niveau d'abstraction suivant est

celui des objets ou régions extraits des images. L'utilisateur peut alors sélectionner des objets dans sa requête, de manière à ce que celle-ci soit plus précise. Les relations spatiales sont prises en compte à ce niveau. Le niveau sémantique est dédié à l'extraction et à la génération de méta-données sémantiques, et peut être réalisé à l'aide d'ontologies (Staab et Studer 2004). Enfin, le niveau connaissance utilise le niveau sémantique pour découvrir des relations cachées entre les objets, de la connaissance et pour résumer et caractériser une grande base d'images.

La fouille dans les images est un domaine de recherche récent (Zhang et al. 2001, Simoff et al. 2002, Djeraba 2002) mais n'est pas encore très développé du fait que l'extraction de connaissances à partir des images reste une tâche difficile. Les techniques classiques de fouille de données (Han et Kamber 2001) ont été largement utilisées pour des données alphanumériques. Cependant, dans un contexte multimédia, les bases de données contiennent un volume important de données numériques (les descripteurs) et de données sémantiques (les annotations). Les techniques classiques de fouille de données ne peuvent donc pas être directement appliquées aux images du fait de leur nature non structurée et de leur grande dimensionnalité (Berrani et al. 2002, Oria et al. 2004). Les images (et plus généralement les données multimédia) représentent de nouveaux défis pour l'apprentissage et la découverte de connaissances. Parmi les techniques de fouille de données, les plus utilisées sont le clustering (Jain et al. 1999), les règles d'association (Agrawal et al. 1994) et les réseaux de neurones (Dreyfus et al. 2002). La plupart des approches de fouille d'images se basent uniquement sur les descripteurs visuels. Notre approche vise à combiner le visuel et le textuel en utilisant des méthodes de clustering et de caractérisation. L'idée est de permettre à l'utilisateur de naviguer du textuel au visuel à travers une ontologie visuelle spécifiquement dédiée à l'application considérée.

L'organisation de cet article est la suivante : le paragraphe 2 décrit la recherche par le contenu en distinguant les phases d'indexation logique et de recherche ; la section 3 décrit l'architecture que nous proposons de mettre en place et d'expérimenter sur des corpus d'images annotées. Enfin, nous concluons sur nos travaux futurs en section 4.

2 Recherche par le contenu

Cette section donne un rapide aperçu de l'état de l'art en matière de recherche d'images. La recherche d'images repose sur deux phases à savoir l'indexation logique et la recherche d'images à proprement parler. La phase d'indexation logique consiste à extraire et à modéliser les méta-données (descripteurs textuels et visuels) associées aux images et à les stocker dans une base de données. La phase de recherche permet à un utilisateur final de retrouver rapidement, facilement et efficacement des images « pertinentes ».

2.1 Indexation logique

La recherche dans une base d'images s'effectue en général à partir de descriptions textuelles (mots-clés, annotations, texte, etc.) et/ou de descriptions visuelles (couleur, forme, texture, relations spatiales). Ces descripteurs doivent donc être modélisés de telle sorte que la recherche d'images pertinentes soit efficace tant sur des bases d'images généralistes (pas de domaine d'application particulier) que spécifiques (visages, empreintes). L'extraction des descripteurs est une étape réalisée en amont de la recherche du fait des traitements et des temps d'exécution qu'elle nécessite. Les descripteurs ainsi modélisés et extraits sont stockés dans la base de données en vue d'une exploitation lors du processus d'interrogation.

En ce qui concerne les descripteurs textuels, ils peuvent être des mots-clés représentant des méta-données sur le contenu ou un modèle de base de données. La recherche par mots-clés est un processus très fortement limité du fait que les relations sémantiques entre les mots-clés ne sont pas prises en considération, comme par exemple le fait qu'un cheval est un animal. Si le mot-clé de la requête est cheval alors les images de chevaux annotées avec le mot-clé animal ne seront pas retournées. Un moyen de pallier cette limitation est d'utiliser une ontologie qui permettra de représenter les liens sémantiques entre les différents objets. D'un point de vue bases de données, de nombreux modèles ont été proposés ainsi que des langages de requêtes. Dans la plupart des cas, la description de la sémantique des images est réalisée manuellement. Le modèle DISIMA (Oria et al. 2001) est basé sur une base de données orientée objets et ICDM (Meharga et Monties 2001) sur une base de données relationnelle objets. Dans le premier prototype, une hiérarchie de classes est définie par l'utilisateur. Une image est composée d'un OID, d'un ensemble de représentations physiques (raster ou vecteur) et d'un contenu (relations spatiales, objets pertinents). Le lien entre un objet détecté dans une image et un objet de la classe des objets pertinents est établi manuellement. Dans le second prototype, le modèle peut être divisé en quatre niveaux d'abstraction : (1) le niveau image contenant les propriétés globales d'une image, (2) le niveau syntaxique qui extrait des caractéristiques locales, (3) le niveau contenu dans lequel les objets syntaxiques sont regroupés, et enfin, (4) le niveau sémantique qui définit des hiérarchies sémantiques. Le principal intérêt de ce modèle réside dans le niveau sémantique qui utilise des relations sémantiques comme la synonymie, l'hyponymie, etc. Les modèles semi-structurés sont également bien adaptés à la modélisation de bases d'images. MPEG-7 (Chang et al. 2001) est un standard de description de contenus multimédias. Cette description est écrite en XML et correspond à une approche semi-structurée.

Les descripteurs visuels quant à eux, résument l'information photométrique de l'image. Dans la mesure où le processus de recherche repose sur ces descripteurs, une grande attention est portée sur leur extraction et modélisation. Dans ce contexte des systèmes visuels, une modélisation est « intéressante » si non seulement elle est fiable, mais si elle est aussi compacte et précise. Un tel objectif ne passe que par une véritable synergie entre les domaines du traitement numérique d'images, du traitement du signal et des mathématiques.

D'une manière générale, ils sont représentés par un vecteur numérique; un ou plusieurs descripteurs pouvant être associés à une image. Deux approches se dessinent pour appréhender le contenu des images. La première consiste à modéliser les propriétés visuelles selon des modélisations ayant une correspondance directe avec des critères psycho-visuels humains. Par exemple, le descripteur texture se décline en termes de critères qualitatifs tels que le contraste, la granularité, la régularité, etc. Pour chaque propriété de nombreuses propositions de modélisations ont été faites tant dans le domaine de l'imagerie que dans celui de la recherche par le contenu. Il faut tout de même noter que le descripteur classique et naturel forme est plus sujet à discussion que les autres. En effet, une étape primordiale et préalable à la modélisation de la forme est la segmentation dont dépend essentiellement la qualité de la modélisation de la forme. Il existe de nombreux algorithmes de segmentation automatique qui donnent de bons résultats avec des images peu complexes ou des images complexes d'un domaine bien particulier sur lequel on a de la connaissance à priori. En revanche la segmentation dans le cadre d'images hétérogènes n'est pas toujours fiable dans le sens où les objets extraits ne correspondent pas forcément à des objets sémantiques du monde réel. C'est pourquoi, certains systèmes privilégient les techniques de segmentation semi-automatiques en vue d'obtenir des régions sémantiques, tandis que d'autres préfèrent avoir recours à des techniques entièrement automatisées afin de privilégier le traitement de gros volumes (au détriment des zones sémantiques). Dans ce dernier cas, une méthode

classique consiste à décomposer l'image en petites zones homogènes (blocs) en termes de texture et/ou couleur afin de déterminer « grossièrement » le contour d'un objet. Chaque bloc « cohérent » est alors caractérisé par sa position, sa forme, sa couleur et sa texture. Pour de plus amples informations sur la modélisation de ces propriétés, on pourra se référer pour la couleur à (Swain et Ballard 1991, Smith et Chang 1996), pour la forme et pour la texture à (Gonzalez et Woods 2002). La seconde approche quant à elle, a recours à des modélisations qui n'ont pas de correspondance directe avec des critères psycho-visuels humains. De ces modélisations résultent une signature de l'image importante et des recherches intéressantes. Par exemple, tandis que dans (Nastar et al. 1998) ces caractéristiques sont calculées au moyen des transformées de Fourier, des ondelettes, etc. dans (Schmid et al. 1998), les caractéristiques extraites appréhendent plutôt l'information photométrique locale de l'image. Enfin, il est à noter que l'estimation de la ressemblance entre deux propriétés est dépendante de la modélisation retenue et qu'elle se fait au travers de distances. Par exemple, le calcul de la similarité entre les descripteurs modélisés sous forme de « signature » est effectué au moyen de la distance de Mahalanobis puisque la distance euclidienne et la distance pondérée ne tiennent pas compte des différentes incertitudes et corrélations. De plus, chaque distance a ses avantages et limites : la distance quadratique apprécie particulièrement bien la similarité entre couleurs, en revanche elle nécessite des temps de calculs non négligeables. Pour de plus amples informations sur l'estimation de la similarité entre propriétés, on pourra se référer à (Niblack et al. 1998, Stricker et al. 1995, Venters et Cooper 2000, Oria et al. 2004).

En résumé, on peut dire que dans de nombreux cas, la modélisation correspond à un vecteur de valeurs numériques résumant en fait sous une autre forme l'information photométrique contenue dans l'image. Ce vecteur qui présente généralement une dimension non négligeable, n'est pas sans poser des problèmes d'indexation physique dans les bases de données puisque les caractéristiques extraites s'avèrent être d'excellentes candidates au support de cette indexation. Malgré son importance, l'indexation physique n'est pas abordée dans cet article (Böhm et al. 2001, Oria et al. 2004).. De plus, le choix de la modélisation à retenir est délicat et est étroitement lié au domaine d'application considéré ainsi qu'aux objectifs visés. C'est pourquoi le groupe MPEG a élaboré MPEG-7 (Chang et al. 2001), un standard de représentation du contenu pour le filtrage, la gestion, le traitement et la recherche d'information multimédia. Des liens évidents existent entre cette interface de description du contenu et les systèmes de recherche par le contenu. Cependant, même si MPEG-7 décrit le contenu des images, il ne spécifie pas comment les caractéristiques sont extraites et comment doit être effectuée la recherche sur ces dernières.

2.2 Recherche

Les systèmes de recherche incluent généralement des outils visuels de recherche permettant aux utilisateurs de définir une requête en dessinant, en sélectionnant des couleurs, des formes, etc. ou encore en sélectionnant une ou plusieurs images d'intérêt dans la base. A travers une interface conviviale et intuitive, les utilisateurs peuvent donc formuler leurs requêtes en exploitant à la fois les descriptions textuelles et visuelles (extraites durant la phase d'indexation logique et stockées dans la base). Ces deux types de métadonnées sont nécessaires pour parvenir à une recherche efficace ; en effet, l'utilisation seule du texte ou de l'information visuelle n'est pas suffisante pour décrire le contenu sémantique des images (la puissance d'expression de chaque descripteur est intrinsèquement limitée). Par exemple, les couleurs et les formes sont bien adaptées pour décrire les aspects visuels d'une région mais ne permettent pas d'exprimer des concepts de haut niveau, contrairement aux annotations textuelles qui permettent ces descriptions de haut niveau mais qui sont faibles pour

représenter le contenu visuel. Ces descripteurs, pris séparément, sont incomplets et inefficaces.

Le processus de recherche s'appuie alors sur une fonction de distance entre les descripteurs et calcule la similarité entre la requête utilisateur et la base d'images. Les images sont ensuite affichées par ordre de similarité décroissante. Du fait de l'imprécision des descripteurs extraits des images, le processus de recherche s'appuie sur l'interrogation probabiliste et le contrôle de pertinence. Ceci signifie que l'utilisateur obtient une liste de résultats ordonnés, et, s'il n'est pas satisfait de ce résultat, peut raffiner sa requête en choisissant, parmi les images retournées, des exemples positifs et négatifs. De nouveaux résultats sont alors obtenus et le processus peut être itéré jusqu'à ce que l'utilisateur soit satisfait du résultat. Le lecteur peut se référer à (Del Bimbo 1999) pour un état de l'art de la recherche par le contenu visuel.

D'un point de vue bases de données, les requêtes sont exprimées en utilisant des extensions de SQL ou OQL. De nouveaux prédicats comme *contains* et *similar* sont introduits. Ces requêtes peuvent être qualifiées de requêtes exactes. Des prédicats flous peuvent être introduits (Dubois et al. 2001) pour exprimer des requêtes imprécises.

La combinaison de descripteurs textuels et visuels s'avère cependant insuffisante, notamment lorsque l'interrogation sémantique prédomine, c'est-à-dire que l'image et son contexte sont nécessaires (comme par exemple la recherche de séquences audiovisuelles traitant du chômage). Cette limitation est connue comme le fossé sémantique entre l'apparence visuelle d'une image et l'idée que l'utilisateur se fait de l'information qu'il recherche, incluant bien évidemment une forte composante sémantique. La recherche par le contenu souffre d'un manque de puissance d'expression du fait que la sémantique n'est pas suffisamment prise en compte. C'est la raison pour laquelle de nombreux travaux de recherche sont actuellement menés sur la sémantique des images.

Les approches actuelles visent essentiellement à propager des annotations à partir de bases d'images partiellement annotées. Dans notre approche, présentée au paragraphe suivant, nous souhaitons construire une ontologie dédiée à l'application à partir des annotations et l'utiliser dans une phase d'exploration de la base d'images.

3 Une architecture intégrant la fouille de données et les ontologies

En vue de supporter des systèmes de recherche d'images plus puissants, une nouvelle architecture est proposée en figure 1. En ce qui nous concerne, nous souhaitons exploiter la synergie de deux approches à savoir la fouille de données et une ontologie visuelle afin de permettre aux utilisateurs d'explorer et d'exploiter au mieux la base d'images.

La fouille de données est un ensemble de méthodes visant à extraire de la connaissance dans un but exploratoire ou décisionnel. Dans notre approche, nous nous situons dans un contexte exploratoire puisque nous cherchons à déterminer un ensemble de clusters et de règles à partir de descripteurs visuels et textuels (métadonnées associées à notre base).

La sémantique peut être exprimée de manière plus ou moins riche, allant de simples taxonomies aux ontologies (Guarino 1995, Staab et Studer 2004). Une taxonomie est un vocabulaire contrôlé organisé sous forme hiérarchique. Un thésaurus est organisé dans un ordre structuré et connu, de telle manière que les relations d'équivalence, homographiques, hiérarchiques et associatives soient clairement identifiées. Wordnet (Miller 1995) en est un exemple, et organise les noms, verbes, adjectifs et adverbes de la langue anglaise en ensembles de synonymes. Une ontologie est un modèle abstrait représentant une

compréhension commune et partagée d'un domaine. Une ontologie est décrite par un ensemble de concepts, de relations entre ces concepts et de propriétés. Elles peuvent être définies de manière plus ou moins formelle, du langage naturel aux logiques terminologiques. Le langage OWL (Web Ontology Language) appartient à cette dernière catégorie.

Dans l'architecture de la figure 1, on retrouve les deux processus que sont l'extraction et la recherche. Le premier processus a pour objectif de créer un résumé de la base d'images. Après extraction et stockage des caractéristiques visuelles et textuelles des images, le système résume la base d'images au moyen de méthodes issues de la fouille de données. Cette étape intitulée « extraction de connaissances » et détaillée dans la figure 2, s'avère être le noyau de notre architecture ; noyau autour duquel s'articulent les processus d'extraction et de recherche. Il nécessite plusieurs méthodes comme le clustering et la caractérisation des clusters sous forme de règles. Alors que le clustering (Jain et al. 1999) est utilisé pour réduire l'espace de recherche, les règles elles, ont pour finalité de caractériser chaque cluster et de permettre la classification automatique de toute nouvelle image dans les clusters auxquels elle doit appartenir. Du fait de leur nature intrinsèque différente, les descriptions textuelles et les descriptions visuelles sont traitées séparément selon des techniques pour lesquelles les mesures et distances sont appropriées à leur spécificité. Ainsi, à partir de chaque ensemble de caractéristiques (comme l'ensemble des couleurs, l'ensemble des mots-clefs, etc.), le système regroupe automatiquement ensemble les images partageant des propriétés similaires en s'appuyant sur le principe des cartes de Kohonen (Kohonen 1995, Dreyfus et al. 2002). Cette étape qui consiste à simplement grouper les objets similaires ensemble est loin d'être simple. La qualité des clusters obtenus dépend souvent du choix des paramètres initiaux ce qui est un problème en soi. De plus, nous souhaitons caractériser chaque cluster au travers d'une représentation plus appropriée que celle du centroïde, à savoir sous forme de règle. Ces règles sont déterminées soit à partir de tous les points du cluster afin d'obtenir les motifs les plus fréquents, soit à partir d'une agrégation des données comme un histogramme médian dans le contexte des clusters couleur (ce qui est représentatif du contenu des clusters). Les règles sont de la forme antécédent \rightarrow conséquent avec une certaine précision où antécédent et conséquent correspondent respectivement à une valeur de caractéristique et à un cluster. La précision quant à elle est fondamentale puisque son rôle est de permettre l'estimation de la qualité des règles induites. Elle repose en fait sur des mesures statistiques. Des méthodes telles que le marquage symbolique (Diday et al. 2000), la découverte de règles d'association (Agrawal et al. 1994, Han et Kamber 2001] etc. sont envisagées.

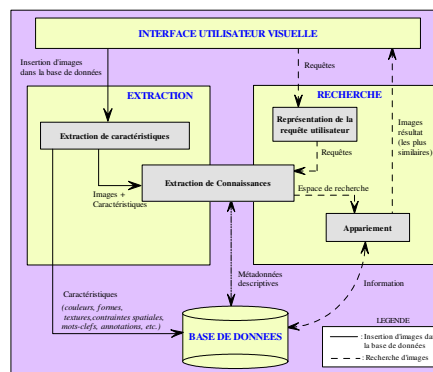


FIG. 1 - Architecture proposée pour les systèmes de recherche d'images

En ce qui concerne le processus de traitement des descriptions textuelles, il se différencie de celui de traitement des descriptions visuelles par le fait qu'il nécessite une phase de pré-traitement afin de réduire le nombre de mots-clés. Ensuite, le clustering peut être réalisé selon des techniques de clustering conceptuel comme Cobweb ou bien encore en appliquant les techniques de nuées dynamiques après transformation des données textuelles en données numériques. L'ensemble des concepts extraits à partir des clusters est alors hiérarchisé à l'aide de connaissances à priori du domaine, d'un expert ou par des méthodes de classification hiérarchique. Dans le cas où nous disposons uniquement de mots-clés, l'ontologie du domaine est primordiale pour traduire les relations sémantiques existant entre les informations textuelles considérées. Dans le cas où l'aspect textuel est plus important et se traduit par des documents associés aux images (comme des pages web par exemple), il est possible d'extraire automatiquement des relations entre les concepts.

De plus, suite au calcul du résumé d'une base, le système doit être en mesure de classer automatiquement de nouvelles images dans les clusters les plus appropriés au moyen des règles de caractérisation. Cette classification d'images dans les « bons » clusters n'est envisageable que si les règles extraites au préalable sont globalement respectées. Dans la négative, cela soulève un problème crucial nécessitant de plus amples travaux : plusieurs solutions peuvent être envisagées (1) la génération d'un cluster « bruit », (2) la prise en considération de cette nouvelle image et surtout de son impact sur le clustering et leur caractérisation, etc. La suppression d'images dans la base de données n'est pas abordée puisqu'en fait elle soulève les mêmes questions que l'insertion de nouvelles images. Enfin, après avoir effectué le clustering pour réduire l'espace de recherche et caractérisé les clusters au moyen de règles, le système sauvegarde les métadonnées descriptives dans la base de données. Ces métadonnées correspondent en fait aux caractéristiques découvertes et partagées par les images appartenant aux mêmes clusters. Elles ont un rôle important puisqu'elles doivent permettre à l'utilisateur de passer du monde « textuel » au monde « visuel » et inversement, c'est-à-dire lui permettre d'explorer et d'exploiter au mieux la base d'images. La navigation se fera au moyen d'une ontologie visuelle : à partir des concepts extraits des mots-clés, une hiérarchie de concepts va être construite et, pour chaque concept, seront associées des images représentatives issues des clusters visuels.

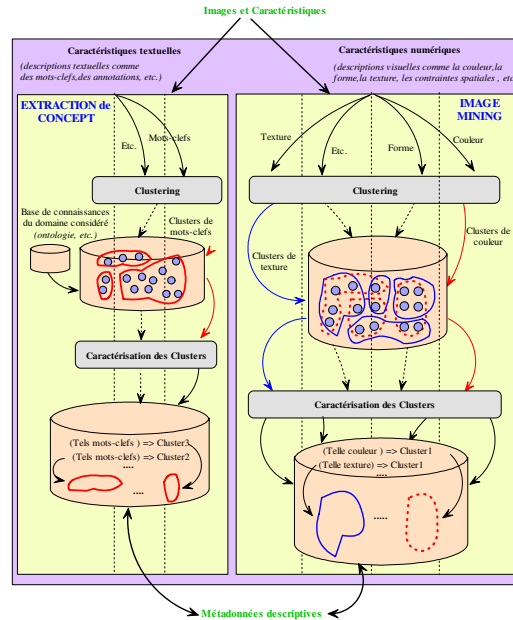


FIG. 2 - Processus « Extraction de Connaissances »

Des premiers résultats de clustering obtenus au moyen des cartes de Kohonen sur une base d'images texturées nous incitent à poursuivre dans cette voie. Ces cartes auto-organisatrices, initialement introduites par Kohonen en 1981, s'attachent à représenter des données multidimensionnelles sur de grands volumes de données. Elles permettent de projeter des données représentées dans un espace de grande dimension dans un espace de faible dimension. De nombreuses applications utilisent les cartes de Kohonen ou des extensions comme par exemple la classification de documents web, la robotique, la recherche d'images par le contenu (Oja et Kaski 2003), etc. Concernant l'aspect textuel, une expérience de découverte d'une ontologie a été réalisée sur un corpus de pages web relatives au domaine du tourisme. Les résultats obtenus sont présentés dans (Karoui et al. 2004). Les clusters obtenus constituent une étape vers la recherche d'images (voire multimédia) plus « intelligente » et cette architecture devrait donner toute son ampleur à la recherche d'information dans les bases d'images voire sur Internet.

4 Conclusion

Ce papier se situe au croisement inévitable du traitement d'image, des bases de données, de la recherche d'informations et de la fouille de données. Il s'inscrit en fait dans la problématique de la recherche d'images au sein d'une base de données. Suite à une présentation des différentes méthodes employées dans les systèmes de recherche actuels et au constat du manque de puissance de ces méthodes pour retrouver efficacement de l'information pertinente comportant des concepts sémantiques, la fouille de données est avancée en vue de pallier leurs limitations en terme d'exploitabilité et d'explorabilité.

L'architecture proposée combine la fouille de données et l'ontologie. Du fait de leur nature intrinsèque différente, les descriptions textuelles et visuelles sont traitées séparément selon des techniques appropriées à leur spécificité. Tandis que nous avons recours au

clustering pour réduire l'espace de recherche, nous utilisons les règles de caractérisation pour décrire chaque cluster et classer une nouvelle image dans les « bons » clusters de la base. Ces techniques favorisent les performances de recherche puisque le système apparie uniquement les caractéristiques sélectionnées avec celles des « bons » clusters de la base. Elles favorisent également l'explorabilité et l'exploitabilité puisque les métadonnées descriptives découvertes et partagées par les images appartenant aux mêmes clusters doivent permettre à l'utilisateur de passer du monde « textuel » au monde « visuel » et inversement, c'est-à-dire naviguer au moyen d'une ontologie visuelle. Des premiers résultats de clustering obtenus au moyen des cartes de Kohonen sur des images texturées nous incitent à poursuivre dans cette voie. Des expérimentations complémentaires sur une base d'images plus complexes sont envisagées. Nous continuons à développer l'architecture proposée et étudions diverses méthodes pour déterminer des métadonnées descriptives et une ontologie visuelle appropriées.

Références

- Agrawal, R. et al (1994), Fast algorithms for mining association rules. International Conference Very Large Data Bases, Santiago, Chili, pp. 487-499.
- Berrani, S., Amsaleg, L. and Gros P. (2002), Recherche par similarité dans les bases de données multidimensionnelles: panorama des techniques d'indexation. RSTI, Ingénierie des systèmes d'information, bases de données et multimédia, 7(5/6), pp 9-44.
- Böhm, C. et al. (2001), Searching in high-dimensional spaces: index structures for improving the performance of multimedia databases. ACM Computing surveys, 33(3).
- Chang, S.F., Sikora, T. and Purl, A. (2001), Overview of the MPEG-7 Standard. IEEE Transactions on Circuits and Systems for Video Technology, special issue on MPEG-7, pp 688-695.
- Del Bimbo, A. (1999), Visual Information Retrieval. Morgan Kaufmann (eds).
- Diday, E., Kodratoff, Y., Brito, P. and Moulet, M. (2000), Induction symbolique numérique à partir de données, (chap. marquage symbolique), Cepadue Ed.
- Djeraba, C. (2002), Association and Content-Based Retrieval. IEEE Transaction on Knowledge and Data Engineering.
- Dreyfus, G., Martinez, J.M., Samuelides, M., Gordon, M.B., Badran, F., Thiria, S. and Hérault, L. (2002), Réseaux de neurones, méthodologie et applications, Ed. Eyrolles.
- Dubois, D., Prade, H., and Sedes, F. (2001), Fuzzy Logic Techniques in Multimedia Databases Querying: A Preliminary Investigation of the Potentials. IEEE TKDE 13(3), pp 383-392.
- Gonzalez, R.C. and Woods, R.E. (2002), Digital image processing. 2nd Ed., Prentice Hall.
- Guarino, N. (1995), Formal Ontology, Conceptual Analysis and Knowledge Representation. International Journal of Human and Computer Studies, 43(5/6), pp 625-640.
- Han, J. and Kamber, M. (2001), Data Mining: Concepts and Techniques, San Francisco, California, Morgan Kaufmann.
- Jain, A.K. and Murty, M.N. and Flynn, P.J. (1999), Data Clustering: A Review. ACM Computing Surveys, 31 (3), pp 264-323.
- Karoui, L., Aufaure, M.A. and Bennacer N. (2004), Ontology Discovery from Web Pages: application to tourism. Workshop on Knowledge Discovery and Ontologies, collocated with ECML/PKDD, Pisa, Italy.
- Kohonen, T. (1995), Self-Organizing Maps. Springer, Berlin.
- Meharga, M.T. and Monties, S. (2001), An Image Content Data Model for Image Database Interrogation. International Workshop on Content-Based Multimedia Indexing, Brescia.

