**Extraction et Gestion de Connaissance**

**Strasbourg, 27 janvier 2009**

# 6ème Atelier Fouille de Données Complexes dans un processus d'extraction de connaissances


**Responsables**

Omar Boussaid (ERIC, Lyon)

Arnaud Martin (ENSIETA - E3I2/EA3876)

## TABLE DES MATIÈRES

**Sixième atelier sur la**
**"Fouille de données complexes dans un processus d'extraction des connaissances"**
**27 janvier 2009 - EGC 2009, Strasbourg, France**

# Présentation

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

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

**Quelques mots sur la fouille de données complexes**

Dans tous les domaines tels que le multi-média, la télédétection, l'imagerie médicale, les bases de données, le web sémantique, la bio informatique et bien d'autres, les données à traiter pour y extraire de la connaissance utilisable sont de plus en plus complexes et volumineuses.

On est ainsi conduit à manipuler des données souvent non structurées :

- issues de diverses provenances comme des capteurs ou sources physiques d'informations variées ;
- représentant la même information à des dates différentes ;
- regroupant différents types d'informations (images, textes) ou encore de natures différentes (logs, contenu de documents, ontologies, etc.).

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

Une liste de thèmes est donnée ci-dessous à titre indicatif.

- Pré traitement, structuration et organisation des données complexes
- Processus et méthodes de fouille de données complexes
- Classification et fusion de données multi-sources
- Retours d'expériences d'extraction de connaissances à partir de données complexes (Web, sciences du vivant, etc.).
- Rôle des connaissances en fouille de données complexes

## Responsables


- Omar Boussaïd (ERIC, Lyon)
  Email Omar.Boussaid@univ-lyon2.fr
  Tel : 04 78 77 23 77
- Arnaud Martin (Laboratoire E3I2, ENSIETA, Brest )
  Email Arnaud.Martin@ensieta.fr
  Tel : 02 98 34 88 84


## Comité scientifique



## Remerciements


Les responsables de l’atelier tiennent à remercier :

- Les auteurs pour la qualité de leurs contributions.
- Les membres du comité de lecture pour leur travail indispensable à la qualité de cet atelier.
- Nicolas Lachiche et Agnès Braud, responsables des ateliers pour EGC 2009.
- Pierre Gançarski, président du comité d’organisation d’EGC 2005.

# Personnalisation collaborative pour l'enrichissement des analyses dans les entrepôts de données complexes

Cécile Favre, Fadila Bentayeb, Omar Boussaid

Université de Lyon (ERIC Lyon 2)
5 av. Pierre Mendès-France
69676 Bron Cedex
{cfavre|bentayeb}@eric.univ-lyon2.fr, omar.boussaid@univ-lyon2.fr
http://eric.univ-lyon2.fr

**Résumé.** Les entrepôts de données XML constituent une bonne alternative pour la représentation, le stockage et l'analyse des données complexes. Le modèle d'un entrepôt de données est classiquement conçu à partir des sources de données disponibles et des besoins d'analyse identifiés au moment de la conception. Or, il s'avère que des besoins d'analyse peuvent émerger, dépendant souvent des connaissances des analystes. Ces connaissances concernent en particulier la manière d'agréger les données. Ainsi, pour répondre aux besoins individuels et tirer profit des connaissances des différents analystes utilisant l'entrepôt de données, nous proposons dans cet article une approche de personnalisation collaborative pour l'enrichissement des possibilités d'analyse de l'entrepôt XML. Cette approche se base sur l'expression des connaissances des analystes sur de nouvelles façons d'agréger les données, permettant à la fois de répondre aux besoins d'analyse individuels émergents et de partager les nouvelles possibilités d'analyses à travers l'enrichissement des hiérarchies de dimension qui guident la navigation dans les données de l'entrepôt XML.

## 1 Introduction

Les entrepôts de données ont pour vocation de permettre l'analyse de données pouvant provenir de différentes sources de données (Kimball, 1996; Inmon, 1996). Cette analyse consiste généralement en une analyse exploratoire grâce à la technologie OLAP (On-Line Analytical Processing). Pour permettre cette analyse, les données sont organisées de façon multidimensionnelle : des faits sont analysés à travers des indicateurs appelés mesures, en fonction de différents axes d'analyse appelés dimensions. Les dimensions peuvent être organisées sous forme de hiérarchies. Ces hiérarchies de dimension permettent d'obtenir différentes vues sur les données avec plusieurs niveaux de granularité, en l'occurrence des données plus ou moins résumées, grâce aux opérateurs OLAP roll-up (forage vers le haut) et drill-down (forage vers le bas).

Depuis quelques années, face au besoin d'analyser des données pouvant être qualifiées de complexes (données du Web, données multimédia, etc), on a vu émerger des entrepôts de données basés sur le langage XML (eXtensible Markup Language) capables de centraliser et d'analyser des données complexes. En effet, XML est approprié pour la structuration de données complexes provenant de différentes sources et soutenues par des formats hétérogènes (Boussaid et al., 2008). XML est un langage qui présente à la fois les données et leur structure (schéma). Leur analyse est rendue possible grâce à l'extension de langages d'interrogation tel que XQuery (Beyer et al., 2005).

XML va aussi permettre de représenter les différentes formes de hiérarchies du modèle de l'entrepôt plus ou moins complexes. Bien que souvent on se limite au cas classique des hiérarchies qualifiées de «strictes», Malinowski et Zimányi (2004) ont proposé une représentation conceptuelle de ces hiérarchies. Notons que nous ne traitons pas ici de la façon d'exploiter ces différentes formes de hiérarchies qui constitue un réel problème.

Que ce soit dans le cadre des entrepôts que l'on qualifiera de classiques, ou de ceux que l'on qualifiera de «complexes», les possibilités d'analyse dépendent finalement du modèle de l'entrepôt de données conçu initialement. Ce modèle est généralement déterminé en fonction des sources de données disponibles d'une part, des besoins d'analyse recensés au moment de la conception du modèle d'autre part. Néanmoins, des besoins d'analyse individuels peuvent émerger, dépendant souvent des propres connaissances des utilisateurs. La caractérisation de ces nouveaux besoins d'analyse au sein de l'entrepôt répond à une certaine personnalisation de l'entrepôt de données voulue par les utilisateurs. L'entrepôt de données doit donc pouvoir s'adapter en prenant en compte des nouveaux besoins utilisateurs. En effet, Y. Ioannidis et G. Koutrika définissent la personnalisation comme «*...providing an overall customized, individualized user experience by taking into account the needs, preferences and characteristics of a user or group of users*» (Ioannidis et Koutrika, 2005).

Au sein d'une organisation, on peut envisager l'utilisateur de façon individuelle. Mais on peut également considérer qu'il appartient à une communauté, la communauté des utilisateurs exploitant l'entrepôt de données de l'organisation en l'occurrence. En effet, au sein d'une organisation, différents acteurs sont amenés à prendre des décisions : à différents niveaux de responsabilité ou sur des «thématiques» différentes (des services différents dans l'entreprise). Cette communauté d'utilisateurs a donc besoin de réaliser des analyses à partir de l'entrepôt de données pour supporter la prise de décision. Ainsi, dans le contexte de cette organisation, et donc de cette communauté d'utilisateurs de l'entrepôt de données, la notion de collaboration émerge.

Il est alors intéressant de combiner les concepts de personnalisation et de collaboration. Ceci a déjà été fait dans le cadre de systèmes de personnalisation proposant des recommandations basées sur le filtrage collaboratif. Dans ce cas, il s'agit de chercher des utilisateurs qui ont les mêmes comportements, préférences, etc. avec l'utilisateur à qui l'on souhaite faire des recommandations. Ensuite, il est possible d'utiliser les informations de ces autres utilisateurs similaires pour calculer une liste de recommandations pour cet utilisateur. Ceci est valable entre autres dans les systèmes de recherche d'informations (Goldberg et al., 1992). Dans ce cas, l'aspect collaboratif est un moyen pour parvenir à une personnalisation basée sur une idée de limitation ; en effet, l'aspect collaboratif permet de s'intéresser aux informations essentielles, pertinentes pour un utilisateur.

Nous pensons qu'il est possible de combiner les concepts de personnalisation et de collaboration d'une manière différente. Plutôt que d'exploiter l'aspect collaboratif pour permettre une personnalisation, nous voulons mettre l'aspect personnalisation au service de l'aspect collaboratif. L'objectif réside alors dans le fait que l'utilisateur puisse répondre à ses propres besoins d'analyse incluant un processus de personnalisation permettant à l'utilisateur, dans ce contexte, d'exprimer ses propres connaissances. Dans ce cas, le concept de personnalisation peut être considéré comme étant étendu puisque la personnalisation n'est pas basée ici sur une opération de restriction, mais plutôt sur une opération d'extension. L'intérêt est alors de pouvoir exploiter les connaissances de cet utilisateur donné, pour que les autres utilisateurs appartenant à la même communauté (au sein d'une organisation donnée par exemple) puissent en tirer profit, dans l'esprit d'un système collaboratif dans lequel chacun apporte sa pierre à l'édifice. Ainsi, à partir d'un entrepôt initial qui constitue une base de travail, assurant l'intégrité des données et de leur chargement par rapport à leur sources, l'aspect collaboratif va se porter sur le développement, l'enrichissement incrémental de nouveaux axes d'analyse à travers la création de nouveaux niveaux de granularité définissant ou enrichissant des hiérarchies de dimension dans l'entrepôt de données complexes.

La suite de cet article est organisée de la façon suivante. Dans la section 2, nous présentons brièvement un état de l'art relatif aux différents aspects évoqués dans notre proposition, à savoir les entrepôts de données complexes, les aspects collaboratifs dans les entrepôts de données et la personnalisation dans ces derniers. Puis nous développons dans la section 3 notre proposition de système de personnalisation collaborative pour l'enrichissement des analyses dans les entrepôts de données XML. Dans la section 4, nous évoquons la mise en œuvre de notre approche avec d'une part les éléments concernant l'implémentation qui est en cours de développement et d'autre part la présentation d'une étude de cas, issu d'un projet mené dans le cadre d'une Action Concertée Incitative avec des collègues linguistes, afin d'illustrer nos propos. Enfin, nous concluons cet article et évoquons les perspectives de ce travail préliminaire dans la section 5.

# 2 État de l'art

Cet article aborde différents domaines. Il se situe dans le cadre des entrepôts de données complexes et propose dans ce contexte, une solution de personnalisation collaborative. C'est pourquoi, nous abordons brièvement ces trois volets dans l'état de l'art avant de positionner notre travail.

## 2.1 Entrepôt de données complexes

À ce jour, les travaux s'intéressant à l'entreposage de données complexes portent essentiellement sur l'exploitation du langage XML pour la structuration et le stockage des données complexes (Boussaid et al., 2008). L'entrepôt de données est finalement constitué d'une collection de documents XML représentant les faits et les dimensions. Différentes approches ont été proposées. Elles peuvent être vues comme des variantes au niveau de l'organisation des données. Pokorný (2002) a proposé un schéma en étoile XML définissant les hiérarchies de dimension comme un ensemble de collections de données XML connectées logiquement, et les faits comme des éléments XML. Golfarelli et al. (2001) proposent de stocker chaque fait dans un document XML comprenant alors les instances et les hiérarchies de dimension. Quant à eux, Hümmer et al. (2003) proposent le modèle nommé XCube qui prévoit de regrouper tous les faits dans un document, et l'ensemble des dimensions dans un autre. Park et al. proposent une plateforme nommée XML-OLAP basée sur un entrepôt de données XML où chaque fait est contenu dans un document XML et où chaque instance de la hiérarchie de dimension est elle-même stockée dans un document XML pour éviter les jointures entre les différents niveaux de la hiérarchie de dimension.

Finalement, les différents travaux diffèrent sur comment sont représentés les faits et les dimensions dans ces documents, et donc, sur le nombre de documents nécessaires au stockage des données.

Une étude de performances des différentes représentations a été conduite par Boukraa et al. (2006). Elle montre que dans le cas d'un schéma en flocon de neige (cas des dimensions hiérarchisées), les meilleures performances sont obtenues lorsque les faits sont représentés dans un seul document XML et que chacune des dimensions est contenue dans un document XML. Ce mode de représentation présente en outre l'avantage d'éviter la duplication des données sur les dimensions dans le cas d'une construction de schéma en constellation dont le principe est de présenter plusieurs faits qui partagent des dimensions. En outre, comme chaque dimension et ses hiérarchies sont représentées dans un document XML, les mises à jour des dimensions sont rendues plus facile que si les dimensions étaient regroupées avec les faits ou stockées dans un seul document. C'est ce point qui nous intéresse plus particulièrement.

## 2.2 Systèmes collaboratifs dans les entrepôts de données

Initialement, les utilisateurs du Web se contentaient de consulter des données mises à leur disposition sur des sites développés par des spécialistes. Par la suite, ils ont pu peu à peu accéder à des technologies pour contribuer eux-mêmes au Web (participation à des forums, création de blogs, contributions dans des sites de type wiki, etc.). Ainsi, l'évolution du Web tend aujourd'hui vers un «Web 2.0», qualifié de «social», «participatif», «collaboratif», etc.

Cet aspect collaboratif est très présent au niveau du Web de façon générale, mais a été assez peu étudié dans un contexte comme celui des entrepôts de données, alors même que c'est un domaine dans lequel cet aspect peut être très intéressant. L'aspect collaboratif doit bien sûr être introduit au niveau où l'interaction entre le système et l'utilisateur est possible. Ainsi, dans le contexte des entrepôts de données, l'analyse est une phase privilégiée. On peut citer les travaux de Cabanac et al. (2007) qui se sont intéressés à la pratique d'annotations collectives dans le contexte des bases de données décisionnelles. Cette pratique permet aux analystes de partager leurs avis sur des analyses : ils réalisent ces analyses, peuvent les commenter et aussi les partager. Citons également les travaux de Aouiche et al. (2008) proposant une visualisation des analyses basée sur les nuages de mots et un partage, une mise à disposition facilitée de ces résultats pour d'autres utilisateurs, les auteurs qualifiant alors leur approche d'OLAP collaboratif.

## 2.3 Personnalisation dans les entrepôts de données

La personnalisation est une thématique abordée depuis déjà assez longtemps dans les domaines de la recherche d'information et des bases de données. Dans le contexte des entrepôts de données, il s'agit d'une thématique émergente. S'inspirant des travaux des domaines de la recherche d'information ou des bases de données, les travaux prennent de plus en plus en compte les spécificités des entrepôts de données. Quels que soient ces domaines, la personnalisation consiste habituellement à exploiter les préférences des utilisateurs pour leur fournir des réponses pertinentes.

Nous pouvons citer les travaux de Bellatreche et al. (2005) qui se sont inspirés des techniques de filtrage d'information en fonction du profil utilisateur pour affiner des requêtes en y ajoutant des prédicats. L'objectif de ces travaux est de pouvoir fournir à l'utilisateur un résultat focalisé sur son centre d'intérêt, tout en prenant en compte des contraintes de visualisation.

Ravat et Teste (2008) proposent une solution pour la personnalisation de la navigation OLAP en exploitant des préférences exprimées par des poids. Dans ce cas, l'utilisateur assigne des poids aux concepts multidimensionnels afin d'obtenir directement les analyses désirées, évitant ainsi des opérations de navigation.

Giacometti et al. (2008) proposent, quant à eux, un système de recommandation d'analyses multidimensionnelles en se basant sur la navigation qu'effectue un utilisateur donné par rapport aux navigations réalisées par les autres utilisateurs.

## 2.4 Positionnement

Notre approche de personnalisation collaborative consiste en la possibilité d'un enrichissement des hiérarchies de dimension via une mise à jour de celles-ci. Vis-à-vis des entrepôts de données XML, dans le cadre de notre approche, l'aspect mise à jour des dimensions est donc crucial. Dans ce cas, compte-tenu des avantages à modéliser les faits dans un seul document XML et chacune des dimensions dans un document XML, nous avons choisi de baser notre approche sur un modèle présentant ces caractéristiques, en l'occurrence sur celui de Mahboubi et al. (2009) que nous détaillerons par la suite.

À travers les différents travaux faits en matière de personnalisation, nous notons un manque afin d'apporter une réponse aux besoins d'analyses individuels. Nous avons apporté une solution à ce problème en proposant une évolution de l'entrepôt de données basée sur l'intégration des connaissances des utilisateurs sur la manière d'agréger les données sous forme de règles pour créer de nouveaux axes d'analyse (Bentayeb et al., 2008). Néanmoins ce travail a été réalisé dans un contexte d'entrepôts de données «classiques».

Notre travail vise alors à étendre ce travail au cas des entrepôts de données complexes, en se focalisant également sur l'aspect collaboratif qui nous paraît tout à fait intéressant dans ce contexte, compte-tenu de la difficulté de concevoir un schéma d'entrepôt de données répondant correctement aux besoins d'analyse de leurs usagers. En effet, bien que le nombre d'usagers d'un entrepôt de données soit réduit par rapport à celui concernant une base de données par exemple, il n'en demeure pas moins qu'il peut être élevé au sein d'une organisation, en particulier dans le cas où cette organisation est structurée hiérarchiquement avec bon nombre de responsables.

Notons que vis-à-vis de cet aspect d'évolution de l'entrepôt de données, nous pouvons distinguer dans la littérature deux types d'approches : la mise à jour du modèle d'une part et la modélisation temporelle d'autre part. La première approche consiste à transformer le schéma de l'entrepôt de données (Hurtado et al., 1999; Blaschka et al., 1999). Ces travaux consistent principalement à proposer des opérateurs adaptés permettant de faire évoluer le schéma de l'entrepôt de données. Dans ce cas, un seul schéma est supporté et l'historique de l'évolution n'est pas préservé. Dans la seconde approche, l'historique des modifications est conservées en exploitant des labels de validité temporelle. Ces labels peuvent être apposés au niveau des instances des dimensions (Bliujute et al., 1998), des liens d'agrégation (Mendelzon et Vaisman, 2000), ou des versions de schéma (Bebel et al., 2004; Body et al., 2003; Morzy et Wrembel, 2004; Ravat et al., 2006). Dans ces entrepôts, chaque version décrit le schéma et les données à une certaine période. Afin de pouvoir analyser ces données, compte-tenu du modèle spécifique, une extension du language SQL est requise. L'inconvénient de ces approches résident également dans le fait qu'elle doivent être mises en œuvre dès la conception de l'entrepôt de données. Ainsi, pour permettre un point de vue collaboratif, nous adoptons une approche de mise à jour de schéma, avec une partie de l'entrepôt qui servira de base que les utilisateurs vont enrichir. La mise à jour permet alors de pouvoir implémenter le processus collaboratif, l'objectif étant un enrichissement incrémental de l'entrepôt, cela ne remet pas en cause les données de l'entrepôt, même si l'historique des modifications n'est pas conservé.

# 3 Personnalisation collaborative dans les entrepôts de données XML

## 3.1 Modèle d'entrepôt de données XML

Comme nous l'avons précisé précédemment, plusieurs modèles d'entrepôt de données XML ont été proposés dans la littérature. Nous basons nos travaux sur le modèle proposé par Mahboubi et al. (2009) qui rend plus facile et plus efficace la mise à jour des dimensions. Ce modèle propose de rassembler les faits dans un document XML, et chacune des dimensions avec ses hiérarchies sont contenues dans un document XML. Un document XML nommé $dw - model.xml$ représente le schéma de l'entrepôt (figure 1). Ensuite, les documents portant le nom $facts_f$.xml contiennent les données sur les faits (figure 2-a), c'est-à-dire les identifiants des dimensions et les mesures. Enfin, les documents $dimension_d.xml$ permettent de stocker les valeurs des attributs décrivant les dimensions et leurs hiérarchies (figure 2-b).

Les éléments auxquels nous nous intéresserons particulièrement dans le cadre de notre personnalisation collaborative sont les niveaux hiérarchiques dans les dimensions. Ainsi, notre approche aura un impact à la fois sur le document contenant la structure de l'entrepôt ($dw - model.xml$), mais également sur les documents contenant les dimensions dont les hiérarchies seront modifiées.

## 3.2 Processus de personnalisation collaborative proposé

À travers cet article, nous voulons poser les bases de notre proposition de personnalisation collaborative. Il s'agit d'exploiter un entrepôt de données initial. Ensuite, une couche collaborative a pour but d'enrichir incrémentalement cet entrepôt initial, au fur et à mesure que la personnalisation répond à de nouveaux besoins individuels qui seront partagés. L'originalité finalement est qu'en voulant répondre à un besoin individuel, par l'expression de ses propres connaissances, l'utilisateur va en même temps collaborer à l'enrichissement des possibilités d'analyse de l'entrepôt de données pour les autres utilisateurs de l'organisation.