- Miller, G.A. (1995), WordNet: A Lexical Database for English. Communications of the ACM, 38(11), pp 39-41.
- Nastar, C. et al, SurfImage: a Flexible Content-Based Image Retrieval System. The 6th ACM International Multimedia Conference (MM'98), Bristol, England, 1998.
- Niblack, W. et al. (1998), The QBIC project: Querying images by content using color, texture and shape. In Proc. SPIE Storage and Retrieval for Image and Video Databases.
- Oja, E. and Kaski, S. (2003), Kohonen Maps, Elsevier, 2nd Ed.
- Oria, V., Özsu, T. and Iglinski, P.J. (2001), Querying Images in the DISIMA DBMS. Proc. of the 7th Int. Workshop on Multimedia Information Systems, Capri, Italy, pp 89-98.
- Oria, V., Li, Y. and Dorai, C. (2004), Multimedia Databases: Analysis, Modeling, Querying and Indexing. In Computer Science and Engineering Handbook, 2nd Ed., CRC Press.
- Schmid, C. et al. Comparing and evaluating interest points. In proceedings of the 6th international conference on Computer vision, Bombay, India, 1998.
- Simoff, S.J., Djeraba, C. and Zaïane, O.R. (2002), Multimedia Data Mining between Promises and Problems. MDM/ KDD2002, ACM SIGKDD Explorations, 4(2).
- Smith, J.R. and Chang, S.F. (1996), Tools and Techniques for Color Image Retrieval. Storage & Retrieval for Image and Video databases IV, SPIE Proceedings, 2670.
- Staab, M. and Studer, R. (eds) (2004), Handbook on Ontologies. Springer.
- Stricker, M. et al. (1995), Similarity of color images. Storage and Retrieval for Image and Video databases III, SPIE Proceedings, 2420.
- Swain, M.J. and Ballard, D.H. (1991), Color indexing. Int. journal of computer vision, 7(1).
- Venters, C.C. and Cooper, M.D. (2000), A Review of Content-Based Image Retrieval Systems. JISC Technology Applications Program.
- Zhang, J., Hsu, W. and Lee, M.L. (2001), Image mining: issues, frameworks and techniques. Second International Workshop on Multimedia Data Mining, San Fransisco, USA.

Summary

Nowadays, image media is omnipresent for various applications. A considerable volume of data has been produced and we need now to develop powerful tools allowing to efficiently retrieve relevant information. At present, CBIR systems suffer from a lack of expressive power because they do not integrate enough semantics. An interesting way we want to explore to fill the semantic gap between visual features and semantics content is image mining. This research field is recent and still remains immature but seems to be a promising issue. This paper presents some preliminary work and ideas about the way to define new descriptors to integrate image semantics. Clustering and obtained class characterization are used to reduce the research space and to produce a summarized view of the image database. As far as the navigation is concerned, it is based on a visual ontology, a powerful and user-friendly tool to retrieve relevant information.

Complexité de l'extraction des connaissances de données : une vision systémique

Walid Ben Ahmed*,***, Mounib Mekhilef*
Michel Bigand**, Yves Page***

*LGI – Laboratoire de Génie Industriel, Ecole Centrale Paris, Grande voie des Vignes 92295
Châtenay-Malabry cedex, France
{walid, mekhilef@lgi.ecp.fr}

**Équipe de Recherche en Génie Industriel, Ecole Centrale de Lille, 59651 Villeneuve
d'Ascq, France
michel.bigand@ec-lille.fr

***LAB (PSA-Renault), Laboratoire d'Accidentologie, de Biomécanique et d'études du
comportement humain, 132, rue des Suisses-92000 Nanterre
yves.page@lab-france.com

Résumé. Les praticiens et les chercheurs dans le domaine d'Extraction de Connaissances de Données (ECD) sont souvent confrontés à des difficultés qui sont relatives à la nature des données, à l'implication de l'opérateur humain et aux aspects algorithmiques. Aujourd'hui, s'il y a un consensus sur la « complexité » du processus d'ECD, ce n'est pas le cas pour la définition et la caractérisation de cette complexité. Définir la complexité de l'ECD, la caractériser et connaître ses sources sont des questions qui animent aujourd'hui la communauté de fouille de données. Dans cet article, pour répondre à ces questions, nous menons une réflexion sur la notion de complexité en ECD en utilisant l'approche systémique, une approche de modélisation de systèmes complexes.

1 Introduction

Aujourd'hui avec l'informatisation des saisies de données (utilisation des codes à barres, informatisation des transactions, etc.) et la puissance des systèmes de collecte de ces données (satellites, ordinateurs, etc.), des grandes Bases de Données (BD) sont construites et ne cessent de s'agrandir. L'exploitation de ces millions de données en management, en administration, en médecine, en géologie, en biologie et dans beaucoup d'autres domaines fait appel à des techniques d'Extraction de Connaissances de Données.

Le processus d'Extraction de Connaissances de Données (ECD) est défini comme : « *un processus d'identification de modèles (ou paradigmes) valables, nouveaux, potentiellement utiles et compréhensibles dans les données* » (Fayyad, Piatetsky-Shapiro et al. 1996). C'est un processus interactif et itératif, impliquant de nombreuses étapes avec des décisions prises par l'utilisateur (Brachman and Anand 1996). Les praticiens et les chercheurs dans le domaine d'ECD sont souvent confrontés à des difficultés qui sont relatives aux trois phases principales de ce processus (i.e. la *préparation des données*, la *phase de data mining* et l'*interprétation des résultats*). Cependant, s'il y a un consensus sur la « complexité » du processus d'ECD, ce n'est pas le cas pour la définition et la caractérisation de cette complexité. Plusieurs facteurs sont généralement considérés comme causes de complexité du

processus d'ECD. Nous citons, par exemple, la quantité de données, la qualité des données (erronées, manquantes, bruyantes, etc.), l'implication de l'opérateur humain, la nature des connaissances à extraire, la complexité des algorithmes utilisés, la multi-disciplinarité du domaine et l'implication de plusieurs points de vue.

Définir la complexité de l'ECD, la caractériser et connaître ses sources sont des questions qui animent aujourd'hui la communauté de fouille de données. Dans cet article, nous menons une réflexion sur la notion de « *complexité* » pour apporter des éléments de réponses à ces questions. Nous utilisons pour cela *l'approche systémique* (à ne pas confondre avec l'approche systématique), qui est une approche de modélisation de systèmes complexes. Dans la première section de cet article, nous présentons un aperçu historique sur les origines de l'approche systémique. Dans la deuxième section, nous présentons ses principes de base. Dans la troisième section, nous utilisons cette approche pour caractériser la complexité du processus d'ECD

2 L'approche systémique : les origines épistémologiques

Aujourd'hui, les termes de la *cybernétique*, *cybernétique du second ordre* (Von Foerster 1995), la *théorie du système général* (Bertalanffy 1969), la *systémique* et la *systémographie* (Le Moigne 1999; Morin and Le Moigne 1999) sont utilisés pour désigner, à peu près, la même approche.

L'*approche cybernétique* ne se focalise pas sur la composition matérielle d'un système, mais insiste sur les interactions entre les composants. L'*observateur* et l'*objet observé* ne sont plus séparés et le *résultat de l'observation* dépend de leur interaction. Il y a eu reconnaissance du fait que toutes nos connaissances sur les systèmes sont basées sur des représentations simplifiées (i.e. des modèles). En cybernétique, Un système n'est plus considéré comme une entité passive qu'on peut observer et manipuler, mais comme un agent qui interagit avec son environnement et avec un autre agent qu'est l'observateur. Ce dernier est lui aussi perçu comme un système cybernétique (i.e. complexe, voir la définition dans le paragraphe 3) qui construit un modèle d'un autre système cybernétique. Il s'agit donc d'appliquer la cybernétique à elle-même ou ce que le fondateur de cette théorie, Heinz von Foerster, appelle la *cybernétique du second ordre* dans son livre « *Cybernetics of Cybernetics* » (Von Foerster 1995).

La systémique (ou la cybernétique) est issue de l'épistémologie constructiviste. Cette épistémologie *reconnaît le caractère relatif de la connaissance et sa dépendance de la construction du sens par les individus en se basant sur leurs expériences et leurs interactions avec leur environnement (contexte)*. Un *système* dans une perspective constructiviste est défini comme une représentation de la réalité perçue par un certain nombre d'individus dans un contexte donné. Un *modèle* est donc une représentation de la réalité et il n'est valide que dans un contexte donné. Les *problèmes* ne sont pas indépendants des perceptions des individus qui les traitent. Il existe alors plusieurs *solutions* et la solution optimale est celle qui est acceptable pour la majorité. Les *méthodes de recherche* ainsi que leurs résultats reflètent donc la perception et l'interprétation des chercheurs. Les *sources d'information* doivent être diversifiées pour couvrir la diversité des opinions. En ce qui concerne les *données*, on accorde plus d'importance au processus de leur collecte qu'aux données elles-mêmes (une synthèse de ces principes peut être trouvée dans (Richard 2003)).

3 L'approche systémique : les principes de base

La systémique distingue les *systèmes complexes* et les *systèmes compliqués*, « *la complexité n'est pas la complication* », nous dit Edgar Morin (Le Moigne 1999; Morin and Le Moigne 1999).

Un **système compliqué** est un système qui est caractérisé par un comportement qui peut être prévu par l'analyse des interactions entre ses composantes, il est déterministe (e.g. un ordinateur). Les approches analytiques sont adaptées à la modélisation de ce type de système. Un **système complexe** ou **système cybernétique** est un système non-déterministe dont le comportement ne peut pas être prévisible par analyse disjonctive de ses différents éléments. Selon Miller (Miller 1995), c'est un système *vivant, évolutif et ouvert à son environnement* avec lequel il est en *interaction continue*. C'est donc un système qui *fonctionne* et se *transforme* en même temps. Son comportement peut être décrit en terme de *feedbacks* et *boucles récursives* (Morin and Le Moigne 1999). Les éléments d'un système complexe sont en interaction réciproque. L'action d'un élément sur un autre entraîne en retour une réponse du second élément vers le premier. On dit alors que ces deux éléments sont reliés par une *boucle de feedback* (ou *boucle de rétroaction*). Les interactions entre les éléments d'un système sont régies selon le principe du *holisme* : « *le tout est supérieur à la somme des parties* ». Les interactions entre les éléments d'un système donnent à l'ensemble des propriétés que ne possèdent pas les éléments pris séparément.

En se basant sur les fondements du constructivisme, l'approche systémique insiste sur *l'inséparabilité* entre *l'observateur* (e.g. analyste), *l'objet observé* (e.g. données) et le *contexte de l'observation* (e.g. objectif de l'analyse). Une tâche telle que *l'extraction des connaissances de données*, dépend donc non seulement des données, mais aussi de la personne qui l'effectue et du contexte (objectif, environnement, etc.). Puisqu'un système complexe est un système qui *existe*, qui *fonctionne* et se *transforme* en même temps et qui a un *objectif* (ou *téléologie*), la systémique propose de conjointre quatre points de vue génériques pour l'analyser et appréhender sa complexité. Nous désignons par *point de vue* « *une position conceptuelle par rapport à un objet, cette position servant à en donner une description particulière* ». Les points de vue systémiques sont (Le Moigne 1999) :

- *Le point de vue ontologique* (le « quoi ») : ce qu'est le système. Le terme *ontologie* est issu du domaine de la philosophie où il signifie « explication systématique de l'existence ». L'axe ontologique représente pour nous les composants du système,
- *Le point de vue fonctionnel* (le « faire ») : ce que fait le système,
- *Le point de vue transformationnel* ou *génétique* (le « devenir ») : comment le système évolue, quels sont les états générés. Ce point de vue décrit l'aspect dynamique, évolutionnel et génétique (dans le sens de la genèse et non celui de l'hérédité) du fonctionnement du système,
- *Le point de vue téléologique* ou *motivationnel* (le « pourquoi ») : quels sont l'objectif et la motivation du système. La téléologie signifie en philosophie « l'étude de la finalité ».

4 L'ECD : un système complexe

Dans cette section, nous commençons par proposer une architecture multi-points de vue pour analyser la complexité du processus d'ECD. Pour cela, nous nous basons sur l'approche systémique. Ensuite, nous appliquons cette architecture à la caractérisation de cette complexité.

4.1 Proposition d'une architecture multi-points de vue pour analyser la complexité de l'ECD

Comme le suggère la systémique, nous assimilons le processus d'ECD à un système complexe (ou système cybernétique). Pour analyser la complexité de ce système, nous proposons une architecture constituée de points de vue génériques et ayant deux niveaux d'abstraction : le premier est composé des trois points de vue « *objet observé* », « *observateur* » et « *contexte de l'observation* ». Le deuxième niveau d'abstraction est composé des quatre points de vue *ontologique*, *fonctionnel*, *transformationnel* et *téléologique*. Chacun des trois points de vue du premier niveau est analysé selon les quatre points de vue du deuxième niveau. La Fig. 1 donne une illustration en UML¹ de cette architecture :

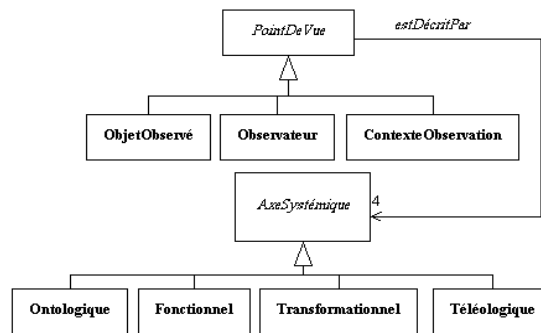


FIG. 1 - Architecture systémique pour l'analyse de la complexité du processus ECD

– **Le point de vue « objet observé ».** Dans un processus d'ECD, ce point de vue concerne les données à traiter. Ci-dessous, nous analysons ce point de vue selon les quatre aspects systémiques :

- *Aspect ontologique* : cet aspect est relatif à la nature des données à traiter. Ces données peuvent être multi-domaines, multi-formes (textuelles, images, vidéo etc.), multi-niveaux de granularité. Elles peuvent être incomplètes, bruitées, aberrantes et incohérentes. Cela a un lien avec le processus de leur collecte que nous abordons dans le point suivant,
- *Aspect fonctionnel* : cet aspect est relatif à la façon dont les données ont été collectées. En d'autres termes, il concerne le processus de la collecte des données. Ces dernières peuvent être issues de mesures automatiques effectuées par des machines (e.g. température). Elles peuvent être aussi issues d'entretiens effectués par des opérateurs humains. Les opérateurs peuvent être issus de disciplines différentes et dans ce cas là, on aura des données multi-sources. Ils peuvent avoir des expériences différentes et/ou des points de vue différents même s'ils sont de la même discipline,
- *Aspect transformationnel et génétique* : cet aspect est relatif à l'évolutivité des données. Le fait qu'elles ne soient pas collectées au même moment peut avoir une influence sur la façon de les traiter et sur la nature des connaissances qu'on peut y trouver,
- *Aspect téléologique* : cet aspect est relatif aux objectifs pour lesquels les données ont été collectées. Sachant que cet objectif peut évoluer au fil du temps, l'objectif de l'étude en

¹ UML : Unified Modeling Language (Booch et al. 1999).

question peut ne pas être prévu pour ces données. Cela affecte les méthodes à utiliser ainsi que les résultats à espérer,

– **Le point de vue « observateur ».** Dans un processus d'ECD, ce point de vue concerne l'opérateur (ou les opérateurs) humain(s) chargé(s) de la tâche d'ECD ainsi que la tâche d'ECD elle-même :

– *Aspect ontologique* : cet aspect est relatif à l'expérience de l'opérateur (statisticiens, expert du domaine, etc.), son domaine et sa discipline. Il est relatif aussi aux ressources (humaines et matérielles) utilisées pour résoudre la tâche,

– *Aspect fonctionnel* : cet aspect est relatif aux différentes fonctions effectuées durant le processus d'ECD. Elles peuvent être regroupées en trois catégories principales : la préparation de données, la phase de data mining et l'interprétation des résultats. La dépendance des différentes tâches a une influence certaine sur le déroulement du processus ECD ainsi que sur les résultats. En effet, les résultats d'une tâche telle que la tâche du data mining peuvent amener à revoir ses inputs qui sont les outputs d'une autre tâche. Il s'agit donc de revoir la tâche précédente, qui est la préparation des données dans ce cas là. Les outputs d'une tâche conditionnent aussi le déroulement de la tâche suivante. C'est le cas par exemple de la relation entre les tâches préparation de données/data mining ou data mining/interprétation des résultats,

– *Aspect transformationnel et génétique* : cet aspect est relatif à la transformation des données tout au long du processus d'ECD. Quant à l'aspect génétique, il est relatif aux résultats générés durant le processus. Ces deux aspects ont une influence sur la perception de l'observateur qui peut elle-même évoluer et donc faire évoluer la façon dont les données sont traitées ou même collectées,

– *Aspect téléologique* : cet aspect concerne l'objectif de l'étude. La sélection des données pour l'étude ainsi que les méthodes d'analyse utilisées dépendent étroitement de cet objectif. Ce dernier peut même évoluer au cours de l'analyse, influencé par exemple, par des résultats intermédiaires,

– **Le point de vue « contexte de l'observation ».** Dans un processus d'ECD, ce point de vue concerne l'environnement dans lequel s'effectue la tâche d'ECD.

– *Aspect ontologique* : cet aspect est relatif à l'environnement physique dans lequel s'effectue le processus d'ECD. Il inclut, entre autres, l'utilisateur final des connaissances à extraire (un apprenti, un expert, un groupe d'experts, une organisation),

– *Aspect fonctionnel* : cet aspect est relatif à la question suivante : « comment ces connaissances seront-elles utilisées ? ». Faut-il les représenter ou traduire sous une forme bien déterminée en vue d'une exploitation précise ? Toutes ces questions induisent des contraintes au niveau du processus d'ECD,

– *Aspect transformationnel* : cet aspect est relatif à l'évolutivité de l'environnement auquel les connaissances sont destinées. Des connaissances extraites de données et qui sont exploitables aujourd'hui, ne le seront pas forcément si l'environnement de leur exploitation change,

– *Aspect téléologique* : cet aspect est relatif à l'objectif global pour lequel les connaissances ont été extraites. Généralement cet objectif est décliné en sous-objectifs au niveau du processus ECD. On peut très bien imaginer qu'un objectif au niveau du contexte n'est pas complètement atteint par le processus ECD. Par exemple, construire des connaissances en accidentologie pour améliorer la sécurité routière ne peut pas être atteint uniquement par l'application de l'ECD sur une base de données d'accidents.

L'architecture que nous venons de présenter sert à présenter les différents points de vue dont il faut tenir compte quand il s'agit d'analyser la complexité de l'ECD. Cependant, nous pensons que cette complexité ne peut pas être analysée d'une façon disjonctive au niveau de ces différents points de vue. Nous ne pouvons pas parler, par exemple, de la complexité de données indépendamment du processus de leur traitement ainsi que des objectifs du traitement. Des données peuvent paraître complexes pour un objectif donné sans l'être pour un autre objectif. Nous ne pouvons pas non plus parler de la complexité d'un processus de traitement indépendamment des données et des objectifs de l'étude, etc. Nous proposons alors, dans la section suivante, la caractérisation de la complexité du processus d'ECD à travers des concepts qui font la conjonction de ces différents points de vue.

4.2 Caractérisation systémique de la complexité de l'ECD

Aucun des aspects systémiques des trois points de vue (i.e. données, tâches effectuées et contexte de l'étude) ne peut être considéré comme la source de la complexité du processus d'ECD s'il est considéré isolément des autres. Mais, c'est la conjonction des différents aspects au niveau de ces trois points de vue qui rend complexe ce processus. Il s'agit précisément des concepts de comportement circulaire tels que l'auto-application, le comportement projectif, les boucles de rétroaction. Dans cette section, nous définissons ces concepts qui sont des caractéristiques de la complexité du processus d'ECD.

4.2.1 Le concept d'auto-application

Le concept d'*auto-application* (*self-application*) est le plus général parmi les concepts de circularité. Son expression mathématique est donnée à travers l'équation suivante : $y = f(y)$. La forme discrète de circularité est exprimée à travers l'équation $y_{t+1} = f(y_t)$. La forme plus générale de ce principe est exprimée à travers la formule suivante : $y = kf(y)^2$. Pour expliquer ce principe, nous prenons l'exemple suivant : $y = \text{un écran TV}$ et $f = \text{une caméra pointée sur cet écran et en même temps transmettant l'image sur lui}$. L'image dans cette situation est cause et effet en même temps, donc $y = f(y)$.

Revenons à notre processus d'ECD que nous assimilons à un système composé par « les données », « l'opérateur humain » et « le contexte de l'étude ». Nous définissons la notion « d'état » de ce système de la manière suivante : à chaque instant t , ce système est défini par $\{\text{le résultat de la transformation des données, la perception de l'opérateur, l'objectif de l'étude}\}$. En définissant ainsi notre système, nous identifions le même *phénomène d'auto-application*. En effet, tout au long de ce processus, l'opérateur applique des tâches (e.g. nettoyer les données, appliquer une technique de data mining, etc.) ce qui génère des transformations et des résultats. Ces derniers ont une influence sur la perception de l'opérateur lui-même car en fonction de ces résultats l'opérateur essaye d'affiner son analyse en définissant de nouvelles tâches (e.g. réutiliser la même technique en changeant les paramètres, utiliser une autre technique, etc.). Il s'agit donc d'un *processus itératif* durant lequel les tâches appliquées dépendent des résultats intermédiaires qu'elles génèrent. Vis-à-vis d'un observateur externe, les tâches sont causes et effet en même temps. Si on pose $y = \text{« l'ensemble des tâches effectuées »}$ et $f = \text{« application d'une tâche »}$, on peut donc écrire $y = f(y)$, ce qui veut dire que le choix des tâches dépend des résultats intermédiaires.

² En algèbre linéaire k représente les valeurs propres de f .

La même chose peut être observée au niveau des données. En effet, les données influencent le choix des techniques appliquées par l'opérateur³ ce qui génère de nouvelles données dont la nature dépend des données initiales et des techniques de traitement. Ainsi, les données sont cause et effet en même temps. En posant $y = \text{« l'ensemble des données »}$ et $f = \text{« tâche de traitement appliquée »}$. Vis-à-vis d'un observateur externe, les données sont les résultats d'eux-mêmes, c'est-à-dire $y = f(y)$.

Le concept d'auto-application peut être généralisé au niveau de tout le système ECD en posant $y = \text{« système ECD »}$ et $f = \text{« l'ensemble des tâches effectuées »}$. Ce système est le résultat de lui-même, donc $y = f(y)$. L'auto-application peut être perçue comme un comportement du système ECD qui essaye de s'adapter à un objectif fixe (celui de l'étude), qui est dans notre cas *« extraire des connaissances des données pour un autre objectif »*. Ce comportement est appelé *un comportement projectif* que nous présentons dans la section suivante.

4.2.2 Le concept de comportement projectif

Une des principales caractéristiques d'un système complexe réside dans le fait qu'il a ses propres objectifs qu'il essaye de réaliser en résistant à toutes les perturbations. Un *comportement projectif* (*goal-directness*) implique une régulation ou un contrôle des perturbations guidé par cet objectif. Ce rôle de contrôle dans un système d'ECD est réalisé essentiellement par l'opérateur humain. En effet, ce dernier choisit ses actions en fonction de l'évolution du système (e.g. résultats intermédiaires), mais aussi en fonction de l'objectif qu'il s'est fixé pour l'étude qu'il effectue. Il essaye d'une façon continue à résister aux perturbations (e.g. bruits au niveau des données) pour atteindre l'objectif de l'étude. La question maintenant est : *pourquoi un comportement projectif est-il source de complexité ?*

- La première réponse est que ce comportement nécessite d'abord de décliner l'objectif de l'étude en sous-objectifs à réaliser au cours du processus. Cela nous amène à un autre concept caractéristique de la complexité : le principe de structure hiérarchique de contrôle (cf. 4.2.3) ;
- La deuxième est qu'il fait appel à des processus de régulation non-linéaires ce qui nous amène à un autre concept caractéristique de la complexité : les boucles de rétroactions (cf. 4.2.4).

4.2.3 Le concept de structure hiérarchique de contrôle

Dans un système complexe, les objectifs sont organisés en hiérarchie. Si une boucle de contrôle ne suffit pas pour réduire les effets d'une perturbation, il faut ajouter une autre (cf. Fig. 2). Un exemple typique dans les organisations est la tendance d'augmenter le nombre des niveaux bureaucratiques. Dans le cas du processus ECD, atteindre l'objectif d'une étude nécessite de préparer les données, appliquer une technique de data mining et interpréter les résultats. Chacune de ces trois tâches est ensuite déclinée en sous-objectifs, etc.

La Fig. 2 représente cette structure hiérarchique. Pour atteindre un objectif, l'opérateur effectue une perception pour faire une représentation interne⁴ des perturbations. Il effectue ensuite, moyennant un processus de traitement d'information, une confrontation de ces

³ Le type de données (texte, image, etc.) ainsi que leur nature (qualitative, quantitative) jouent un rôle dans le choix des techniques de traitement.

⁴ interne au système ECD.

Complexité de l'extraction des connaissances de données : une vision systémique

perceptions avec les objectifs du système ECD suite à quoi il décide des actions. Ces dernières, moyennant un processus dynamique de transformation, modifient l'effet des perturbations. Cette boucle est exécutée jusqu'à l'atteinte d'un résultat satisfaisant pour le régulateur (i.e. l'opérateur dans notre cas). Si cette boucle de contrôle ne suffit pas pour atteindre l'objectif, il faut en ajouter une autre. Cependant, plus le nombre de couches d'hierarchie de contrôle est important, plus les bruits sur les perceptions et les actions à entreprendre par l'opérateur sont importants. Il est donc préférable de maximiser la capacité de régulation d'une seule couche d'hierarchie et réduire le nombre de couches.

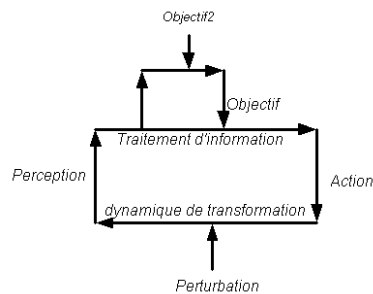


FIG. 2 - Une structure hiérarchique de contrôle

L'opérateur est confronté aux questions suivantes : Quel est le nombre de boucles optimal ? Comment traduire un objectif global d'une étude donnée en sous-objectifs en tenant compte des autres composantes du système (i.e. données et contexte) ? Comment gérer le conflit entre les sous-objectifs ? Les réponses à ces questions dépendent des données traitées, du contexte de l'étude et de l'opérateur. Ce dernier est amené sans cesse à effectuer des choix. Mais, dès qu'une décision est prise, des fonctions sont effectuées et des transformations sont induites, l'état du système change et de nouvelles décisions qui peuvent être contradictoires aux premières doivent être prises. Ainsi, une décision dépend finalement d'elle-même ce qui nous renvoie encore au principe d'auto-application. Ce comportement est assuré par ce qu'on appelle les boucles de rétroaction que nous présentons dans la section suivante.

4.2.4 Le concept de boucle de rétroaction

Le principe d'auto-application que nous avons présenté dans la section 4.2.1 et traduit par la formule $y = kf(y)$, peut être utilisé pour analyser la déviation d'un système par rapport à un état donné y_0 (e.g. un état d'équilibre). En d'autres termes la déviation $\Delta y = (y - y_0)$ à un instant $(t + \Delta t)$ dépend de la déviation à l'instant t . Cela est exprimé à travers la formule suivante : $\Delta y(t + \Delta t) = k\Delta y(t)$.

Si on pose « y_0 = l'état du système quand il atteint son objectif », le comportement projectif (cf. 4.2.2) paraît comme la tentative d'atteindre y_0 , donc de supprimer la déviation Δy . Ainsi, la déviation de l'état y du système par rapport à la situation objective y_0 dépend d'elle-même, c'est-à-dire : $(y - y_0)_{(t + \Delta t)} = k(y - y_0)_t$. Il s'agit de ce qu'on appelle le principe de rétroaction ou de feedback.

Dans le cadre du processus d'ECD, le principe de feedback exprime bien l'interdépendance entre les différentes tâches du processus, i.e. préparation de données, data mining et interprétation des résultats. En effet, selon l'interprétation des résultats de la fouille, on décide de refaire ou pas l'étape de data mining (e.g. rappliquer la technique de data mining en changeant des paramètres,

appliquer une nouvelle technique, etc.). On peut choisir aussi de revenir à l'étape de préparation des données (e.g. sélectionner d'autres données, refaire le nettoyage, etc.). Il s'agit bien donc, conformément à la définition du processus d'ECD donnée dans (Brachman and Anand 1996), d'un *processus itératif*.

Les feedbacks peuvent être négatifs⁵ et tendent vers la stabilisation du système quand une déviation positive (par rapport à y_0) génère une déviation négative. Prenons comme exemple y =nombre de variables à considérer dans une étude de fouille de données. Soit y_0 le nombre optimum. Si l'opérateur augmente y au-dessus du seuil y_0 , la connaissance extraite diminue⁶ ce qui conduit l'opérateur à réduire y . Cela va faciliter l'interprétation et augmenter la connaissance extraite. Si y continue à diminuer et passe au-dessous du seuil y_0 , la connaissance extraite diminue aussi ce qui va amener l'opérateur à augmenter y . Ainsi le système ECD, grâce à son aspect itératif, oscille autour de la position y_0 qui correspond en quelque sorte à un état d'équilibre.

Les feedbacks peuvent être aussi positifs⁷ et tendent dans ce cas vers l'amplification quand une déviation positive (par rapport à y_0) génère une déviation positive. Prenons l'exemple précédent, mais avec un autre mode de régulation se basant sur la règle suivante : « \forall l'état de y par rapport à y_0 , Si (connaissance diminue), Alors (augmenter le nombre de variables y) ». Dans ce cas, la boucle de rétroaction devient positive. Elle est traduite par : *Nombre de variables augmente \Rightarrow Connaissance diminue \Rightarrow Nombre de variables augmente \Rightarrow Connaissances diminue, etc.*

Ce sont essentiellement les feedbacks négatifs qui assurent la stabilité et la convergence d'un processus d'ECD. Le rôle de l'opérateur peut donc être perçu comme celui d'un régulateur utilisant ce type de feedbacks.

5 Conclusion & perspectives

Nous avons montré dans cet article que la complexité d'un processus d'ECD doit être analysée selon différents points de vue. En se basant sur une approche de modélisation de systèmes complexes (i.e. la *systémique*) nous avons défini une architecture d'analyse de la complexité de l'ECD. Cette architecture est composée de deux niveaux d'abstraction : le premier est composé des trois points de vue « *objet observé* », « *observateur* » et « *contexte de l'observation* » qui correspondent respectivement aux « *données* », « *l'opérateur humain* » et « *le contexte de l'étude* ». Le deuxième niveau d'abstraction est composé des quatre points de vue *ontologique*, *fonctionnel*, *transformationnel* et *téléologique*. Chacun des trois points de vue du premier niveau est analysé selon les quatre points de vue du deuxième niveau. Les sources de complexité du processus d'ECD ont été analysées selon cette architecture.

Nous avons montré ensuite qu'aucun des aspects systémiques des trois points de vue ne peut être considéré comme la source de la complexité du processus d'ECD s'il est considéré isolément des autres. Mais, c'est la conjonction des différents aspects au niveau de ces trois points de vue qui rend complexe ce processus. Nous avons défini alors des concepts faisant cette conjonction et permettant la caractérisation de la complexité du système ECD. Il s'agit des concepts d'*auto-application*, de *comportement projectif*, de *structure hiérarchique de contrôle* et de *boucles de rétroaction*.

⁵ Le feed-back est négatif quand k dans la formule $y=kf(y)$ est négatif.

⁶ Car le bruit et la difficulté d'interprétation augmentent.

⁷ Le feed-back est positif quand k dans la formule $y=kf(y)$ est positif.

Le travail dans ce papier a eu pour objectif de comprendre la complexité d'un système d'ECD en la caractérisant. Mais, cet objectif n'est pas une fin en soit car l'objectif final est de *contrôler cette complexité*. En effet, dans la réalité, le choix des actions dans un processus d'ECD n'est ni opéré complètement à l'aveuglette, ni de façon complètement déterministe. Il dépend d'un processus d'apprentissage issu d'expériences antérieures. Les feedbacks aident le système d'ECD à ajuster sa réponse. De nouvelles questions surgissent alors et méritent une réflexion profonde : Quels sont les différents modes de contrôle de la complexité du système d'ECD ? Quel est le rôle de l'apprentissage ? Quel est le rôle des connaissances du domaine ?

Références

- Bertalanffy, L. v. (1969). General system theory: foundations, development, applications. New York, George Braziller.
- Booch, G., R. J., et al. (1999). Unified Modelling Language User Guide, Addison Wesley Professional.
- Brachman, R. and T. Anand (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Advances in Knowledge Discovery and Data Mining, 37-58, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, Calif.
- Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications Of The ACM **39**(11): 27-34.
- Le Moigne, J.-L. (1999). La modélisation des systèmes complexes, Dunod.
- Miller, J. G. (1995). Living Systems, University Press of Colorado.
- Morin, E. and J. L. Le Moigne (1999). L'intelligence de la complexité. Paris, L'Harmattan.
- Richard, H. (2003). Approaches to Inquiry - Key Concepts, ICRA Learning Materials - Methods.
- Von Foerster, H. (1995). The Cybernetics of Cybernetics (2nd edition). Minneapolis, Future Systems Inc.

Summary

Patricians and researchers in knowledge Discovery in Databases (KDD) have to address many issues related to the data nature, the implication of human operator and algorithmic aspects. Nowadays, there is a consensus that the KDD process is complex. However, there is discordance when questions deal with the definition and characterization of this complexity. To answer these questions, we use the systemic approach, which is a complex system modeling approach.

Une représentation des arborescences pour la recherche de sous-structures fréquentes

Federico Del Razo Lopez, Anne Laurent, Maguelonne Teisseire

LIRMM - Université Montpellier II
161 rue Ada 34392 Montpellier cedex 5
{delrazo,laurent,teisseire}@lirmm.fr

Résumé. La recherche de structures fréquentes au sein de données arborescentes est une problématique actuellement très active qui trouve de nombreux intérêts dans le contexte de la fouille de données comme, par exemple, la construction automatique d'un schéma médiateur à partir de schémas XML. Dans ce contexte, de nombreuses propositions ont été réalisées mais les méthodes de représentation des arborescences sont très souvent trop coûteuses. Dans cet article, nous proposons donc une méthode originale de représentation de ces données. Les propriétés de cette représentation peuvent être avantageusement utilisées par les algorithmes de recherche de structures fréquentes (sous-arbres fréquents). La représentation proposée et les algorithmes associés ont été évalués sur des jeux de données synthétiques montrant ainsi l'intérêt de l'approche proposée.

1 Introduction

L'explosion du volume de données disponible sur internet conduit aujourd'hui à réfléchir sur les moyens d'interroger les grosses masses d'information afin de retrouver les informations souhaitées. Les utilisateurs ne pouvant pas connaître les modèles sous-jacents des données qu'ils souhaitent accéder, il est donc nécessaire de leur fournir les outils automatiques de définition de schémas médiateurs. Un schéma médiateur fournit une interface permettant l'interrogation des sources de données par l'utilisateur au travers de requêtes. L'utilisateur pose alors ses requêtes de manière transparente à l'hétérogénéité et la répartition des données.

XML étant maintenant prépondérant sur internet, la recherche de moyens d'intégration de tels schémas est indispensable. Si les recherches permettant l'accès aux données quand un schéma d'interrogation est connu sont maintenant bien avancées (Xyleme, 2001), les recherches concernant la définition automatique d'un schéma médiateur restent incomplètes et sont donc non satisfaisantes (Tranier et al., 2004). Dans le but de proposer une approche permettant de répondre à cette dernière problématique, nous nous focalisons sur la recherche de sous-structures fréquentes au sein d'une base de données de schémas XML. Une sous-structure fréquente est un sous-arbre se trouvant dans *la plupart* des schémas XML considérés. Cette proportion est examinée au sens d'un *support* qui correspond à un nombre minimal d'arbres de la base dans lesquels doit se retrouver le sous-arbre pour être considéré comme *fréquent*. Une telle recherche

est complexe dans la mesure où il est nécessaire de traduire l'ensemble des schémas en une structure aisément manipulable. Cette transformation des données conduit parfois à doubler ou tripler la taille de la base initiale dès lors que l'on souhaite utiliser des propriétés spécifiques permettant d'améliorer le processus de fouille. Il n'existe pas de solution efficace à ce problème alliant une représentation compacte à des propriétés intéressantes. L'objet de cet article est la définition d'une nouvelle structure répondant à cet objectif puisque les propriétés de la représentation proposée peuvent permettre une génération des candidats et un élagage aussi performants que les approches de référence (Asai et al., 2002, Zaki, 2002, Termier et al., 2002).

La section 3 présente le cœur de notre proposition définissant une nouvelle méthode de représentation des données arborescentes ainsi qu'un aperçu de la méthode mise en œuvre pour la génération et l'élagage des candidats. La section 4 présente les résultats des premières expérimentations menées. Enfin, la section 5 conclut et présente les principales perspectives associées à nos travaux.

2 Travaux connexes

2.1 Définitions préliminaires

Les travaux les plus significatifs concernant l'extraction de sous-arbres fréquents se trouvent dans (Kuramochi et Karypis, 2001, Asai et al., 2002, Yan et Han, 2002, Zaki, 2002, Termier et al., 2002). Les définitions suivantes sont inspirées de ces travaux.

Un *arbre* est un graphe connexe sans cycle. Il est composé d'un ensemble de nœuds reliés par des arcs tels qu'il existe un nœud particulier nommé *racine* et tel que tous les autres nœuds hormis la racine sont composés d'un ensemble de sous-arbres. On parle d'*arbre ordonné* si l'ordre des sous-arbres importe, et d'*arbre non ordonné* sinon.

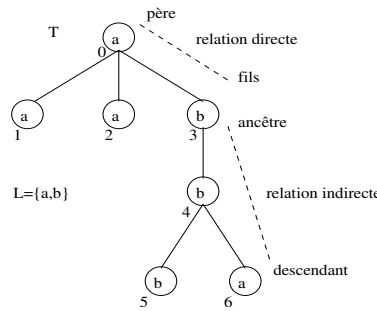


FIG. 1 – *Arbre*

Soit un alphabet d'étiquettes $\mathcal{L} = \{a, b, c, \dots\}$, on considère un *arbre étiqueté et ordonné* noté $T = \{r, N, B, L, F\}$ où : r est la racine de l'arbre, N est l'ensemble des nœuds, B est l'ensemble des arêtes tel que $B \subseteq V^2$, $(L : N \rightarrow \mathcal{L})$ est une fonction qui associe une étiquette aux nœuds dans N et F est une relation d'ordre de droite à

gauche entre frères.

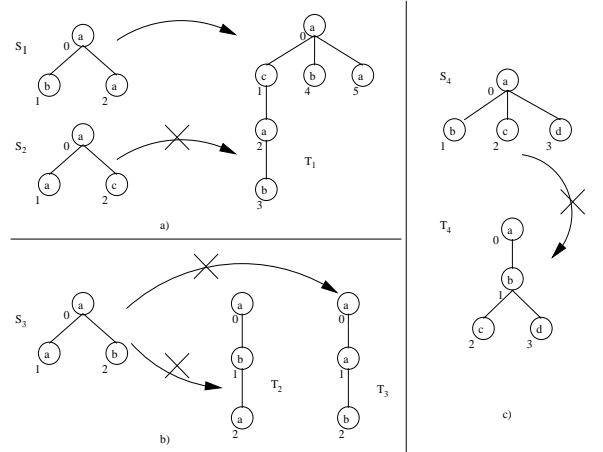


FIG. 2 – *Inclusion et non inclusion de S dans T*

Plusieurs nœuds pourront avoir la même étiquette, ce qui introduit un problème de polysémie (une même étiquette renvoie à plusieurs nœuds différents). La *taille* de T , notée $|T|$, est le nombre de nœuds dans T . l'*ordre* de T correspond au nombre de branches dans T .

Chaque nœud n est associé à un numéro unique i , correspondant à sa position par un parcours en profondeur d'abord. n_i fait alors référence au i -ème nœud en utilisant un schéma de numération ($i = 0 \dots |T| - 1$).

Pour chaque couple $(x, y) \in N \times N$, il peut exister une relation *directe* entre x et y si x est le père de y ou bien une relation *indirecte* si x est un ancêtre de y (il faut alors suivre plusieurs arcs successifs pour aller de x à y). La figure 1 montre une relation directe (père-fils) entre les nœuds 0 et 3 de T , et une relation indirecte (ancêtre-descendant) entre les nœuds 3 et 6.

On note $S \preceq T$ le fait que le sous-arbre S est *inclus* dans l'arbre étiqueté ordonné T (voir figure 2). Dans le cadre d'une relation indirecte *ancêtre-descendant*, $S \preceq T$ si et seulement si les suivantes conditions sont satisfaites :

1. $N_S \subset N_T$
2. Pour toutes les branches $b_S = (x, y) \in B_S$, x est ancêtre de y dans T
3. Pour toutes les branches $b_T = (x, y) \in B_T$, x est ancêtre de y dans S
4. Pour toutes les branches $b_{1S} = (x, y_1) \in B_S$ et $b_{2S} = (x, y_2) \in B_S$ telles que $y_1 \prec_F y_2$, $y_1 \prec_F y_2$ dans T

2.2 Les représentations existantes

L'algorithme *TREEMINER* (Zaki, 2002) propose une méthode d'extraction de sous-arbres fréquents. De même que dans notre approche, ces travaux reposent sur une

Structure de représentation des arborescences

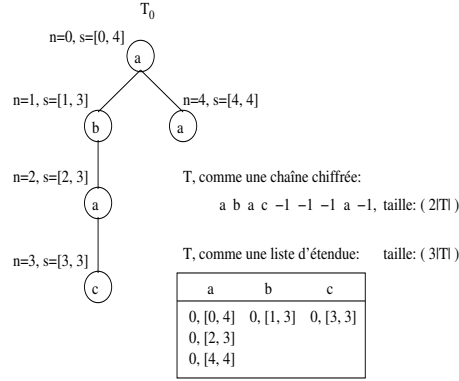


FIG. 3 – Représentation proposée dans (Zaki, 2002)

représentation originale des arbres facilitant la gestion des candidats. Néanmoins, la représentation verticale adoptée aboutit à stocker trois fois la taille d'un arbre comme l'indique la figure 3.

L'approche proposée dans (Asai et al., 2002) est dédiée aux arbres ordonnés et adopte une structure permettant des performances très intéressantes. Mais cette représentation, illustrée figure 4, conduit également à tripler la taille de la base afin de stocker les informations nécessaires.

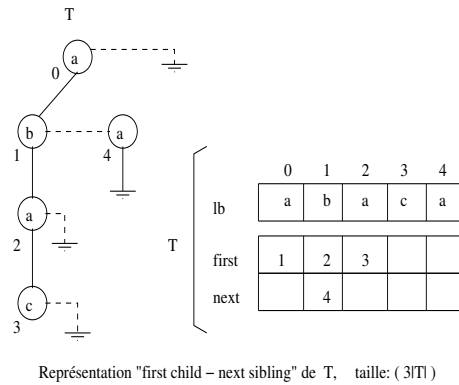


FIG. 4 – Représentation proposée dans (Asai et al., 2002)

Des proposition récentes se basent sur une représentation efficace des arbres mais néanmoins, elles n'utilisent pas des propriétés aussi intéressantes que les travaux précédents afin d'améliorer les traitement des structures candidates. Nous pouvons citer (Wang et al., 2004) dont la structure est illustrée figure 5 et (Chi et al., 2004), (Chi et al., 2003) proposant une représentation des arbres illustrée figure 6.

L'objectif est alors de proposer à la fois une représentation peu coûteuse de la base

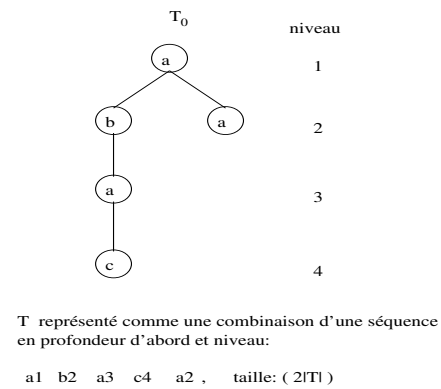


FIG. 5 – Représentation proposée dans (Wang et al., 2004)

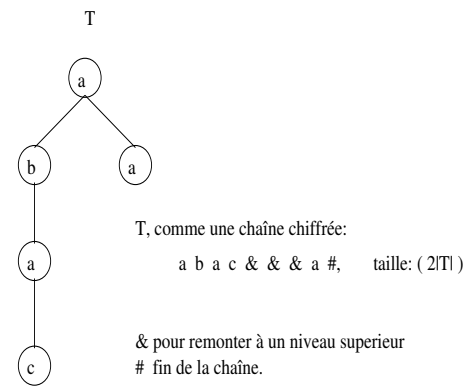


FIG. 6 – Représentation proposée dans (Chi et al., 2004, Chi et al., 2003)

et des propriétés intéressantes pour améliorer le processus de fouille de données. C'est dans ce contexte que se situe notre proposition décrite au paragraphe suivant.

3 Proposition

Dans cet article, nous proposons une nouvelle approche permettant l'extraction efficace de sous-arbres fréquents au sein d'une base de données arborescentes. Notre proposition s'appuie sur une méthode de représentation des arbres originale qui permet de générer de manière efficace les sous-arbres candidats puis d'élaguer les sous-arbres non fréquents (après calcul du support).

3.1 Représentation des arbres

Pour la représentation d'un arbre T , nous profitons de la propriété suivante : chaque nœud dans l'arbre possède un seul parent. Nous proposons d'utiliser deux vecteurs pour représenter un arbre comme indiqué dans (Weiss, 1998). Le premier, nommé st , conserve la position du père de chaque nœud. Les nœuds de l'arbre sont numérotés en profondeur d'abord, la racine de T correspondant à l'index 0 et ayant une valeur $st[0] = -1$ pour indiquer que la racine n'a pas de père. Les valeurs $st[i]$, $i = 0, 1, \dots, k-1$ correspondent alors aux positions du père des nœuds i , comme illustré sur la figure 7.

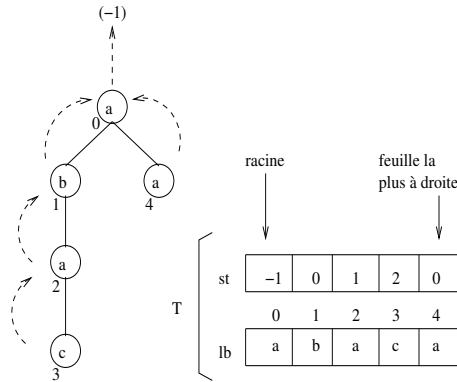


FIG. 7 – Représentation d'un arbre

Cette représentation permet de retrouver en temps constant le père d'un nœud. De plus, elle permet la localisation directe de la *feuille la plus à droite* par rapport à l'index k . En parcourant l'arbre, on peut bâtir toutes les relations directes *père-fils* entre nœuds.

Le deuxième vecteur, nommé lb , est utilisé pour enregistrer les étiquettes de l'arbre avec $lb[i]$, $i = 0, 2, \dots, k-1$ représentant l'étiquette de chaque nœud $n_i \in T$.

La structure adoptée permet une représentation des arbres peu coûteuse puisqu'elle se réduit à $2|T|$. De plus elle possède des propriétés intéressantes, évoquées au paragraphe suivant, pouvant être utilisées lors de la recherche de sous structures fréquentes.

3.2 Génération et élagage des candidats

Les candidats de taille 1 sont obtenus en parcourant tous les nœuds des arbres de la base de données. Chaque nœud voit son support incrémenté lors de ce parcours et seuls sont conservés les nœuds dont le support final est supérieur au support minimal défini par l'utilisateur. La base de données est alors transformée pour ne conserver que les nœuds fréquents, comme illustré par la figure 8.

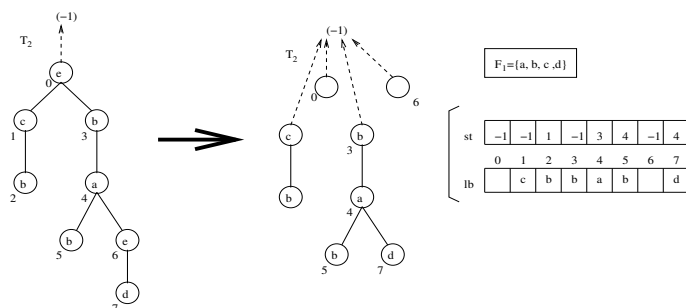


FIG. 8 – Transformation de la base de données après génération de F_1

Les candidats de taille 2 sont générés en combinant deux à deux tous les fréquents de taille 1.

La base de données est alors mise à jour, en modifiant les arbres de la racine, des sommets et des feuilles afin de ne conserver que les liens entre les nœuds fréquents. Les figures 9, 10 et 11 illustrent ce processus, en considérant $\sigma = 7$ et $F_2 = \{a - d, a - b\}$.

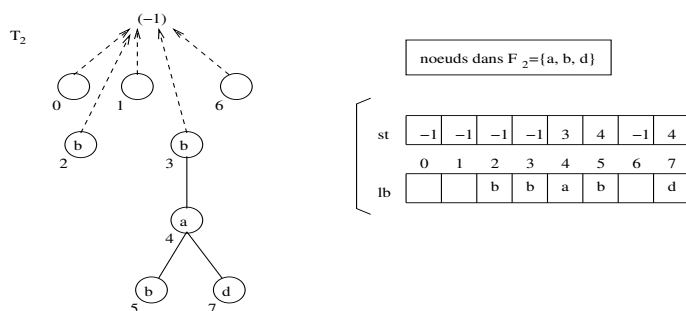


FIG. 9 – Transformation de la base - racine

La génération des candidats de taille $k \geq 3$ s'effectue de la même manière que dans les approches classiques de type Apriori (Agrawal et Srikant, 2002), par combinaison

Structure de représentation des arborescences

des fréquents de taille $k - 1$.