Pour permettre une personnalisation collaborative des analyses dans les entrepôts de données XML, nous proposons un processus au sein duquel les utilisateurs ont bien évidemment une place centrale (figure 3). Chaque utilisateur de la communauté dans l'organisation peut avoir des besoins spécifiques en termes d'analyse, nécessitant l'ajout ou l'enrichissement des hiérarchies de dimension de l'entrepôt. Ainsi, chaque utilisateur peut exprimer ses propres connaissances pour créer un nouveau niveau de granularité. Un module permet l'acquisition des connaissances sous forme de règles de type si-alors, correspondant à une phase participative. Un module d'évolution de l'entrepôt permet ensuite de prendre en compte ces règles pour faire évoluer l'entrepôt de données, en l'occurrence les documents XML adéquats. Enfin, un module d'analyse permet à l'utilisateur qui a exprimé ses connaissances, de faire l'analyse correspondant à ses propres besoins, mais ce module permet également l'accès à ces mêmes analyses pour les autres utilisateurs de la communauté, dans un esprit collaboratif où chacun enrichit l'entrepôt pour lui et pour les autres.

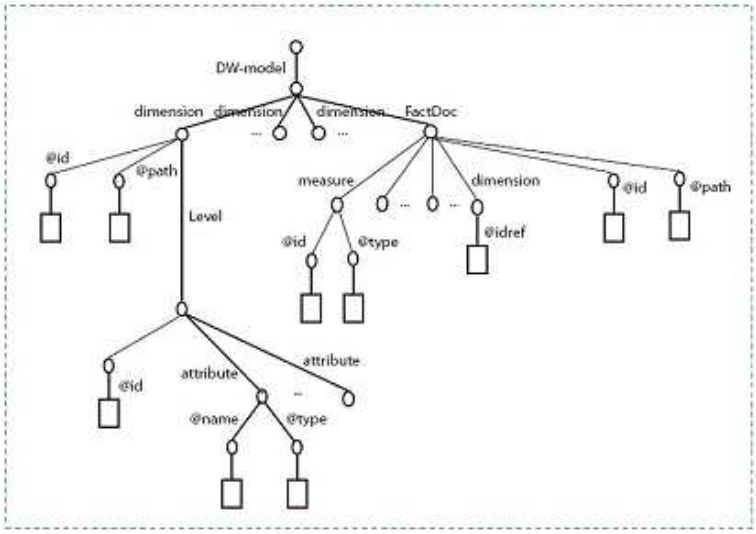



FIG. 1 – *Structure du graphe* $dw - model.xml$.

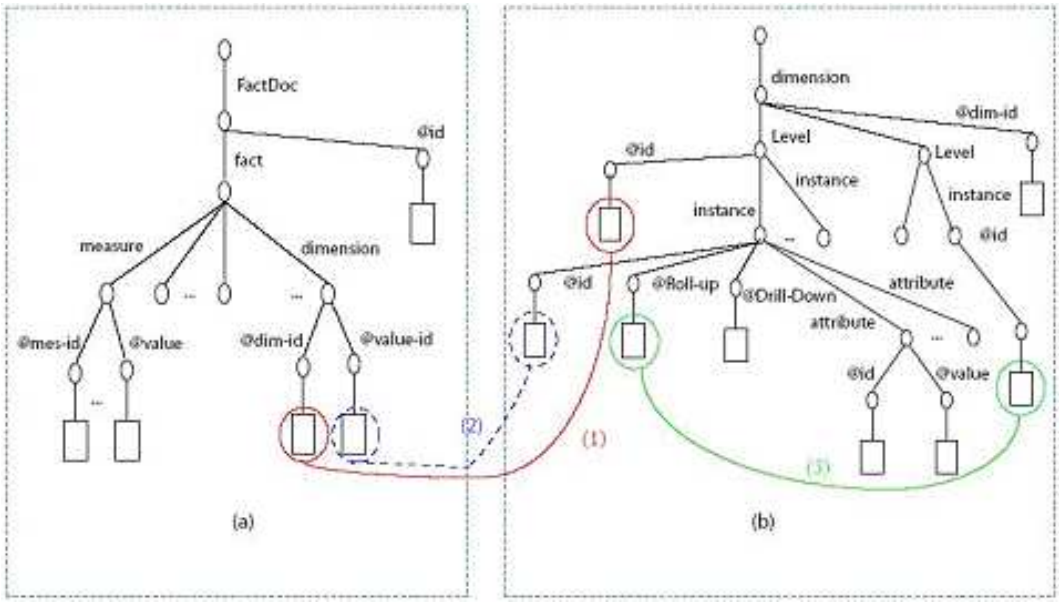



FIG. 2 – *Structure des graphes* $facts_f.xml$ *(a) et* $dimension_d.xml$ *(b).*

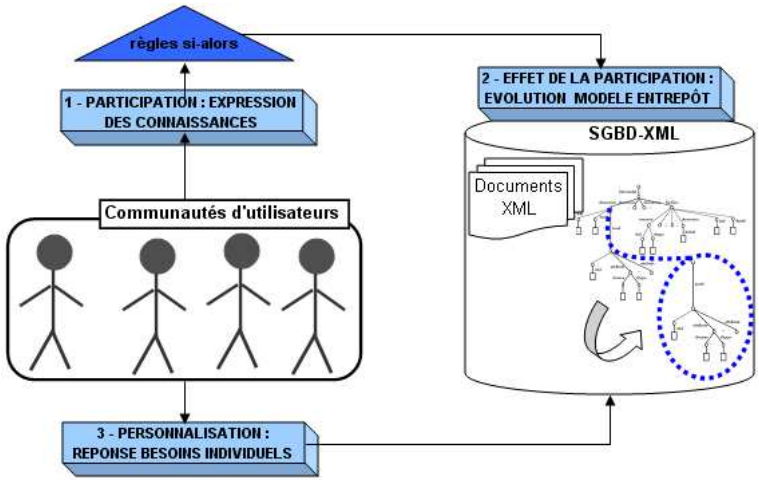



FIG. 3 – *Processus de personnalisation collaborative dans les entrepôts de données XML.*

Reprécisons que les utilisateurs disposent de l'entrepôt de données initial. Celui-ci a été conçu à partir des sources de données et d'un ensemble de besoins d'analyse globaux, correspondant à des besoins communs pour l'ensemble des utilisateurs, recensés au moment de la conception. Cet entrepôt initial permet de garantir l'intégrité des données « de base» (assurant le maintien de la cohérence de la phase de chargement des données dans l'entrepôt). Il appartient ensuite à chaque utilisateur d'ajouter les niveaux d'analyse dont il a besoin et d'en faire profiter les autres usagers.

## 3.3 Participation des analystes

Dans ce contexte d'entrepôt de données, la participation des analystes pour une personnalisation collaborative s'effectue dans la phase d'acquisition des connaissances qui traduit finalement l'expression d'un besoin individuel d'analyse, nécessitant un processus pouvant être qualifié de personnalisation.

La connaissance que nous considérons ici concerne la façon d'agréger les données. Ainsi, nous souhaitons représenter ces connaissances de façon simple pour les analystes. Nous avons alors choisi de représenter ces connaissances sous forme de règles dites d'agrégation qui sont des règles de type si-alors («if-then»).

Rappelons que le langage XML renferme à la fois la structure et les données elles-mêmes. Il s'agit pour les analystes d'exprimer leurs connaissances pour ajouter de nouveaux niveaux de granularité. Cet ajout a un effet à la fois sur la structure et les données elles-mêmes.

Ainsi, pour permettre l'acquisition des connaissances, nous proposons de la découper en deux étapes : expression des connaissances structurelles sur la création du nouveau niveau, expression des connaissances sur les données de ce nouveau niveau.

De ce fait, les règles d'agrégation sont de deux types : règle structure et règle données. Pour créer un niveau de granularité, il faut une règle structure et un ensemble de règles données. En effet, la règle structure permet de définir la structure des liens d'agrégation : quel niveau est créé, quels attributs le caractérisent, à quel(s) niveau(x) est-il relié, à partir de quels attributs le lien d'agrégation est défini, etc. Les règles données, quant à elles, instancient la règle structure, c'est-à-dire qu'elles définissent les liens d'agrégation au niveau des données. Cette formalisation s'inspire de celle que nous avons proposée dans (Bentayeb et al., 2008).

Soit $EL$ le niveau de hiérarchique au dessus duquel sera construit le niveau à créer. Soit $\{EA_i, i = 1..m\}$ l'ensemble des m attributs parmi les $m'$ attributs de $EL$, sur lesquels seront basées les conditions définissant les groupes d'instances. Soient GL le nouveau niveau et $GA_j, j = 1..n$ l'ensemble des n attributs du nouveau niveau $GL$. La règle structure notée $SR$ est définie comme suit :

$$SR: \ if\ ConditionOn(EL,\ \{EA_i,\ i = 1..m\})\ then\ Generate(GL,\ \{GA_j,\ j = 1..n\})$$

La règle $SR$ est instanciée par différentes règles données.

Une règle donnée est basée sur un ensemble $T$ de $z$ termes de règles, notés $RT_x$ tels que :

$$T = \{RT_x, 1 \leq x \leq z\} = \{EA_x op_x \{ens|val\}_x\}$$

où $EA_x$ est un attribut du niveau existant, $op_x$ est un opérateur, et $\{ens|val\}_x$ est soit un ensemble de valeurs, soit une valeur (selon l'opérateur utilisé), ces valeurs appartenant au domaine de définition de $EA_x$.

Soient $q$ le nombre d'instances de $GL$, un ensemble $R = \{r_d, d = 1..q\}$ de $q$ règles données doivent être définies. Notons $v_{dj}$ la valeur de l'attribut généré $GA_j$ dans la règle données $r_d$.

La clause «si» (if) est basée sur une composition de conjonctions ou de disjonctions de termes de règle, la clause «alors» (then) définit les valeurs des attributs (autrement dit les instances du nouveau niveau).

Une règle données $r_d$ est définie de la façon suivante :

$$r_d : if\ RT_1\ AND|OR\ ...\ AND|OR\ RT_x\ AND|OR\ ...\ AND|OR\ RT_z$$

$$then\ GA_1 = v_{d1}\ AND\ ...\ AND\ GA_j = v_{dj}\ AND\ ...\ AND\ GA_n = v_{dn}$$

Ainsi, la participation des utilisateurs réside dans l'expression de ces règles d'agrégation (structure et données) représentant leurs connaissances sur la façon d'agréger les données et traduisant leurs propres besoins d'analyse.

### 3.4 Exploitation de la participation des analystes

Une fois les connaissances exprimées, il s'agit d'exploiter cette participation, en l'occurrence, faire évoluer l'entrepôt de données en fonction des règles exprimées.

Compte-tenu des spécificités du stockage XML, l'évolution de l'entrepôt n'est pas une tâche facile. En effet, nous devons prendre en compte le fait que la structure et les données sont renfermées dans les documents, même si le schéma de l'entrepôt est représenté dans un document bien identifié $dw - model.xml$. De plus, nous devons prendre en compte l'organisation des informations sous forme d'arbre.

La création d'un nouveau niveau nécessite non seulement la création de ce niveau lui-même avec les données adéquates, mais aussi les liens avec le ou les autres niveaux (attribut @Drill-Down si le niveau est ajouté à la fin d'une hiérarchie, attributs @Drill-Down et @Roll-Up si le niveau est inséré entre deux niveaux existants).

Pour considérer cette évolution, nous devons l'envisager pour deux documents : le document représentant le modèle ($dw - model.xml$) et le document de la dimension qui est concerné par l'ajout d'un niveau ($dimension_d.xml$) puisque chaque dimension est représentée dans un document XML.

Notons que le document $facts.xml$ n'est pas modifié puisqu'il s'agit d'une partie fixe qui assure l'intégrité par rapport aux données sources et au processus d'alimentation.

Nous pouvons résumer les différentes opérations pour exploiter le résultat de la participation d'un analyste donné comme suit :

1. dans le document $dw - model.xml$ :
   (a) ajouter le nœud correspond au nouveau niveau, en extrayant les informations dans la règle structure
2. dans le document $dimension_d.xml$ :
   (a) ajouter l'élément représentant le niveau
   (b) exploiter les règles données pour ajouter les éléments nécessaires
   (c) mettre à jour les propriétés roll-up
   (d) si le niveau a été inséré entre deux niveaux existants, mettre à jour les propriétés drill-down du niveau supérieur

## 4 Mise en œuvre de notre approche

### 4.1 Éléments d'implémentation

Notre approche de personnalisation collaborative pour les analyses dans les entrepôts XML est actuellement en cours de développement.

Une interface web permet l'interaction avec les utilisateurs. Ainsi, cette interface va aider l'utilisateur à exprimer ses connaissances, autrement dit les règles (structure et données), de façon intuitive. L'utilisateur est en effet guidé pour choisir les éléments dans l'interface, l'aidant ainsi à exprimer ses connaissances et donc à participer à l'enrichissement des possibilités d'analyse de l'entrepôt XML. Cette interaction est développée grâce à des scripts PHP.

L'évolution de l'entrepôt de données XML nécessite des mises à jour au sein des documents XML qui sont exploités pour stocker l'entrepôt. Ainsi, nous devons développer le programme requis pour mettre à jour les documents XML.

Xupdate est un langage de requête XML dédié à la modification de données XML. Il s'agit d'une spécification de XML :DBInitiative. C'est un langage de mise à jour XML utilisé pour modifier le contenu XML en déclarant quels changements doivent être opérés sur la syntaxe XML. Différentes opérations élémentaires peuvent être combinées pour réaliser l'évolution que nous avons présentée pour apporter une personnalisation collaborative dans les entrepôts de données XML. Ces opérations élémentaires sont par exemple l'insertion d'un élément, l'insertion d'un attribut, modifier un attribut.

Une autre alternative pour notre implémentation serait de considérer le recours au DOM (Document Object Model). En effet, DOM est un modèle objet standard indépendant de toute plateforme et de tout language pour représenter des formats relatifs au HTML ou au XML. Ainsi, il serait possible d'utiliser des scripts PHP avec du DOM pour réaliser l'évolution. En particulier, mentionnons la méthode DomNode :appendChild qui permet d'insérer un nouvel élément dans un document et la méthode DomNode.setNodeValue qui met à jour un nœud et ses propriétés.

Nous sommes actuellement en train d'étudier ces deux alternatives d'implémentation.

## 4.2 Étude de cas

Pour illustrer notre approche, considérons le projet issu d'une Action Concertée Incitative avec des collègues linguistes. Ce projet, nommé CLAPI[1] pour «Corpus de Langues Parlées en Interaction», traite de l'intégration, du stockage, de la gestion et de l'analyse de corpus de langues parlées en interaction (Aouiche et al., 2003). Un corpus comprend des enregistrements audio et/ou vidéo d'interactions de la vie courante comme par exemple le déroulement d'un cours dans une salle de classe.

Chaque intervenant dans un enregistrement est identifié avec un pseudonyme et peut apparaître dans plusieurs interactions. Afin d'être exploité par les linguistes, les enregistrements sont reportés sous forme de texte dans des transcriptions. Ces transcriptions sont actuellement modélisées en XML : les «tokens» qui sont les formes orales d'un mots retranscrites comme «h'llo» pour «hello» et les phénomènes d'interaction tels que les pauses, les rires, les chevauchements de parole, etc. Les linguistes disposent d'une interface web dédiée pour exploiter ces transcriptions.

L'analyse multidimensionnelle sur ces transcriptions est très pertinente pour les linguistes. Citons par exemple l'intérêt pour eux d'observer les fréquences de tokens en fonction de la place de ceux-ci dans la transcription (début, milieu, fin) et l'âge de l'intervenant. Cet exemple simplifié correspond à l'instanciation du document $dw - model.xml$ présentée dans la figure 4.

```
<?xml version="1.0" encoding="utf-8">
<DW-model>
   <dimension id="time-d" path="dim-time.xml">
      <Level id="location-in-transcription">
         <attribute name="location" type="string" />
      </Level>
   </dimension>
   <dimension id="speaker-d" path="dim-speaker.xml">
      <Level id="speaker">
         <attribute name="sex" type="boolean" />
      </Level>
   </dimension>
   <dimension id="transcription-d" path="dim-transcript.xml">
      <Level id="token">
         <attribute name="term" type="string" />
      </Level>
      <Level id="transcription">
         <attribute name="transcription-name" type="string" />
      </Level>
   </dimension>
   <FactDoc id="facts" path="facts.xml">
      <measure id="frequency" type="real" />
      <dimension idref="time-d" />
      <dimension idref="speaker-d" />
      <dimension idref="transcription-d" />
   </FactDoc>
</DW-model>
```

FIG. 4 – *Document* $dw - model.xml$ *exemple.*

Ce schéma initial a été conçu à partir des données de la base CLAPI et en fonction des besoins d'analyse identifiés. Mais un linguiste peut avoir besoin d'agréger les fréquences en groupant certains emplacements des tokens. Il souhaite par exemple savoir si certains tokens apparaissent davantage à l'extrémité (début ou fin) qu'au milieu des interactions.

[1] http ://clapi.univ-lyon2.fr

Même si ce besoin n'a pas été exprimé initialement, le processus de personnalisation collaborative va non seulement permettre de répondre à ce besoin d'analyse individuel, mais également de faire profiter de cette possibilité d'analyse aux autres linguistes exploitant le système.

Le linguiste va donc formuler les règles d'agrégation traduisant son besoin d'analyse, en l'occurrence la façon d'agréger les données : une règle structure et les règles données correspondantes.

Règle structure :

$$(SR)\ if\ ConditionOn(location-in-transcription,\ \{location\})$$
$$then\ Generate(group-of-location,\ \{group-location\})$$

Règles données :

$$(r_1)\ if\ location\ in\ \{'begin',\ 'end'\}\ then\ group-location\ =\ 'extreme'$$
$$(r_2)\ if\ location\ not\ in\ \{'begin',\ 'end'\}\ then\ group-location\ =\ 'middle'$$

Grâce à ces règles, l'entrepôt de données XML peut évoluer à travers la modification du document $dw-model.xml$ d'une part, du document $dim-time.xml$ d'autre part puisqu'il correspond au document de la dimension qui va être enrichie d'un niveau hiérarchique. Le document $dw-model.xml$ est modifié pour inclure le nouveau niveau (parties en gras dans la figure 5).

```
<?xml version="1.0" encoding="utf-8">
<DW-model>
   <dimension id="time-d" path="dim-time.xml">
      <Level id="location-in-transcription">
         <attribute name="location" type="string" />
      </Level>
      <Level id="group-of-location-in-transcription">
         <attribute name="location-group" type="string" />
      </Level>
   </dimension>
   <dimension id="speaker-d" path="dim-speaker.xml">
      <Level id="speaker">
         <attribute name="sex" type="boolean" />
      </Level>
   </dimension>
   <dimension id="transcription-d" path="dim-transcript.xml">
      <Level id="token">
         <attribute name="term" type="string" />
      </Level>
      <Level id="transcription">
         <attribute name="transcription-name" type="string" />
      </Level>
   </dimension>
   <FactDoc id="facts" path="facts.xml">
      <measure id="frequency" type="real" />
      <dimension idref="time-d" />
      <dimension idref="speaker-d" />
      <dimension idref="transcription-d" />
   </FactDoc>
</DW-model>
```

FIG. 5 – *Document $dw-model.xml$ exemple mis à jour.*

Le document $dim-time.xml$ est également mis à jour pour représenter le nouveau niveau et ses instances, ainsi que les liens d'agrégation requis (parties en gras dans la figure 6).

Ainsi, le linguiste va pouvoir connaître les fréquences des tokens en fonction des groupes d'emplacements qu'il a défini, obtenant ainsi une réponse à ses propres besoins d'analyse. Et le niveau ainsi créé va pouvoir être exploité pour différentes analyses (détail des fréquences par groupe d'emplacements et par sexe du locuteur par exemple), par le linguiste qui en avait besoin mais également par les autres linguistes.

## 5 Conclusion et perspectives

Dans cet article, nous avons posé les bases d'un système de personnalisation collaborative pour l'enrichissement des analyses dans les entrepôts de données complexes XML. Nous avons explicité les principes de cette proposition,

```
<?xml version="1.0" encoding="utf-8">
<dimension dim-id="time-d">
     <Level id="location-in-transcription">
       <Instance id="begin" Roll-up="extreme">
         <attribute id="location" value="begin">
       </Instance>
       <Instance id="middle" Roll-up="middle">
         <attribute id="location" value="middle">
       </Instance>
       <Instance id="end" Roll-up="extreme">
         <attribute id="location" value="end">
       </Instance>
     </Level>
     <Level id="group-of-location-in-transcription">
       <Instance id="extreme" Drill-Down={"begin","end"}>
         <attribute id="location-group" value="extreme">
       </Instance>
       <Instance id="middle" Roll-up="middle">
         <attribute id="location-group" value="middle">
       </Instance>
     </Level>
</dimension>
```

FIG. 6 – *Document* $dim - time.xml$ *mis à jour.*

décrit le processus centré utilisateurs pour permettre la réalisation de cette proposition, nous avons détaillé les deux principaux modules pour atteindre l'objectif, à savoir la participation des utilisateurs (l'acquisition des connaissances) et l'exploitation de cette participation (l'évolution incrémentale de l'entrepôt). L'implémentation de notre approche est en cours de développement, mais nous avons indiqué quelques éléments la concernant, et nous avons illustré notre approche par une étude de cas issu d'une collaboration avec des linguistes.

Cet article présente un travail préliminaire dans le domaine de la personnalisation collaborative dans les entrepôts de données complexes ouvrant de nombreuses perspectives. Dans l'immédiat, il s'agit, à partir d'un recensement des hiérarchies complexes inspiré par la modélisation des hiérarchies établie par Malinowski et Zimányi (2004), d'aider à la saisie des règles exprimant les connaissances. Ainsi les règles seront le fondement du système, il s'agit alors de guider l'utilisateur dans l'expression de ses connaissances et d'adapter la vérification des règles saisies. Aini la formalisation devra également être enrichie par rapport aux différentes propriétés pour prendre en compte les différents types de hiérarchie.

Dans notre approche, l'apect collaboratif se traduit par un enrichissement incrémental de l'entrepôt mettant finalement à disposition des utilisateurs de la communauté les suggestions individuelles. En d'autres mots, il s'agit d'un partage des possibilités d'analyse créées individuellement. Il serait alors intéressant d'envisager l'aspect collaboratif en y introduisant l'échange de points de vue. Ceci peut être intéressant en particulier dans le cas de points de vue divergents. Nous avons d'ores et déjà proposé d'avoir recours au versionnement pour traduire des points de vue différents sur un même niveau (Bentayeb et al., 2008), mais il serait sans doute pertinent d'envisager d'autres solutions.

Par ailleurs, ce principe de personnalisation collaborative pourrait être étendu. En effet, dans notre cas, les nouveaux axes sont accessibles par les autres usagers du système. Mais nous pourrions aller au-delà en envisageant un système de recommandation. Des travaux commencent à émerger sur cet aspect de recommandation dans les entrepôts de données (Giacometti et al., 2008). Il s'agit alors de pouvoir exploiter l'aspect collaboratif que nous avons introduit. Nous pensons également que le concept de profil pourrait être pertinent. En effet, dans le contexte d'une organisation telle qu'une entreprise, la notion de métier peut être intéressante comme base pour la recommandation.

Enfin, Bentayeb (2008) a proposé l'utilisation d'une méthode de fouille de données, celle des K-means en l'occurrence, pour permettre la construction de nouveaux niveaux de granularité. Il serait sans doute pertinent de s'intéresser à des méthodes de fouille de données dans les documents XML pour faire émerger de nouveaux axes d'analyse. Ceci aurait pour objectif, entre autres, d'exploiter des connaissances dont les utilisateurs ne disposent pas.

## Références

Aouiche, K., F. Bentayeb, O. Boussaid, et J. Darmont (2003). Conception informatique d'une base de données multimédia de corpus linguistiques oraux : l'exemple de clapi. In *36ème Colloque International de la Societas Linguistica Europaea, Lyon, France*, pp. 11–12.

Aouiche, K., D. Lemire, et R. Godin (2008). Collaborative OLAP with Tag Clouds - Web 2.0 OLAP Formalism and Experimental Evaluation. In *4th International Conference on Web Information Systems and Technologies (WEBIST 08), Funchal, Madeira, Portugal*, pp. 5–12.

Bebel, B., J. Eder, C. Koncilia, T. Morzy, et R. Wrembel (2004). Creation and Management of Versions in Multiversion Data Warehouse. In *19th ACM Symposium on Applied Computing (SAC 04), Nicosia, Cyprus*, pp. 717–723.

Bellatreche, L., A. Giacometti, P. Marcel, H. Mouloudi, et D. Laurent (2005). A Personalization Framework for OLAP Queries. In *8th ACM International Workshop on Data Warehousing and OLAP (DOLAP 05), Bremen, Germany*, pp. 9–18.

Bentayeb, F. (2008). K-means based Approach for OLAP Dimension Updates. In *10th International Conference on Enterprise Information Systems (ICEIS 08), Barcelona, Spain*, Volume DISI, pp. 531–534.

Bentayeb, F., C. Favre, et O. Boussaïd (2008). A User-driven Data Warehouse Evolution Approach for Concurrent Personalized Analysis Needs Integrated Computer-Aided Engineering. *Journal of Integrated Computer-Aided Engineering 15*(1), 21–36.

Beyer, K., D. Chambérlin, L. Colby, F. Özcan, H. Pirahesh, et Y. Xu (2005). Extending XQuery for Analytics. In *24th International Conference on Management of Data (SIGMOD 05), Baltimore, Maryland, USA*, pp. 503–514.

Blaschka, M., C. Sapia, et G. Höfling (1999). On Schema Evolution in Multidimensional Databases. In *1st International Conference on Data Warehousing and Knowledge Discovery (DaWaK 99), Florence, Italy*, Volume 1676 of *LNCS*, pp. 153–164.

Bliujute, R., S. Saltenis, G. Slivinskas, et C. Jensen (1998). Systematic Change Management in Dimensional Data Warehousing. In *3rd International Baltic Workshop on Databases and Information Systems, Riga, Latvia*, pp. 27–41.

Body, M., M. Miquel, Y. Bédard, et A. Tchounikine (2003). Handling Evolutions in Multidimensional Structures. In *19th International Conference on Data Engineering (ICDE 03), Bangalore, India*, pp. 581–591.

Boukraa, D., R. B. Messaoud, et . Boussaïd (2006). Proposition d'un modèle physique pour les entrepôts XML. In *Atelier Systèmes Décisionnels (ASD 06) en conjonction avec 9th Maghrebian Conference on Information Technologies (MCSEAI 06), Agadir, Morocco*.

Boussaid, O., J. Darmont, F. Bentayeb, et S. Loudcher (2008). Warehousing Complex Data from the Web. *International Journal of Web Engineering and Technology 4*(4), 408–433.

Cabanac, G., M. Chevalier, F. Ravat, et O. Teste (2007). An Annotation Management System for Multidimensional Databases. In *9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 07), Regensburg, Germany*, Volume 4654 of *LNCS*, pp. 89–98.

Giacometti, A., P. Marcel, et E. Negre (2008). A Framework for Recommending OLAP Queries. In *11th ACM International Workshop on Data Warehousing and OLAP (DOLAP 08), Napa Valley, California, USA*, pp. 73–80.

Goldberg, D., D. Nichols, B. M. Oki, et D. Terry (1992). Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM 35*(12), 61–70.

Golfarelli, M., S. Rizzi, et B. Vrdoljak (2001). Data Warehouse Design from XML Sources. In *4th ACM International Workshop on Data Warehousing and OLAP (DOLAP 01), Atlanta, Georgia, USA*, pp. 40–47.

Hümmer, W., A. Bauer, et G. Harde (2003). XCube : XML for data warehouses. In *6th International Workshop on Data Warehousing and OLAP (DOLAP 03), New Orleans, Louisiana, USA*, pp. 33–40.

Hurtado, C. A., A. O. Mendelzon, et A. A. Vaisman (1999). Updating OLAP Dimensions. In *2nd ACM International Workshop on Data Warehousing and OLAP (DOLAP 99), Kansas City, Missouri, USA*, pp. 60–66.

Inmon, W. H. (1996). *Building the Data Warehouse*. John Wiley & Sons.

Ioannidis, Y. et G. Koutrika (2005). Personalized systems : models and methods from an IR and DB perspective. In *31st International Conference on Very Large Data Bases (VLDB 05), Trondheim, Norway*, pp. 1365–1365.

Kimball, R. (1996). *The Data Warehouse Toolkit*. John Wiley & Sons.

Mahboubi, H., J. C. Ralaivao, S. Loudcher, O. Boussaid, F. Bentayeb, et J. Darmont (2009). *X-WACoDa: An XML-based approach for Warehousing and Analyzing Complex Data*. Advances in Data Warehousing and Mining. IGI Publishing.

Malinowski, E. et E. Zimányi (2004). OLAP Hierarchies: A Conceptual Perspective. In *16th International Conference on Advanced Information Systems Engineering (CAiSE 04), Riga, Latvia*, Volume 3084 of *LNCS*, pp. 477–491.

Mendelzon, A. O. et A. A. Vaisman (2000). Temporal Queries in OLAP. In *26th International Conference on Very Large Data Bases (VLDB 00), Cairo, Egypt*, pp. 242–253.

Morzy, T. et R. Wrembel (2004). On Querying Versions of Multiversion Data Warehouse. In *7th ACM International Workshop on Data Warehousing and OLAP (DOLAP 04), Washington, Columbia, USA*, pp. 92–101.

Park, B.-K., H. Han, et I.-Y. Song. XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. In *7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 05), Copenhagen, Denmark*.

Pokorný, J. (2002). XML Data Warehouse: Modelling and Querying. In *5th Baltic Conference (BalticDB&IS 06), Tallin, Estonia*, pp. 267–280.

Ravat, F. et O. Teste (2008). Personalization and OLAP Databases. *Annals of Information Systems, Numéro spécial New Trends in Data Warehousing and Data Analysis 3*.

Ravat, F., O. Teste, et G. Zurfluh (2006). A Multiversion-Based Multidimensional Model. In *8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 06), Krakow, Poland*, Volume 4081 of *LNCS*, pp. 65–74.

## Summary

The XML data warehouses are a good alternative for the representation, storage and analysis of complex data. The model of a data warehouse is classically designed from the available data sources and analysis needs identified during the conception. It turns out that the analysis needs may emerge, depending on knowledge of analysts. This knowledge may concern new ways to aggregate data. Thus, to provide an answer to individual analysis needs and take advantage of knowledge of different analysts using the data warehouse, we propose in this paper a collaborative personalization for enrichment opportunities for analysis of XML data warehouse. This approach is based on the expression of knowledge analysts on how to aggregate the data, allowing the sharing of new possibilities for analysis through the enrichment of dimension hierarchies that drive the navigation in the XML data warehouse.

# Système Automatique de reconnaissance de cibles radar : Problématique de l'extraction de la forme

Mohamed Nabil Saidi*, **, Abdelmalek Toumi*
Brigitte Hoeltzener*, Ali Khenchaf*, Driss Aboutajdine**

* Ecole Nationale Supérieure d'Ingénieurs des Etudes et Techniques d'Armement(ENSIETA)
2 rue François Verny 29806 Brest Cedex 9, France
{saidimo, hoelzbr, toumiab, Ali.Khenchaf}@ensieta.fr
**LRIT, Université Mohammed V- Agdal Faculté des sciences de Rabat, B.P.1014 Maroc
Aboutaj@fsr.ac.ma

**Résumé.** Nous présentons dans ce papier un système intelligent de automatique de cibles radar non-coopératives en environnements incertains (ex. : terrestre, maritime). Nous utilisons le processus d'extraction de connaissance à partir de données adapté au domaine radar. Nous nous intéressons dans ce travail à la phase de préparation de données radar notamment l'extraction et la modélisation de la forme à partir des images radar dites ISAR (Inverse Synthetic Aperture Radar)

## 1 Introduction

La reconnaissance/identification automatique de cibles radar trouve de nombreuses applications en environnement aérien, terrestre ou maritime. Par exemple la croissance des aéroports et des différents types d'avion deviennent de plus en plus importante dans le monde. Il s'avère donc nécessaire d'introduire des techniques originales et automatiques de reconnaissance de cibles à partir des signaux radar. Cependant le caractère non coopératif des cibles conduit à intégrer davantage des connaissances au sein du système d'identification. D'autre part, le volume important des données radar implique l'intégration de l'homme aux différentes étapes du processus de reconnaissance. C'est ainsi que la méthodologie choisie, et la mieux adaptée à la conception de la chaîne de reconnaissance se base sur le processus d'extraction de connaissance à partir des données (processus ECD).

Nous introduisons alors le processus ECD adapté au domaine radar (Cf. figure 1), qui est un processus interactif et itératif, constitué principalement de trois grandes phases, allant de l'acquisition de données jusqu'à la classification et l'évaluation en passant par la phase de préparation de données.

Dans ce papier, nous nous intéressons à la phase de préparation de données pour la classification, et cela en se basant sur les images radar à haute résolution (Images ISAR) construites à partir des signaux fréquentiels en utilisant l'algorithme de la transformée de Fourier 2D (Toumi et al., 2008b) (Toumi, 2007).

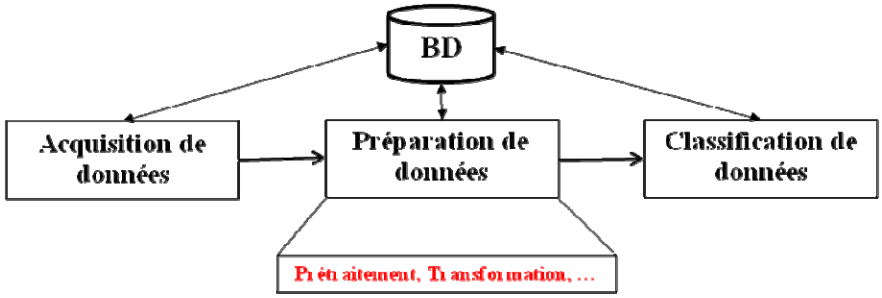



FIG.1 - *Processus ECD adapté au système de reconnaissance de cibles.*

## 2 Acquisition et Prétraitement de données

### 2.1 Acquisition de données

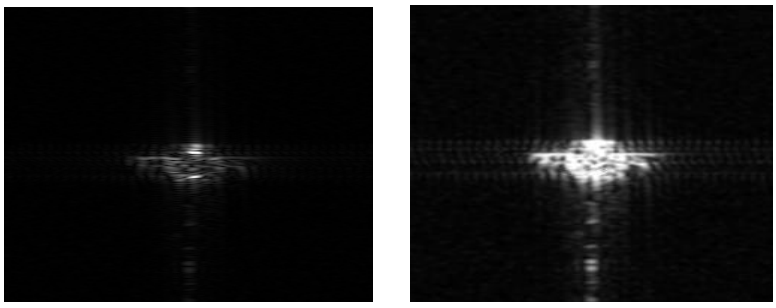
La phase d'acquisition des données radar présente une influence non négligeable sur la qualité des données, et donc sur les performances globales de la fonction de reconnaissance. C'est pour çà que nous avons eu recours à des données expérimentales acquises en laboratoire dans une chambre anéchoïque (du laboratoire E3I2 de l'ENSIETA) de dimensions finies simulant un espace libre, et sur des maquettes au 1/48ième.

Pour l'acquisition des signaux radar, le mode « Inverse[1]» est adopté pour produire les images de cibles présentant un mouvement de rotation. Ce mode d'acquisition est basé sur une analyse du signal reçu en fonction du temps et de la fréquence Doppler. L'image est obtenue en appliquant l'analyse temporelle qui fournit la position des points brillants suivant l'axe de visée du radar et par l'analyse de la fréquence Doppler qui fournit la position des points brillants suivant l'axe azimutal.

## 2.2 Prétraitement de données

Les signaux obtenus par les systèmes d'émission-réception cohérente comme le radar à ouverture synthétique inverse présentent un bruit (*Speckle*) assez biaisant. Le *Speckle* est un bruit multiplicatif dû aux superpositions constructives ou destructives des réflexions élémentaires. Il confère à l'image un aspect granulaire commun aux systèmes d'imagerie utilisant une source d'éclairement cohérente (laser, onde électromagnétique – radar-, …).

Dans le but de rendre les images mieux interprétables, et conserver l'information pertinente contenue dans les images ISAR, l'étape de prétraitement s'avère incontournable. Pour cela, il existe de nombreux types de filtrage permettant d'améliorer la qualité des images ISAR. Chaque filtres possède ses propres particularités : certains permettent de lisser l'image, d'autres permettent de diminuer le bruit ou encore d'améliorer l'interprétation visuelle des images. L'utilisation des différents filtres dépend donc de l'application envisagée. Les principaux filtres utilisés en imagerie radar qui permettent de diminuer le *speckle* sont : le filtre de Lee (Lee, 1981), le filtre de Kuan et le filtre de Frost (Frost et al., 1982). La figure 2 illustre l'utilisation d'un filtre de rehaussement suivi du filtre de Frost.



*a)* *b)*

FIG 2 – *a) Image Originale. b) Image prétraitée (rehaussée+filtre de Frost).*

# 3 Extraction de la forme

Afin de préparer les données à la phase de reconnaissance, l'extraction des descripteurs invariants aux transformations géométriques a été réalisée. Le choix dans ce cadre de travail est porté sur les contours comme éléments d'analyse. Pour le cas des images ISAR à traiter, les techniques classiques de détection de contour fondées sur des techniques dérivatives ne fournissent que des ensembles de contours non fermés révélant un certain nombre de disparités, d'où la difficulté de reconstruire une forme (contours fermés) générale de la cible à partir de tels contours (Toumi, 2007).

Nous avons donc utilisé trois techniques d'extraction de contours qui sont : la ligne de partage des eaux (LPE), le flux du vecteur gradient (FVG) (dérivée des contours actifs) et au final, la combinaison des deux précédentes techniques. Nous exposons dans la suite les différentes techniques.

## 3.1 Ligne de partage des eaux

La ligne de partage des eaux est l'une des techniques les plus répandues de morphologie mathématique, elle utilise la terminologie de la géographie qui définit la LPE comme la crête qui forme entre deux bassins versants dans une image considérée comme une surface topographique (Roerdink et Meijster, 2001).

Dans le cas des images ISAR, chaque objet (point brillant) dans l'image correspond généralement à un minimum du gradient morphologique (Toumi et al., 2008) et son contour correspond aux LPE de ce gradient (Cf. figure 3).

[1] L'acquisition en mode inverse produit des images d'objets présentant un mouvement de rotation par rapport au radar.

Cependant, l'application directe de la LPE sur le gradient des images ISAR, produit une sur-segmentation de l'image due à certains minima locaux qui contribuent à générer des LPE non significatives. Pour palier ce problème, nous avons procédé au renforcement des variations de niveaux de gris en passant par l'image simplifiée, appelée image mosaïque (Toumi et al., 2008).

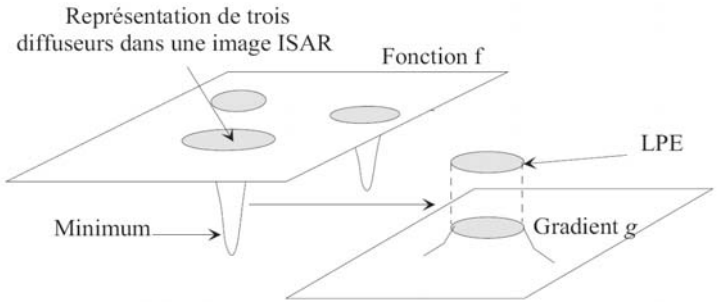
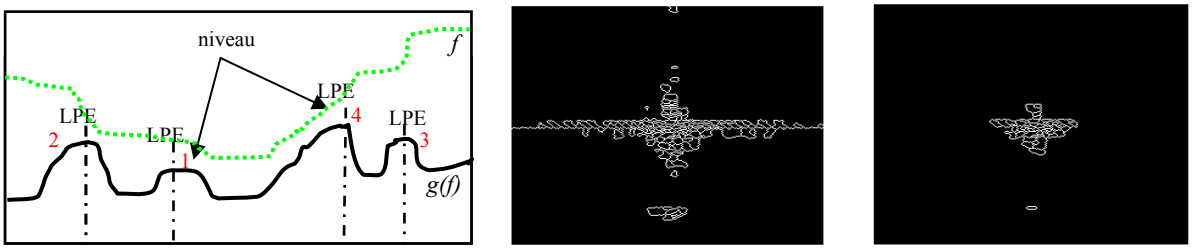



FIG 3 – *Principe de détection de contour par le gradient morphologique.*

L'objectif est donc d'éliminer l'influence de minima locaux non significatifs pour ne conserver que ceux associés aux zones d'intérêts. L'image mosaïque peut s'interpréter comme un graphe sur lequel sont évalués des arcs. Chaque bassin versant de l'image représente un des sommets du graphe et les différences de niveaux de gris entre ces bassins versants ($f_i$) permettent de mesurer la longueur de l'arrête séparant deux sommets : $h(C_{ij}) = |f_i - f_j|$. Le graphe valué de l'image mosaïque est appelée *gradient mosaïque* (Toumi et al,. 2008). Ce gradient est la fonction *h* définie sur tous les arcs de l'image et ses niveaux de gris correspondent à la différence de valeur existante entres deux composantes connexes. En éliminant les premiers niveaux, nous arrivons à garder l'information nécessaire révélant la forme générale de la cible (cf. figure 4.b).




a) *Principe du gradient-mosaïque.* b) *Suppression des niveaux dans le gradient mosaïque.*

FIG. 4 – *LPE du gradient mosaïque.*

Une dernière phase, pour extraire la forme générale de la cible, consiste à éliminer les lignes de partage des eaux internes et ne garder que celles qui se trouvent à la frontière externe (Toumi et al., 2008).