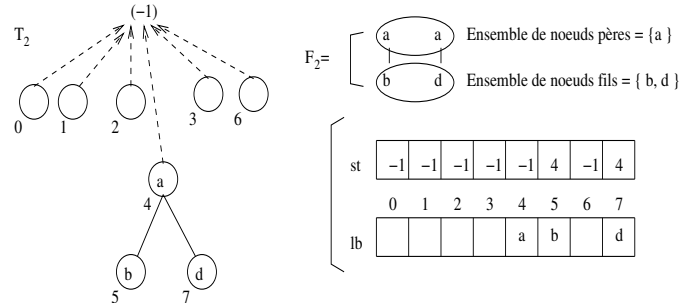


FIG. 10 – Transformation de la base - sommets

L'originalité de notre approche réside dans l'utilisation de notre représentation pour l'élagage des candidats non fréquents. Le calcul du support de chaque candidat consiste à compter le nombre d'arbres de la base qui contiennent ce sous-arbre candidat. Pour ce faire, pour chaque arbre de la base, il s'agit de chercher les *points d'ancrage* sur lesquels la racine du sous-arbre à tester peut s'instancier. Pour chaque point d'ancrage trouvé, on cherche alors à instancier l'ensemble des nœuds de l'arbre candidat au sein de l'arbre courant testé. On note que nous cherchons une instanciation *exacte* du candidat au sein des arbres de la base. Si tous les nœuds du candidat ont été trouvés, l'arbre supporte le candidat. Le support de la structure candidate est alors incrémenté. Ce qui n'est pas le cas si tous les nœuds de l'arbre ont été parcourus sans trouver l'ensemble des nœuds du candidat.

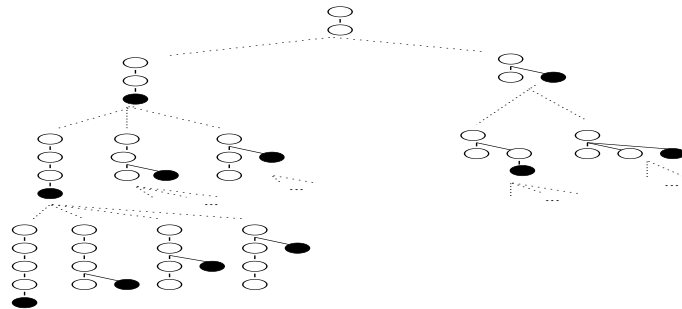


FIG. 11 – Transformation de la base - feuilles

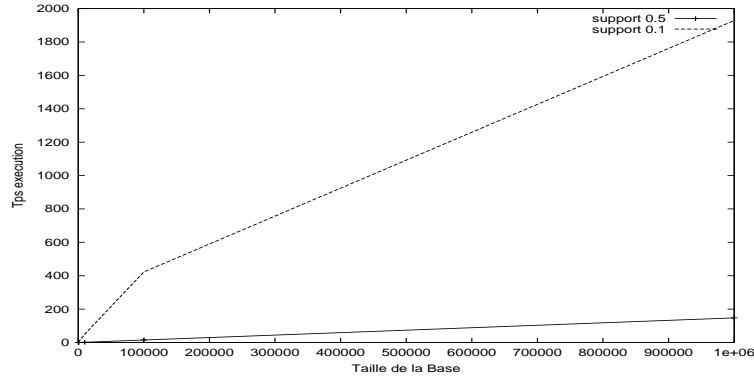


FIG. 12 – Résultats expérimentaux : temps de calcul

4 Expérimentations

Des expérimentations sont menées sur les données issues du générateur de schémas XML développé par Alexandre Termier (Termier et al., 2002). Les résultats, illustrés figure 12, soulignent l'intérêt de notre approche, concernant les temps de calcul.

5 Conclusion et perspectives

Dans cet article, nous proposons une approche originale de représentation de données arborescentes dont les propriétés permettent une extraction efficace de sous-arbres fréquents. Les premières expérimentations réalisées sur des données synthétiques laissent envisager des résultats très prometteurs par rapport aux approches de référence. Notre objectif est donc de poursuivre dans cette voie et d'optimiser les différents algorithmes associés.

Ces travaux seront également utilisés dans le cadre de la médiation de données, les sous-arbres fréquents extraits servant de support à la construction automatique d'un schéma médiateur. Une telle solution peut également être adoptée dans le cadre de la fouille de données en ligne (data streams) pour le traitement à la volée de données XML. Cette perspective permettra de traiter les gros volumes de données transitant sur internet de manière efficace et rapide.

Références

- Agrawal, R. and Srikant, R. (2002). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB Conference*, Santiago, Chile.
- Asai, T., Abe, K., Kawasoe, S., Arimura, H., and Sakamoto, H. (2002). Efficient substructure discovery from large semi-structure data. In *2nd Annual SIAM Symposium on Data Mining, SDM2002*, Arlington, VA, USA. Springer-Verlag.

Chi, Y., Yang, Y., and Muntz, R. R. (2003). Indexing and mining free trees. In *International Conference on Data Mining 2003 (ICDM2003)*.

Chi, Y., Yang, Y., and Muntz, R. R. (2004). Cmtreeminer : Mining both closed and maximal frequent subtrees. In *The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)*.

Kuramochi, M. and Karypis, G. (2001). Frequent subgraph discovery. In *IEEE International Conference on Data Mining (ICDM)*.

Termier, A., Rousset, M.-C., and Sebag, M. (2002). Treefinder, a first step towards XML data mining. In *IEEE Conference on Data Mining (ICDM)*, pages 450–457.

Tranier, J., Baraer, R., Bellahsene, Z., and Teisseire, M. (July, 7th - 9th 2004). Where's Charlie : Family based heuristics for peer-to-peer schema integration. In *Proceedings of the 8th International Database Engineering and Applications Symposium (IDEAS '04)*, Coimbra, Portugal.

Wang, C., Yuan, Q., Zhou, H., Wang, W., and Shi, B. (May 2004). Chopper : An efficient algorithm for tree mining. *Journal of Computer Science and Technology*, 19 :309–319.

Weiss, M. A. (1998). *Data Structures And Algorithm Analysis In C*.

Xyleme, L. (2001). A dynamic warehouse for xml data of the web. In *IEEE Data Engineering Bulletin*.

Yan, X. and Han, J. (2002). gspan : Graph-based substructure pattern mining. In *IEEE Conference on Data Mining (ICDM)*.

Zaki, M. (2002). Efficiently mining frequent trees in a forest. In *ACM-SIGKDD'02*.

Summary

Mining frequent subtrees from tree databases is currently a very active research topic. This research has many interests, including for instance mediator schema building from XML schemas. In this framework, many works have been proposed. However, the way the data are represented is often very memory-consuming. In this paper, we propose a new method to represent such data. We show that this representation has many properties, which can be used in order to enhance subtree mining algorithms. Experiments are led to assess our proposition (data representation and its associated algorithms).

Classement d'objets incomplets dans un arbre de décision probabiliste

Lamis Hawarah, Ana Simonet, Michel Simonet*

*TIMC-IMAG

Institut d'Ingénierie et de l'information de Santé

Faculté de Médecine – IN3S

38700 LA TRONCHE

{[Lamis.Hawarah](mailto:Lamis.Hawarah@imag.fr), [Ana.Simonet](mailto:Ana.Simonet@imag.fr), [Michel.Simonet](mailto:Michel.Simonet@imag.fr)}@imag.fr

<http://www-timc.imag.fr>

Résumé. Nous présentons une approche probabiliste pour déterminer les valeurs manquantes des objets incomplets pendant leur classement dans les arbres de décision. Cette approche est dérivée de la méthode d'apprentissage supervisé appelée Arbres d'Attributs Ordonnés proposée par Lobo et Numao en 2000, qui construit un arbre de décision pour chacun des attributs, selon un ordre croissant en fonction de l'information mutuelle entre chaque attribut et la classe. Notre approche étend la méthode de Lobo et Numao d'une part en prenant en compte les dépendances entre les attributs pour la construction des arbres d'attributs, et d'autre part en fournissant un résultat de classement d'un objet incomplet sous la forme d'une distribution de probabilités (au lieu de la classe la plus probable).

1 Introduction

Le problème des valeurs manquantes est un problème connu dans le domaine de la fouille de données, où, dans la base d'apprentissage, on rencontre des objets ayant des valeurs manquantes pour certains attributs. Nous étudions ce problème dans le cadre des arbres de décision. Un arbre de décision est construit à partir d'un ensemble d'apprentissage selon l'approche divide-and-conquer (Quinlan 1993). Une fois l'arbre construit, il est utilisé pour classer de nouveaux objets. Pour cela on parcourt l'arbre en commençant par la racine et en suivant les branches correspondant aux valeurs de l'objet, jusqu'à une feuille. La classe associée à cette feuille est la classe de cet objet. Les arbres de décision sont confrontés au problème des données manquantes, à la fois lors de leur construction et lors du classement d'objets. Lors de la construction, l'existence de valeurs manquantes pose problème pour le calcul de gain d'information, nécessaire au choix de l'attribut test, ainsi que pour la partition de l'ensemble d'apprentissage selon l'attribut test choisi. Le classement d'un objet avec des valeurs manquantes soulève également des problèmes lorsqu'un nœud correspondant à un attribut manquant est rencontré dans le parcours de l'arbre. Dans ce travail nous nous intéressons exclusivement au second problème, c'est-à-dire le classement d'objets incomplets.

Les méthodes qui traitent les valeurs manquantes dans les arbres de décision, remplacent un attribut manquant par une seule valeur, qui peut être la valeur la plus probable (Kononenko et al. 1984) ou la plus similaire (Breiman et al. 1984), etc. Ce type d'approche présente l'inconvénient d'oublier les autres valeurs possibles. Notre approche vise à une détermination probabiliste des valeurs manquantes, en prenant en compte les dépendances entre l'attribut manquant et les autres attributs de l'objet, ce qui permet d'utiliser le maximum de l'information contenue dans l'objet pour le calcul des valeurs manquantes. De plus, nous voulons un résultat sous la forme d'une distribution de probabilités plutôt que la valeur la plus probable, ce qui donne une information plus fine. Parce que les arbres de décision sont

capables de déterminer la classe d'une instance à partir des valeurs de ses attributs, on peut les utiliser pour déterminer les valeurs d'un attribut inconnu (qui joue alors le rôle de la classe) à partir des attributs dont il dépend. Dans notre travail, nous nous sommes intéressés aux méthodes qui utilisent les arbres de décision pour trouver l'attribut manquant, et en particulier à la méthode des Arbres d'Attributs Ordonnés (Lobo 1999) et (Lobo 2000).

Dans cet article, nous rappelons les principales méthodes qui traitent le problème des valeurs manquantes dans un arbre de décision, et nous détaillons la méthode des Arbres d'Attributs Ordonnés, qui sert de base à notre approche. Nous présentons ensuite notre extension à cette approche (Hawarah et al. 2004). Enfin, nous montrons comment, pour chaque attribut manquant, nous calculons une distribution de probabilités afin d'obtenir un résultat de classement probabiliste.

2 Etat de l'art

Plusieurs méthodes ont été proposées pour traiter le problème des valeurs manquantes dans un arbre de décision (White et al. 1997) et (Quinlan 1989) lors de la phase de construction de l'arbre, comme la méthode de majorité (Kononenko et al. 1984) et la méthode de Shapiro, décrite par (Quinlan 1986). L'utilisation de l'arbre pour classer un objet avec des valeurs manquantes a aussi fait l'objet de quelques études, comme l'approche probabiliste de C4.5 (Quinlan 1993), la méthode *Lazy decision tree* (Friedman et al. 1996), et la méthode *surrogate splits* proposée par (Breiman et al. 1984), qui consiste à utiliser un autre attribut, appelé l'attribut de substitution, pour décider quelle branche (gauche ou droite) choisir pour continuer le classement. En général, lorsqu'un attribut manquant est envoyé dans un sous-arbre d'un nœud en suivant une branche déterminée par l'attribut de substitution, cela revient à compléter cet attribut manquant par la modalité¹ qui étiquette la branche choisie. Nous nous sommes intéressés à une méthode particulière, les *Arbres d'Attributs Ordonnés* (Lobo 1999) et (Lobo 2000), que nous expliquons en détail dans la section suivante.

2.1 Les Arbres d'Attributs Ordonnés

Les Arbres d'Attributs Ordonnés (AAO) sont une méthode d'apprentissage supervisé proposée par Lobo et Numao pour traiter le problème des valeurs manquantes, à la fois dans les phases de construction et de classement (Lobo 1999) et (Lobo 2000). L'idée générale de cette méthode est de construire un arbre de décision, appelé arbre d'attribut, pour chaque attribut dans la base en utilisant un sous-ensemble d'apprentissage contenant les instances ayant des valeurs connues pour cet attribut. Pour un attribut donné, son arbre d'attribut est un arbre de décision dont les feuilles représentent les valeurs de cet attribut. Ces arbres sont construits selon un ordre de construction croissant en fonction de l'Information Mutuelle (IM)² entre chaque attribut et la classe (Shannon 1949). L'arbre d'attribut est utilisé pour

¹ Les arbres produits par la méthode CART sont des arbres binaires, où tous les tests étiquetant les nœuds de décision sont binaires. Le nombre de tests à explorer va dépendre de la nature des attributs. A un attribut binaire correspond un test binaire. A un attribut qualitatif ayant n modalités, on peut associer autant de tests qu'il y a de partitions en deux classes, soit 2^{n-1} tests binaires possibles. Enfin, dans le cas d'attributs continus, il y a une infinité de tests envisageables. Dans ce cas, on découpe l'ensemble des valeurs possibles en segments.

² L'Information Mutuelle mesure la force de la relation entre deux attributs ou entre un attribut et la classe. L'IM entre deux attributs catégoriels X et Y est définie comme suit:

$$IM(X,Y) = - \sum_{x \in D_X} P(x) \log_2 P(x) + \sum_{y \in D_Y} P(y) [\sum_{x \in D_X} P(x|y) \log_2 P(x|y)]$$

déterminer la valeur de l'attribut pour des instances où elle est inconnue. Il est utilisé dans deux cas distincts : 1) lors de la construction de l'arbre de décision, pour déterminer la valeur de l'attribut pour les instances de la base d'apprentissage où cet attribut est inconnu; 2) lors du classement d'instances incomplètes, pour déterminer la valeur de l'attribut lorsque celle-ci est manquante.

Après avoir calculé l'IM entre chaque attribut et la classe, les attributs sont ordonnés par ordre croissant d'IM. Le premier arbre d'attribut construit est celui qui correspond à l'attribut ayant l'IM minimale. Il est représenté par un seul nœud-feuille avec sa valeur la plus probable dans la base d'apprentissage. Pour les autres attributs, on fournit, à partir de l'ensemble d'apprentissage initial, le sous-ensemble d'apprentissage qui contient les instances ayant des valeurs connues pour cet attribut. Ces instances sont décrites seulement par les attributs qui ont déjà été traités (c'est à dire les attributs pour lesquels on a déjà construit les arbres d'attributs et déterminé leurs valeurs manquantes dans la base d'apprentissage). L'algorithme utilisé pour la construction est un algorithme standard de construction d'un arbre de décision. Lors d'un classement, les valeurs des attributs inconnus de l'objet sont calculées successivement, par ordre d'IM croissante. Nous présentons dans la Fig.1 la méthode AAO en utilisant un exemple pris de (Quinlan 1993); les attributs sont ordonnés par ordre croissant en fonction de l'IM : *Température*, *Vent*, *Humidité*, *Temps*. Les arbres sont construits en utilisant l'algorithme ID3 (Quinlan 1986) et le logiciel Weka³. Le nombre de cas sur chaque nœud est indiqué entre parenthèses.

Dans cet exemple, on suppose qu'il n'y a pas de valeurs manquantes dans la base d'apprentissage initial, mais les arbres sont construits à partir d'une base d'apprentissage complète et ils sont utilisés seulement pendant le classement d'objet ayant des valeurs manquantes.

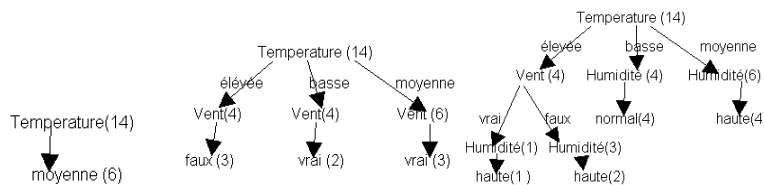


FIG. 1 - Arbres d'Attributs Ordonnés pour *Température*, *Vent*, *Humidité*.

Le premier arbre construit selon de la méthode AAO est donc l'arbre de *Température* ; il est composé d'un seul nœud, ayant pour valeur *moyenne*, qui est la valeur la plus probable. Selon l'ordre de construction imposé, l'arbre de *Vent* est construit en utilisant seulement l'attribut *Température*. Les arbres pour les attributs *Humidité* et *Temps* sont ensuite construits dans cet ordre⁴. Nous rappelons que le fait d'associer la valeur la plus probable à une feuille élimine les autres valeurs possibles. Par exemple, cette méthode remplace l'attribut *Température* par la valeur *moyenne* dans tout objet dont l'attribut *Température* est inconnu. D'autre part, pour l'arbre de *Vent* et pour la valeur *basse* de l'attribut *Température*, la base d'apprentissage dispose de deux cas où *Vent* est *faux* et deux cas où *Vent* est *vrai*. Dans cette

D_x et D_y sont les domaines des attributs catégoriels X et Y . $P(x)$ et $P(y)$ sont les probabilités de $x \in D_x$ et $y \in D_y$, respectivement. $P(x|y)$ est la probabilité conditionnelle que X prenne la valeur x sachant que Y est connu et prend la valeur y .

³ www.cs.waikato.ac.nz/ml/weka/index.html

⁴ Dans cet exemple simple destiné à illustrer le fonctionnement de la méthode AAO et les extensions proposées dans la suite de l'article, on n'a pas présenté tous les arbres construits. Par exemple, l'arbre de *Temps* est construit en utilisant les attributs *Température*, *Vent* et *Humidité*, mais il n'est pas présenté ici.

situation, ID3 a choisi arbitrairement la valeur *vrai*. Enfin, nous rappelons que l'ordre imposé par cette méthode ne garantit pas que l'arbre d'un attribut soit construit à partir des attributs dont il dépend. L'étude qui a été faite par Lobo et Numao (Lobo 2001) a montré que les relations entre attributs d'une base d'apprentissage doivent vérifier certaines conditions⁵ pour que la méthode AAO soit applicable.

Pour traiter le problème de valeurs inconnues de manière probabiliste en prenant en compte les dépendances entre les attributs, nous faisons deux propositions : les Arbres d'Attributs Ordonnés Probabilistes (AAOP), qui étendent les AAO par la prise en compte de la distribution des fréquences des classes dans les feuilles, et les Arbres d'Attributs Probabilistes (AAP), qui sont construits en utilisant les dépendances entre attributs.

2.2 Vers une approche probabiliste

Un arbre de décision idéal est un arbre où toutes les instances arrivant à une feuille appartiennent à la même classe. Cette catégorie d'arbres, rarement rencontrée dans le monde réel, est représentée par un arbre semblable à celui de la Fig.2. Plus généralement, un arbre de décision se présente sous une forme où les instances d'une feuille appartiennent à plusieurs classes. Dans ce cas, classiquement on associe à chaque feuille la classe la plus probable. (Breiman et al. 1984) ont proposé de construire des *Class Probability Trees* où on associe à chaque feuille F la probabilité de chaque classe J sur F; $P(J|F)$ avec $J=1,...,n$. Ainsi, dans le cas du diagnostic en médecine, pour un patient qui pourrait avoir trois maladies m_1, m_2, m_3 , il serait préférable d'estimer les probabilités relatives d'avoir m_1, m_2, m_3 au lieu de lui affecter une seule maladie, et ceci même s'il n'y a pas d'information manquante.

Quinlan a également proposé d'utiliser les probabilités pour traiter le problème des valeurs manquantes dans les phases de construction et de classement (Quinlan 1986), (Quinlan 1990) et (Quinlan 1993). Son approche consiste à associer un poids à chaque valeur d'un attribut. Pour un attribut connu, le poids est 1 pour la valeur de l'attribut et 0 pour toutes ses autres valeurs. Pour un attribut inconnu, le poids associé à chacune de ses valeurs est sa fréquence dans le sous-ensemble d'apprentissage correspondant au nœud de cet attribut. Dans ce cas, lors de la construction de l'arbre de décision, Quinlan associe à chaque feuille F la classe la plus probable, notée C_j . Cependant, il conserve le nombre total de cas arrivant à F ainsi que des couples (C_i, nb_i) où nb_i est le nombre de cas appartenant à la classe C_i ($C_j \neq C_i$) qui arrivent à F. Ces informations lui permettent, lors d'un classement d'un objet avec des valeurs manquantes, de calculer la probabilité de chaque classe.

Selon (Quinlan 1990), pour un seul chemin (une seule règle de classement) de la racine de l'arbre jusqu'à une feuille F passant par les branches $B_1, B_2, ..., B_L$, où chaque branche correspond à une valeur d'attribut test (le résultat d'un nœud), la probabilité qu'un objet E arrive à une feuille F (c'est à dire qu'il passe par les branches $B_1, B_2, ..., B_L$) est :

$$P_E(F) = P_E(B_1, B_2, ..., B_L) = P_E(B_1) * P_E(B_2|B_1) * P_E(B_3|B_1, B_2) * ... * P_E(B_L|B_1, B_2, ..., B_{L-1})$$

Si la valeur de chaque attribut est connue, chacune des probabilités précédentes est 0 ou 1. Si la valeur de l'attribut correspondant à la branche B_i n'est pas connue, sa probabilité est calculée par $P(B_i|B_1, B_2, ..., B_{i-1})$ à partir de l'ensemble d'apprentissage initial, i.e., la proportion des cas (instances) arrivés au $i^{ième}$ test qui prennent la branche B_i .

Parce qu'un cas E avec des valeurs manquantes peut appartenir à plusieurs feuilles, la probabilité que le cas E appartienne à une classe C est : $\sum_C P_E(F) P(C|F)$

⁵ Les attributs qui ont des relations faibles avec la classe devraient avoir des relations fortes avec le reste des attributs dans la base. En général, une base d'apprentissage dont les corrélations entre les attributs sont fortes est plus favorable pour l'application de cette méthode.

Par exemple, si on veut classer une instance où *Temps* est *enseleillé* et *Humidité* est inconnue, la probabilité que cette instance appartienne à la classe A est (Fig.2) :

$$P(A) = P(A \setminus \text{enseleillé, haute}) * P(\text{enseleillé}) * P(\text{haute} \setminus \text{enseleillé}) \\ + P(A \setminus \text{Pluvieux, vrai}) * P(\text{Pluvieux, vrai}) = 1 * 1/3 * 5 = 0.6.$$

La partie $P(A \setminus \text{Pluvieux, vrai}) * P(\text{Pluvieux, vrai})$ est égale à 0 car le *Temps* est *enseleillé* alors $P(\text{Pluvieux}) = 0$.

Remarquons que la probabilité de *haute* est calculée sachant seulement *enseleillé*, car le *Temps* est le nœud-père de *Humidité*. Ainsi, le fait que *Humidité* dépend de *Température* n'est pas pris en compte. En prenant en compte la corrélation qui existe entre *Humidité* et *Température*, le résultat sera vraisemblablement meilleur.

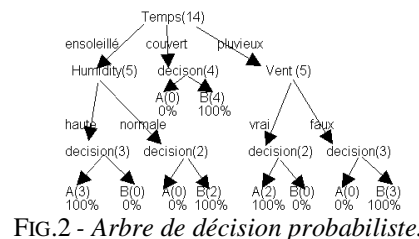


FIG.2 - Arbre de décision probabiliste.

3 Approche probabiliste

Nous étudions le problème des valeurs manquantes pendant le classement. Nous étendons la méthode des Arbres d'Attributs Ordonnés (Lobo 1999) et (Lobo 2000) en construisant pour chaque attribut un arbre de décision probabiliste au lieu d'un arbre de décision classique (§3.1). Pour cela, nous nous inspirons de (Breiman et al. 1984) (Quinaln 1990) et (Quinlan1993) pour la construction des arbres d'attributs probabilistes. Une deuxième extension, que nous appelons Arbres d'Attributs Probabilistes, est proposée afin de prendre en compte les dépendances entre les attributs au lieu de l'ordre d'IM croissante, lors de la construction des arbres d'attributs (Hawarah et al. 2004).

3.1 Arbres d'Attributs Ordonnés Probabilistes

Cette première proposition est une extension de la méthode Arbres d'Attributs Ordonnés (Lobo 1999) et (Lobo 2000). Elle consiste à construire pour chaque attribut un arbre d'attribut selon la méthode de Lobo. Cependant, contrairement à Lobo, qui, en suivant la méthodologie classique, associe à chaque feuille la valeur la plus probable, nous proposons de conserver dans chaque feuille d'un arbre d'attribut la distribution des fréquences des valeurs de l'attribut en question. Cette distribution de probabilités permet de déterminer le classement probabiliste des valeurs d'un attribut manquant. En conséquence, elle permet le classement probabiliste d'un objet avec des attributs manquants. On appelle cette proposition Arbres d'Attributs Ordonnés Probabilistes (AAOP). Le résultat du classement permettant de déterminer une valeur manquante est une distribution de probabilités des valeurs de l'attribut. En conséquence, le classement d'un objet incomplet en utilisant les AAOPs est une distribution probabiliste de classe au lieu d'une seule valeur de classe. La Fig.3 montre les AAOPs de *Température*, *Vent*, *Humidité* :

Classement d'objets incomplets dans un arbre de décision probabiliste

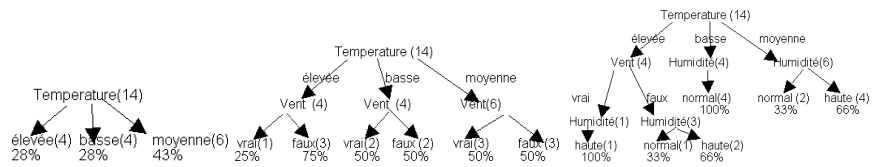


FIG.3 - Arbres d'Attributs Ordonnés Probabilistes pour Température, Vent, Humidité.