## 3.2 Flux du vecteur gradient

Les contours actifs (*Snakes*), sont utilisés intensivement en traitement d'images, particulièrement pour localiser les contours d'objets et le suivi des objets en mouvement. Nous avons utilisé une dérivée des modèles déformables, qui introduit une nouvelle énergie externe, appelée *flux de vecteurs gradients* (FVG). Elle a été introduite en 1997 dans (Xu et Prince, 1997) pour traiter les problèmes associés à l'initialisation de la détection des zones de forte concavité. Le FVG est calculé comme une diffusion des vecteurs gradients d'une carte de contours d'une image en niveaux de gris. Voir pour plus de détail (Xu et Prince, 1998).

## 3.3 Combinaison de LPE et FVG

Afin de faciliter la phase d'initialisation pour le FVG et augmenter sa rapidité de détection, nous l'avons initialisé par la forme extraite par la LPE puis délattée. Par ailleurs, la dilatation est appliquée sur la forme initialement extraite par la LPE afin d'assurer un encadrement total de la zone délimitée par la forme initiale. C'est ainsi que la forme dilatée représente le contour initial du FVG. La figure ci-dessous illustre le fonctionnement de cette combinaison.

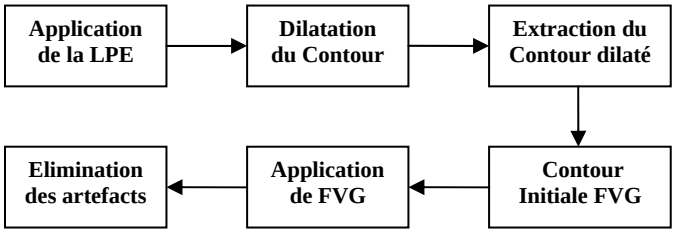



FIG. 5 - *Processus de combinaison de la LPE et le FVG.*

## 3.4 Evaluation des résultats

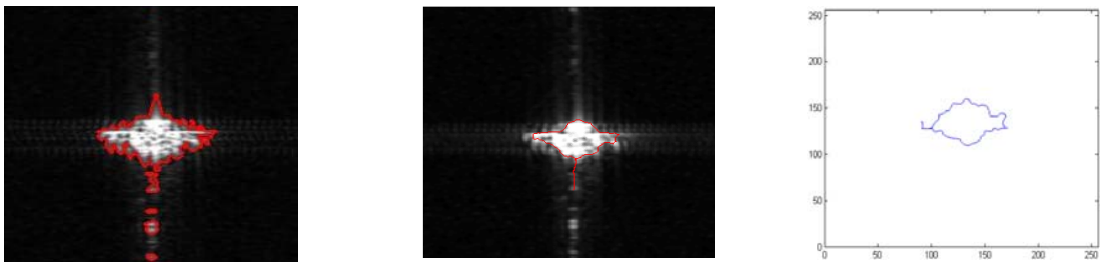
L'évaluation qualitative et quantitative à ce niveau de ces techniques reste prématurée sans une validation en termes de performances globales de la fonction de reconnaissance. Au niveau visuel, (Cf. figure 6), nous constatons que l'incertain sur les contours l'est moins dans la technique de la LPE que pour le cas des deux autres approches. Ces deux dernières présentent deux problèmes majeurs :

1. les contours détectés ne sont pas toujours fermés,
2. la présence des artefacts qui nécessitent un post traitement pour les éliminer.

En absence d'une vérité terrain, nous utilisons des critères statistiques pour évaluer la qualité des formes extraites par la LPE et le FVG. Ces critères sont proposés dans (Chabrier et al. 2004) et utilisées pour l'évaluation des méthodes de segmentation en régions. Nous pouvons classer ces critères en deux grandes familles : les critères de contraste et les critères d'adéquation à un modèle. Les premiers critères recherchent une variabilité inter-région, alors que les seconds recherchent une uniformité en intensité à l'intérieur des régions. Pour cela, nous reposons sur une image segmentée en deux régions qui représentent respectivement l'objet (cible) et le fond. Nous nous orientés dans ce papier, sur les techniques d'évaluation basées sur le contraste (les contrastes de Levine et Nazif et les contrastes de Zeboudj) (Philipp-Foliguet et Guigues, 2001). Les résultats obtenus sont présentés dans le tableau 1 où nous constatons que la LPE a des indices relativement élevés par rapport aux indices obtenus pour la FVG.

| Critére d'évaluation | LPE | FVG |
|---|---|---|
| Contraste inter-région de Levine et Nazif | 0.0969 | 0.0847 |
| Contraste intra-région de Levine et Nazif | 0.9201 | 0.9031 |
| Combinaison disparité intraclasse et interclasse | 0.5009 | 0.5009 |
| Contraste de Zeboudj | 0.546 | 0.420 |

TAB. 1 – *Comparaison des critères d'évaluation pour la LPE et le FVG.*



*a) LPE.* *b) FVG.* *c) Fusion de LPE et FVG.*

FIG. 6 - *Extraction de la forme par combinaison de la LPE et le FVG.*

# 4 Descripteurs de Formes

Le problème de la description de la forme est largement traite dans la littérature. La modélisation de la forme présentée dans cette section est basée sur la transformée de Fourier qui est un outil adapte à une modélisation satisfaisante des formes. En effet, une modélisation satisfaisante des formes doit être précise et invariante à un certain nombre de transformations géométriques susceptibles de se manifester dans les images ISAR (changement d'échelle, rotation, translation).

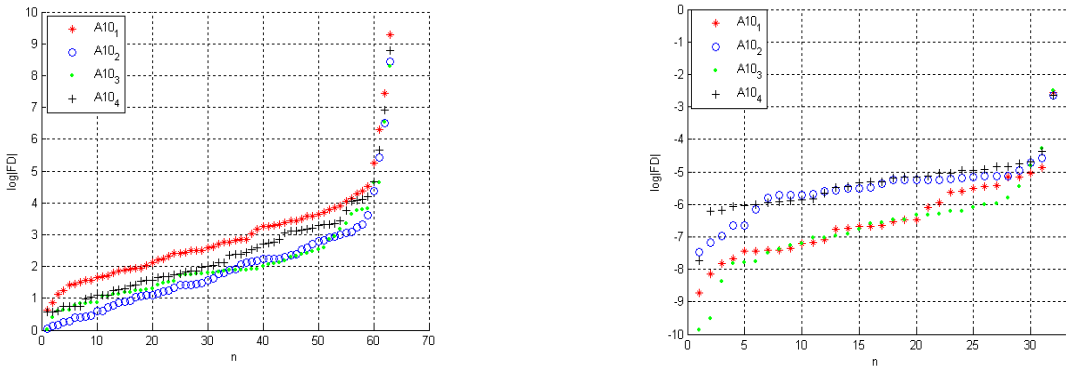
## 4.1 Descripteurs de Fourier

Les descripteurs de Fourier constituent un outil classique et largement utilisé pour la description des contours. Nous les avons choisis pour caractériser la forme. Ce choix sur les descripteurs de Fourier (Sarfraz, 2006) est motivé par leur simplicité d'implémentation et leur rapidité de calcul.

Avant d'appliquer la transformation de Fourier sur la forme extraite par les méthodes citées précédemment, de la cible, la fonction de forme a été normalisée en 64 points (L=64). En résultat, la signature obtenue pour chaque image est constituée de L/2 descripteurs de Fourier en utilisant la distance centroide comme signature de la forme. Le vecteur de distance centroïde *r* peut être calculé en utilisant l'équation suivante :

$$r(t) = \left( [x(t) - x_c]^2 + [y(t) - y_c]^2 \right)^{1/2} \text{ Avec } x_c = \frac{1}{L}\sum_{i=0}^{L-1} x(i) \quad , \quad y_c = \frac{1}{L}\sum_{i=0}^{L-1} y(i)$$

*x* et *y* représente respectivement les vecteurs d'abscisses et d'ordonnées des points de contours. La figure 5 présente les descripteurs de Fourier des 4 formes de la cible A10 (avec différents angles de rotation).




*a) Signature par coordonnées complexes.* *b) Signature par distance centroïde.*

FIG. 5 – *Descripteurs de Fourier.*

D'après les résultats illustrés dans la figure ci-dessus, nous pouvons tirer une première constatation que les descripteurs de Fourier des 4 formes de la cible A10 calculés par la signature des coordonnées complexes présentent une disparité moins importante que pour ceux calculés à partir du vecteur des distances centroïdes. Cependant, il reste prématuré de tirer des conclusions à ce niveau sur la pertinence de ces deux représentations de la forme. Par conséquent, l'étape de classification se montre très utile afin de valider nos résultats en termes de taux de reconnaissance en utilisant chacune des primitives extraites.

# 5 Recherche et classification des images

## 5.1 Calcul de similarité/dissimilarité

Généralement, le processus de recherche d'image (ou aussi classification d'images) se base sur l'image requête (image exemple) pour effectuer une recherche en la comparant avec toutes les images de la base de données. La comparaison entre les images (images requête et les images de la base de données) est réalisée par l'intermédiaire d'une mesure de similarité/dissimilarité entre les signatures qui décrivent le contenu visuel des images. Au final, les images qui présentent le plus/moins de similarité/dissimilarité sont sélectionnées, ordonnées et présentées a l'utilisateur.

Dans le cadre de ce présent travail, nous utilisons la distance de Hausdorff afin de mesurer la dissimilarité entre les ensembles de descripteurs de Fourier. Nous présentons dans la figure 6 les résultats obtenus caractérisant la distance de Hausdorff entre une forme requête et toutes les autres formes de la même cible et cela pour les deux techniques citées précédemment. Nous pouvons constater d'après la figure 6 que la LPE présente des meilleurs résultats que ceux donnés par le FVG.

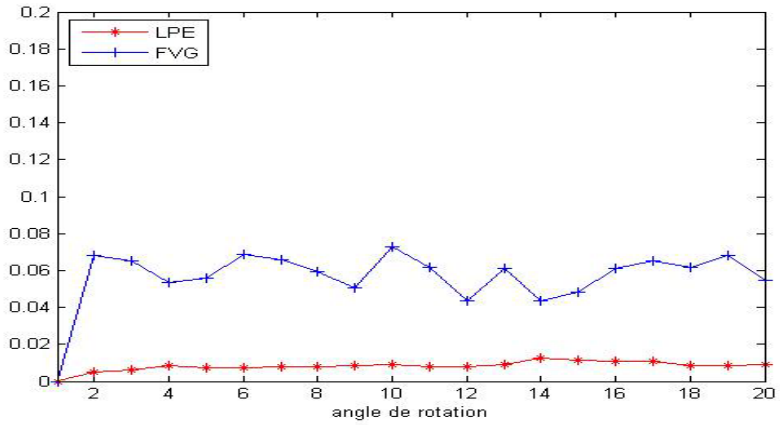



FIG. 6 – *Dissimilarité entre la forme A10 de 0° et les formes de A10 de différents angle de rotation, extraites par la LPE et le FVG.*

## 5.2 Résultats

Afin de valider les résultats présentés dans la section précédente, l'étape de classification est réalisée à base de la machine à vecteurs support (SVM). Nous avons utilisé pour cela, le noyau RBF. Les tests sont réalisés sur une base de données de 810 images représentant 5 cibles (162 images pour chaque cible). Chaque forme est représentée par 31 descripteurs de Fourier. Les taux moyens de bonne classification obtenus pour les deux technique de segmentation (extraction de contours) sont donnés par le tableau 2 en fonction de la méthode utilisée (validation croisée) pour la sélection de la base de test/d'apprentissage.

| **Méthodes d'extraction de forme** | **Taux moyen de bonne classification** | |
|---|---|---|
| | ***Validation Croisée V-2*** | ***Validation Croisée V-3*** |
| LPE (Sans Prétraitement) $(C = 10000, \gamma = 15)$ | 56.34% | 58.12% |
| LPE (Avec Prétraitement) $(C = 100, \gamma = 150)$ | 85.67% | 88.39% |
| Flux du Vecteur Gradient (FVG) $(C = 1000, \gamma = 300)$ | 47.37% | 48.27% |

TAB. 2 – *Résultats de classification en utilisant SVM de noyau RBF.*

Nous constatons d'après le Tableau 2 que le taux moyen de bonne classification (88.39%) des données extraites par la LPE avec un prétraitement sur les images ISAR présente un meilleur résultat relativement au taux de bonne classification (48.27%) obtenu en utilisant le FVG. Par ailleurs, le prétraitement a montré son intérêt par l'amélioration considérable sur le taux moyen de bonne reconnaissance en utilisant la LPE sans et avec le prétraitement.

# 6 Conclusion

Dans ce papier, nous nous sommes focalisés particulièrement sur l'étape de préparation de données, notamment l'extraction de la forme de la cible et sa modélisation. Les résultats obtenus peuvent nous aider à tirer quelques conclusions comparatives entres les différentes techniques d'extraction de contours depuis les images ISAR. La LPE présente un meilleur outil pour extraire des contours fermés afin de préparer les données pour la phase de reconnaissance. Il serait toutefois important de confirmer les résultats en termes de robustesse au bruit et à la volumétrie de données. Pour ce qui est de la signature de la forme, l'utilisation des coordonnées complexes nous semble donner plus de précision que la signature avec les distances controïdes, Cependant, cela reste à valider en termes de taux de bonne classification.

Quand à l'étape de reconnaissance, nous avons pu comparer les performances globales du système où nous avons pu constater que les données prétraitées extraites par la LPE donnent de meilleurs résultats relativement aux autres techniques.

Au final, nos travaux en cours portent sur l'utilisation d'autres techniques de classification, ainsi que l'utilisation d'autres signatures afin d'améliorer le taux de reconnaissance du système.

# Références

Chabrier, S., B. E., H. Laurent, C. Rosenberger, et P. Marché (2004). Unsupervised Evaluation of Image Segmentation: Application to multi-spectral images. *International Conference on Pattern Recognition*, (3), pages 576-579, Cambridge.

Frost, V. S., J. A. S., A. Josephine, K. S. Shanmu-gan et J.C. Holtzman (1982). A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-4, No. 2.

Roerdink, J.B.T.M. et A. Meijster (2001) The Watershed Transform: Definitions, Algorithms and Parallelization strategies. In IOS press, Fundamenta Informaticae. Vol. 41:187-228.

Lee, J.S. (1981). Speckle Analysis and Smoothing of Synthetic Aperture Radar Images. *Computer Graphics and Image Processing*, Vol. 17:24-32.

Philipp-Foliguet, S., et L. Guigues (2001). Evaluation de la segmentation d'images : état de l'art, nouveaux indices et comparaison. *Rapport technique.*

Saidi, M. N., B. Hoeltzener, A. Toumi, A. Khenchaf, et D. Aboutajdine (2008). Automatic recognition of ISAR Images: Target shapes features extraction. *International Conference on Information & Communication Technologies: from Theory to Applications*, ICTTA'08. Damascus, Syria.

Sarfraz M. (2006). Object Recognition using Fourier Descriptors: Some Experiments and Observations. *International Conference on Computer Graphics, Imaging and Visualisation*. CGIV'06.

Toumi, A., B. Hoeltzener, et A. Khenchaf (2008). Hierarchical segmentation on ISAR image for target recognition. *International Journal of Computational Intelligence Research, special issue*.

Toumi, A., B. Hoeltzener, et A. Khenchaf (2008b)., Traitements des signaux radar pour la reconnaissance/identification de cibles aériennes, à apparaitre dans RNTI : *revue des nouvelles technologies, RNTI-C-2 Classification : points de vue croisés.*

Toumi, A. (2007). Intégration des bases de connaissances dans les systèmes d'aide à la décision : Application à l'aide à la reconnaissance de cibles radar non-coopératives. Thèse de doctorat de l'université de Bretagne occidentale.

Xu, C. et J. L. Pince, (1997). Gradient Vector Flow: A New External Force for Snakes. *IEEE Proc. Conf. on Comp. Vis. Patt. Recogn*. CVPR' 97.

Xu, C. et J. L. Prince (1998). Snakes, shapes, and gradient vector flow: Special issue on partial differential equations and geometry-driven diffusion in image processing and analysis. *IEEE transactions on image processing.* vol. 7, no3, pp. 359-369.

## Summary

This paper presents intelligent systems for automatic non-cooperative targets recognition in uncertain environment. We use a process of knowledge discovery form data (KDD process) which has been adapted to radar field. We present results and simulations in the data preparation step of the KDD process, in particular, a shape extraction from ISAR images (Inverse Synthetic Aperture Radar) and the computation of feature vectors.

# Exploration temporelle de données archéologiques imprécises : graphe d'antériorité.

Cyril de Runz, Eric Desjardin

CReSTIC-SIC
IUT de Reims Châlons Charleville
Rue des Crayères, BP 1035, 51687 Reims Cedex 2
{cyril.de-runz,eric.desjardin}@univ-reims.fr,
http://crestic.univ-reims.fr

**Résumé.** Dans cet article, nous proposons une nouvelle technique d'exploration temporelle d'un ensemble d'objets archéologiques dont les périodes d'activité sont représentées par des nombres flous. Pour cela, en se basant sur la définition d'un indice d'antériorité entre deux nombres flous, on construit un graphe orienté pondéré dont les sommets seront les objets archéologiques. À l'aide de ce graphe d'antériorité, nous déterminons le potentiel d'antériorité, de postériorité ainsi que la position temporelle de l'objet associé au sommet dans l'ensemble des objets de la base. L'information dégagée par ce graphe met en lumière les rapports temporels entre objets. Nous avons appliqué cette démarche aux données portant sur les tronçons de rues datant de l'époque Romaine et trouvés à l'époque romaine.

## 1 Introduction

Dans un Système d'Information Géographique (SIG) dédié à l'archéologie, l'analyse exploratoire de données cherche à dégager des relations et des corrélations afin d'extraire de nouvelles connaissances entre les objets d'une base de données spatiotemporelles. Par exemple, dans de Runz et al. (2008), nous avons proposé une méthode pour l'extraction de représentants d'un ensemble d'objets archéologiques dans un SIG. Dans cet article, nous explorons la base de données afin de déterminer le positionnement temporel de chaque objet. Cette exploration est riche d'information pour l'archéologue. C'est d'ailleurs une démarche largement utilisée pour l'étude stratigraphique d'un chantier archéologique par le diagramme de Harris[1].

Le positionnement temporel des objets fournit de nouvelles connaissances sur l'information portée par ces objets. Ces nouvelles connaissances facilitent la compréhension des données archéologiques et donc l'expertise. Dans cet article, nous positionnerons temporellement, dans une base de données spatiotemporelles dédiée à l'archéologie, chaque objet archéologique par rapport aux autres objets, dont les périodes d'activité sont représentées par des nombre flous.

Comme les comparaisons de deux nombres flous sont souvent non transitives (voir Wang et al. (1995)), celles-ci ne sont pas directement utilisées pour le classement de nombres flous dans un ensemble. Cependant, une alternative consiste à positionner chaque nombre flou individuellement dans l'ensemble. Celle-ci a conduit à la définition de nombreuses techniques permettant de positionner un nombre flou relativement à un ensemble de nombres flous et non vis-à-vis de chaque nombre.

Parmi ces méthodes, on retrouve l'approche proposée par Kerre (1982). Pour Kerre, après avoir déterminé le maximum de l'ensemble selon le principe d'extension défini dans Zadeh (1965), le positionnement d'un élément correspond à sa distance au maximum des éléments. Une autre approche est celle de Jain qui considère les nombres selon leur intersection à un ensemble flou appelé « maximisant » construit sur l'union des supports des nombres flous à positionner (Jain (1977)). Dans la démarche proposée dans cet article, nous proposons de regarder le positionnement de chaque objet de la base de données via sa capacité à être antérieur ou/et postérieur à chaque objet de la base.

Pour cela, nous utilisons l'indice d'antériorité proposé dans (de Runz et al. (2006, 2008)). En utilisant cet indice sur un ensemble d'objets archéologiques dont les périodes d'activité sont représentées par des nombres flous, nous proposons de définir un graphe orienté pondéré dont les sommets représentent les tronçons de rues et dont les arcs ont pour coût l'indice d'antériorité de l'origine de l'arc par rapport à la destination.

[1] Diagramme de Harris : méthode de représentation qui à l'aide de traits et de cases modélise les rapports qu'entretiennent les couches stratigraphiques entre elles.