Ces arbres sont utilisés pendant le classement d'objets avec des valeurs d'attributs manquantes. Ainsi, si on classe un objet où *Temps* est *ensoleillé*, *Vent* est *faux*, *Température* est *élevée* mais *Humidité* est inconnue, les probabilités des valeurs de l'attribut *Humidité* sont calculées à partir de son arbre d'attribut dans la Fig.3 (à droite). Dans notre exemple, la distribution de probabilités des valeurs de l'attribut *Humidité* sont : *normal* avec la probabilité 0.33 et *haute* avec la probabilité 0.66. Contrairement aux arbres d'attributs ordonnés de Lobo, les AAP ont l'avantage de permettre d'aboutir à des résultats probabilistes en utilisant un calcul identique à celui proposé par (Quinlan 1990). Cependant, de notre point de vue, ces AAPs continuent de poser problème car les attributs pris en compte pour la construction d'un arbre d'attribut sont choisis en fonction de leur IM par rapport à la classe. En particulier, seuls les attributs ayant, par rapport à la classe, un IM inférieur à celui de l'attribut courant, pour lequel on construit l'arbre d'attribut, sont pris en compte. En conséquence, il n'y a aucune garantie que ces attributs dépendent de l'attribut courant. Or, ce sont ces attributs qui détiennent l'information la plus pertinente pour le calcul de la distribution de probabilités de cet attribut. Les Arbres d'Attributs Probabilistes (AAP), présentés ci-dessous, constituent notre deuxième proposition d'extension des arbres d'attributs de Lobo. Contrairement aux AAP, les AAP prennent en compte les dépendances entre attributs.

3.2 Arbres d'Attributs Probabilistes

La méthodologie des arbres d'attributs probabilistes (AAP) que nous proposons ici est une méthodologie qui, pour chaque attribut, construit un arbre d'attribut probabiliste en utilisant les attributs dont il dépend. Afin de déterminer les dépendances entre les attributs, nous calculons l'IM entre chaque couple d'attributs de la base. En effet, l'IM entre deux attributs est la réduction de l'incertitude sur un attribut sachant l'autre. Ainsi, pour un attribut A_i , les attributs dont il dépend sont calculés par l'expression :

$$\text{Dep}(A_i) = \{A_j \mid \text{IM}(A_i, A_j) > 0.01^6\}$$

Dans une deuxième étape, un arbre de décision probabiliste est construit pour chaque attribut en prenant en compte les attributs dont il dépend. L'application de cette méthodologie à la base extraite de (Quinlan 1993) détermine que :

$$\text{IM}(\text{Humidité}, \text{Température}) = 0.37465, \quad \text{IM}(\text{Humidité}, \text{Temps}) = 0.02074$$

$$\text{IM}(\text{Temps}, \text{Température}) = 0.2377, \quad \text{IM}(\text{Température}, \text{Vent}) = 0.039$$

$$\text{IM}(\text{Humidité}, \text{Vent}) = 0, \quad \text{IM}(\text{Vent}, \text{Temps}) = 0.005$$

En conséquence, les dépendances prises en compte sont :

$$\text{Dep}(\text{Humidité}) = \{\text{Température}, \text{Temps}\}$$

$$\text{Dep}(\text{Temps}) = \{\text{Température}, \text{Humidité}\}$$

$$\text{Dep}(\text{Température}) = \{\text{Humidité}, \text{Temps}, \text{Vent}\}$$

$$\text{Dep}(\text{Vent}) = \{\text{Température}\}$$

⁶ On a choisi le degré de dépendance 0.01 arbitrairement.

L'arbre d'attribut probabiliste pour l'attribut *Humidité* est donné dans la Fig.4 et l'arbre de *Temps* est donné dans la Fig.5 :

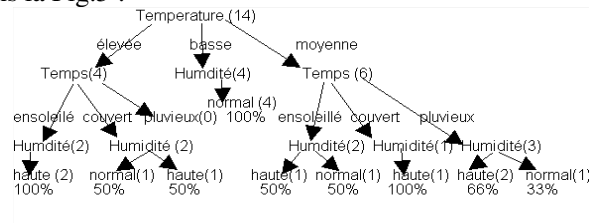


FIG. 4 - Arbre d'Attribut Probabiliste de Humidité.

Si *Température* est *moyenne* et *Temps* est *ensoleillé*, alors la probabilité que *Humidité* soit *haute* est 0.5 et la probabilité qu'elle soit *normale* est 0.5. On remarque ici que ces probabilités sont calculées en prenant en compte les valeurs de *Température* et *Temps*. Cette approche est meilleure que AAOP car elle prend en compte les dépendances qui existent entre les attributs connus dans l'objet à classer et l'attribut dont la valeur est manquante. Les contraintes (Lobo 2001) que les attributs d'une base d'apprentissage devraient vérifier pour que la méthode AAO soit applicable n'ont pas de raison d'être dans les AAPs car il n'y a pas d'ordre de construction imposé.

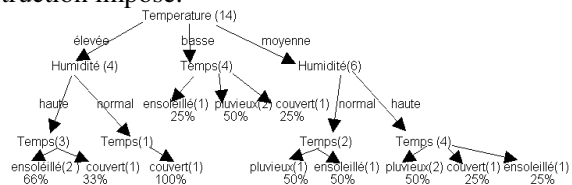


FIG. 5 - Arbre d'Attribut Probabiliste de Temps.

3.3 Problèmes rencontrés avec l'approche AAP

L'utilisation des AAP pose un certain nombre de problèmes :

1) *Cycle*. Lorsque deux attributs mutuellement dépendants sont manquants, le problème du calcul se pose. Pour casser ce cycle, nous proposons la solution suivante : on calcule d'abord la probabilité de l'attribut le moins dépendant de la classe à partir de son arbre d'attribut ordonné probabiliste AAOP, puis les distributions de probabilités de l'autre attribut à partir de son arbre d'attribut probabiliste AAP.

2) *Feuilles indéterminées*. On rencontre ce problème quand aucune instance n'est associée à une feuille. Dans ce cas, on ne sait pas quelle classe on doit associer à cette feuille. Par exemple, prenons l'AAP de l'*Humidité* donné dans la Fig.4. Aucune instance ayant pour l'attribut *Température* la valeur *élevée* et le *Temps* la valeur *pluvieux* n'a été retrouvée dans la base d'apprentissage. Dans ce cas, nous utilisons l'AAOP de l'attribut *Humidité*, ce qui fournit les probabilités suivantes : $P(\text{Humidité} = \text{haute}) = 0.66$, $P(\text{Humidité} = \text{normale}) = 0.33$

4 Le classement probabiliste

Pour classer un nouvel objet, on parcourt l'arbre de décision final de la racine jusqu'à une feuille en suivant les branches correspondant aux valeurs de l'objet à classer. Si on rencontre un attribut inconnu, on appelle son arbre d'attribut probabiliste pour récupérer la distribution

Classement d'objets incomplets dans un arbre de décision probabiliste

de probabilité de ses valeurs. Ce résultat est utilisé dans l'arbre de décision final pour trouver la distribution de probabilité de la classe.

Nous présentons le processus de classement à l'aide d'un exemple : étant donné un objet où *Température* a pour valeur *élevée* et *Vent* a pour valeur *faux*, la probabilité que l'objet appartient à la classe A est calculée à partir de l'arbre de décision donné dans la Fig.2, ce qui correspond à la probabilité totale de A :

$$P(A) = P(A \setminus \text{ensoleillé, haute}) * P(\text{ensoleillé, haute}) + P(A \setminus \text{pluvieux, vrai}) * P(\text{pluvieux, vrai})$$

Comme *Vent* a pour valeur *faux*:

$$P(\text{Vent} = \text{vrai}) = 0, \quad P(A \setminus \text{pluvieux, vrai}) * P(\text{pluvieux, vrai}) = 0$$

$P(A \setminus \text{ensoleillé, haute}) = 1$ C'est la probabilité de la classe A sur cette feuille, tous les cas arrivant à cette feuille appartiennent à la classe A (voir Fig.2). On a alors :

$$P(A) = P(A \setminus \text{ensoleillé, haute}) * P(\text{ensoleillé, haute}) = P(\text{ensoleillé, haute})$$

Pour calculer $P(\text{ensoleillé, haute})$ on va distinguer trois cas :

1) Si *Temps* est *ensoleillé* et *Humidité* est manquante, on a :

$$P(\text{ensoleillé, haute}) = P(\text{haute}) * P(\text{ensoleillé} \setminus \text{haute})$$

$$P(\text{ensoleillé} \setminus \text{haute}) = 1 \quad \text{car } P(\text{ensoleillé}) = 1$$

$P(\text{haute})$ est calculé à partir de son AAP donné dans la Fig.4 ; La racine de l'arbre est *Température*, et sa valeur dans l'objet à classer est *élevée*. En descendant dans l'arbre et en suivant la branche correspondant à la valeur *ensoleillé* pour l'attribut *Temps*, on arrive à une feuille où $P(\text{Humidité} = \text{haute}) = 1$, $P(\text{Humidité} = \text{normal}) = 0$

$$P(A) = P(\text{ensoleillé, haute}) = P(\text{haute}) * P(\text{ensoleillé} \setminus \text{haute}) = 1 * 1 = 1$$

2) Si *Temps* est inconnu et *Humidité* est *haute* :

$$P(\text{ensoleillé, haute}) = P(\text{ensoleillé}) * P(\text{haute} \setminus \text{ensoleillé})$$

$$P(\text{haute} \setminus \text{ensoleillé}) = 1 \quad \text{car } P(\text{haute}) = 1$$

$P(\text{ensoleillé})$ est calculé à partir de son AAP construit en utilisant *Température*, *Humidité*.

L'arbre est donné dans la Fig.5 :

$$P(\text{ensoleillé}) = P(\text{ensoleillé} \setminus \text{haute, élevé}) * P(\text{haute, élevé}) = 0.66$$

$$P(\text{haute, élevé}) = 1 \quad \text{car } P(\text{haute}) = 1 \text{ et } P(\text{élevé}) = 1$$

$$P(A) = P(\text{ensoleillé, haute}) = P(\text{ensoleillé}) * P(\text{haute} \setminus \text{ensoleillé}) = 0.66 * 1 = 0.66$$

3) Si *Temps* et *Humidité* sont inconnus, on a un cycle:

$$P(\text{ensoleillé, haute}) = P(\text{haute}) * P(\text{ensoleillé} \setminus \text{haute})$$

Humidité étant l'attribut le moins dépendant de la classe, on calcule la probabilité que *Humidité* soit *haute* à partir de son AAP donné dans la Fig.3 :

$$P(\text{haute}) = P(\text{haute} \setminus \text{faux, élevé}) * P(\text{faux, élevé}) = 0.66$$

$$P(\text{faux, élevé}) = 1 \quad \text{car on sait que } \text{Vent est faux, et } \text{Température est élevée.}$$

La probabilité que *Temps* soit *ensoleillé* sachant que *Humidité* est *haute* est calculée à partir de son AAP donné dans la Fig.5 :

$$\begin{aligned} P(\text{ensoleillé} \setminus \text{haute}) &= \sum P(\text{ensoleillé} \setminus \text{haute}, A_i) * P(A_i \setminus \text{haute}) \\ &= P(\text{ensoleillé} \setminus \text{haute, élevée}) * P(\text{élevée} \setminus \text{haute}) \\ &\quad + P(\text{ensoleillé} \setminus \text{haute, moyenne}) * P(\text{moyenne} \setminus \text{haute}) \\ &= P(\text{ensoleillé} \setminus \text{haute, élevée}) * P(\text{élevée} \setminus \text{haute}) = 0.66 * 1 = 0.66 \end{aligned} \quad (1)^7$$

Car *Température* est *élevée* alors $P(\text{élevée} \setminus \text{haute}) = P(\text{élevée}) = 1$, $P(\text{moyenne} \setminus \text{haute}) = 0$

$$P(A) = P(\text{ensoleillé, haute}) = P(\text{haute}) * P(\text{ensoleillé} \setminus \text{haute}) = 0.66 * 0.66 = 0.4356$$

Nous calculons la probabilité que l'objet appartienne à la classe B de la même manière que précédemment. Nous pouvons également calculer cette probabilité comme suit :

⁷ Pour prouver que $P(B \setminus C) = \sum P(B \setminus A_i, C) * P(A_i \setminus C)$ donné dans la relation (1) :

$$\begin{aligned} P(B) &= \sum P(B \setminus A_i) * P(A_i) = \sum P(B, A_i) \rightarrow P(B \setminus C) = \sum P(B, A_i \setminus C) = \sum P(B, A_i, C) / P(C) \\ &= \sum P(C) * P(A_i \setminus C) * P(B \setminus A_i, C) / P(C) = \sum P(B \setminus A_i, C) P(A_i \setminus C) \end{aligned}$$

$$P(B) = 1 - P(A)$$

5 Conclusion et Perspectives

Dans le monde réel, un objet incomplet peut potentiellement appartenir à plusieurs classes et devrait donc être associé à plusieurs feuilles dans l'arbre de décision. Dans un domaine critique comme la médecine, prendre une seule décision quand il y a manque d'information peut être dangereux. Notre approche consiste à utiliser la notion de probabilité pour résoudre le problème des valeurs manquantes dans les données. Nous avons proposé de remplacer une valeur manquante par une distribution de probabilités et un objet incomplet par une distribution de probabilités de classe.

La première expérimentation concernant la construction des arbres d'attributs probabilistes a été faite sur deux étapes : nous avons étendu l'algorithme ID3 (Quinlan 1986) qui construit des arbres de décision sans élagage pour avoir un algorithme qui construit des arbres de décision probabilistes (ID3-Probabiliste) et nous avons développé un programme en Java qui utilise ID3-Probabiliste pour construire à partir d'une base d'apprentissage complète un arbre de décision probabiliste (AAP) et un arbre de décision ordonné probabiliste (AAOP) pour chaque attribut dans la base, ainsi que l'arbre de décision probabiliste final qui correspond à la base entière.

A partir de ces arbres, on peut déduire les relations entre les attributs. Par exemple, si les attributs sont indépendants comme dans la base contact-lenses (Blake 1998), chacun de ses arbres est un seul nœud-feuille avec sa distribution de probabilité. Par contre, chaque arbre construit à partir de cette base selon AAO est un seul nœud-feuille avec sa valeur la plus probable. Dans le cas, où tous les valeurs sont équiprobables, l'algorithme utilisé par AAO (comme ID3 ou C4.5) choisit une valeur aléatoirement. D'autre part, nous remarquons que quelques feuilles contiennent une valeur de classe avec la probabilité 1, mais le nombre d'instances arrivant à cette feuille est faible (inférieur à 5). Contrairement à d'autres tests statistiques, il n'existe pas pour les arbres de décision de seuil reconnu sur le nombre d'individus nécessaires pour que le résultat soit significatif. Selon (Labarere et al. 2003), il est souvent considéré en informatique que 5 individus par feuille sont suffisants pour la valider alors qu'en médecine il serait nécessaire d'avoir au moins 20 individus par feuille, faute de quoi le résultat risque de ne pas être significatif. Dans notre cas, nous trouvons que l'augmentation du degré de dépendance peut éliminer quelques attributs non significatifs ce qui conduit également à augmenter le nombre de cas par feuille. Une étude est en cours pour valider, avec des experts du domaine, les résultats de notre approche appliquée à une base de données médicale sur l'apnée du sommeil.

Références

- Blake C.L. and Merz C.J. (1998): UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
- Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984): Classification and Regression Trees, Wadsworth and Brooks.
- Kononenko I., Bratko I. and Roskar E. (1984): Experiments in Automatic Learning of Medical Diagnostic Rules, Technical Report, Jozef Stefan Institute, Ljubljana, Yugoslavia.
- Friedman J.H., Kohavi R. and Yun Y. (1996): Lazy Decision Trees, AAAI.
- Hawarah L., Simonet A. and Simonet M. (2004): Une approche probabiliste pour le classement d'objets incomplets dans un arbre de décision, EGC 2004, poster.

Classement d'objets incomplets dans un arbre de décision probabiliste

- Hawarah L., Simonet A. and Simonet M. (2004): A probabilistic approach to classify incomplete objects using decision trees, Spain, DEXA. Lecture Notes in Computer Science 3180 pp. 549-558.
- Labarere J., Bosson J-L. and Robert C. (2003): Utilisation des arbres d'induction en épidémiologie : Principes et exemple d'application à l'analyse d'une enquête de pratiques de prévention de la maladie thrombo-embolique veineuse. Congrès épidémiologie et biométrie, Lille.
- Lobo O.O. and Numao M. (1999): Ordered estimation of missing values, Pacific-Asia Conference on Knowledge Discovery and Data Mining.
- Lobo O.O. and Numao M. (2000): Ordered estimation of missing values for propositional learning, Japanese Society for Artificial Intelligence, JSAI, vol.15, no.1.
- Lobo O.O. and Numao M. (2001): Suitable Domains for Using Ordered Attribute Trees to Impute Missing Values. IEICE TRANS. INF. & SYST., Vol.E84-D, NO.2.
- Quinlan J.R. (1989): Unknown attribute values in induction. Proc. Sixth International Machine Learning Workshop, Morgan Kaufmann.
- Quinlan J.R. (1986): Induction of decision trees. Machine Learning, 1, pp.81-106.
- Quinlan J.R. (1990): Probabilistic decision trees, in Machine Learning: an Artificial Intelligence Approach, ed.Y.Kodratoff, vol.3, Morgan Kaufmann, San Mateo, pp.140-152.
- Quinlan. J.R (1993) : C4.5 Programs for Machine Learning, Morgan Kaufmann.
- Shannon C.E., Weaver W. (1949): Théorie Mathématique de la communication, les classiques des sciences humaines.
- White A.P. Liu W.Z., Thompson S.G. and Bramer M.A. (1997): Techniques for Dealing with Missing Values in Classification. LNCS 1280, pp. 527-536.

Fouille de données du génome à l'aide de modèles de Markov cachés

Sébastien Hergalant * **, Bertrand Aigle *
Pierre Leblond*, Jean-François Mari**

*Laboratoire de Génétique et Microbiologie, UMR-UHP-INRA, IFR 110,
54506 Vandœuvre-lès-Nancy, France
{bertrand.aigle,pierre.leblond}@nancy.inra.fr,

**LORIA UMR-CNRS 7503, 54506 Vandœuvre-lès-Nancy, France
{hergalan,jfmari}@loria.fr
<http://www.loria.fr/~jfmari/ACI/>

Résumé. Nous décrivons un processus de fouille de données en bioinformatique. Il se traduit par la spécification de modèles de Markov cachés du second-ordre, leur apprentissage et leur utilisation pour permettre une segmentation de grandes séquences d'ADN en différentes classes qui traduisent chacune un état organisationnel et structural des motifs d'ADN locaux sous-jacents. Nous ne supposons aucune connaissance *a priori* sur les séquences que nous étudions. Dans le domaine informatique, ce travail est dédié à la définition d'observations structurées (les k-d-k-mers) permettant la localisation en contexte d'irrégularités, ainsi qu'à la description d'une méthode de classification utilisant plusieurs classifieurs. Dans le domaine biologique, cet article décrit une méthode pour prédire des ensembles de gènes co-régulés, donc susceptibles d'avoir des fonctions liées en réponse à des conditions environnementales spécifiques.

1 Introduction

L'accumulation des séquences issues des projets de séquençage oblige la mise en œuvre de méthodes de fouille de données efficaces pour comprendre les mécanismes impliqués dans l'expression, la transmission et l'évolution des gènes. Nous nous intéressons aux modèles stochastiques et méthodes classificatoires permettant de prédire les séquences promotrices et autres petites séquences régulatrices chez les bactéries. Une manière de cerner notre ignorance vis à vis des motifs et segments d'ADN impliqués dans les mécanismes décrits plus haut est de modéliser l'évolution et la structuration du génome par des processus stochastiques capables d'apprentissage statistique nécessitant un minimum de connaissances *a priori*. Ces modèles stochastiques sont utilisés comme révélateurs d'organisations locales remarquables qu'un expert doit interpréter.

Nous nous intéressons à la localisation de sites de fixation de protéines. Ces sites de fixation – appelés TFBS (*Transcription Factor Binding Sites*) ou encore promoteurs transcriptionnels – sont constitués de trois séquences adjacentes de nucléotides :

$$N_x - N_y - N_z \quad \text{avec} \quad N \in \{A, C, G, T\} \\ 3 \leq x, z \leq 9 \\ 0 \leq y \leq 25$$

N_x et N_z sont susceptibles d'être altérées par quelques substitutions tandis que N_y – appelé espaceur – est une chaîne de composition et de taille variable. L'ensemble N_x — N_y — N_z se situe en amont d'un gène et sert de site de fixation pour une protéine qui vient réguler son expression. Un régulon est un ensemble de gènes possédant nécessairement le même TFBS en amont.

2 Matériel et méthodes

2.1 Définition des HMM2

La modélisation stochastique est une approche mathématique pour prendre en compte la variabilité inhérente aux processus issus du vivant comme le sont la reconnaissance de la parole, ou la segmentation du génome. Un modèle stochastique particulier – le modèle de Markov caché (HMM pour *Hidden Markov Model*) – représente la suite des nucléotides par deux processus stochastiques : l'un caché, prenant ses valeurs sur un ensemble d'états et qui est une chaîne de Markov, l'autre visible prenant ses valeurs parmi les observations physiques : la séquence de nucléotides constituant l'ADN. La variabilité est capturée par la supposition que les observations ne sont pas uniquement associées aux états mais dépendent d'une densité définie sur chaque état. En reprenant les notations introduites par Churchill et Mury (Mury, 1997), nous définissons plus formellement un *HMM2* comme un HMM du type M2-M0 de la façon suivante :

- $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$, un ensemble fini comprenant N états ;
- \mathcal{A} la matrice donnant les probabilités de transitions entre états : $\mathcal{A} = (a_{ijk})$ pour un *HMM2*, avec la contrainte :

$$\sum_k a_{ijk} = 1 \quad \forall i, j;$$

- $b_i()$ les lois des densités associées aux états s_i .

La matrice \mathcal{A} est initialisée avec un ensemble de valeurs qui permettent de définir la topologie du graphe des transitions entre états : quelles sont les transitions autorisées ? aller simple ($a_{ijk} > 0$, $a_{kji} = 0$), aller-retour ($a_{ijk} > 0$, $a_{kji} > 0$), bouclage ($a_{ijj} \neq 0$), etc.

La modélisation du génome par des HMM est ainsi fondée sur deux principes : (i) le génome peut être découpé en segments par une chaîne de Markov et (ii) chaque segment est la réalisation d'un processus stationnaire représenté par une densité de probabilité sur l'espace des observations.

En modélisant de la sorte les segments composant le chromosome, on ignore la réalité de la constitution génétique comme résultat d'un processus organisé mais on peut utiliser des algorithmes (Boyer et al., 1990) d'apprentissage et de reconnaissance rapides. Cette façon de procéder est complémentaire d'une approche analytique et explicative fondée sur un mécanisme de raisonnement. En mesurant précisément par une probabilité ce que l'on qualifie au premier abord de hasardeux, on diminue l'indéterminisme de notre perception du processus et on peut faire apparaître des comportements explicables, donc prévisibles, qui pourront être réutilisés dans un mécanisme de raisonne-

ment ; ce mécanisme d'extraction et de réutilisation est un des principes de la fouille de données.

2.2 Les k-mer

Les observations les plus simples modélisées par les densités $b_i()$ sont les 4 nucléotides : $A_1 = \{A, C, G, T\}$. Si on considère les 16 paires possibles de nucléotides, on définit alors A_2 comme l'ensemble des 2-mer, et ainsi de suite pour l'ensemble de k -mer A_k constitué des 4^k séquences de k nucléotides. La suite d'ADN peut être vue comme une chaîne de k -mer qui se recouvrent – on appelle l le décalage compté en nombre de bases entre deux k -mer successifs dans le chromosome – et qui constitue la suite des observations du HMM.

Par exemple si on traite la suite : TAGGCTAGGTG, avec $k=4$ et $l=1$, la chaîne observée est : TAGG-AGGC-GGCT-GCTA-CTAG-TAGG-AGGT-GGTG (1). Avec $k=4$ et $l=4$, elle devient TAGG-CTAG (2).

Nous définissons aussi les $k_1 - d - k_2 - mers$ ($1 < k_1 + k_2 < 7$) constitués de deux k -mer espacés par d nucléotides. Il y a $4^{k_1+k_2}k_1 - d - k_2 - mers$ différents ; les d bases constituant l'espaceur n'intervenant pas dans la définition. Par exemple, la séquence TAGGCTAGGTG peut être vue comme une séquence de 2-3-2-mers. Dans ce cas, avec un décalage de 2, nous observons la chaîne TATA-GGGG-CTTG, qui est différente des deux autres chaînes (1) et (2). Cet article montre que les k - d - k -mers sont adaptés à la recherche de sites de fixation constitués par deux segments non adjacents de nucléotides situés en amont de gènes.

2.3 Estimation du *HMM2*

Dans cette section, nous décrivons la méthode qui permet de spécifier le *HMM2* qui rende le mieux compte des données au sens du maximum de vraisemblance. Nous ne contrôlons que la topologie du *HMM2*. Les matrices \mathcal{A} et \mathcal{B} (respectivement, les probabilités de la chaîne de Markov et les paramètres des densités) sont estimées par l'algorithme EM (Dempster et al., 1977).

Le but de la modélisation stochastique est d'élaborer un indice capable d'aider l'expert – le bioinformaticien – dans son travail de fouille. On désire calculer un indice concis et expurgé du bruit qui rende compte de l'organisation locale des nucléotides. Le *HMM2* est utilisé comme une machine à segmenter et à classer. Nous nous servons de la probabilité *a posteriori* de la transition $s_i \rightarrow s_j \rightarrow s_k$ entre $t - 1$ et $t + 1$ pendant l'observation de tout le chromosome. La figure 1 représente la probabilité de la transition “boucle” sur un état du *HMM2* dans Artemis (logiciel d'annotation de génome (Rutherford et al., 2000)) en même temps que l'annotation du génome reprenant les résultats des bases de données d'annotation issues du Web. Sur cet exemple, les pics de la probabilité sont des indices permettant à l'expert de retrouver une information connue : la présence de site de fixation d'une protéine régulatrice devant un gène.

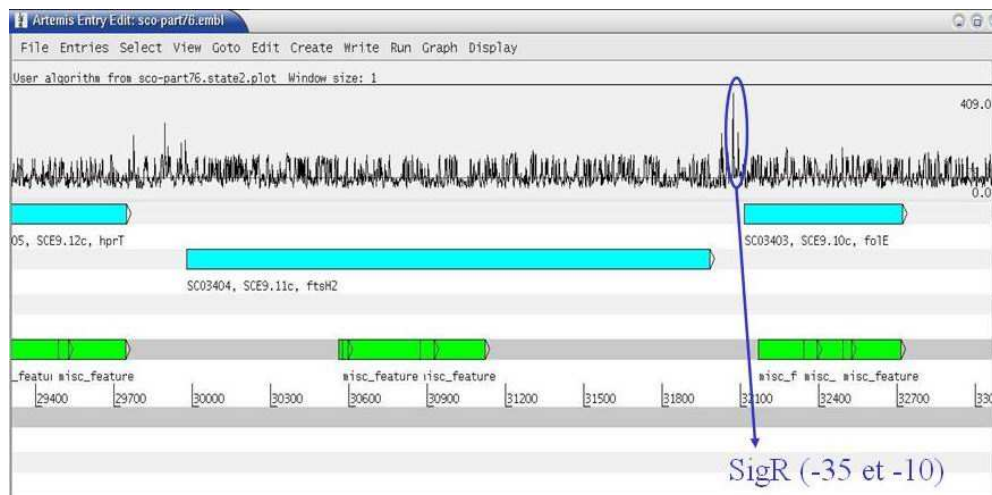


FIG. 1 – Représentation d’une probabilité de segmentation donnée par le *HMM2* dans *Artemis*

2.4 Spécification de la topologie du *HMM2*

Nous nous limitons à l’étude de *HMM2* ergodiques, c’est à dire dans lesquels toutes les transitions entre états sont possibles. Il s’agit donc d’identifier le nombre d’états susceptibles de donner la classification la plus utile pour l’expert.

La distance entre deux états est représentée par la divergence entre les deux densités $b_i()$ et $b_j()$.

$$div(i, j) = \int_x b_i(x) \ln \frac{b_i(x)}{b_j(x)} dx + \int_x b_j(x) \ln \frac{b_j(x)}{b_i(x)} dx$$

Nous utilisons des *HMM2* de 3 à 6 états utilisant des observations du type *k-mer* ($k = 2, 3, 4, 5, 6$) et/ou 2-d-2, 3-d-3 avec $0 \leq d \leq 25$. Le déplacement entre deux *k-mer* est toujours de $l = 1$. A partir d’un *HMM2* de trois états dans lequel toutes les transitions sont équiprobables et les densités des lois uniformes, nous utilisons l’algorithme Forward-backward pour obtenir une estimation selon le maximum de vraisemblance. La divergence entre toutes les paires d’états est calculée. Si la plus grande divergence est inférieure à un seuil donné, nous recommençons l’apprentissage avec un modèle possédant un état supplémentaire. Nous cherchons à faire apparaître ainsi des états captant des irrégularités locales dans la distribution des nucléotides donc bien différents d’un état rendant compte d’une organisation “moyenne”. Une situation dans laquelle on n’arrive pas à faire apparaître un état différent des autres est une situation d’échec de classification par les *HMM2*. Nous tentons alors, un autre apprentissage en envisageant une autre définition des observations. Cette situation n’a pas été rencontrée sur les corpus étudiés.

Deux *HMM2*, appelés *HMM2+* et *HMM2-* sont construits : l’un pour le sens direct, l’autre pour le sens complémentaire inverse du chromosome. (les deux brins sont symétriques et s’enroulent pour former la double hélice d’ADN constituant le génome).

L'ensemble des programmes est écrit en C++ et dérive du logiciel¹ CARROTAGE (Le Ber et al., 2004) utilisé initialement en agronomie pour la recherche de successions de cultures.

2.5 Recherche automatique des pics

La recherche des pics de probabilité *a posteriori* pour un état donné se fait en utilisant une fenêtre glissante de 200 pb (paires de bases) se recouvrant de 100 pb avec la fenêtre voisine. Dans chaque fenêtre sont calculées des statistiques (moyenne, variance, min et max). La construction du modèle a été faite de telle sorte qu'un état au moins possède des fluctuations de cette probabilité. Un pic est défini par sa valeur maximum qui doit être significativement supérieure à la moyenne calculée dans la fenêtre et par sa longueur p qui doit vérifier l'inégalité : $2k - 1 < p < 12$, k étant la taille des *k-mer* observés. La valeur 12 correspond à la largeur maximum d'un pic recherché. En effet, les *HMM2* – bien que supérieur aux *HMM1* (Mari et al., 1997) dans ce domaine – ont une piètre capacité à modéliser les durées de longs segments. Sous chaque pic, nous extrayons un motif qui doit être ensuite classé.

2.6 Clustering des motifs

La classification des motifs se fait à l'aide de deux programmes effectuant des alignements entre chaînes. Le premier (MULTALIN (Corpet, 1988)) effectue une classification hiérarchique de motifs sous-jacents aux pics fondée sur la distance calculée entre deux motifs. Ce programme autorise l'alignement de séquences courtes sans pénalités aux extrémités. Le nombre maximum de substitutions de nucléotides peut être spécifié et les trous interdits. Le résultat de MULTALIN est un ensemble de classes représentées par leur consensus. Le deuxième, FASTA effectue une classification incrémentale sur les paires de motifs espacés comme le montre l'algorithme 1.

Algorithm 1 Algorithme de recherche de TFBS.

```

for all classe trouvée par MULTALIN do
  construit le consensus de cette classe
  recherche dans le génome entier les occurrences de cette séquence
end for
Sélectionne les paires d'occurrences  $(s_1, s_2)$  espacées de  $d$  bp. Appelle ce triplet  $s_1 + s_d + s_2$ 
Recherche dans tout le génome les occurrences de  $s_1 + s_d + s_2$ . Ceci élimine les  $(s_1, s_2)$  non espacés convenablement
Effectue un alignement par FASTA entre les paires de séquences  $s_1 + s_d + s_2$  pour permettre un regroupement de séquences légèrement dégénérées.
Sélectionne celles qui apparaissent plus de 3 fois
Sélectionne  $s_1 + s_d + s_2$  trouvée dans les régions intergéniques et/ou à moins de 600 bp en amont d'une ORF
Regroupe dans une classe tous les gènes trouvés en aval d'une séquence  $s_1 + s_d + s_2$ 

```

¹Licence publique Gnu

3 Expérimentations

3.1 Le génome d'étude

Les *Streptomyces* sont des bactéries filamenteuses du sol qui revêtent un intérêt économique fort compte tenu de l'importance des produits de leur métabolisme secondaire. Elles sont en effet la source principale d'antibiotiques parmi les micro-organismes. L'espèce *Streptomyces coelicolor* A3(2) présente un chromosome linéaire de 8,7 méga paires de bases (8,7 Mb), qui se caractérise par un taux global en bases G + C de 72%. Chez *S. coelicolor*, environ 12% des 7825 gènes prédits (ou ORF pour *Open Reading Frame*) se répartissent dans différentes familles de régulateurs transcriptionnels. Les facteurs sigma y représentent une classe particulièrement importante, avec 65 gènes prédits chez *S. coelicolor* et 60 chez *S. avermitilis*. Moins de dix facteurs sigma ont été étudiés jusqu'à présent chez les *Streptomyces*, et seulement quelques régulons (ensembles de gènes co-régulés) ont été définis par des approches expérimentales en biologie. De même, un nombre très faible de régulateurs transcriptionnels (activateurs ou répresseurs) a été étudié jusqu'alors. La très grande majorité des motifs nucléotidiques sur lesquels agissent ces facteurs de transcription reste donc à définir.

Notre approche bio-informatique vise à extraire et classer les motifs nucléotidiques présents dans les régions intergéniques des génomes des bactéries. Cette approche exhaustive permettra de définir des motifs régulateurs en faisant abstraction du bruit notamment généré par la superposition de motifs dans les gènes à régulation complexe. La validation biologique expérimentale permettra ensuite de définir les gènes co-régulés.

L'algorithme d'extraction des irrégularités locales fait apparaître 4 fois plus de pics dans les régions intergéniques que dans les gènes. Bon nombre de pics se situe au voisinage des sites de fixation des protéines intervenant dans la régulation de l'expression des gènes. Ce sont des séquences d'ADN situées en amont des gènes, organisées en paires espacées, et qui sont spécifiquement reconnues par les protéines régulatrices (les facteurs de transcription comme les facteurs sigma) en question. Chaque régulateur transcriptionnel possède sa séquence d'ADN cible qui lui est propre. Celle-ci est définie sous la forme d'un consensus. Les gènes co-régulés par un facteur de transcription possèdent donc tous la même séquence en amont. Après l'extraction de ces pics, notre travail a consisté à classer les segments nucléotidiques sous-jacents pour voir dans quelles mesures ils pourraient être des TFBS connus ou à découvrir.

3.2 Validation de la méthode sur un régulon connu : SigR

Chez *S. coelicolor*, SigR est un des 65 facteurs sigma prédits ou connus. Il est impliqué à un niveau clé dans les mécanismes de réponse au stress oxydant, en co-régulant (positivement ou négativement) 30 gènes de cette bactérie. Ceci a été montré expérimentalement (Paget et al., 2001). La séquence consensus reconnue par SigR est $GGGAAT - N_{18} - GTTN$ (structure type $s_1 + s_d + s_2$). En nous basant sur ces résultats, nous avons testé nos algorithmes sur 3 génomes bactériens du groupe des actinomycètes : *S. coelicolor*, *S. avermitilis* et *Mycobacterium tuberculosis*, proche parente des deux premières et pathogène pour l'homme (vecteur de la tuberculose). Nous utilisons pour cela des *HMM2* à 3-mers ou 3-d-3-mers ($0 \leq d \leq 25$), qui représentent le

mieux la réalité du code biologique et fournissent les meilleurs résultats en termes quantitatifs (ni trop, ni trop peu de pics) et qualitatifs (localisation, spécificité et largeur des pics).

En amont des 30 gènes régulés par SigR chez *S. coelicolor*, 84 motifs sont extraits. Parmi ceux-ci, 47 décrivent les motifs types s_1 et s_2 . Les 30 motifs s_1 trouvés permettent d'inférer la position des 30 promoteurs reconnus par SigR (table 1). Le motif s_2 n'est pas toujours extrait. Ceci est dû au fait que la séquence est trop petite pour pouvoir être modélisée de façon convenable par un *HMM2* utilisant des *3-mers* dans leur définition.

Sur la figure 1, nous observons une segmentation globalement homogène le long de la séquence, interrompue par la présence d'hétérogénéités locales dans les régions intergéniques. Cadrons notre attention sur la région intergénique en amont du gène *folE*. Celui-ci est un des 30 gènes contrôlés par SigR. Deux des trois pics décelables dans cette région se trouvent aux positions des motifs s_1 et s_2 (appelés respectivement boîtes -35 et -10). Le dernier pic a une signification inconnue.

Parmi les 37 pics ne décrivant pas les promoteurs reconnus par SigR, 22 décrivent d'autres promoteurs pour ces 30 gènes ou d'autres motifs de fixation dont le sens biologique est démontré.

Génome	Nombre de gènes	s_1	s_2	TFBS détectés
<i>S. coelicolor</i>	30 (SigR)	30	17	30
<i>S. avermitilis</i>	23 (SigR)	23	15	23
<i>M. tuberculosis</i>	13 (SigH)	13	10	13

TAB. 1 – Détection par *HMM2* de promoteurs connus chez 3 bactéries actinomycètes.

Chez *S. avermitilis*, les promoteurs SigR n'ont pas été définis mais un test de similarité de séquences permet de proposer un jeu de 23 gènes homologues à ceux régulés par SigR chez *S. coelicolor*. L'hypothèse de départ est que les mécanismes de régulation sont vraisemblablement comparables entre génomes voisins. Cette modélisation permet de détecter les 23 promoteurs de type SigR correspondant (cf. table 1).

Chez *M. tuberculosis*, SigH est le régulon homologue à SigR (Manganelli et al., 2002). Il comporte au moins 13 gènes possédant en amont une séquence reconnue par SigH (très similaire de la séquence de type SigR). Ces 13 promoteurs sont également extraits avec cette méthode (table 1).

Enfin, les motifs s_1 et s_2 , pris isolément, sont largement distribués sur les génomes étudiés. Cependant, ces *HMM2* ne réagissent qu'en présence de la paire de motifs convenablement espacés, dans un environnement nucléotidique intergénique spécifique que l'on pourrait qualifier de "région promotrice". C'est cette propriété remarquable qui fait tout l'attrait de cette méthode.

3.3 Classification des motifs extraits de *S. coelicolor*

Pour cette étude, les modèles *HMM2+* et *HMM2-* utilisent des *3-mers* avec $l = 1$. Nous considérons une portion de 1/8 du génome de *S. coelicolor* (1,15 Mb) pour laquelle 3000 motifs intergéniques ont été extraits à partir des 2 *HMM2*. Les classes trouvées

par la méthode de clustering décrite dans cet article représentent 229 pics/3000 et permettent de prédire :

- Des promoteurs reconnus par des facteurs sigma connus (SigB, WhiG, SigR).
- Des promoteurs reconnus par des régulateurs transcriptionnels qui font actuellement l'objet d'expérimentations biologiques encore non publiées (régulon PhoR / PhoP, définition d'une classe plus large pour le régulon SigR). Les résultats obtenus *in vitro* confirment la validité de cette méthode *in silico*.
- Des promoteurs potentiels reconnus par des régulateurs transcriptionnels hypothétiques et déjà prédits par d'autres méthodes de recherche de motifs promoteurs (Li et al., 2002; Studholme et al., 2004).
- Cinq régulons potentiels entièrement nouveaux. Deux d'entre eux sont potentiellement régulés par deux facteurs sigma hypothétiques pour lesquels il est possible de déduire une fonction biologique. Une des deux classes de gènes situés en aval des séquences extraites serait impliquée dans les processus de sporulation et de développement de *S. coelicolor*. L'autre contrôlerait plusieurs régulateurs transcriptionnels différents pour fournir une réponse plus généralisée.

L'apport de ces connaissances est primordiale pour le biologiste, qui voit son panel de conditions expérimentales restreint, et peut ainsi entreprendre de tester ces résultats.

4 Conclusions et Perspectives

Nous proposons ici une nouvelle méthodologie sans *a priori* pour construire et utiliser des *HMM2* permettant de localiser des sites de fixation de protéines régulatrices dans les séquences d'ADN des procaryotes. Sur trois génomes de bactéries actinomycètes, l'étude de régulons connus a montré la capacité qu'ont ces *HMM2* à décrire de courts motifs d'ADN (5 à 12 pb) riches en sémantique biologique. La classification de ces motifs permet également de prédire des nouvelles classes de gènes co-régulés, potentiellement testables par le biologiste.

Cependant, cette méthode repose sur une classification hiérarchique elle-même basée sur l'alignement multiple de séquences courtes. Ceci est mal réalisé par les méthodes courantes d'alignement de séquences, dont le degré de similarité est calculé à partir de scores obtenus pour un nombre limité de symboles. La classification utilisée est séquentielle et chaque classifieur utilise des données propres. Dans cette optique, une meilleure définition des consensus de classe, ainsi qu'une fusion de résultats classificatoires sont envisagées. A plus long terme, l'adaptation de la méthode à des génomes de structuration différentes sera une démarche importante à mettre en œuvre.

Références

- Boyer, A., Martino, J. D., Divoux, P., Haton, J.-P., Mari, J.-F., and Smaili, K. (1990). Statistical Methods in Multi-Speaker Automatic Speech Recognition. *Applied Stochastic Models and Data Analysis*, 6(3) :143–155.
- Corpet, F. (1988). Multiple Sequence Alignment with Hierarchical Clustering. *Nucl. Acids Res*, 16(22) :10881–10890.

- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum-Likelihood From Incomplete Data Via The EM Algorithm. *Journal of Royal Statistic Society, B (methodological)*, 39 :1 – 38.
- Le Ber, F., Mari, J.-F., Benoît, M., Mignolet, C., and Schott, C. (2004). Carrotage, a software for mining land-use data. In *Fourth International Workshop on Environmental Applications of Machine Learning - EAML'2004, Bled, Slovenia*.
- Li, H., Rhodius, V., Gross, C., and Siggia, E. D. (2002). Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. USA*, 99(18) :11772–11777.
- Manganelli, R., Voskuil, M. I., Schoolnik, G. K., Dubnau, E., Gomez, M., and Smith, I. (2002). Role of the extracytoplasmic-function sigma factor sigma(H) in *Mycobacterium tuberculosis* global gene expression. *Molecular Microbiology*, 45(2) :365–374.
- Mari, J.-F., Haton, J.-P., and Kriouile, A. (1997). Automatic Word Recognition Based on Second-Order Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 5 :22 – 25.
- Mury, F. (1997). *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN*. Thèse de doctorat, Université René Descartes, Paris V.
- Paget, M. S. B., Molle, V., Cohen, G., Aharonowitz, Y., and Buttner, M. J. (2001). Defining the disulphide stress response in *Streptomyces coelicolor* A3(2) : identification of the σ^R regulon. *Molecular Microbiology*, 42(4) :1007–1020.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M., and Barrell, B. (2000). Artemis : sequence visualization and annotation. *Bioinformatics*, 16(10) :944–945.
- Studholme, D. J., Bentley, S. D., and Kormanec, J. (2004). Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor*. *BMC Microbiol.*, 4(1) :14.

Summary

We describe a data mining process in Bioinformatics based on second-order hidden Markov models. After an automatic, unsupervised training, these models perform a segmentation of a whole genome in several classes that contain specific genetic features. We do not make any *a priori* assumption on the genetic organisation. This work is devoted to the definition of structured motifs (the k-d-k mers) that allow the localization of irregularities. It also describes a method to predict co-regulated sets of genes.

Exploration visuelle d'images IRMf basée sur des Gaz Neuronaux Croissants

Jerzy Korczak*, Jean Hommet*, Nicolas Lachiche*, Christian Scheiber**

*LSIIT, CNRS, Illkirch

<jjk,hommet,lachiche>@lsiit.u-strasbg.fr

**CHU Lyon

christian.scheiber@chu-lyon.fr

Résumé. Les algorithmes actuels de fouille de données ne supportent que de façon très limitée les mécanismes de guidage et d'engagement d'expert dans le processus de découverte. Dans cet article, nous présentons une nouvelle approche interactive de fouille des images IRMf, guidée par les données, permettant l'observation du fonctionnement cérébral. La discrimination des voxels d'image du cerveau qui présentent une réelle activité est en général très difficile à cause d'un faible rapport signal sur bruit et de la présence d'artefacts. L'exploration de donnée visuelle se focalise sur l'intégration de l'utilisateur dans le processus de découverte de connaissance en utilisant des techniques de visualisation efficaces, d'interaction et de transfert de connaissances. Dans ce travail, nous montrons sur les données réelles, que l'exploration visuelle permet d'accélérer le processus d'exploration d'images IRMf et aboutit à de meilleurs résultats dotés d'une confiance accrue.

1 Introduction

Les techniques modernes d'imagerie cérébrale, comme l'Imagerie par Résonance Magnétique fonctionnelle (IRMf), offrent la possibilité d'enregistrer en même temps l'activité de l'ensemble du cerveau. C'est une force, mais cela génère une masse de données considérable (environ 300000 voxels, "pixels tridimensionnels", pour lesquels on recueille entre 100 et 1000 observations). Du point de vue de la fouille de données, le cerveau est un objet complexe par excellence. En général, la discrimination des voxels d'image du cerveau qui présentent une réelle activité est très difficile à cause d'un faible rapport signal sur bruit et de la présence d'artefacts. Les premiers tests des algorithmes actuels de fouille dans ce domaine ont montré que leurs performances et leurs qualités de reconnaissance sont faibles (Sommer and Wichert, 2003). En raison de la difficulté qu'il y a à manipuler de telles quantités d'informations, l'essentiel des études ne cherchent pas à les explorer, mais s'en servent pour tester un modèle par le biais de statistiques univariées effectuées en chacun des points. C'est le principe de logiciels de traitement tels que Statistical Parametric Mapping (SPM) (Friston et al., 1995), AFNI (Cox, 1996) ou BrainVoyager (Goebel, 1997) qui consiste à mettre en évidence les voxels plus actifs dans une condition par rapport à une autre.

Sur le plan international, un grand nombre de recherches méthodologiques sont en cours pour mettre en évidence les variations qui ont du sens. On peut regrouper celles-ci en deux grandes familles. La plus commune est l'approche par les statistiques

multivariées comme les MANCOVA, PCA, PLS, analyse canonique ou plus récemment l'ICA (Beckmann and Smith, 2003). La seconde famille d'approche regroupe toutes les méthodes de fouille de données, partant des techniques de clustering jusqu'aux algorithmes génétiques en passant par les réseaux neuronaux (Sommer and Wichert, 2003).

La fouille de donnée visuelle se focalise sur l'intégration de l'utilisateur, un expert-médecin, dans le processus de découverte de connaissance en utilisant des techniques de visualisation efficaces, d'interaction et de transfert de connaissances. Dans cet article, nous présentons une nouvelle approche interactive de fouille des images IRMf, guidée par les données. L'originalité de notre approche tient au fait que nous n'allons pas seulement mettre en oeuvre des techniques de fouille de données encore non appliquées au domaine de l'imagerie fonctionnelle, mais surtout que celles-ci vont être étendues par l'injection de connaissance à priori dans les mécanismes de fouille du système et l'interactivité, intégrant directement l'expert-médecin dans le processus de découverte et d'apprentissage de concepts pour mettre en évidence les zones fonctionnelles du cerveau et leur organisation.

Les algorithmes actuels de fouille de données ne supportent que de façon très limitée les mécanismes de guidage et d'engagement d'expert dans le processus de découverte. Les algorithmes de classification classiques requièrent des paramètres à définir à l'initialisation qui influencent de manière importante le résultat final. A contrario, l'approche visuelle autorise une gestion dynamique des paramètres de classification en intégrant l'utilisateur dans le processus de découverte de connaissance. Le principe de base consiste à visualiser dynamiquement l'apparition des classes, filtrer et détailler les plus prometteuses et affiner le processus de découverte. Les présentations visuelles augmentent la perception de l'utilisateur sur le déroulement de la classification et permettent d'arriver plus rapidement à des résultats concluants. Les régions intéressantes sont repérées immédiatement par l'expert qui peut agir sur les classes découvertes avec une précision désirée.

L'article est structuré de la façon suivante. Dans le chapitre 2 nous présentons les caractéristiques des données IRMf. Dans le chapitre 3 nous décrivons notre approche de fouille interactive d'images IRMf en particulier les techniques d'engagement d'expert dans le processus d'exploration d'images IRMf. Le chapitre 4 détaille les aspects d'interactivité de l'algorithme de classification neuronale : Gaz Neuronaux Croissants. Nous concluons dans le dernier chapitre en dégageant quelques perspectives.

2 IRMf comme objet complexe

L'activité cérébrale qui se caractérise par des phénomènes électriques rapides, peut être mesurée directement avec les techniques d'électro-encéphalographie (EEG) et de magnéto-encéphalographie (MEG). Si ces deux techniques ont une très bonne résolution temporelle, leur résolution spatiale par contre est quasi-nulle. Pour pouvoir savoir où et quand a lieu l'activité cérébrale, on a recours à l'Imagerie par Résonance Magnétique fonctionnelle (IRMf). Cette technique est une mesure indirecte car elle n'est pas sensible à l'activité même des neurones, mais à la consommation d'oxygène qui en dépend. L'effet BOLD (eng. *Blood Oxygenation Level Dependent*) est le phénomène qui permet

la mesure de cette consommation, il se traduit par une évolution du signal RMN induite à la fois par des variations de concentration d'oxygène et des variations de débit sanguin. L'IRMf consiste donc en une IRM optimisée sur l'effet BOLD. La mesure de l'activité d'un cerveau revient à enregistrer des images IRM à la suite les unes des autres pendant que ce cerveau travaille. Un tel enregistrement est donc composé d'une série d'images en trois dimensions d'un même cerveau immobile, et dont les variations d'intensité d'une image sur l'autre sont liées avec l'activité de ce dernier. Les zones actives sont ainsi localisées aussi bien spatialement que temporellement avec des résolutions typiques de 3 mm et 3 s. A ces échelles, le signal IRMf d'un voxel représente l'activité d'un agrégat de millions de neurones moyenné sur un intervalle de temps très grand devant le signal isolé d'un seul neurone.

L'IRMf permet de localiser précisément chez un sujet des zones liées à un fonctionnement ou à un dysfonctionnement donné. Le cerveau étant toujours en activité et le rapport signal sur bruit d'IRMf étant très faible, cette technique a recours systématiquement à un paradigme pour faire ressortir un effet escompté. Le paradigme est une représentation du fonctionnement de l'ensemble de l'acquisition. Intégrant la séquence de stimuli ainsi que les paramètres d'acquisition, le paradigme doit être optimisé de manière à maximiser les contrastes entre les différentes conditions du test afin de relever au mieux leurs effets.

Les données IRMf suivent principalement cinq grands axes de complexité. Elles sont à la fois des données en trois dimensions spatiales plus une dimension temporelle, elles peuvent correspondre à plusieurs sujets ou bien à l'historique d'un même patient, enfin elles peuvent être combinées avec des connaissances anatomiques ou plus généralement médicales. De plus, une acquisition IRMf génère une série de 100 à 1000 images IRM, formant une série IRMf de taille comprise entre 25 Mo et 1 Go.

Les bruits durant l'acquisition IRMf sont nombreux et très variés. On trouve tout d'abord les bruits physiques instrumentaux classiques comme par exemple des dérives ou des parasites électroniques qui sont fréquents dans les conditions instrumentales extrêmes de l'IRM où il s'agit de capter des signaux RF très faibles au sein de champs magnétiques très intenses. Autre source de bruit, les mouvements du sujet qui génèrent des signaux parasites, ce sont non seulement les déplacements de la tête et des yeux, mais aussi des déplacements internes induits par les pulsations cardiaques à proximité des veines. Enfin les activités cérébrales sans relation avec le paradigme comme des souvenirs ou des envies sont également générateurs de bruit.