La capacité d'antériorité d'un objet est déterminée par la somme des coûts des arcs sortants du sommet associé à l'objet. La capacité de postériorité d'un objet est définie par la somme des coûts des arcs entrants dans le sommet. L'indice temporel d'un objet est le différentiel entre sa capacité de postériorité et sa capacité d'antériorité.

Le positionnement temporel de chaque objet dans un ensemble correspond au rang obtenu par la valeur de son indice temporel dans l'ensemble des valeurs de l'indice temporel de tous les objets. Ce graphe permet donc une représentation synthétique et formelle des structures temporelles de l'échantillon.

Dans le cadre du programme SIGRem (Pargny et Piantoni (2005)), nous proposons de construire un tel graphe sur les données portant sur les tronçons de rues romaines dont les périodes d'activité ont préalablement été modélisées par des nombres flous[2]. Ces données sont stockées dans la base *BDFRues*. Par la construction de ce graphe sur les objets de *BDFRues*, nous proposons donc de déterminer le positionnement temporel de chaque objet. À partir de ces positions temporelles, nous extrayons de *BDFRues* l'objet le plus antérieur ou l'objet le plus postérieur.

Afin d'illustrer cet article, le cas de trois objets spatiotemporels $A_1$, $A_2$ et $A_3$, dont les composantes temporelles sont respectivement représentées par les nombres flous $A_1.fDate = (0, 48, 68)$, $A_2.fDate = (22, 22, 57, 57)$ et $A_3.fDate = (11, 33, 81)$, dont les fonctions d'appartenance sont présentées figure 1, sera considéré.

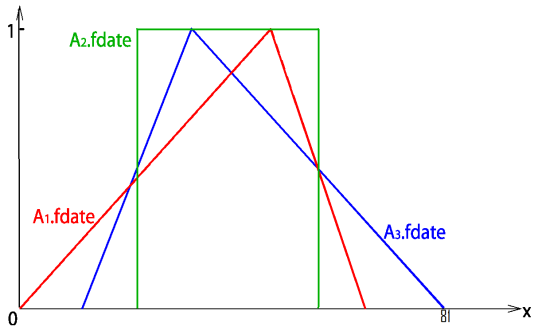



FIG. 1 – *Fonctions d'appartenance $A_1.fdate$, $A_2.fdate$ et $A_3.fdate$ de respectivement $A_1.fDate$, $A_2.fDate$ et $A_3.fDate$*

Ainsi, après avoir présenté la démarche classique pour le positionnement d'un nombre flou dans son ensemble, nous étudierons via la construction du graphe, la position temporelle de chaque objet par l'intermédiaire de sa capacité d'antériorité et de sa capacité de postériorité dans l'ensemble des objets. Nous terminerons cet article par l'application de notre démarche sur les données archéologiques contenues dans *BDFRues*.

## 2 Rangement de nombres flous dans un ensemble

Considérons un ensemble $\Omega$ de $n$ nombres flous $\{F_1, \ldots, F_n\}$. Les méthodes usuelles de classement les rangent par comparaison des valeurs associées à chacun d'entre eux et calculées selon un indice. Celles-ci sont obtenues à l'aide d'une ou plusieurs valeurs de référence ($val_{ref}$) définies sur l'ensemble des nombres flous (Wang et Kerre (2001a)). Dans $\Omega$, la position d'un nombre flou $F_i$ correspond au rang que confère la valeur de l'indice pour $F_i$ au regard des valeurs de l'indice pour tous les autres nombres flous.

Classiquement on retrouve deux grands types de méthodes de rangement. Dans le premier type, les techniques (*c.f.* Jain (1977) et ses dérivées) se basent sur un ensemble flou de référence, appelé ensemble flou maximisant $\Omega$. Dans le second type (*c.f.* Kerre (1982) et ses dérivées), les méthodes utilisent l'ensemble flou maximum sur $\Omega$, au sens de l'extension de Zadeh.

### 2.1 Approche de Jain et ses dérivées

Dans l'approche de Jain (1977), considérant une valeur $k > 0$ donnée, on définit un ensemble flou $F^k_{max}$ maximisant $\Omega$ dont la fonction d'appartenance $f^k_{max}$ est :

$$f^k_{max}(x) = \left(\frac{x}{x_{max}}\right)^k,$$

[2]La modélisation floue des objets archéologiques a été présentée dans de Runz et al. (2007).

avec $k \in \mathbb{R}^{+*}$, $x \in \bigcup_{i=1}^{n} Support(F_i)$, $x \geq 0$, $x_{max} = \sup(\bigcup_{i=1}^{n} Support(F_i))$.

Cela est valable si : $\min(\bigcup_{i=1}^{n} Support(F_i)) \geq 0$. Si ce n'est pas le cas on procède à un changement de variable.

L'indice de Jain d'un nombre flou $F_i$ ($J_\Omega^k(F_i)$) est alors la hauteur de la *t-norme* proposée par Zadeh entre $F_i$ et $F_{max}^k$ :

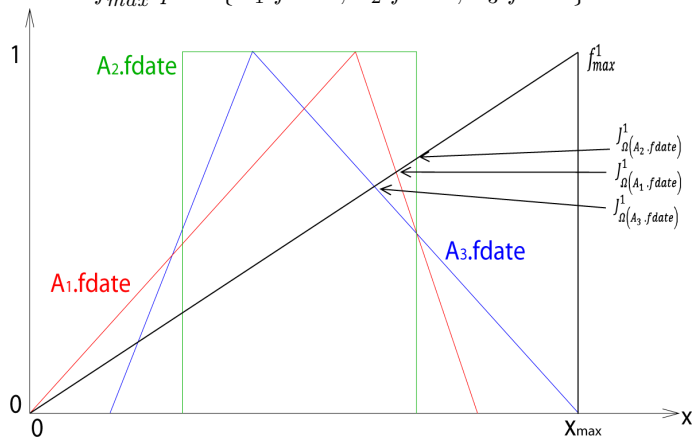
$$J_\Omega^k(F_i) = Hauteur(F_i \wedge F_{max})$$

Soient $F_i$ et $F_j$ appartenant à $\Omega$, si $J_\Omega^k(F_i)$ est plus grand que $J_\Omega^k(F_j)$ alors $F_i$ aura un plus haut rang que $F_j$.

Différentes approches dérivées de celles de Jain ont été proposées dans la litterature. Par exemple, l'approche de Chen (1985) complète celle de Jain par l'ajout d'une valeur de référence : l'ensemble flou minimisant $\Omega$. A l'aide des ensembles flous maximisant et minimisant $\Omega$, il détermine une utilité gauche et une utilité droite d'un nombre flou sur lesquelles il base le calcul de l'indice pour le nombre flou considéré.

Pour les objets $A_1$, $A_2$ et $A_3$ dont les périodes d'activité sont illustrées dans la Figure 1, $\Omega$ est composé de $A_1.fDate$, $A_2.fDate$ et $A_3.fDate$. La fonction d'appartenance $f_{max}^1$ de l'ensemble flou $F_{max}^1$ maximisant $\Omega$ pour $k = 1$ est illustrée dans la figure 2.

FIG. 2 – $f_{max}^1$ *pour* $\{A_1.fdate, A_2.fdate, A_3.fdate\}$ *avec* $k = 1$




Les valeurs de l'indice de Jain pour les représentations des périodes d'activité des objets $A_1$, $A_2$ et $A_3$, pour $\Omega = \{A_1, A_2, A_3\}$ et pour $k = 1$ sont :

$$J_\Omega^1(A_1.fDate) = 0.67, \quad J_\Omega^1(A_2.fDate) = 0.71, \quad J_\Omega^1(A_3.fDate) = 0.63.$$

Donc $J_\Omega^1(A_3.fDate) < J_\Omega^1(A_1.fDate) < J_\Omega^1(A_2.fDate)$. Le rangement selon Jain pour $k = 1$ sera donc dans l'ordre croissant des rangs $A_3.fDate$, $A_1.fDate$, et $A_2.fDate$. Les objets sont donc positionnés dans l'ordre croissant de la manière suivante : $A_3$, $A_1$ et $A_2$.

## 2.2 Approche de Kerre et ses dérivées

La proposition de Kerre est la suivante : si on prend l'ensemble des nombres flous à classer, et que l'on compare les distances séparant chaque nombre flou du maximum de l'ensemble, alors on peut ranger les nombres flous par ordre de distance décroissante ; le plus petit nombre aura la distance au maximum la plus grande.

Pour un ensemble $\Omega$ de $n$ nombres flous $\{A_1, A_2, \ldots, A_n\}$, Kerre propose de ranger ces nombres flous par la comparaison de leur distance de Hamming au maximum de l'ensemble $\Omega$ selon le principe d'extension Kerre (1982).

La distance de Hamming entre deux nombres flous $F$ et $G$ de fonctions d'appartenance $f$ et $g$ est définie dans Wang et Kerre (2001b) comme suit :

$$D_H(F, G) = \int |f(x) - g(x)| dx.$$

La distance de Hamming entre $A_i$, avec $i \in [1, n]$, et $\widetilde{max}(A_1, A_2, \ldots, A_n)$ est appelée indice de Kerre $K_\Omega(A_i)$. Ainsi, l'indice de Kerre de $A_i$ dans $\{A_1, A_2, \ldots, A_n\}$ est obtenu de la manière suivante :

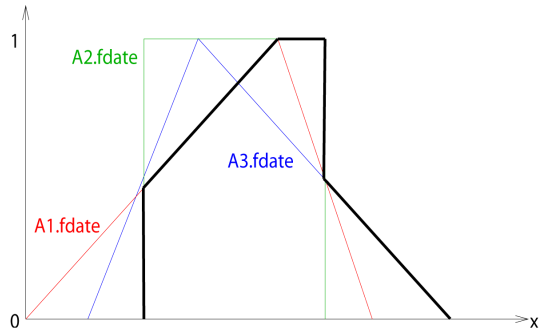
$$K_\Omega(A_i) = D_H\left(A_i, \widetilde{max}(A_1, A_2, \ldots, A_n)\right) \tag{1}$$

Pour Kerre, un nombre flou est plus petit qu'un autre dans un ensemble donné si et seulement si son indice de Kerre est supérieur à celui de l'autre. C'est à dire $A_i \preccurlyeq A_j$ par rapport à $\Omega = \{A_1, A_2, \ldots, A_n\}$, avec $(i, j) \in [1, n]$, si et seulement si $K_\Omega(A_i) \geq K_\Omega(A_j)$.

Depuis, différentes approches se sont basées sur l'évaluation de la proximité de chaque nombre flou $F_i$ au maximum ou au minimum de $\Omega$. Ainsi, par exemple, en 1987, Wang (voir dans Wang et Kerre (2001a)) proposa d'utiliser d'autres méthodes de quantification de la proximité au maximum.

Pour les objets $A_1$, $A_2$ et $A_3$ dont les représentations des périodes d'activité sont illustrées dans la Figure 1, $\Omega$ est composé de $A_1.fDate$, $A_2.fDate$ et $A_3.fDate$. La fonction d'appartenance du maximum sur $\Omega$ selon le principe d'extension de Zadeh est représentée Figure 3.

FIG. 3 – $\widetilde{max}(A_1.fDate, A_2.fDate, A_3.fDate)$




Les valeurs de l'indice de Kerre pour les représentations des périodes d'activité des objets $A_1$, $A_2$ et $A_3$ et pour $\Omega = \{A_1.fDate, A_2.fDate, A_3.fDate\}$ sont :

$$K_\Omega(A_1.fDate) = 11.1, \quad K_\Omega(A_2.fDate) = 13.1, \quad K_\Omega(A_3.fDate) = 10.7.$$

Ainsi $K_\Omega(A_1.fDate) < K_\Omega(A_2.fDate)$ et $K_\Omega(A_1.fDate) > K_\Omega(A_3.fDate)$. À partir de ces comparaisons, le rangement, selon Kerre, des objets par l'intermédiaire de leurs périodes d'activité est, dans l'ordre croissant des rangs $A2.fDate$, $A1.fDate$, $A3.fDate$.

Comme on peut le voir, dans ces méthodes, le rangement dans l'ensemble ne se fait pas par comparaison deux à deux des nombres flous mais par le calcul d'un indice reposant sur la définition de valeurs de référence sur l'ensemble des nombres à comparer. Une des causes à cela est que les méthodes de comparaison deux à deux de nombres flous sont le plus souvent non transitives (Wang et al. (1995)).

Dans l'approche proposée ci-après, l'idée est d'utiliser l'indice d'antériorité *Ant* (voir de Runz et al. (2006, 2008)) plutôt qu'une méthode de comparaison deux à deux donnant une décision binaire. Ainsi, au regard des objets archéologiques, dont les périodes d'activité sont représentées par un nombre flou, la quantification de l'antériorité permet d'étudier le positionnement temporel de chaque objet à l'aide de la construction d'un graphe orienté pondéré (le graphe d'antériorité). Ce graphe permet l'exploration schématique des données selon l'information temporelle.

## 3 Graphe d'antériorité

L'analyse exploratoire proposée dans ce chapitre se base sur la construction d'un graphe orienté pondéré, appelé graphe d'antériorité, à partir de l'indice d'antériorité *Ant*. Ce dernier quantifie une relation binaire, définie sur les nombres flous et interprétée comme antériorité, entre deux nombres flous $F$ et $G$ quelconques de la manière suivante :

$$Ant(F, G) = \begin{cases} \frac{K_{\{F,G\}}(F)}{K_{\{F,G\}}(F) + K_{\{F,G\}}(G)} & \text{si } K_{\{F,G\}}(F) + K_{\{F,G\}}(G) > 0, \\ 1 & \text{si } K_{\{F,G\}}(F) + K_{\{F,G\}}(G) = 0; \end{cases}$$

avec $\Omega = \{F, G\}$.

L'ensemble des nombres flous à comparer, la relation d'antériorité et les valeurs de l'indice d'antériorité fournissent les sommets, les arcs et les coûts du graphe d'antériorité.

La construction du graphe d'antériorité ayant pour objectif l'observation du positionnement temporel de chaque objet d'une base de données archéologiques par rapport aux autres, les nombres flous mis en relation sont les représentations des périodes d'activité des objets archéologiques. Ainsi on apparie temporellement les objets archéologiques les uns aux autres et on quantifie les paires obtenues à l'aide d'une représentation schématique.

## 3.1 Construction du graphe

Considérons un ensemble d'éléments $E$, une relation binaire $\mathcal{R}$ sur $E$ (*i.e.* un sous-ensemble de $E \times E$) et une application *App* sur cette relation binaire prenant valeur dans $\mathbb{R}$.

Un graphe orienté pondéré $G_{\mathcal{R}}(L_S, L_A, L_C)$ est une représentation schématique constituée d'un ensemble $L_S$ de sommets, d'un ensemble $L_A$ d'arcs reliant les sommets deux à deux, et d'un ensemble $L_C$ de coûts associés aux arcs.

Soit deux éléments $A$ et $B$ de $E$ reliés par la relation binaire $\mathcal{R}$, on associe, dans $G_{\mathcal{R}}(L_S, L_A, L_C)$, à $A$ le sommet $S_A$ et à $B$ le sommet $S_B$ :

$$A \in E \Leftrightarrow S_A \in L_S.$$

L'arc $(S_A, S_B)$ représente alors le fait que $A\mathcal{R}B$ :

$$A\mathcal{R}B \Leftrightarrow (S_A, S_B) \in L_A.$$

Le coût $C(S_A, S_B)$ d'un arc $(S_A, S_B)$ est égal à $App(A, B)$ :

$$(S_a, S_b) \in L_A \Leftrightarrow (C(S_A, S_B) = App(A, B) \text{ et } C(S_A, S_B) \in L_C).$$

L'analyse exploratoire de cette partie se base sur la construction d'un graphe orienté pondéré à partir de l'indice d'antériorité et de l'ensemble $\Omega$ d'objets d'une base de données archéologiques dont les périodes d'activité sont définies et représentées par des nombres flous. Ce graphe sera noté $G_{Ant}(L_S, L_A, L_C)$ et sera appelé graphe d'antériorité.

La construction de ce graphe suit la démarche suivante. La relation $\mathcal{R}$ est une relation qui relie deux à deux tous les objets de $\Omega$ :

$$\forall A_i, A_j \in \Omega,\ A_i\mathcal{R}A_j \text{ et } A_j\mathcal{R}A_i.$$

À chaque objet $A_i$ de $\Omega$ de période d'activité $A_i.fDate$, nous associons un sommet $S_{A_i}$ du graphe $G_{Ant}(L_S, L_A, L_C)$. Ainsi, le cardinal de $L_S$ est égal à celui de $\Omega$. Soit deux objets $A_i$ et $A_j$ de $\Omega$ représentés respectivement par les sommets $S_{A_i}$ et $S_{A_j}$ du $G_{Ant}(L_S, L_A, L_C)$, l'arc $(S_{A_i}, S_{A_i})$ représente l'antériorité possible de $A_i$ à $A_j$ au regard de leurs périodes d'activité. Le coût $C(S_{A_i}, S_{A_j})$ de l'arc $(S_{A_i}, S_{A_j})$ est égal à $Ant(A_i.fDate, A_j.fDate)$ et représente la quantification de l'antériorité de $A_i$ à $A_j$.

Pour l'exemple des objets dont les périodes d'activité sont représentées dans la figure 1, $\Omega = \{A_1, A_2, A_3\}$ et les valeurs de l'indice d'antériorité pour l'ensemble des couples de $\Omega$ sont :

$$\begin{array}{ll} Ant(A_1.fDate, A_2.fDate) = 0.44, & Ant(A_2.fDate, A_1.fDate) = 0.56, \\ Ant(A_1.fDate, A_3.fDate) = 0.51, & Ant(A_3.fDate, A_1.fDate) = 0.49, \\ Ant(A_2.fDate, A_3.fDate) = 0.50, & Ant(A_3.fDate, A_2.fDate) = 0.50. \end{array}$$

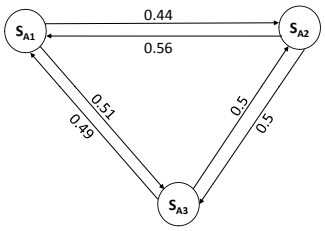
Le graphe d'antériorité $G_{Ant}(L_S, L_A, L_C)$ est alors représenté dans la figure 4, dans lequel un sommet $S_{A_i}$ correspond à l'objet $A_i$ ayant pour période d'activité $A_i.fDate$.

Pour chaque sommet du graphe $G_{Ant}(L_S, L_A, L_C)$ ainsi construit, la somme des coûts des arcs sortants, celle des coûts des arcs entrants et leur différence sont des valeurs particulières et ont une signification importante.

## 3.2 Capacité d'antériorité, capacité de postériorité et positionnement d'un objet

Soient $S_{A_i}$ et $S_{A_j}$ deux sommets de $G_{Ant}(L_S, L_A, L_C)$ correspondant respectivement aux objets $A_i$ et $A_j$, le coût $C(S_{A_i}, S_{A_j})$ correspond à la valeur de l'indice d'antériorité $Ant(A_i.fDate, A_j.fDate)$. Ainsi, la somme

FIG. 4 – *Graphe d'antériorité pour* $\Omega = \{A_1, A_2, A_3\}$




des coûts des arcs sortants de $S_{A_i}$ correspond à la capacité d'antériorité $CapAnt_\Omega(A_i)$ de $A_i$ relativement à l'ensemble des nombres flous $\Omega$ :

$$CapAnt_\Omega(A_i) = \sum_{\substack{S_{A_j} \in L_S \\ (S_{A_i}, S_{A_j}) \in L_A}} C(S_{A_i}, S_{A_j}).$$

Pour l'exemple où $\Omega = \{A_1, A_2, A_3\}$, d'après le graphe d'antériorité présenté Figure 4, les capacités d'antériorité des objets de $\Omega$ sont les suivants :

$$CapAnt_\Omega(A_1) = 0.95, \quad CapAnt_\Omega(A_2) = 1.06, \quad CapAnt_\Omega(A_3) = 0.99.$$

L'indice $Ant(A_j.fDate, A_i.fDate)$ entre deux nombres flous $A_j.fDate$ et $A_i.fDate$ a été présenté comme l'indice d'antériorité de $A_j$ à $A_i$. On peut aussi le voir comme l'indice de postériorité de $A_i$ à $A_j$. Ainsi, la somme des coûts des arcs entrants sur le sommet $S_{A_i}$ de $G_{Ant}(L_S, L_A, L_C)$ associé à l'objet $A_i$ de $\Omega$ peut être assimilée à la capacité de postériorité $CapPost_\Omega$ de $A_j$ relativement à l'ensemble des nombres flous $\Omega$ :

$$CapPost_\Omega(A_i) = \sum_{\substack{S_{A_j} \in L_S \\ (S_{A_j}, S_{A_i}) \in L_A}} C(S_{A_j}, S_{A_i}).$$

Pour l'exemple où $\Omega = \{A_1, A_2, A_3\}$, d'après le graphe présenté Figure 4, les capacités de postériorité des objets de $\Omega$ sont les suivants :

$$CapPost_\Omega(A_1) = 1.05, \quad CapPost_\Omega(A_2) = 0.94, \quad CapPost_\Omega(A_3) = 1.01.$$

L'indice temporel d'un objet $A_i$ doit permettre de définir sa position temporelle. Son calcul doit donc prendre en considération à la fois sa capacité à être postérieur et sa capacité d'antériorité. C'est pourquoi, l'indice temporel $IndTemp_\Omega(A_i)$ de $A_i$ est déterminé comme suit :

$$IndTemp_\Omega(A_i) = CapPost_\Omega(A_i) - CapAnt_\Omega(A_i).$$

Les valeurs de l'indice temporel des objets $A_1$, $A_2$ et $A_3$, dont les fonctions d'appartenance des périodes d'activité sont représentées dans la figure 1 et pour le graphe d'antériorité présenté dans la figure 4 (avec $\Omega = A_1, A_2, A_3$), sont :

$$IndTemp_\Omega(A_1) = 0.10, \quad IndTemp_\Omega(A_2) = -0.12, \quad IndTemp_\Omega(A_3) = 0.02.$$

À l'aide de cet indice, nous déterminons la position temporelle $PosTemp(A_i)$ de $A_i$ et $PosTemp(A_j)$ de $A_j$ dans l'ensemble $\Omega$ en suivant le principe suivant :

$$\text{Si } IndTemp_\Omega(A_i) > IndTemp_\Omega(A_j), \text{ alors } PosTemp(A_i) > PosTemp(A_j).$$

La position temporelle de $A_i$ correspond au rang de $A_i$ dans la liste des objets de $\Omega$ ordonnée selon les valeurs de l'indice temporel.