Le protocole expérimental d'IRMf classique consiste à définir un paradigme du fonctionnement du cerveau pour une expérience donnée. Cela consiste à définir dans le même temps, le programme des tâches à soumettre au cerveau, aussi bien qu'un modèle de la réponse hémodynamique pour chacune de ces tâches. La réponse mesurée expérimentalement est ensuite comparée statistiquement avec ce modèle proposé *ab initio*. La méthode statistique est performante mais ne peut conclure en dehors du modèle préfixé. Avec cette méthode les résultats doivent être forcément anticipés, ce qui n'est pas toujours possible. Le concept de fouille de donnée, peut se révéler utile en complément ou en remplacement de la méthode classique lorsqu'il est délicat de prévoir ce qui va se passer durant l'acquisition. La démarche proposée ici est d'utiliser des méthodes d'exploration de données au sein d'un système interactif pour faire ressortir

les zones actives sans avoir recours à un modèle.

3 Engagement d'expert dans le processus de découverte

La fouille interactive d'images IRMf n'est pas basée sur un modèle mais au contraire est une approche guidée par les données qui de plus intègre directement un expert dans le processus de découverte. L'architecture de notre système de fouille d'images IRMf et le processus d'exploration sont détaillés dans (Korczak et al., 2005). Ce processus est composé de cinq phases que sont l'acquisition et la sélection des données, leurs pré-traitements, la classification, l'extraction de règles et de concepts, enfin la validation. Les données IRMf sont issues de l'acquisition de 100 à 1000 images IRM séquentielles d'un cerveau, l'enregistrement ayant lieu alors que ce dernier effectue une série de tâches programmées. La préparation des données consiste à choisir des attributs mettant en forme des vecteurs de manière à établir une relation de distance sur les données. Une acquisition de m images de taille n voxels, génère un ensemble de données de n individus. Un individu est un point dans un espace à m dimensions dont chaque coordonnée est l'écart de l'intensité du voxel considéré par rapport à sa moyenne sur la série.

Le système permet de faire une fouille de données interactivement afin de pouvoir optimiser l'expérience et la renouveler aussitôt sur le sujet. Les images IRMf peuvent être classifiées à partir de séries complètes ou en cours d'acquisition avec l'engagement de l'expert médecin dans le processus de découverte. Dans ce dernier mode, en fonction des résultats obtenus, l'expert-médecin peut modifier l'acquisition d'un point de vue purement technique par des paramètres géométriques ou temporels (résolution, zoom, etc). Cette réaction sera influencée par un retour d'information uniquement lié à l'état des signaux. Les algorithmes de classification et d'explication symbolique des classes aident l'expert-médecin à comprendre les classes générées et à modifier l'expérience directement au niveau du paradigme. Ces modifications sont suscitées par les classifications générées par le système de fouille de donnée, et choisies par l'expert en fonction d'hypothèses sur le fonctionnement cérébral ou à partir de données connues comme des données anato-fonctionnelles préalablement saisies dans la base. Pour arriver à ce stade, le système doit être très performant, en retournant des informations à un niveau d'interprétation suffisamment élevé. Le système est capable d'apprendre et stocker des connaissances acquises pendant les expérimentations et de s'en resservir comme connaissances préalables dans de futurs examens. Cette imagerie fonctionnelle interactive demande une analyse très rapide de résultats préliminaires pendant l'acquisition, cette analyse n'est envisageable que par une méthode spécifique qui s'insère dans la démarche actuelle.

Le travail présenté dans cet article est centré essentiellement sur la phase d'exploration, dont la démarche se veut interactive et dynamique autour d'algorithmes de classification non supervisée. La classification consiste à décrire des voxels en les distribuant en un nombre limité de classes homogènes. Ces classes regroupent les voxels qui ont des caractéristiques et comportements similaires. La classification est dite "dirigée par les données" car le processus n'utilise pas de connaissance externe aux données et

ne dispose que de leur description pour extraire des informations sur la structure de l'ensemble.

Les méthodes de classification non supervisée (Bock and Diday, 2000) ont déjà été appliquées à l'IRMf à travers plusieurs méthodes dérivées de statistique, de logique floue ou encore de classification neuronale et hiérarchique (Dimitriadou et al., 2004), (Goutte et al., 1999), (Moller et al., 2001), l'Analyse en Composante Principale (ACP) (Andersen et al., 1999), (Lai and Fang, 1999), et l'Analyse en Composantes Indépendantes (ACI) (Esposito et al., 2002).

Dans nos travaux, nous nous sommes plus particulièrement focalisés sur les méthodes neuronales : l'algorithme de Kohonen (Kohonen, 1982), (Fischer and Hennig, 1999) (Ngan and Hu, 1999) et l'algorithme des Gaz Neuraux Croissants (GNC) (Fritzke, 1995). Suite à nos expérimentations, nous favorisons l'algorithme GNC qui s'adapte bien à l'exploration visuelle interactive d'images IRMf. Dans cet article, nous ne considérerons donc que cet algorithme. L'originalité et l'intérêt des GNC résident dans le fait que le nombre de classes n'est pas fixé à l'avance contrairement à la plupart des autres méthodes. Le nombre de classes peut tout aussi bien augmenter que diminuer durant l'exécution, rendant l'algorithme particulièrement flexible. Nous avons rendu la procédure de classification interactive, de manière à engager l'expert dans le processus de découverte. L'utilisateur dispose de retours et de contrôles sur l'évolution de la classification, pouvant ainsi suivre et intervenir pour orienter le déroulement. Cette exploration visuelle en IRMf dispose des outils de vision 3D, des retours statistiques et un algorithme de classification suffisamment souple et dynamique pour supporter l'interaction [Fig. 1]. Nous avons étendu le système de visualisation SLICER (<http://www.slicer.org/>) en y intégrant des méthodes de classification ainsi que des outils d'exploration d'images IRMf. Les classes produites par l'algorithme sont évaluées par l'expert qui ne retient que les classes pertinentes. Chaque classe correspond à un ensemble de voxels, régions du cerveau, ayant la même réponse hémodynamique au cours du temps. Les caractéristiques des classes sont affichées par le système durant le processus d'exploration ; la figure [Fig. 1] montre un exemple de classe relevant d'une activation de type bloc associée aux statistiques sur le taux d'échanges de voxels entre classes informant sur la stabilité et la convergence de l'algorithme. Cette réponse peut être caractérisée explicitement par la construction de règles. Ces règles combinent des motifs temporels observés avec des informations spatiales telles que l'activité des voxels ou des régions voisines, en plus des connaissances du domaine, comme l'atlas des fonctions des régions du cerveau. Ces motifs peuvent être synchronisés avec le paradigme, par exemple pour découvrir l'interaction entre les régions du cerveau utilisées pour la mémoire visuelle. D'autres motifs temporels peuvent être indépendants de tout paradigme, par exemple pour mettre en évidence la succession d'activations de régions du cerveau typiques de l'hallucination.

4 Classification interactive avec GNC

Afin d'illustrer l'exploration interactive d'images, nous avons choisi GNC parmi les algorithmes implantés dans notre système (Korczak et al., 2005). Dans cet algorithme, des classes peuvent apparaître ou disparaître au cours du temps sous l'influence d'er-

Exploration visuelle d'images IRMf basée sur des Gaz Neuraux Croissants

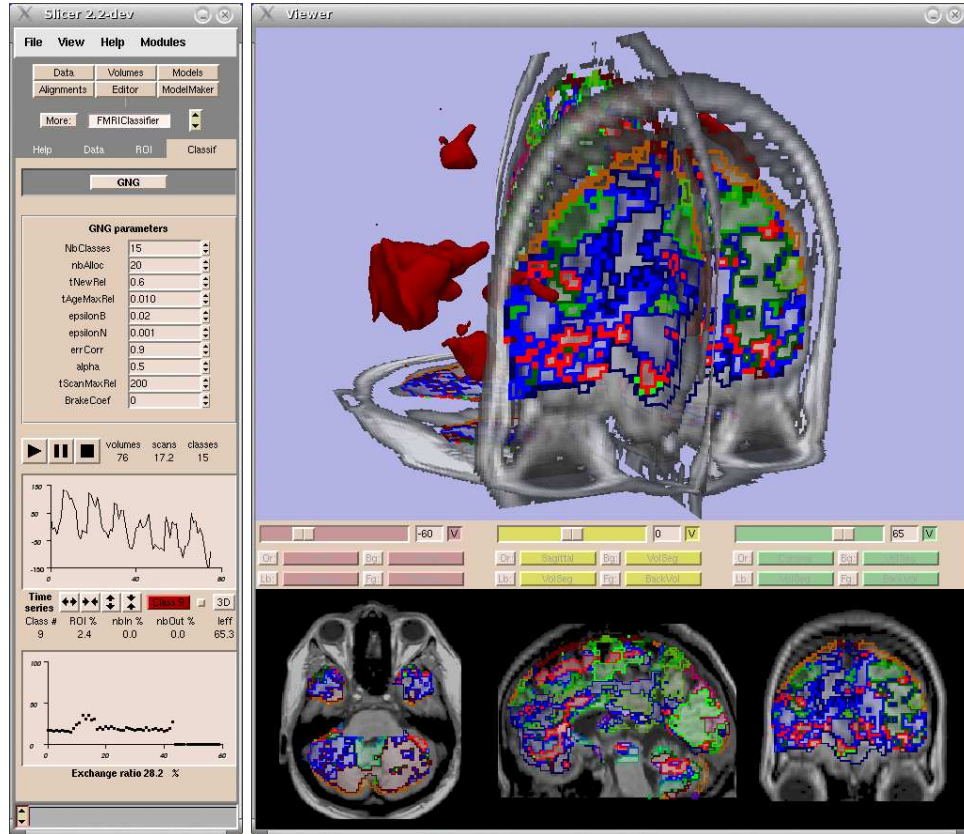


FIG. 1 – Exemple d'exploration interactive d'images IRMf

reurs évaluées. Ceci donne à GNC une souplesse qui lui permet de s'adapter à une démarche interactive, dans laquelle l'évolution de la classification est également influencée par un expert. Il s'agit d'un algorithme dynamique de réseau de neurones inter-connectés. Les connexions entre classes ont la propriété de vieillir et disparaissent lorsqu'elles atteignent un âge maximal prédéfini. Cette propriété est la cause de la disparition des classes qui sont éliminées lorsqu'elles ne sont plus connectées. L'apparition des classes se fait à intervalles de temps régulier en insérant une nouvelle classe auprès de la classe qui présente la plus grande erreur. L'erreur d'une classe est une évaluation basée sur la somme des distances de cette classe aux individus qui l'ont déclarée gagnante. Ainsi le réseau se renforce auprès des classes qui parcourent les plus grandes distances ou qui réunissent le plus d'individus. Nos contributions portent sur deux aspects de classification, notamment sur la réduction de l'espace des données et une aide au paramétrage de l'algorithme assisté par des outils statistiques et de fouille de données visuelle.

L'algorithme des Gaz Neuraux Croissants adapté aux données IRMf a été spécifié

dans (Korczak et al., 2005). En bref, la modification des classes se fait itérativement à l'échelle de la donnée ; on adapte la classe la plus proche ainsi que les classes voisines en fonction de la dernière donnée tirée. En modifiant de la sorte plusieurs classes simultanément, on accélère le processus et l'on donne une cohésion à l'ensemble des classes. De plus, les adaptations sont réglables par des coefficients qui prennent en compte le degré de voisinage. La contre partie de cette organisation en réseau de classes est qu'elle peut générer des classes vides.

Nous introduisons une pause dans l'algorithme durant laquelle le GNC est dans un état cohérent. Ceci permet d'effectuer des mesures sur cet état, de le modifier, de le sauver ou encore de le restaurer. Cette pause est ainsi invoquée systématiquement à intervalles de temps réguliers pour effectuer des mesures absolues sur la classification et alimenter les outils de visualisation. Elle est aussi utilisée à la demande lorsque l'expert intervient sur le cours de la classification. Ce dernier a donc accès et peut modifier tout point de la classification durant l'exécution. Tout peut être modifié sans avoir besoin d'arrêter et de relancer le processus. Les paramètres de l'algorithme ainsi que l'espace des données peuvent donc être optimisés et varier durant la classification. Quatre niveaux d'interaction peuvent être discernés. Le niveau minimum correspond à un retour de mesures sur l'état courant et permet à l'utilisateur de pouvoir décider quand interrompre le processus. Le deuxième niveau est la modification du volume des données, ce niveau est rendu possible lorsque l'adaptation des classes est indépendante de l'historique de la classification. Le troisième niveau est accessible aux algorithmes qui disposent de paramètres non nécessairement fixés. Dans ces conditions l'utilisateur peut les faire varier pour les optimiser en cours d'exécution. Au delà de ces trois niveaux d'interaction, une gestion de sauvegarde et de restauration des états de la classification constitue un dernier aspect d'interaction avec la classification.

L'expert-médecin peut suivre l'évolution de la classification au cours du temps à partir d'outils visuels alimentés à intervalles de temps réguliers par l'algorithme. Les informations fournies sont de trois ordres : statistiques, temporelles et spatiales. Les statistiques portent sur la classification en tant que telle, quant aux informations temporelles et spatiales, elles sont liées à la nature de l'IRMf. Les informations statistiques décrivent l'évolution de la classification par des mesures sur les classes générées. L'utilisateur peut suivre à partir de graphiques, l'évolution des erreurs, des inerties intra-classe et inter-classes, du nombre de voxels réunis au sein de chaque classe et du nombre de voxels qui changent de classe par unité de temps. L'ensemble de ces informations renseigne notamment sur la dispersion des classes, la convergence et la stabilité de la classification [Fig. 1, partie gauche].

Les informations temporelles et spatiales véhiculées dans chaque classe, représentent l'évolution des signaux IRMf au cours du temps et leurs localisations dans le cerveau. Les signaux sont affichés sous forme de courbes dans des graphiques où l'on peut également faire figurer le paradigme, soit par la séquence de stimuli, soit par un modèle de la réponse lorsqu'il existe. La visualisation spatiale est effectuée simultanément de deux manières, à partir de trois coupes perpendiculaires et directement en trois dimensions [Fig. 1, partie droite].

La classification en cours alimente un volume où les couleurs des voxels sont associées aux classes. Cette vision 3D de la classification peut être superposée suivant les

cas, à un des volumes de la série d'acquisition, à une IRM structurale du sujet ou à un volume normalisé faisant office d'atlas anatomique.

Toutes ces informations sont rafraîchies régulièrement durant la classification permettant à l'utilisateur de suivre et de guider en direct l'évolution du processus. Il opère ainsi une visualisation active, en affichant les aspects qu'il juge opportuns et en intervenant sur le cours de la classification. Les leviers dont il dispose pour influencer le processus sont les paramètres de l'algorithme et le volume des données qu'il peut modifier quand il le désire.

Dans le mode d'exploration, l'expert-médecin peut facilement optimiser l'algorithme, par l'observation en direct des mesures statistiques et de leurs variations induites par des modifications de paramétrage. Ainsi, le réglage de l'algorithme peut être convenablement effectué sans avoir besoin d'arrêter et de recommencer sans cesse. D'autre part, les paramètres peuvent être adaptés en fin de classification pour stabiliser ou affiner un résultat.

Dans la découverte de régions actives du cerveau, les moyens d'intervention de l'expert-médecin se situent au niveau du réglage du nombre de classes nécessaires et de la sélection des régions explorées. Ainsi, l'augmentation du nombre de classes se fait naturellement, et sa réduction peut être réalisée directement par l'utilisateur qui a la possibilité d'éliminer des classes. Notons au passage que la suppression d'une classe est accompagnée d'une gestion appropriée des connexions afin de maintenir la cohérence du réseau. En définitive pour adapter au mieux le nombre de classes aux données IRMf, l'expert-médecin fixe la valeur du nombre maximal de classes en observant l'apparition des nouvelles classes durant l'exploration des images. Il peut par exemple décider de stopper cette croissance après l'apparition d'une classe escomptée ou bien revenir en arrière par l'élimination d'une classe de son choix.

Pour augmenter la vitesse d'exécution et diminuer la complexité des résultats, l'utilisateur a toujours intérêt à limiter au maximum le volume des données. L'espace d'exploration peut être ajusté au moyen de plusieurs techniques. Un outil inclus dans la visualisation 3D, permet de sélectionner des volumes d'intérêt ou de désintérêt. L'espace peut aussi être restreint à des structures anatomiques. Dans tous les cas un seuillage permet d'éliminer l'espace qui entoure le crâne. Et dans le cas d'images normalisées, il est possible de limiter les recherches à la matière grise. Les résultats intermédiaires de classification sont aussi utilisés pour la sélection des données. Ainsi, l'expert-médecin peut poursuivre une classification en focalisant son intérêt sur certaines classes après avoir éliminé l'ensemble des voxels des autres classes. L'augmentation de l'espace de recherche, également possible, est parfois désiré. Le cas se produit lorsque l'on veut étendre une classification à d'autres zones, par exemple après avoir optimisé certains paramètres de l'algorithme sur une région volontairement réduite pour gagner du temps.

Des expérimentations ont été effectuées sur un jeu de données caractéristique : une série IRMf de tests auditifs. Ces données proviennent du site de l'institut de recherche londonien "The Wellcome Department of Imaging Neuroscience" ¹. Les résultats obtenus (Korczak et al., 2005) sont encourageants. Le système identifie les régions pertinentes et facilite le guidage interactif de processus de fouille.

¹<http://www.fil.ion.ucl.ac.uk/spm/>

5 Conclusion

Dans cet article, nous avons présenté une approche d'exploration visuelle d'images médicales avec un engagement d'expert dans le processus de découverte. Les séries d'images 3D IRMf ont été considérées comme objets complexes en prenant en compte le volume, la dimension temporelle, les relations spatiales et le bruit. Nous avons détaillé la partie visualisation de notre système de fouille de séquences d'images 3D IRMf. Dans notre approche, nous avons mis en avant l'implication d'un expert-médecin dans le processus de classification d'une part, et d'autre part, l'intégration d'un algorithme de classification capable de s'adapter aux spécificités de l'IRMf avec de faibles connaissances préalables. La méthode d'exploration visuelle est indépendante de l'algorithme de classification utilisé. Plusieurs algorithmes ont été étudiés : K-means, LBG, SOM et GNC ; le modèle des Gaz Neuraux Croissants s'est avéré plus robuste et reproductible dans les mêmes conditions de fonctionnement et surtout plus adapté à l'exploration visuelle interactive. Tous ses paramètres peuvent être modifiés en cours d'exécution y compris le paramètre du nombre de classes. Cet aspect dynamique est déterminant pour l'exploration interactive. L'outil de fouille de données visuelle qui en découle permet notamment d'optimiser les paramètres de l'algorithme, de réduire l'espace de recherche, de présenter le processus de découverte sous plusieurs angles et participe ainsi à la diminution de la complexité des objets concernés pour fournir à l'expert-médecin des classifications de régions cervicales plus compréhensibles. Les premiers résultats sur des données de type bloc sont encourageants et nous permettent d'envisager la poursuite de ces travaux vers des expériences de type événementiel. Il s'avère que l'exploration visuelle permet d'accélérer le processus d'exploration d'images IRMf et aboutit à de meilleurs résultats dotés d'une confiance accrue.

Remerciements. Les auteurs remercient K. Friston et G. Rees pour les données de test SPM, ainsi que les étudiants de l'Université Louis Pasteur de Strasbourg, France, H. Hager, P. Hahn, V. Meyer, J. Schaeffer et O. Zitvogel, pour leur participation dans la phase initiale de réalisation du projet.

Références

- Andersen, A. H., Gash, D. M., and Avison, M. J. (1999), Principal component analysis of the dynamic response measured by fMRI : a generalized linear systems framework. *Magnetic Resonance Imaging*, 17 :795–815, 1999.
- Beckmann, C. and Smith, S. M. (2003), Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. on Medical Imaging*, 2003.
- Bock, H. and Diday, E. (2000), *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*, Springer Verlag, 2000.
- Cox, R. W. (1996), AFNI : Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29 :162–173, 1996.
- Dimitriadou, E., Barth, M., Windischberger, C., Hornick, K., and Moser, E. (2004), A quantitative comparison of fMRI cluster analysis. *AI in Medicine*, 31 :57–71, 2004.
- Esposito, F., Formisano, E., Seifritz, E., Goebel, R., Morrone, R., Tedeschi, G., and

- DiSalle, F. (2002), Spatial independent component analysis of fMRI time-series : To what extent do results depend on the algorithm used? *Human Brain Mapping*, 16 :146–157, 2002.
- Fischer, H. and Hennig, J. (1999), Neural network-based analysis of MR time series. *Magnetic Resonance in Medicine*, 41 :124–131, 1999.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. J. (1995), Statistical parametric maps in functional imaging : A general linear approach. *Human Brain Mapping*, 2 :189–210, 1995.
- Fritzke, B. (1995), A growing neural gas network learns topologies. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 625–632. MIT Press, Cambridge MA, 1995.
- Goebel, R. (1997), *Brain Voyager : Ein Programm zur Analyse und Visualisierung von Magnetresonanztomographiedaten*. T. Plesser and P. Wittenburg, Forschung und Wissenschaftliches Rechnen, 1997.
- Goutte, C., Toft, P., Rostrup, E., Nielsen, F. A., and Hansen, A. K. (1999), On clustering fMRI time series. *NeuroImage*, 9 :298–310, 1999.
- Kohonen, T. (1982), Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43 :59–69, 1982.
- Korczak, J., Scheiber, C., Hommet, J., and Lachiche, N. (2005), Fouille interactive de séquences d'images 3D d'IRMf. *RNTI*, 2005.
- Lai, S. H. and Fang, M. (1999), A novel local PCA-based method for detecting activation signals in fMRI. *Magnetic Resonance Imaging*, 17 :827–836, 1999.
- Moller, U., Ligges, M., Grunling, C., Georgiewa, P., Kaiser, W. A., Witte, H., and Blanz, B. (2001), Pitfalls in the clustering of neuroimage data and improvements by global optimization strategies. *NeuroImage*, 14 :206–218, 2001.
- Ngan, S. C. and Hu, X. (1999), Analysis of functional magnetic resonance imaging data using self-organizing mapping with spatial connectivity. *Magnetic Resonance in Medicine*, 41 :939–946, 1999.
- Sommer, F. T. and Wichert, A. (2003), *Exploratory Analysis and Data Modeling in Functional Neuroimaging*. The MIT Press, 2003.

Summary

Current data mining algorithms do not contain sophisticated guiding and engagement of expert facilities. In this paper a new interactive data-driven approach of brain fMRI images is presented. Discrimination of the image voxels of the brain that represent real activity is, in general, very difficult because of a weak signal-to-noise ratio and of the presence of artifacts. Visual exploration combines the expert involvement in brain mining process, interactive visualisation techniques, and the use of domain knowledge. As an example of unsupervised classification an algorithm of Growing Neural Gas has been developed and tested on sequences of fMRI images. The results of the tests have shown a strong influence on the classifier performances by the number of classes, signal-to-noise ratio, and volumes of activated and explored zones. The interface allows not only to accelerate the classification process but also to increase the user confidence in generated hypotheses.

Mise en évidence d'invariants dans une population de cas chirurgicaux

Mélanie Raimbault*, Ricco Rakotomalala**
Xavier Morandi*,, Pierre Jannin*

*Laboratoire IDM, 2 avenue du Pr. Léon Bernard, 35043 Rennes
pierre.jannin@univ-rennes1.fr
<http://idm.univ-rennes1.fr>

**Laboratoire ERIC, 5 avenue Pierre Mendès France, 69676 Bron
Ricco.Rakotomalala@univ-lyon2.fr
<http://eric.univ-lyon2.fr>

***Hôpital Pontchaillou, Service de neurochirurgie, 35000 Rennes

Résumé. Ces dernières années, les progrès en informatique et en imagerie numérique ont fait émerger une nouvelle discipline, la chirurgie assistée par ordinateur. Les systèmes de chirurgie assistée par ordinateur contribuent à l'amélioration du déroulement des procédures chirurgicales. Un des objectifs à long terme de nos travaux est de proposer des solutions d'amélioration de ces systèmes, basées sur les connaissances du chirurgien quant au déroulement de la procédure, par l'utilisation d'un modèle générique qui permet de capturer et de représenter ces connaissances. Cet article présente une méthodologie d'exploitation d'un ensemble de cas chirurgicaux décrits à l'aide de ce modèle générique, par des algorithmes issus de l'extraction de connaissance à partir de données, afin de mettre en évidence des invariants dans les descriptions structurées du déroulement des cas chirurgicaux. Il détaille en outre les difficultés rencontrées de par notamment le caractère complexe des données étudiées.

1 Introduction et contexte

Les continus progrès de l'informatique, l'amélioration des techniques d'imagerie numérique et la banalisation de l'usage des ordinateurs au sein de l'univers chirurgical, ont participé à l'émergence d'un nouveau domaine : la chirurgie assistée par ordinateur. Un éventail d'expertises et de compétences en biologie, en médecine, en sciences de l'ingénieur et en sciences informatiques participe à l'amélioration constante de ce domaine de recherche (Shahidi et al 2001). Les bénéfices apportés par l'utilisation de tels systèmes ont déjà été mis en évidence dans de nombreuses disciplines chirurgicales, en particulier la neurochirurgie, et notamment la neurochirurgie guidée par l'image ou neuronavigation. Ces systèmes contribuent à rendre la chirurgie plus sûre et moins invasive.