Pour l'exemple de la figure 1, avec $\Omega = \{A_1, A_2, A_3\}$, d'après le graphe présenté figure 4, les positions temporelles des objets de $\Omega$ sont dans l'ordre croissant : $A_2$, $A_3$, $A_1$. Cet ordre n'est donc pas le même que celui obtenu par l'approche de Kerre ni le même que celui issu de la démarche de Jain. Toutefois, grâce aux valeurs de l'indice d'antériorité, la méthode proposée définit un indice de position temporelle d'un individu vis-à-vis des autres dans un ensemble. On obtient une vision pondérée de la structure temporelle de l'ensemble.

Les valeurs de l'indice temporel et les positions temporelles des objets donnent des informations importantes sur les relations temporelles entre objets.

### 3.3 Analyse des objets selon le graphe d'antériorité

À l'aide du graphe d'antériorité, on peut extraire trois objets particuliers : l'objet le plus antérieur, l'objet le plus postérieur et l'objet temporellement médian.

L'objet le plus vieux dans le cadre applicatif, c'est-à dire le plus antérieur, noté $PA$, est celui dont la valeur de l'indice temporel est la plus petite dans l'ensemble des valeurs de l'indice temporel des objets de $\Omega$. La position temporelle de l'objet le plus antérieur est la position temporelle minimale des objets de $\Omega$ :

$$PosTemp(PA) = \min_{A_i \in \Omega} (PosTemp(A_i))$$

L'objet le plus récent dans le cadre applicatif, c'est-à dire le plus postérieur, noté $PP$, est celui dont la valeur de l'indice temporel est la plus grande dans l'ensemble des valeurs de l'indice temporel des objets de $\Omega$. La position temporelle de l'objet le plus antérieur est la position temporelle maximale des objets de $\Omega$ :

$$PosTemp(PP) = \max_{A_i \in \Omega} (PosTemp(A_i))$$

Pour les objets de la figure 1, l'objet $A_2$ est le plus antérieur.

Grâce à l'approche par rang (position temporelle), il est trivial de définir l'objet temporellement médian, noté $TM$. Il est celui dont la valeur de l'indice temporel est médiane à l'ensemble des valeurs de l'indice temporel pour les objets de $\Omega$.

Pour les objets de l'exemple, on a :

$$PA = A_2, \quad PP = A_1, \quad TM = A_3.$$

De plus, on peut considérer qu'un objet ayant un indice temporel négatif peut être considéré, par rapport à l'ensemble des objets, comme un objet « plutôt antérieur » tandis qu'un objet ayant un indice temporel positif correspond à un objet « plutôt postérieur ». Ceux ayant un indice d'antériorité nul ont une capacité d'antériorité égal à celle de postériorité. Ils occupent une position intermédiaire spécifique dans l'ensemble : nous les nommerons « anté-postérieurs ». Nous proposons ainsi une classification des objets selon qu'ils soient « plutôt antérieurs », « plutôt postérieurs » ou « anté-postérieurs ».

Dans l'exemple de la figure 1, $\Omega = \{A_1, A_2, A_3\}$, il n'y a pas d'élément anté-postérieur, $A_2$ est « plutôt antérieur » dans $\Omega$ et $\{A_1, A_3\}$ forme l'ensemble des objets « plutôt postérieurs » dans $\Omega$.

La construction du graphe d'antériorité est une approche originale pour le rangement d'objets archéologiques dont les périodes d'activité sont imparfaites et sont représentées par des nombres flous. Ce graphe donne une vision globale des relations chronologiques entre les objets archéologiques. Il donne de nombreuses indications en vue de la classification des objets archéologiques, de l'analyse à l'échelle locale (le chantier de fouille) et globale (la ville) et donc de la généralisation. Dans le cadre du projet SIGRem, les périodes d'activité des objets représentant les tronçons de rues trouvés à Reims et datant de l'époque romaine et stockées dans *BDFRues*, étant représentées par des nombres flous, ces objets sont dans la suite explorés afin d'en dégager leurs positions temporelles.

## 4 Application aux données archéologiques

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

Au-delà de l'élaboration de la cartographie archéologique de la cité des Rèmes[3], le projet *SIGRem* Pargny et Piantoni (2005), soutenu par la région Champagne Ardenne, l'état et la ville de Reims, et cadre applicatif de ce travail, porte sur la mise en place d'un Système d'Information Géographique (SIG) pluridisciplinaire Piantoni (2005). Il relève d'une ambition scientifique puisant ses outils conceptuels dans la recherche fondamentale, ses méthodes opérationnelles dans les technologies informatiques en matière d'analyse spatiale et son application pratique dans la mise en valeur des données archéologiques recueillies durant les trentes dernières années.

Dans cette partie, nous proposons d'appliquer le processus exploratoire proposé dans cet article sur la base de données *BDFRues*, partie intégrante du projet SIGRem. Cette base est dédiée aux éléments de rues romaines

[3]Cité des Rèmes : Reims et ses environs à l'époque romaine

à Reims, *BDFRues*. Les objets, qui y sont stockés, ont chacun une période d'activité représentée par un ensemble flou convexe et normalisé.

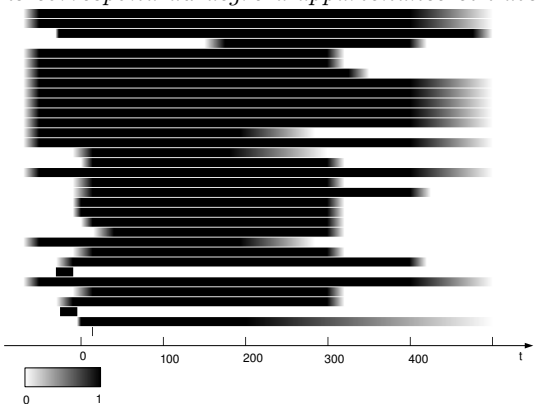
Ainsi dans en premier temps, nous présenterons la base de données *BDFRues*. Ensuite, afin d'obtenir une grille de lecture des résultats plus globale, nous présenterons les rangs des objets obtenus par le classement des représentations de leurs périodes d'activité selon les valeurs des indices de Jain et de Kerre. Enfin, nous exposerons les résultats issus de la construction du graphe d'antériorité.

## 4.1 A propos de *BDFRues*

Les données archéologiques sont des données spatio-temporelles, ce qui diffère des cas classiques des données géographiques. Quelques études, telles que Dragicevic et Marceau (2000), s'approchent conceptuellement de notre cadre de travail. Dans la base de données sur les rues de Durocortorum[4], les tronçons de rues sont caractérisés notamment par une période d'activité.

La datation de la période d'activité des objets est généralement issue d'interprétations ou d'estimations dépendantes de l'environnement de la découverte (lieux de fouilles, stratigraphie, comparaison aux objets se situant dans la même pièce...). De plus, la codification linguistique de périodes temporelles n'a pas toujours la même représentation. Par exemple l'estimation du début du Bas Empire varie selon les experts entre 193 et 284 après J.C. Elle est donc largement imprécise.

FIG. 5 – *Périodes floues d'activité des objets de BDFRues (Chaque ensemble flou -chaque période- est représenté par une "bande". Le niveau de gris correspond au degré d'appartenance et l'abscisse au temps).*




Nous représentons les périodes d'activité par des ensembles flous convexes et normalisés (généralement des intervalles flous). On peut ainsi prendre en compte cette imprécision. Une représentation visuelle de ces ensembles est proposée dans la Figure 5.

## 4.2 Rangs des objets selon leurs périodes d'activités pour Kerre et Jain

En utilisant Kerre pour classer les représentations des périodes d'activité des objets de *BDFRues* on obtient la figure 6. Dans cette figure, relativement à l'ensemble des nombres flous associés aux périodes d'activité des objets de *BDFRues*, plus le rang d'un objet est grand plus le nombre flou associé à la période d'activité de l'objet est grand selon Kerre (1982).

En utilisant l'indice de Jain avec $k = 1$ pour classer les nombres flous associés aux périodes d'activité des objets de *BDFRues* on obtient la figure 7. Dans cette figure plus le rang d'un objet est grand plus le nombre flou associé à sa période d'activité est grand selon le classement des indices de Jain (1977).

On peut remarquer que certains rangs attribués aux objets selon l'indice de Jain diffèrent de ceux attribués selon l'indice de Kerre. L'interprétation de ces rangs est difficile, car il n'y a pas de sémantique associée à ces indices. A contrario, les positions temporelles obtenues à l'aide de la construction du graphe d'antériorité ont une forte interprétabilité qui sera illustrée dans la suite.

FIG. 6 – *Rangs, déterminés à l'aide de Kerre, des 33 objets de BDFRues selon leurs périodes d'activité*

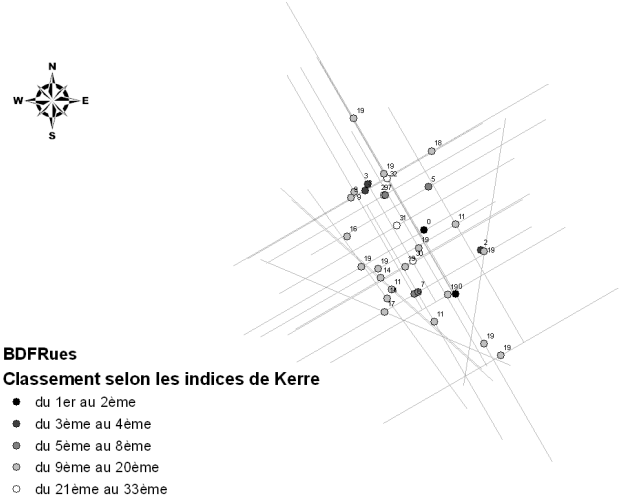
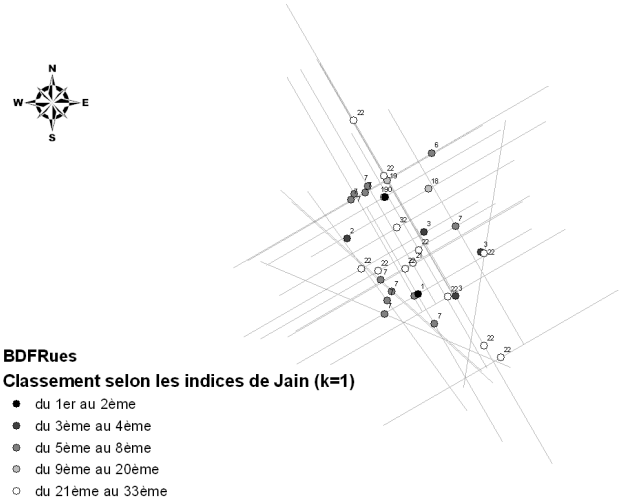



FIG. 7 – *Rangs, déterminés à l'aide de Jain ($k = 1$), des 33 objets de BDFRues selon leurs périodes d'activité*




## 4.3 Positionnement temporel des objets selon le graphe d'antériorité

En construisant le graphe d'antériorité sur les données de *BDFRues*, on obtient la figure 8. Dans cette figure, plus le rang d'un objet est grand, plus la position temporelle est élevée, c'est-à-dire plus sa capacité de postériorité à tous les autres est grande.

On peut remarquer que les positions temporelles obtenues à l'aide du graphe d'antériorité, et les rangs obtenus à l'aide des approches de Kerre ou de Jain présentent dans certains cas des divergences. Cependant, les positions temporelles ont une interprétabilité plus forte que les rangs issus de Kerre ou de Jain, ce qui permet de définir une classification temporelle des objets selon leur positionnement temporel.

La position temporelle d'un objet archéologique est obtenue à l'aide de l'indice temporel de l'objet lui-même issu des capacités d'antériorité et de postériorité. Ces capacités de l'objet sont calculées à l'aide de l'indice d'antériorité de l'objet à chacun des autres objets archéologiques. Cet indice quantifie l'antériorité d'un objet à un autre. Ainsi, la position temporelle d'un objet reflète la relation temporelle de l'objet à l'ensemble des autres.

Par exemple, de ces positions temporelles, on peut extraire les trois objets particuliers suivants : le plus antérieur, le plus postérieur et le médian. Ces objets particuliers sont identifiables dans la figure 9.

[4]Durocortorum : Reims à l'époque romaine

FIG. 8 – *Positions temporelles des 33 objets de BDFRues selon leurs périodes d'activité*

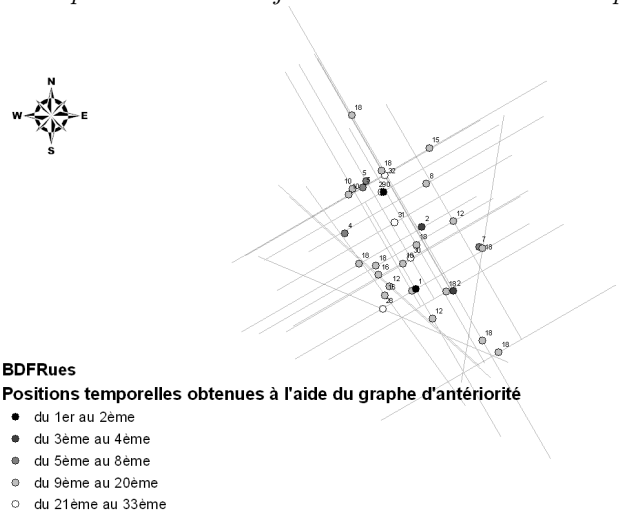
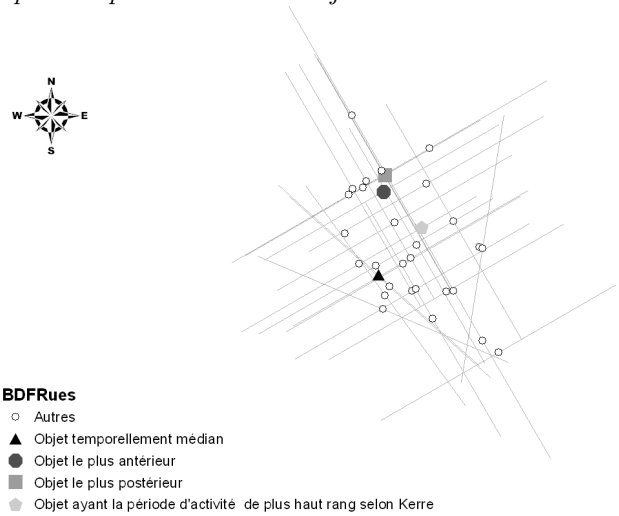



FIG. 9 – *Positions temporelles particulières des objets de BDFRues selon leurs périodes d'activité*




La figure 10 présente les objets de *BDFRues* selon les trois classes : « plutôt postérieurs », « plutôt antérieurs » et « anté-postérieurs ».

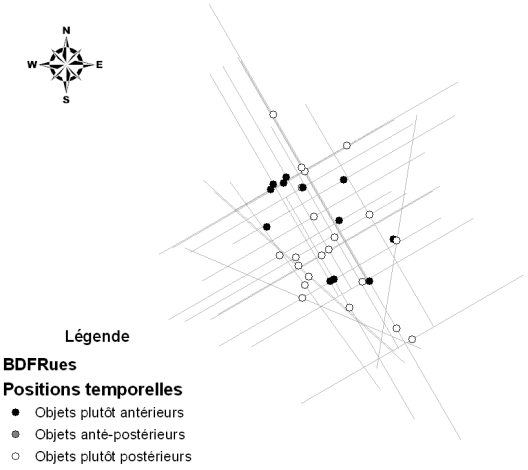
Pour les données de *BDFRues*, la classe des objets « plutôt postérieurs » regroupe deux fois plus d'objets que celle des objets « plutôt antérieurs ». La classe des objets « anté-postérieurs » est vide.

On peut supposer que les objets « plutôt antérieurs » ont des valeurs de l'indice d'antériorité avec les autres objets proches de 1. Cela expliquerait le fort différentiel entre le cardinal de la classe des objets « plutôt postérieurs » avec celui de la classe des objets « plutôt antérieurs ». Cette hypothèse est d'ailleurs vérifiée puisque, sachant qu'il y a trente trois objets :

- la valeur de l'indice temporel de l'élément le plus antérieur est de $-29.8$ ;
- la plus grande valeur d'indice temporel des éléments « plutôt antérieurs » est $-5.4$ ;
- plus de 72% des objets « plutôt postérieurs » ont une valeur d'indice temporel supérieure à 6.

La bipolarisation des valeurs est donc pertinente. Cependant, on peut remarquer que trois objets ont une valeur de l'indice temporel supérieur à 0 mais proche de 0 . Une définition plus vague de la notion *anté-postérieure*

FIG. 10 – *Objets de BDFRues classés en « plutôt postérieurs », « plutôt antérieurs » et « anté-postérieurs »*




aurait sûrement permis de détecter ces éléments. Mais cette nouvelle définition demande un paramétrage fortement dépendant des données. Ce paramétrage n'entre pas dans la démarche exploratoire proposée dans cette partie et donc dans cet article.

De plus, par le recoupement de la figure 10 avec la figure 9, on remarque que l'objet temporellement médian appartient aux objets « plutôt postérieurs » à l'instar de l'objet le plus postérieur. L'objet le plus antérieur appartient à la classe des « plutôt antérieurs ».

Ainsi, les valeurs de l'indice temporel des objets de *BDFRues* obtenues à partir du graphe d'antériorité permettent de classer les objets archéologiques selon l'antériorité aux objets de la base. Ces positions permettent de définir des classes d'objets et d'extraire des objets particuliers. Ce graphe permet donc, par une organisation schématique structurée, de dégager de nouvelles connaissances temporelles sur les informations contenues dans *BDFRues*. Ces nouvelles connaissances sont riches d'information pour l'expertise archéologique.

# 5 Conclusion

Dans cet article, nous avons proposé un processus exploratoire ayant pour but l'analyse des relations entre objets selon la temporalité de l'information archéologique disponible. Pour cela, on construit un graphe orienté pondéré à partir des objets de la base et des valeurs de l'indice d'antériorité les reliant deux à deux. De ce graphe, nous dégageons l'indice temporel de chaque objet afin d'en déterminer sa position temporelle dans l'ensemble et l'analyse associée (« plutôt antérieur »,« plutôt postérieur », « anté-postérieur »). Les positions, relativement au temps, des objets archéologiques dans une base de données facilitent la compréhension des relations liant les objets de la base dans le SIG.

Dans leurs expertises, les archéologues évaluent les objets qu'ils sont susceptibles de trouver dans un site tant d'un point de vue fonctionnel que temporel. Ils ont notamment besoin d'étudier les relations temporelles entre les objets stockés dans les bases de données afin de (i) regarder si la logique temporelle est respectée, et (ii) dégager une évolution temporelle de la cité. Dans cet objectif, ils peuvent utiliser le graphe d'antériorité.

En perspective, nous avons l'intention d'étudier différents modes de visualisation 3D des informations portées à la fois par le graphe d'antériorité (positions temporelles des objets) et par les localisations 2D des objets afin de pouvoir les croiser pour l'analyse spatiotemporelle des données. De plus, nous souhaitons détecter grâce à une réduction du graphe les possibles incohérences temporelles dues au changement d'échelle : les périodes d'activité sont estimées à l'échelle du site, le processus exploratoire se fait à l'échelle de la ville.

# Remerciements

Nous tenons à remercier le Service Régional d'Archéologie de Champagne-Ardenne et le centre rémois de l'Institut National de Recherche en Archéologie Préventice pour nous avoir permis d'accéder à leurs données.

Nous tenons de même à souligner la contribution de Dominique Pargny, ingénieur d'études au laboratoire GEGENA, de Frédéric Piantoni, Maître de Conférences au laboratoire HABITER, et de Michel Herbin, Professeur au CReSTIC, au projet SIGRem, porté par l'Université de Reims Champagne-Ardenne, dans le contexte duquel c'est fait ce travail.

## Références

Chen, S.-H. (1985). Ranking fuzzy numbers with maximizing set and minimizing set. *Fuzzy Sets and Systems 17*, 113–129.

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

de Runz, C., E. Desjardin, M. Herbin, et F. Piantoni (2006). A new Method for the Comparison of two fuzzy numbers extending Fuzzy Max Order. In *Information Processing and Managment of Uncertainty in Knowledge-Based Systems - IPMU'06*, Paris, France, pp. 127–133. Editions EDK.

de Runz, C., E. Desjardin, F. Piantoni, et M. Herbin (2007). Using fuzzy logic to manage uncertain multi-modal data in an archaeological GIS. In *International Symposium on Spatial Data Quality - ISSDQ'07*, Enschede, Pays-Bas.

de Runz, C., E. Desjardin, F. Piantoni, et M. Herbin (2008). Anteriority index for managing fuzzy dates in archaeological GIS. *Soft Computing*. Soumis le 30-10-2007, en révision majeure depuis le 09-06-2008.

Dragicevic, S. et D. J. Marceau (2000). An application of fuzzy logic reasoning for GIS temporal modeling of dynamic processes. *Fuzzy Sets and Systems 113*, 69–80.

Jain, R. (1977). A procedure for multiple-aspect decision making using fuzzy set. *Internat. J. Systems Sci. 8*, 1–7.

Kerre, E. E. (1982). The use of fuzzy set theory in electrocardiological diagnostics. In M. Gupta et E. Sanchez (Eds.), *Approximate Reasoning in Decision-Analysis*, pp. 277–282. North-Holland Publishing Company.

Pargny, D. et F. Piantoni (2005). SIGRem : un Système d'Information Géographique pour l'archéologie en Champagne-Ardenne. In *colloque Archéologie en Champagne-Ardenne*. INRAP.

Piantoni, F. (2005). Le SIGRem. Problématique et méthodologie. In *séminaire de recherche du CIRAR*.

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

Wang, X. et E. E. Kerre (2001b). Reasonable properties for the ordering of fuzzy quantities (II). *Fuzzy Sets anf Systems 118*, 387–405.

Wang, X., E. E. Kerre, B. Cappelle, et D. Ruan (1995). Transitivity of Fuzzy Orderings Based on Pairwise Comparis. *The Journal of Fuzzy Mathematics 3*(2), 455–463.

Zadeh, L. A. (1965). Fuzzy Sets. *Information Control 8*, 338–353.

## Summary

In this paper, we propose a new temporal data mining method considering a set of archæological objects which are temporally represented with fuzzy numbers. Our method uses an index which quantifies the anteriority between two fuzzy numbers for the construction of a weighted oriented graph. The vertices of the graph correspond to the archæological objects. Using this anteriority graph, we determine the potential of anteriority, of posteriority and the temporal position of each object. The information obtained designs the temporal relations between objects and allows us to extract the most anterior object and the most posterior object. We apply this approach to the roman streets discovered in Reims.

# Regroupement de données multi-représentées: une approche par k-moyennes floues

Jacques-Henri Sublemontier, Guillaume Cleuziou,
Matthieu Exbrayat, Lionel Martin

Laboratoire d'Informatique Fondamentale d'Orléans
Université d'Orléans
B.P. 6759 - 45067 ORLEANS Cedex 2
{prenom.nom}@univ-orleans.fr

**Résumé.** Nous nous intéressons dans cette étude au regroupement de données multi-représentées, i.e. des données décrites par plusieurs sources d'information (ensembles d'attributs ou matrices de proximités). Ce domaine d'étude trouve ses principales applications en recherche d'information, en biologie ou encore en chimie. Il s'agit alors de proposer un cadre méthodologique permettant la recherche d'une classification réalisant un consensus entre les différentes représentations. Dans ce cadre, la fusion des informations issues de chacune des sources est nécessaire. Cette fusion peut être réalisée en amont du processus de classification (fusion a priori), en aval (fusion a posteriori) ou pendant le processus (approche collaborative). Nous nous inspirons du travail récent de Bickel et Sheffer visant à étendre les modèles de mélanges au cas des données multi-représentées (Co-EM) et proposons un modèle déterministe de classification floue Co-FKM généralisant à la fois les approches collaboratives, de fusion a priori et a posteriori. Les expérimentations proposées valident l'étude sur un jeu de données adapté.

## 1 Introduction

Nous nous intéressons dans cette étude à la classification non-supervisée de données complexes. Dans un processus d'extraction des connaissances, la complexité des données peut être liée par exemple au volume de données à traiter, à leur nature (numérique, symboliques, mixte), à des aspects temporels ou multi-sources. C'est précisément cette dernière particularité que nous traitons dans cette étude en considérant des données que nous qualifierons de "multi-représentées" pour spécifier que plusieurs représentations (ensembles d'attributs ou matrices de proximité) sont disponibles pour un même ensemble d'individus. Ce phénomène est commun à de nombreuses applications du monde réel. En Recherche d'information un document multimédia pourra être décrit par ses contenus textuels, hyper-textuels, images, audio et vidéo le cas échéant ; il peut être souhaitable de tirer parti de l'ensemble de ces sources d'information pour en extraire une connaissance utile et précise. En bioinformatique l'analyse d'un gène donne lieu à plusieurs niveaux de description tels que l'expression de ce gène, sa localisation, son profil phylogénétique, son comportement biochimique ; encore une fois le croisement de ces différentes caractéristiques permettrait certainement de mieux traiter les données

génétiques. De nombreuses autres situations se prêtent naturellement à l'étude de données multi-représentées. Citons pour compléter les données biologiques (plusieurs niveaux de description pour les molécules), les données médicales (radiographies, comptes-rendus médicaux, analyses biologiques d'un patient), les données textuelles (descriptions lexicales, morphosyntaxiques et sémantiques de la langue), etc.

Dans le cadre de la classification non-supervisée il s'agit d'organiser l'ensemble des individus en classes d'individus similaires (Jain et al. (1999),Tan et al. (2005)) de manière à réaliser un consensus sur un ensemble des représentations. Deux stratégies naturelles pour procéder au regroupement sur données multi-représentées consisteraient à fusionner des informations en amont (fusion a priori) ou en aval (fusion a posteriori) d'un processus traditionnel de clustering.

La fusion a priori reviendrait à concaténer l'ensemble des descripteurs ou à combiner les matrices de proximités de chaque représentation (e.g. Heer et Chi (2002)) : cette manière de procéder conduit à mélanger les niveaux de description et risque de gommer les caractéristiques propres à chaque niveau de représentation, d'autre part l'écueil de la dimensionalité s'en trouve renforcée ; nous montrerons cependant qu'en pratique cette méthode peut conduire à de bons résultats.

La fusion a posteriori reviendrait cette fois à réaliser plusieurs processus de regroupement sur chaque représentation indépendamment puis à définir une manière de combiner ces organisations en un seul schéma final de classification. Dans cette approche, l'indépendance de chaque processus conduira à des organisations parfois éloignées qui seront d'autant plus difficile à concilier dans l'étape ultime ; cette approche est peu utilisée et nous observerons en pratique les faibles performances qu'elle engendre.

Nous choisissons d'explorer une troisième voie consistant à réaliser la fusion des informations au cours du processus de classification : approche collaborative. Pour cela nous nous inspirons du travail récent de Bickel et Scheffer (2005) visant à étendre les modèles de mélanges au cas des données multi-représentées (Co-EM). Nous proposons un modèle de classification floue qui généralise à la fois les approches collaboratives, de fusion a priori et a posteriori. Le choix d'un modèle flou permet d'envisager une solution algorithmique convergente (contrairement au Co-EM actuel) et plus simple autant en terme de paramétrage que d'extension possible au clustering à noyau.

L'article se décompose en quatre parties principales : nous présentons en section 2 quelques méthodes collaboratives pour le traitement de données multi-représentées avant d'introduire en section 3 le modèle CoFKM sur lequel s'appuie cette étude ; la section 4 présente une expérimentation préliminaire du modèle illustrant le comportement de l'algorithme, sa performance et son caractère généralisant sur un jeu de données adapté. Enfin la section 5 présente en conclusion quelques pistes d'amélioration du modèles CoFKM.

## 2 État de l'art sur la classification de données multi-représentées

Nous présentons dans cette section un aperçu des différents modèles de classification proposés pour traiter les données multi-représentées. Nous nous focalisons sur les modèles guidés par un critère objectif prenant en compte les différentes représentations.

Dans le cas particulier de données multimédia, Bekkerman et Jeon (2007) utilisent le modèle récent des champs de markov aléatoires combinatoires (Comraf : *Combinatorial Markov*

*random field* décrit plus en détails dans Bekkerman et al. (2006)) pour rechercher la partition des individus qui explique le mieux (*Most Probable Explanation*) les informations observées dans chaque représentation. Sous l'hypothèse d'indépendance des représentations, le critère objectif proposé estime l'"explication" d'une partition $\mathcal{P}$ en faisant intervenir pour chaque représentation $r$, l'information mutuelle entre l'ensemble des groupes de $\mathcal{P}$ et l'ensemble $V_r$ des descripteurs de la représentation :

$$Q_{comraf}(\mathcal{P}) = \sum_{r=1}^{|R|} w_r I(\mathcal{P}, V_r) \tag{1}$$

Dans le critère (1), la pondération $w_r$ permet de donner plus ou moins d'influence à une représentation à partir d'informations a priori. D'autre part, on notera que le calcul de l'information mutuelle $I(\mathcal{P}, V_r)$ nécessite de disposer d'une distribution de probabilité sur $V_r$ ; si pour certains types de données, par exemple multimédia, une telle distribution est naturelle (distribution d'un mot dans un texte, proportion de pixels d'une certaine couleur dans une image, etc.) cette contrainte n'est malheureusement pas satisfiable dans la plupart des applications. Enfin, la mise en oeuvre algorithmique proposée par Bekkerman et Jeon (2007) repose sur un parcours stochastique de l'espace des partitions, guidé par le critère objectif (1).

Wiswedel et Berthold (2007) ont récemment proposé un modèle flou s'appliquant cette fois à tout type de données numériques sur lesquelles on dispose d'un vecteur de descriptions. Leur approche opère simultanément sur l'ensemble des représentations (ou "univers parallèles"), et permet de moduler la contribution d'un individu aux groupes dans chaque représentation, par une pondération des individus dans ces différentes représentations. L'intuition sous-jacente est que tous les individus n'ont pas la même contribution au regroupement dans toutes les représentations. Le critère d'inertie qu'ils proposent de minimiser prend la forme suivante :

$$Q_{paralleluniverse} = \sum_{r=1}^{|R|} \sum_{x_i \in \mathcal{X}} v_{i,r}^{\alpha} \sum_{k=1}^{K} u_{i,k,r}^{\beta} d_r(x_{i,r}, c_{k,r})^2 \tag{2}$$

$v_{i,r}$ modélise la contribution de l'individu $x_i$ dans la représentation $r$, $u_{i,k,r}$ le degré d'appartenance de $x_i$ au groupe $k$ dans la représentation $r$, $\alpha$ et $\beta$ des paramètres de flou, $c_{k,r}$ correspondant au prototype du groupe $k$ dans la représentation $r$, et $d_r$ la distance utilisée dans la représentation $r$. Ainsi modélisé, l'objectif visé par Wiswedel et Berthold (2007) consiste davantage à extraire des informations fortes dans chaque représentation (*local patterns*) plutôt qu'à aboutir à une classification consensus sur toutes les représentations.

La classification guidée par la recherche d'un consensus est illustrée par les approches de Pedrycz (2002) puis Bickel et Scheffer (2005) qui, dans des cadres applicatifs différents, proposent une démarche analogue consistant à pénaliser le critère objectif par un terme de désaccord entre les représentations.

Pedrycz (2002) utilise la méthode des $k$-moyennes floues (critère d'inertie) pour classer des données d'une source (ici représentation) en utilisant des informations d'autres sources sur ces mêmes données, sans accéder aux valeurs des descripteurs dans ces dernières sources (pour des raisons de confidentialité). Le critère objectif pénalisé s'écrit :

$$Q_{collab.}(r) = \sum_{k=1}^{K} \sum_{x_i \in \mathcal{X}} u_{i,k,r}^2 d(x_{i,r}, c_{k,r})^2 + \sum_{r'=1}^{|R|} \alpha_{r,r'} \sum_{k=1}^{K} \sum_{x_i \in \mathcal{X}} (u_{i,k,r} - u_{i,k,r'})^2 d(x_{i,r}, c_{k,r})^2 \quad (3)$$

On observera que ce critère (3) s'applique à une représentation $r$ sur laquelle on cherche à minimiser un premier terme d'inertie floue traditionnel, pénalisé par le second terme qui mesure le désaccord avec les autres représentations $r'$ en recourant seulement aux degrés d'appartenances $u_{i,k,r'}$ plutôt qu'aux données $x_{i,r'}$ directement, ceci afin de respecter la confidentialité des sources. Dans (3), le paramètre $\alpha_{r,r'}$ modélise une information de degré de collaboration souhaitable entre représentations (information donnée a priori).

Enfin, Bickel et Scheffer (2005) ont choisi le formalisme des modèles de mélanges pour proposer une méthode (Co-EM) de type collaboratif recherchant une classification consensus. L'algorithme associé à Co-EM vise à maximiser le critère objectif suivant :

$$Q_{Co-EM} = \sum_{r=1}^{|R|} Q_{EM}(r) - \eta\Delta \quad (4)$$

Dans (4), $Q_{EM}(r)$ désigne le critère de log-vraisemblance traditionnel, dans la représentation $r$ ; $\eta$ est un paramètre permettant de donner plus ou moins d'importance au terme de désaccord $\Delta$ ressemblant à une divergence de Kullback-Leibler entre les distributions de probabilités a posteriori, sur toutes les paires de représentations :

$$\Delta = \frac{1}{|R|-1} \sum_{r \neq r'} \sum_{x_i \in \mathcal{X}} \sum_{k=1}^{K} P(k|x_{i,r}, \Theta_r^t) \log \frac{P(k|x_{i,r}, \Theta_r)}{P(k|x_{i,r'}, \Theta_{r'})} \quad (5)$$

Bickel et Scheffer (2005) montrent que le critère (4) présente la bonne propriété de pouvoir se réécrire comme une somme de critères objectifs locaux pouvant être maximisés dans chaque représentation de manière indépendante. En revanche, leur critère ne peut pas être globalement optimisé. En conséquence, la convergence de l'algorithme associé ne peut être assurée. Bickel et Scheffer (2005) proposent une annulation (progressive) du paramètre $\eta$, autrement dit, une suppression de l'aspect collaboratif, pour garantir cette convergence.

# 3 Le modèle CoFKM : Co-$k$-moyennes floues

L'approche que nous proposons ici est une extension de la méthode des k-moyennes floues (FKM Bezdek (1981)) qui vise à obtenir, dans chaque représentation, une organisation spécifique, et qui comme dans Co-EM, favorise des organisations "proches" sur les différentes représentations, par l'introduction d'un terme de désaccord.

Rappelons que l'objectif de FKM (Bezdek (1981)) est de minimiser un critère d'inertie pondérée (par des degrés d'appartenance) :

$$Q_{FKM} = \sum_{k=1}^{K} \sum_{x_i \in X} u_{i,k}^{\beta} ||x_i - c_k||^2$$

avec $\forall x_i \in X, \sum_{k=1}^{K} u_{i,k} = 1$, où les variables du problème sont les centres des groupes ($c_k$) et les degrés d'appartenance ($u_{i,k}$ : degré d'appartenance de l'individu $i$ au groupe $k$, $\beta$ un paramètre de flou).

## 3.1 Le critère à optimiser

Étant donné un ensemble $R$ de représentations, nous pouvons noter $Q_{FKM}(r)$ le critère ci-dessus dans la représentation $r \in R$. Dans chaque représentation, les individus sont décrits par un vecteur appartenant à $\mathbb{R}^{N_r}$, $N_r$ étant la dimensionalité de la représentation $r$.

Nous proposons une approche collaborative, basée sur FKM, qui consiste à minimiser les valeurs des $Q_{FKM}(r)$ de chaque représentation, tout en pénalisant les divergences d'organisation entre chaque couple de représentations. Dans ce cadre, le critère à minimiser peut donc s'écrire :

$$\begin{aligned} Q_{CoFKM} &= \left(\sum_{r\in R} Q_{FKM}(r)\right) + \eta\Delta \\ &= \sum_{r\in R}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r}^{\beta}||x_{i,r} - c_{k,r}||^2 + \eta\Delta \end{aligned}$$

à condition de réaliser une normalisation sur les représentations, afin que les inerties dans chaque représentation soient du même ordre de grandeur. Cette normalisation est réalisée dans chaque représentation $r$ :

– en réduisant chaque variable (variance égale à 1),

– en affectant à chaque variable (après réduction) un poids égal à $N_r^{-1/2}$.

Dans l'expression de $Q_{CoFKM}$, $\Delta$ est un terme de désaccord : si toutes les représentations produisent la même organisation, ce terme doit être nul. Les centres $c_{k,r}$ n'étant pas comparables d'une représentation à l'autre, nous proposons :

$$\Delta = \frac{1}{|R|-1}\sum_{r\neq r'}\sum_{x_i\in X}\sum_{k=1}^{K}(u_{i,k,r'}^{\beta} - u_{i,k,r}^{\beta})||x_{i,r} - c_{k,r}||^2$$

Il s'agit donc de sommer, pour chaque couple de représentations $(r, r')$, un écart entre les organisations obtenues dans les représentations $r$ et $r'$. En effet, en notant $d_{i,k,r} = ||x_{i,r} - c_{k,r}||$, on peut réécrire cette expression en sommant sur les couples $(r, r')$ tels que par exemple $r > r'$. On obtient ainsi :

$$\Delta = \frac{1}{|R|-1}\sum_{r>r'}\sum_{x_i\in X}\sum_{k=1}^{K}(u_{i,k,r'}^{\beta} - u_{i,k,r}^{\beta})(d_{i,k,r}^2 - d_{i,k,r'}^2)$$

Dans la mesure où $u_{i,k,r}$ est d'autant plus grand que $d_{i,k,r}$ est petit, l'expression obtenue pour un couple de représentation $(r, r')$ peut donc être représentation comme une mesure de divergence entre les organisations obtenues dans les deux représentations $r$ et $r'$.

Finalement, l'expression de $Q_{CoFKM}$ peut être réécrite sous une forme plus simple :

$$Q_{CoFKM} = \sum_{r \in R} \sum_{x_i \in X} \sum_{k=1}^{K} u_{i,k,r,\eta} ||x_{i,r} - c_{k,r}||^2 \quad (6)$$

avec $u_{i,k,r,\eta} = (1-\eta) u_{i,k,r}^{\beta} + \frac{\eta}{|R|-1} (\sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta})$ (les détails de cette réécriture sont donnés en annexe), $\eta$ est un paramètre qui module l'importance du désaccord. Ce critère s'écrit donc comme une inertie pondérée, où dans chaque représentation, le poids $u_{i,k,r,\eta}$ est une moyenne pondérée entre les poids usuels ($u_{i,k,r'}^{\beta}$) de chaque représentation.