Un des objectifs de nos travaux est de proposer une solution permettant l'amélioration d'une ou plusieurs phases du processus de neurochirurgie guidée par l'image, par la modélisation des connaissances du neurochirurgien quant à son expérience, acquise pendant sa formation et tout au long de sa pratique chirurgicale. Notre démarche

générale de modélisation s'inscrit sur plusieurs niveaux. Notre but est de capturer et de représenter les connaissances des neurochirurgiens quant à la préparation et à la réalisation des procédures chirurgicales, dans le cadre restreint d'une neurochirurgie guidée par l'image. Cette représentation nous permettra de disposer d'une information *a priori* sur la chirurgie à réaliser et par là même nous permettra d'améliorer la préparation et la réalisation de l'intervention.

Contrairement à d'autres spécialités chirurgicales, il est difficile de modéliser spontanément la neurochirurgie à ciel ouvert, de par sa variabilité inter-patient et inter-chirurgical. La catégorisation en grandes familles de procédures neurochirurgicales reste toutefois pour nous un but important à atteindre : plus la description *a priori* de la structure d'une famille de procédures chirurgicales sera spécifique et détaillée, plus le neurochirurgien pourra adapter son geste en connaissance de cause. Nous avons donc dans un premier temps construit un modèle générique, qui définit les bases permettant de structurer, de décrire et d'étudier les procédures chirurgicales, et qui nous servira d'ontologie du domaine étudié (Jannin et al 2003). Le principe du modèle générique est de décomposer la procédure chirurgicale d'un patient donné en une séquence d'étapes principales définissant le scénario chirurgical de l'intervention. Le modèle assigne à chaque étape une liste de structures, représentées par des entités extraites des images multimodales préopératoires (i.e. anatomiques et/ou fonctionnelles et/ou pathologiques), et qui sont nécessaires au bon déroulement de l'étape. Le rôle de chaque structure dans une étape est identifié à partir d'une liste de valeurs prédéfinies. La figure 1 présente le modèle générique d'une procédure chirurgicale sous forme de diagramme de classe UML.

Nous avons ensuite décrit des cas concrets d'interventions cliniques, à l'aide de ce modèle générique. Nous pouvons ainsi comparer ces descriptions structurées entre elles pour extraire des ressemblances et des invariants. En appliquant notre méthodologie sur des données de cas chirurgicaux, nous désirons montrer qu'il est possible de prévoir au moins partiellement, à l'aide de méthodes issues de l'Extraction de Connaissances à partir de Données ou ECD (Hand et al 2000), le déroulement d'une intervention chirurgicale, à partir de certains paramètres ou variables qu'il nous faudra choisir et déterminer. Dans le cadre de l'étude décrite dans cet article, nous allons étudier uniquement un type de procédure chirurgicale : la chirurgie de tumeur supratentorielle (TST). La chirurgie des tumeurs intraparenchymateuses a un rôle essentiel dans le traitement des tumeurs intracrâniennes. Selon le type et la localisation de la tumeur, elle suffit souvent à traiter certaines tumeurs bénignes, qui répondent mal à d'autres formes de traitement comme la radiothérapie ou la chimiothérapie. Dans le cas de tumeurs malignes, la chirurgie contribue à une réduction tumorale importante, améliore la survie du patient et sa qualité de vie.

Dans cet article, nous allons tout d'abord présenter la méthodologie employée pour passer du modèle générique aux variables exploitées par les algorithmes d'ECD. Nous allons ensuite présenter les méthodes d'ECD que nous avons utilisées, avant de détailler les résultats obtenus, et de discuter de la pertinence de ces résultats et des difficultés rencontrées, dues au caractère complexe des données.

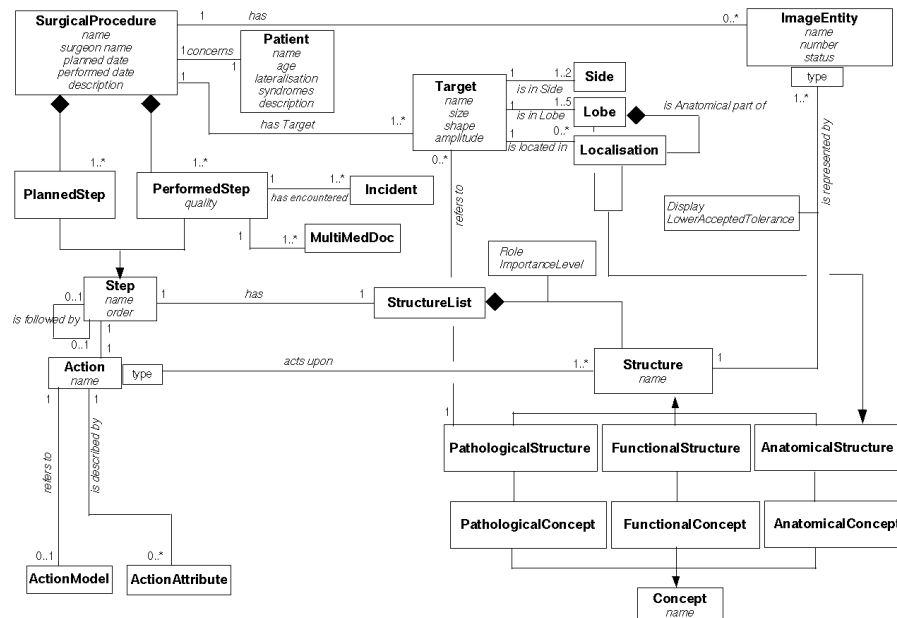


FIG. 1 – Diagramme de classe UML présentant la structuration des données.

2 Matériel et méthodes

2.1 Du modèle vers les variables

Les données étudiées sont 159 cas de chirurgies de tumeur supratentorielle ou TST, qui décrivent sous une forme structurée et hiérarchisée (fichiers XML contraints par une DTD) le déroulement de procédures chirurgicales réalisées sur un groupe de patients droitiers, opérés au service de neurochirurgie de l'Hôpital Pontchaillou de Rennes, entre 1995 et 2003. Le recueil des données a été fait sous forme d'interview du neurochirurgien pendant la préparation de l'intervention, ou sous forme d'interview du neurochirurgien s'appuyant sur le compte-rendu opératoire de l'intervention.

Ces interviews ont été décomposées à la volée sous forme structurée en utilisant des formulaires. Les patients étudiés ont été choisis explicitement selon certains critères de correspondance au domaine d'étude. Ils répondent à la pathologie étudiée et ne sont pas sujets à interprétation pour la saisie ou traduction sous forme structurée. Deux chirurgiens principaux ont participé à l'étude : un chirurgien expérimenté et un chirurgien novice, qui fut pendant un temps l'élève du chirurgien expérimenté.

Les cas chirurgicaux contenus dans l'ensemble d'étude sont décrits par les différents concepts du modèle générique. Chacun de ces concepts correspond à une variable qu'il est possible d'étudier. Nous les répartissons en deux catégories définies de la façon suivante :

- Les variables de catégorie I correspondent aux concepts du modèle générique

décrivant le cas chirurgical. Ces concepts représentent les informations sur la nature pathologique de la tumeur que l'intervention vise à enlever, la localisation de cette tumeur (lobe, gyrus), sa profondeur, l'hémisphère dans lequel elle est située, etc ;

- Les variables de catégorie II correspondent aux concepts du modèle générique décrivant le déroulement de la procédure chirurgicale. Ces concepts représentent les informations sur les différentes étapes réalisées, les actions correspondantes ainsi que les attributs et modèles d'action, les structures anatomiques, pathologiques ou fonctionnelles dont la connaissance a été jugée nécessaire pour réaliser l'étape en question, etc.

Les valeurs des variables de catégorie I, appelées dans notre contexte variables prédictives, sont connues avant la réalisation de la procédure chirurgicale. Ce sont ces valeurs qui vont déterminer, dans une mesure que nous cherchons à quantifier, le déroulement de la procédure chirurgicale. Les valeurs des variables de catégorie II, ou variables à prédire, ne sont connues qu'une fois la procédure chirurgicale réalisée. Ce sont ces valeurs que nous cherchons à prédire à partir de la connaissance des variables de catégorie I.

Notre problématique consiste à faire des choix de représentation plate pour transcrire nos données structurées et hiérarchiques, afin de pouvoir appliquer à ce tableau des algorithmes issus de l'ECD. Notre niveau d'individu statistique est le patient, ce qui veut dire que nous devons toujours ramener sur une même ligne les informations concernant le même patient. Si des informations concernant le même patient sont réparties sur plusieurs lignes, les calculs statistiques réalisés ne peuvent plus être ramenés au patient et l'interprétation des résultats n'est pas possible.

Tout le problème revient en fait à trouver une représentation plate pour chacune des variables qui minimise la perte d'information, sans pour autant noyer l'étude dans un nombre trop élevé de variables différentes et sans trop fragmenter l'échantillon. Il n'est évidemment pas intéressant d'étudier des répartitions avec un faible nombre d'occurrences. De plus, les variables devront être le plus décorréélées possible, afin de ne pas noyer l'information dans du bruit de fond et de ne récolter que des évidences lors de l'extraction.

Presque toutes les variables de catégorie I (à l'exclusion de TargetLobe et TargetLoc dont nous parlons plus bas) sont des variables nominales. Leur transformation ou recodage sous forme plate est donc immédiate et ne pose pas de problème particulier.

TargetLobe décrit le ou les lobes dans lesquels peut être située la pathologie cible, dans notre cas la tumeur. Notre échantillon présente des cas chirurgicaux dont la tumeur peut être située jusqu'à dans trois lobes différents, parmi les 4 lobes cérébraux : temporal, frontal, occipital et pariétal. La solution de traduction dans le cas de TargetLobe reste donc simple : la variable est remplacée par quatre variables binaires de présence/absence (TemporalLobe, FrontalLobe, OccipitalLobe, et ParietalLobe). Toutefois, il faut noter une limitation à cette représentation, induite par la relation de dépendance entre les quatre variables fondée sur leur voisinage anatomique.

TargetLoc décrit plus précisément, au niveau gyrus et lorsque c'est possible, au niveau pars de gyrus, la localisation de la tumeur : par exemple, pour un cas chirurgical de notre ensemble d'étude, la tumeur d'un patient située dans le lobe temporal au

niveau TargetLobe, peut être située plus précisément à la fois dans le pars postérieur du gyrus temporal supérieur et dans le pars postérieur du gyrus temporal intermédiaire. Il existe autant de valeurs possibles pour TargetLoc qu'il existe de gyrus, de pars de gyrus ou de sillon cortical. Nous n'avons pas à l'heure actuelle trouvé de représentation plate satisfaisante pour TargetLoc. Le faible nombre d'individus dans notre échantillon ne nous permet pas d'utiliser la même technique que pour TargetLobe : on obtiendrait alors actuellement 32 variables présence/absence (et la liste n'est pas exhaustive), avec pour chacune un faible nombre d'individus ayant la variable présente.

Les variables de catégorie II sont les variables qui posent problème quant à leur traduction sous forme tabulaire. Il existe un problème de dépendance fonctionnelle entre les différentes variables de catégorie II de par leur nature et leur structure hiérarchique. Les données représentées par les variables de catégorie II sont complexes. Nous avons fait certains choix de représentation des variables de catégorie II d'après nos réflexions suite à l'analyse en statistique descriptive opérée sur l'échantillon. Nous n'avons pas cherché à traduire sous forme tabulaire toutes les variables de catégorie II issues du modèle générique, mais uniquement dans un premier temps celles qui nous paraissaient pertinentes par rapport aux questions posées et adaptées à notre contexte et à notre échantillon. Par exemple, nous avons vu précédemment lors de l'étude de notre échantillon que 100% des cas chirurgicaux présentaient comme première étape une étape de positionnement du patient. Par conséquent, une variable de catégorie II indiquant la présence ou l'absence d'une étape de positionnement du patient n'est pas intéressante dans notre contexte particulier.

En outre, il convient de remarquer que les variables de catégorie II ne sont pas à l'heure actuelle utilisables en tant que variables actives. Compte-tenu de leur dépendance fonctionnelle et des limitations inhérentes à nos choix de représentation sous forme tabulaire, les variables de catégorie II seront utilisées de manière illustrative : elles n'interviendront pas dans la classification des données mais seront disponibles à la consultation pour apporter des informations supplémentaires sur les classes obtenues.

Une fois établies les règles de transformation, le passage du formalisme XML à la représentation plate s'est effectué à l'aide de scripts de traduction automatique.

2.2 Méthodes utilisées

Notre objectif est de caractériser la description de la procédure chirurgicale à partir de la description de la pathologie et du patient, c'est à dire caractériser les variables de catégorie II à partir des variables de catégorie I. Nous avons utilisés deux méthodes différentes, une méthode de classification, et une méthode de prédiction. Avec la première méthode, nous allons chercher à catégoriser, avec la deuxième méthode, nous allons chercher à prédire.

La première approche est l'approche la plus facile à mettre en œuvre compte tenu du faible nombre d'individus dans l'échantillon. Elle consiste à effectuer une classification *mixte* de l'échantillon en utilisant les variables de catégorie I comme variables actives, et les variables de catégorie II comme variables illustratives. La classification *mixte* consiste en une classification de l'échantillon par l'algorithme *K-means* afin d'identifier le nombre de classes présentes dans l'échantillon, puis en une classification des individus dans les classes par classification ascendante hiérarchique. En étudiant ensuite les

caractéristiques particulières des individus répartis dans chaque classe, il sera alors possible de définir des règles de description de la classe, ainsi que des règles de prédiction des valeurs des variables de catégorie II à partir des valeurs des variables de catégorie I.

La deuxième approche utilisée consiste à prédire certaines variables de catégorie II à partir des variables de catégorie I, en utilisant des arbres de décisions basés sur l'algorithme CART (Zighed et al 2000). Se pose ici le problème du choix des variables de catégorie II que l'on souhaite prédire et de ce qu'elles représentent au niveau du déroulement de la procédure chirurgicale. En outre, se pose encore et toujours le problème du petit nombre d'individus et de la distribution asymétrique donc très instable de ces individus dans l'échantillon.

3 Résultats

3.1 Première méthode

Pour la première méthode, nous avons utilisé une classification par l'algorithme *K-means* afin de nous donner une idée du nombre de classes présentes dans notre échantillon de 159 cas chirurgicaux. Les variables de catégorie I sont utilisées comme variables actives et les variables de catégorie II sont uniquement intégrées en tant que variables illustratives, c'est à dire qu'elles n'interviennent pas dans l'algorithme de classification. On obtient le dendrogramme qui nous indique qu'il est possible de partitionner notre échantillon en 2, 3, 4, 6 ou 9 classes.

Nous avons ensuite utilisé une classification ascendante hiérarchique en fixant à 6 le nombre de classes que l'on désire obtenir. La répartition des individus dans chaque classe est détaillée dans le tableau 1. Pour chaque classe, on donne l'effectif et les variables de catégorie I les plus représentatives. Par variable représentative de la classe, on entend le couple (*variable, valeur*) pour lequel le pourcentage d'apparition dans la classe est sensiblement plus élevé que pour l'échantillon global. Les variables en italique sont les variables de catégorie II, elles n'ont qu'un rôle illustratif.

3.2 Deuxième méthode

Dans le cadre de la deuxième méthode, nous nous sommes intéressés au positionnement du patient. Le positionnement du patient est une problématique récurrente très importante en neurochirurgie (Sevach et al 1992). Cette position doit être choisie avant l'intervention, afin d'obtenir le meilleur angle de visibilité de la région d'intérêt et de minimiser la déformation cérébrale peropératoire (ou *brainshift*, terme anglo-saxon) en tenant compte des contraintes anesthésiques. Des discussions avec les neurochirurgiens et une étude manuelle préalable nous ont indiqué que le positionnement dépend de la localisation de la tumeur. Nous avons donc construit l'arbre de décision nous permettant de prédire le positionnement du patient à partir de la localisation de la tumeur, au niveau du lobe. Les éléments de notre ensemble sont les cas chirurgicaux décrits par les attributs LobeTemporal, LobeParietal, LobeOccipital et LobeFrontal. Nous avons utilisé l'algorithme CART, avec un échantillon d'apprentissage composé de 106 cas chi-

urgicaux, et un ensemble de test composé de 53 cas chirurgicaux. L'arbre de décision obtenu est illustré par la figure 2.

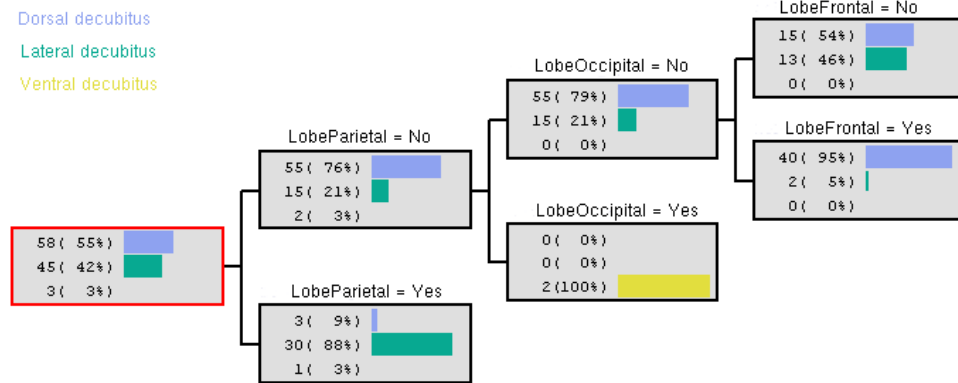


FIG. 2 – Arbre de décision pour les cas chirurgicaux décrits par quatre attributs *LobeTemporal*, *LobeParietal*, *LobeOccipital* et *LobeFrontal*, et pouvant appartenir à trois classes dorsal, latéral ou ventral décubitus.

Les règles de décision correspondant à cet arbre sont les suivantes :

SI *LobeParietal* = No ET *LobeOccipital* = No ET *LobeFrontal* = No ALORS Dorsal.

SI *LobeParietal* = No ET *LobeOccipital* = No ET *LobeFrontal* = Yes ALORS Dorsal.

SI *LobeParietal* = No ET *LobeOccipital* = Yes ALORS Ventral.

SI *LobeParietal* = Yes ALORS Lateral.

4 Discussion

Nous avons présenté des résultats obtenus par des méthodes d'extraction automatique sur un ensemble de cas chirurgicaux. Ces résultats peuvent apparaître facilement prédictibles et restent évidemment peu concluants vu le petit nombre d'individus peuplant l'ensemble étudié. Toutefois, ces résultats sont intéressants dans le sens où ils démontrent la pertinence d'une telle méthodologie : il est possible de mettre en évidence des invariants dans le déroulement d'une procédure chirurgicale à l'aide de méthodologies issues de l'ECD.

Les résultats de la première méthode nous apportent une classification de notre échantillon en six classes distinctes. Les variables de catégorie II ont été utilisées de façon illustrative, les classes obtenues représentent donc une agglomération des cas chirurgicaux par ressemblance entre la description du profil du cas, c'est à dire des variables de catégorie I. On observe toutefois en étudiant les variables de catégorie II

de chaque classe qu'il existe une ressemblance au sein des variables de catégorie II, c'est à dire dans la description du déroulement de la procédure chirurgicale. On peut donc émettre l'hypothèse que deux cas chirurgicaux avec des profils identiques (forme, nature, amplitude et localisation de la tumeur) auront deux déroulements de procédure (positionnement du patient, nombre d'étapes, ordre des étapes, etc.) sensiblement identiques.

Les résultats de la deuxième méthode nous proposent des règles de décision permettant de prédire le positionnement du patient d'après la connaissance de la localisation de la tumeur dans un ou plusieurs lobes. Cet arbre de décision présente une erreur de prédiction ou coût de mauvais classement de 0,18. Quoique cette erreur soit tout à fait acceptable, elle est en partie due, à notre avis, au manque de précision dans la définition de la localisation de la tumeur. Toutefois, la construction d'un arbre de prédiction du positionnement du patient à partir de la localisation de la tumeur dans un ou plusieurs gyri n'est pas envisageable à l'heure actuelle vu la taille de notre échantillon et la distribution de la variable TargetLoc.

Nous nous sommes aussi interrogés sur les biais introduits dans nos données. Il faut faire une distinction entre les deux visions différentes que le neurochirurgien peut avoir de l'intervention : la procédure idéale et la procédure réelle soumise aux contraintes techniques, humaines et matérielles. Les différences se définissent en terme d'outils disponibles ou non, de contexte opératoire (si l'intervention est planifiée de longue date, il sera possible de faire bénéficier le patient d'une série complète d'examens alors que s'il s'agit d'une intervention en urgence, il n'y a souvent le temps que pour un scanner). Les moyens techniques à la disposition du chirurgien influencent beaucoup la préparation et par conséquent, le déroulement de la procédure. Ce biais et cette difficulté à distinguer la procédure idéale de la procédure réelle pourraient expliquer les cas chirurgicaux atypiques remarqués dans les résultats, et qui induisent une erreur de classification ou une erreur de prédiction.

Comme nous l'avons vu précédemment, le passage sous forme tabulaire de nos données n'est pas trivial. La transformation de données hiérarchisées, issues de la vision du monde chirurgical, en un tableau plat propice à l'ECD est un problème important : ce passage n'est pas automatique et reste largement tributaire des interprétations et des simplifications que l'on peut faire. Il est nécessaire de garder un contrôle sur la perte d'information qui en découle, ainsi que sur les hypothèses implicites que l'on introduit, car la pertinence et la qualité des résultats obtenus par les méthodologies d'ECD sont fortement dépendants des choix faits lors de cette transformation. La traduction de données structurées sous forme tabulaire introduit un biais : comme nous l'avons souligné précédemment, les colonnes du tableau qui représentent les variables décrivant les individus doivent être indépendantes entre elles. Or, les données structurées sont liées entre elles par des relations hiérarchiques, elles ne sont donc pas indépendantes.

5 Conclusion

Dans ce article, nous avons cherché à mettre en évidence des invariants dans le déroulement des interventions, en utilisant un ensemble de cas chirurgicaux. Cet ensemble contient les descriptions de procédures chirurgicales, structurées en XML d'après

les concepts et relations du *modèle générique*. L'utilisation de ce formalisme nous donne la possibilité de définir des métriques de comparaison entre les différents cas chirurgicaux. Nous avons vu que la mise en œuvre de méthodes d'extraction permettant de mettre en évidence les invariants n'est pas triviale, de par notamment le caractère complexe de nos données. Les problèmes à résoudre sont encore nombreux et cette partie du travail semble promettre de nombreuses perspectives de recherche dans les années à venir.

Malgré le faible nombre de cas chirurgicaux dans notre ensemble d'étude et la trivialité apparente de certains résultats, ces études sont pertinentes car basées sur des cas concrets. Elles cherchent à démontrer la faisabilité de notre méthodologie et la possibilité de formaliser de manière explicite des connaissances sur le déroulement de la procédure chirurgicale.

Références

- Shahidi R., Clarke L., Bucholz RD., Fuchs H., Kikinis R., Robb RA. et Vannier MW. (2001), White paper : Challenges and Opportunities in Computer-Assisted Interventions, *Computer Aided Surgery*, 6, pp. 176-181.
- Jannin P., Raimbault M., Morandi X., Riffaud L. et Gibaud B. (2003), Models of Surgical Procedures for Multimodal Image-Guided Neurosurgery, *Computer Aided Surgery*, 8 :2, pp. 98-106.
- Hand D., Manilla H. et Smyth P. (2000), *Principles of Data Mining*, Éditions MIT Press.
- Lebart I., Morineau A. et Piron M. (2000), *Statistique exploratoire multidimensionnelle*, Éditions Dunod.
- Zighed D. et Rakotomalala R. (2000), *Graphes d'induction : apprentissage automatique et data-mining*, Éditions Hermès.
- Sevach I., Cohen M. et Rappaport ZH. (1992), Patient positioning for the operative approach to midline intracerebral lesions : technical note, *Neurosurgery*, 31 :1, pp. 154-155.

Summary

These last years, a new discipline emerged from the recent progress in data processing and numerical imaging : computer assisted surgery. Computer assisted surgery systems contribute to improve the course of surgical procedures. A long-term objective of our work is to propose new solutions for improving these systems, by basing these solutions on the surgeon knowledge of the surgical procedure and by using a generic model that capture and represent this particular knowledge. This article presents a methodology of exploitation of surgical cases described using this generic model, by using data-mining algorithms, in order to highlight invariants in the structured descriptions of surgical cases. Moreover, it details some of the difficulties encountered because of the complexity of the studied data.

Mise en évidence d'invariants dans une population de cas chirurgicaux

Variable	% Groupe	% Global
Classe 1 effectif : 34		
LobeTemporal = Yes	100	40.88
LobeParietal = No	100	71.07
TargetRef = Malignant tumor	97.06	76.73
LobeFrontal = No	94.12	55.97
TargetAmplitude = Subcortical	76.47	49.69
<i>SkinIncisionForm = ? shape</i>	47.06	13.21
Classe 2 effectif : 37		
LobeParietal = No	100	28.93
TargetRef = Malignant tumor	94.59	76.73
<i>SkinIncisionForm = Horseshoe shape</i>	91.89	41.51
LobeFrontal = No	86.49	55.97
<i>PatientPosition = Lateral decubitus</i>	86.49	40.25
Classe 3 effectif : 8		
LobeOccipital = Yes	100	5.03
<i>PatientPosition = Ventral decubitus</i>	75	4.40
LobeFrontal = No	100	55.97
Classe 4 effectif : 38		
LobeFrontal = Yes	100	44.03
<i>PatientPosition = Dorsal decubitus</i>	100	55.35
LobeParietal = No	100	71.07
LobeTemporal = No	92.11	59.12
<i>SkinIncisionForm = Arciform</i>	84.21	42.14
Classe 5 effectif : 25		
TargetRef = Benign tumor	100	23.27
TargetShape = Regular	96	64.15
<i>TransgyralApproach = None</i>	88	63.52
<i>NumberOfSteps = Six</i>	84	45.28
Classe 6 effectif : 17		
TargetAmplitude = Cortical	100	11.32
<i>TransgyralApproach = None</i>	100	63.52
TargetShape = Regular	94.12	64.15
<i>NumberOfSteps = Six</i>	76.47	45.28

TAB. 1 – Répartition des individus dans les 6 classes obtenues après classification.