## 3.2 Recherche d'une solution optimale

Comme dans le cas de FKM, nous souhaitons obtenir une solution qui minimise le critère global (6), avec pour chaque représentation $r$ et chaque individu $x_i \in X$ : $\Sigma_k u_{i,k,r} = 1$. Pour résoudre ce problème d'optimisation sous contraintes, nous pouvons considérer le lagrangien du problème :

$$L(C, U, \lambda) = Q_{CoFKM} + \sum_{r \in R} \sum_{x_i \in X} \lambda_{r,i} (1 - \sum_{k=1}^{K} u_{i,k,r})$$

où $C$ représente la matrice des centres dans chaque représentation, $U$ la matrice des degrés d'appartenance de chaque objet à chaque centre, dans chaque représentation, et $\lambda$ est le vecteur des coefficients de Lagrange. Si $(C^*, U^*)$ est une solution (localement) optimale, une condition nécessaire est que le gradient du lagrangien soit nul, i.e. les dérivées partielles par rapport aux variables $u_{i,k,r}$ et $c_{k,r}$ sont nulles. Ces dérivées partielles s'écrivent :

$$\begin{aligned} \frac{\partial L}{\partial u_{i,k,r}} &= (1-\eta)\beta u_{i,k,r}^{\beta-1} ||x_{i,r} - c_{k,r}||^2 + \frac{\eta}{|R|-1} (\sum_{\overline{r}} \beta u_{i,k,r}^{\beta-1} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2) - \lambda_{r,i} \\ \frac{\partial L}{\partial c_{k,r}} &= -2 \sum_{x_i \in X} u_{i,k,r,\eta} (x_{i,r} - c_{k,r}) \end{aligned}$$

Comme pour FKM, nous proposons un algorithme qui itère deux étapes d'optimisation après une initialisation aléatoire de $k$ centres parmi l'ensemble des individus (les mêmes dans chaque représentation) :

- mettre à jour les centres $c_{k,r}$ en considérant les valeurs des degrés $u_{i,k,r}$ constantes
- mettre à jour les degrés $u_{i,k,r}$ en considérant les valeurs des centres $c_{k,r}$ constantes

Les équations $\frac{\partial L}{\partial c_{k,r}} = 0$ et $\frac{\partial L}{\partial u_{i,k,r}} = 0$ entraînent respectivement :

$$
\begin{aligned}
c_{k,r} &= \frac{\sum_{x_i \in X} u_{i,k,r,\eta} x_{i,r}}{\sum_{x_i \in X} u_{i,k,r,\eta}} \\
u_{i,k,r} &= (\frac{\lambda_{r,i}}{\beta})^{1/(\beta-1)} ((1-\eta)||x_{i,r} - c_{k,r}||^2 + \frac{\eta}{|R|-1} \sum_{\overline{r}} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2)^{1/(1-\beta)}
\end{aligned}
$$

En sommant sur $k$ l'expression des $u_{i,k,r}$ et avec la contrainte $\Sigma_{k=1}^{K} u_{i,k,r} = 1$, on obtient finalement pour $u_{i,k,r}$ :

$$
u_{i,k,r} = \frac{((1-\eta)||x_{i,r} - c_{k,r}||^2 + \frac{\eta}{|R|-1} \sum_{\overline{r}} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2)^{1/(1-\beta)}}{\Sigma_{k=1}^{K} ((1-\eta)||x_{i,r} - c_{k,r}||^2 + \frac{\eta}{|R|-1} \sum_{\overline{r}} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2)^{1/(1-\beta)}}
$$

Ainsi, à chaque étape de calcul, on détermine la valeur optimale de $c_{k,r}$ (resp. $u_{i,k,r}$) pour des valeurs de $u_{i,k,r}$ (resp. $c_{k,r}$) fixées. Pour cet algorithme, le critère (6) est donc décroissant à chaque étape de calcul, ce qui assure sa convergence (vers un optimum local).

### 3.3 Règle d'affectation

Nous avons proposé un moyen optimal de mettre à jour systématiquement les centres de groupes, ainsi que les degrés d'appartenance des individus aux groupes, à la manière de FKM. Néanmoins, il faut garder à l'esprit que ce que nous obtenons au final est une matrice d'appartenance (individus $\times$ groupes) dans chaque vue. Afin d'obtenir un unique regroupement, il est nécessaire d'établir une règle d'affectation prenant en compte les différents degrés d'appartenance obtenus. Nous avons choisi une règle simple nous semblant la plus naturelle qui consiste à calculer un degré global $\hat{u}_{i,k}$ comme étant le produit sur les représentations, des degrés locaux, et à affecter les individus aux groupes pour lesquels ils ont le plus fort degré d'appartenance global :

$$
\hat{u}_{i,k} = \prod_{r=1}^{|R|} u_{i,k,r}
$$

Un problème se pose néanmoins dans ce contexte, et il s'adresse à la façon de faire le lien entre les groupes dans les différentes vues. Nous considérons ici qu'un même groupe est identifié par une même étiquette $k \in [1..K]$ dans les différentes représentations. Pour palier au risque de considérer deux groupes $k_1$ et $k_2$ différents alors qu'ils contiendraient les mêmes objets, nous considérons une initialisation bien particulière. Nous choisissons comme centres initiaux, ceux qui correspondent aux groupes d'une partition établie commune pour toutes les représentations.

## 3.4 Comparaisons

Nous montrons dans un premier temps que notre modèle CoFKM généralise le modèle FKM appliqué à la concaténation des représentations (fusion a priori), puis dans un second temps, qu'il généralise également un modèle simple de fusion a posteriori, où le modèle FKM est appliqué indépendamment sur chaque représentation. Pour finir, nous argumenterons sur les améliorations théoriques apportées par notre modèle, par rapport à Co-EM.

Considérons les expressions des $c_{k,r}$ et $u_{i,k,r}$ de la section précédente, dans lesquelles les termes correspondants aux différentes représentations ont les mêmes coefficients, i.e. $(1-\eta) = \frac{\eta}{|R|-1}$ soit $\eta = \frac{|R|-1}{|R|}$.

Dans ce cas,

$$u_{i,k,r,\eta} = \frac{1}{|R|} u^{\beta}_{i,k,r} + \frac{1}{|R|} \sum_{\overline{r}} u^{\beta}_{i,k,\overline{r}} = \frac{1}{|R|} \sum_{r'} u^{\beta}_{i,k,r'}$$

par conséquent $u_{i,k,r,\eta}$ ne dépend pas de $r$, et l'expression de $c_{k,r}$ correspond exactement à celle obtenue par FKM à partir de la concaténation des représentations.

De la même manière, $||x_{i,r} - c_{k,r}||^2 + \sum_{\overline{r}} ||x_{i,\overline{r}} - c_{k,\overline{r}}||^2$ est égal à la distance entre $x_i$ et $c_k$ pour la concaténation des représentations et de nouveau, l'expression de $u_{i,k,r}$ correspond exactement à celle obtenue par FKM à partir de la concaténation des représentations. Finalement, CoFKM peut donc être vu comme une généralisation de FKM appliqué à la concaténation des représentations, dans laquelle il est possible d'accentuer la nécessité de consensus en fixant $\eta$ à une valeur différente de $\frac{(|R|-1)}{|R|}$. Nous avons observé empiriquement que la positivité du désaccord dépend de la valeur attribuée à $\eta$. Si $\eta > \frac{(|R|-1)}{|R|}$, alors le désaccord exprimé devient négatif. Nous suggérons donc, pour rester cohérent en théorie[1], de choisir $0 \leq \eta \leq \frac{(|R|-1)}{|R|}$.

Considérons maintenant le modèle CoFKM, pour lequel on fixe $\eta = 0$. Le critère à optimiser se réécrit alors simplement comme une somme sur chaque représentation, du critère FKM local à cette représentation :

$$\begin{aligned} Q_{CoFKM_{\eta=0}} &= \left( \sum_{r \in R} Q_{FKM(r)} \right) \\ &= \sum_{r \in R} \sum_{x_i \in X} \sum_{k=1}^{K} u^{\beta}_{i,k,r} ||x_{i,r} - c_{k,r}||^2 \end{aligned}$$

qui correspond bien à optimiser les inerties locales, indépendamment dans chaque représentation, à l'aide du modèle FKM. La fusion a posteriori est réalisée par la règle d'affectation de notre modèle CoFKM (produit des degrés d'appartenances). On constate alors que notre approche collaborative généralise également une approche simple de fusion a posteriori, en choisissant $\eta = 0$.

[1] Un désaccord négatif ayant peu de sens.

Si on compare maintenant notre approche à l'approche Co-EM, on constate que l'on s'abstrait du principal défaut théorique éprouvé par Co-EM. Co-EM ne converge pas autrement que par un artifice, l'annulation de $\eta$. Cela est principalement dû au fait que Bickel et Scheffer (2005) n'optimisent pas globalement leur critère, ils réalisent une succession d'optimisations locales qui influencent le résultat global sans l'optimiser, et sans nécessairement l'améliorer. l'algorithme CoFKM quant à lui converge quelle que soit la valeur de $\eta$, puisque l'on est capable de mettre à jour les paramètres du critère de manière à optimiser à chaque étape ce dernier.

# 4 Expérimentations

Nous validons notre approche sur un jeu de données adapté : *multiple features*[2]. Nous observerons les apports empiriques de notre approche au regard de Co-EM, ainsi que les modèles FKM et EM appliqués à la concaténation des représentations, ainsi qu'aux représentations prises une à une.

## 4.1 Présentation des données

Le jeu de données "multiple features" correspond à un ensemble de $2000$ caractères manuscrits (numérisées en images binaires) décrits par $6$ représentations différentes. $10$ classes sont à retrouver (les caractères sont compris entre $0$ et $9$), et il y a $200$ individus par classe. Les différentes représentations sont les suivantes :

- mfeat-fou : 76 coefficients de Fourier des formes des caractères,
- mfeat-fac : 216 corrélations de profils
- mfeat-kar : 64 coefficients de Karhunen-Love,
- mfeat-pix : 240 moyennes de pixel dans des fenêtres $2 \times 3$,
- mfeat-zer : 47 moments de Zernike,
- mfeat-mor : 6 descripteurs morphologiques.

Afin de répartir au mieux l'importance des diverses représentations, nous avons réalisé des pré-traitements statistiques classiques en analyse de données, que sont le centrage, la réduction et la pondération des variables (cf. section 3.1).

## 4.2 Résultats empiriques

L'évaluation des approches de regroupement reste sujet à débats. Néanmoins, pour valider les approches, il est nécessaire de fixer une mesure pertinente afin de comparer les divers algorithmes proposés dans la littérature. Nous avons choisi ici, parce que les données nous le permettaient, d'utiliser l'information externe des étiquettes de classes (les chiffres) pour calculer le F-score, ou F-mesure permettant de comparer la classification obtenue avec la classification de référence. La formule de calcul du F-score est la suivante[3] :

$$\textit{F-score}(\beta) = \frac{(\beta^2 + 1) \times \textit{ Précision.Rappel}}{\beta^2 \times \textit{Précision} + \textit{Rappel}}$$

[2] disponible à l'adresse *http ://archive.ics.uci.edu/ml/*)

[3] $\beta$=1 dans les résultats présentés.

| | Précision (%) | Rappel (%) | FScore (%) |
|---|---|---|---|
| CoFKM | 91.95 | 92.07 | **92.01** |
| CoEM(v1) | 68.44 | 73.96 | 71.07 |
| CoEM(v2) | 27.04 | 64.71 | 37.87 |
| CoEM(v3) | 78.80 | 82.83 | 80.73 |
| FKMconcat | 90.20 | 91.75 | 90.93 |
| EMconcat(v1) | 55.57 | 64.03 | 59.44 |
| EMconcat(v2) | 20.11 | 70.21 | 31.03 |
| EMconcat(v3) | 71.90 | 85.47 | 77.78 |

| | FScore (%) | | | | | |
|---|---|---|---|---|---|---|
| | fac | fou | kar | mor | pix | zer |
| FKM | 66.88 | 33.65 | 23.03 | 55.69 | 71.52 | 42.36 |
| EM(v1) | 61.99 | 44,33 | 58.81 | 47.93 | 44.39 | 36.78 |
| EM(v2) | 21.18 | 18.13 | 19.25 | 38.42 | 21.75 | 18.55 |
| EM(v3) | 21.18 | 18.13 | 19.25 | 38.42 | 21.75 | 18.55 |

TAB. 1 – *Comparaisons des modèles collaboratifs et de fusion a priori en utilisant les* 6 *représentations et résultats sur chaque vue indépendamment.*

Les résultats que nous obtenons correspondent à une moyenne de 100 exécutions pour lesquelles les méthodes sont comparées avec les mêmes initialisations. Un récapitulatif est présenté dans les tableaux 1. On constate d'une part, que CoFKM surpasse les approches appliquées à une unique représentation (quelque soit cette représentation) ; d'autre part, que CoFKM surpasse également l'approche Co-EM pour un mélange de gaussiennes.

Nous avons constaté lors des tests, que l'utilisation de l'approche co-EM générale n'était pas du tout efficace. Nous avons alors testé à la place, différents modèles parcimonieux, tels que :

- (v1) des matrices de variances/covariances diagonales,
- (v2) des matrices de la forme $\sigma_k.I$, différentes pour chaque composante du mélange,
- (v3) des matrices de la forme $\sigma.I$, identiques pour chaque composante du mélange.

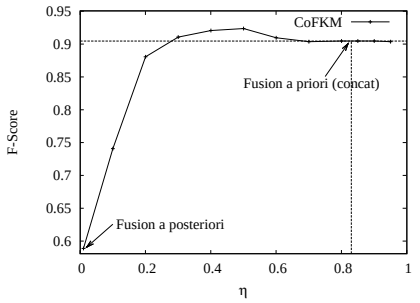



FIG. 1 – *Influence du paramètre $\eta$ sur le modèle CoFKM.*

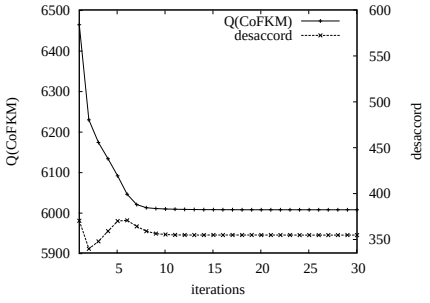



FIG. 2 – *Optimisation du critère objectif de CoFKM.*

La courbe (Fig.2) présente le comportement typique de CoFKM (au fil des itérations) sur une initialisation. On peut observer la minimisation systématique du critère $Q_{CoFKM}$, ainsi que l'évolution oscillante puis stabilisée du terme de désaccord.

Nous avons également exploré les divers résultats obtenus selon différentes valeurs de $\eta$ pour souligner l'intérêt que peut avoir le choix d'un "bon" $\eta$. La courbe (Fig.1) présente la performance du modèle selon différentes valeurs de $\eta$. On observe que l'on peut dépasser assez nettement les résultats obtenus par la simple concaténation des représentations en choisissant convenablement $\eta$. De manière empirique nous proposerons de choisir $\eta = \frac{|R|-1}{2.|R|}$ à égale distance d'un modèle de fusion a posteriori ($\eta = 0$) et d'un modèle de fusion a priori ($\eta = \frac{|R|-1}{|R|}$). Il s'agit du choix effectué pour les expérimentations précédentes.

## 5 Conclusion et perspectives

Nous avons centré cette étude sur la problématique de classification non-supervisée (ou clustering) sur des données complexes de type “multi-représentées”. Nous avons alors présenté les différentes alternatives proposées dans la littérature et choisi de poursuivre l’amorce proposée par Pedrycz (2002) puis Bickel et Scheffer (2005) pour les méthodes de clustering collaboratif. Le modèle CoFKM que nous avons défini présente de bonnes propriétés puisqu’il généralise différentes solutions de fusion, permet de lui associer une solution algorithmique efficace (convergente) et se compose de peu de paramètres (moins sensible au paramétrage). Les premiers résultats expérimentaux viennent confirmer les observations théoriques précédentes.

Les orientations futures de ce travail se situent tant au niveau théorique que pratique. En effet nous envisageons de proposer d’autres formalisations du désaccord, d’étudier le signe de ces termes relativement au nombre de représentations et de proposer une version “noyau” du modèle CoFKM (cf. Kulis et al. (2005)) pour permettre un traitement semi-supervisé de ces données. Enfin nous complèterons les expérimentations avec d’autres jeux de données adaptés tel que le corpus WebKb sur lequel les tests en cours semblent à nouveau confirmer l’intérêt du modèle CoFKM.

## Références

Bekkerman, R. et J. Jeon (2007). Multi-modal clustering for multimedia collections. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*.

Bekkerman, R., M. Sahami, et E. G. Learned-Miller (2006). Combinatorial markov random fields. In *ECML*, pp. 30–41.

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.

Bickel, S. et T. Scheffer (2005). Estimation of mixture models using co-em. In *Proceedings of the ICML Workshop on Learning with Multiple Views*.

Heer, J. et H. Chi (2002). Mining the structure of user activity using cluster stability. In *in Proceedings of the Workshop on Web Analytics, SIAM Conference on Data Mining (Arlington VA*. ACM Press.

Jain, A. K., M. N. Murty, et P. J. Flynn (1999). Data clustering : a review. *ACM Computing Surveys 31*(3), 264–323.

Kulis, B., S. Basu, I. Dhillon, et R. Mooney (2005). Semi-supervised graph clustering : a kernel approach. In *ICML ’05 : Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA, pp. 457–464. ACM.

Pedrycz, W. (2002). Collaborative fuzzy clustering. *Pattern Recogn. Lett. 23*(14), 1675–1686.

Tan, P.-N., M. Steinbach, et V. Kumar (2005). *Introduction to Data Mining*, Chapter 8. Addison Wesley.

Wiswedel, B. et M. R. Berthold (2007). Fuzzy clustering in parallel universes. *Int. J. Approx. Reasoning 45*(3), 439–454.

## Annexe : Simplification de l'expression du critère de $Q_{CoKMF}$

$$
\begin{aligned}
Q_{CoKMF} &= \left(\sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r}^{\beta}||x_{i,r}-c_{k,r}||^2\right)+\eta\Delta \\
&= \left(\sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r}^{\beta}||x_{i,r}-c_{k,r}||^2\right)-\eta\frac{1}{|R|-1}\sum_{r\neq r'}\sum_{x_i\in X}\sum_{k=1}^{K}(u_{i,k,r}^{\beta}-u_{i,k,r'}^{\beta})||x_{i,r}-c_{k,r}||^2 \\
&= \left(\sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r}^{\beta}||x_{i,r}-c_{k,r}||^2\right)-\eta\sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r}^{\beta}||x_{i,r}-c_{k,r}||^2 \\
&- \frac{\eta}{|R|-1}\sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K}(\sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta})||x_{i,r}-c_{k,r}||^2 \\
&= \sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K}(u_{i,k,r}^{\beta}||x_{i,r}-c_{k,r}||^2-\eta(u_{i,k,r}^{\beta}||x_{i,r}-c_{k,r}||^2) \\
&+ \frac{\eta}{|R|-1}(\sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta})||x_{i,r}-c_{k,r}||^2) \\
&= \sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K}(1-\eta)u_{i,k,r}^{\beta}||x_{i,r}-c_{k,r}||^2+\frac{\eta}{|R|-1}(\sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta})||x_{i,r}-c_{k,r}||^2 \\
&= \sum_{r=1}^{|R|}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r,\eta}||x_{i,r}-c_{k,r}||^2
\end{aligned}
$$

avec $u_{i,k,r,\eta} = (1-\eta)u_{i,k,r}^{\beta}+\frac{\eta}{|R|-1}(\sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta})$

## Summary

This paper deals with clustering for multi-view data, i.e. data described with several sets of variables or proximity matrices. Many application domains can be concerned by this problematic, for instance Information Retrieval, biology or chemestry. The aim of this research field is to propose a theoretical and methodological framework allowing the search of clustering schemes that perform a consensus between the different views. This requires to merge information from each source. This fusion can be performed before (a priori fusion), after (a posteriori fusion) or during the clustering process (collaborative approach). We draw one's inspiration from the recent work of Bickel and Sheffer extending the mixture-models clustering approach in order to deal with multi-view data (Co-EM) and we present a general model for fuzzy clustering that generalizes in the same time a priori and a posteriori fusion approaches. The model is validated with first experiments on a suitable dataset.

# Initialisation des masses d'évidence par les Okm pour la théorie des fonctions de croyance. Application aux bioprocédés.

Yann Permal*, Sébastien Danichert*
Guillaume Cleuziou**, Sébastien Régis*

*Laboratoire G.R.I.M.A.A.G.
Université des Antilles et de la Guyane
campus de Fouillole 97110 Pointe-à-Pitre
yann.permal,sdaniche@etu.univ-ag.fr
sregis@univ-ag.fr,
http://grimaag.univ-ag.fr
**LIFO, Université d'Orléans
Rue Léonard de Vinci 45067 Orléans Cedex 2
guillaume.cleuziou@univ-orleans.fr
http://www.univ-orleans.fr/lifo/Members/cleuziou/

**Résumé.** Dans cet article nous présentons une méthode d'initialisation des masses d'évidence des fonctions de croyance par la méthode des Overlapping k-means (Okm). Les Okm permettent de générer des masses d'évidence sur des unions de singletons. La méthode est présentée et une première expérimentation sur des données réelles est présentée.

## 1 Introduction

La théorie des fonctions de croyances connait un succès croissant depuis quelques décénnies dans les domaines de la reconnaissance des formes et de la fusion d'informations. Ce succès peut s'expliquer en partie par le fait que cette théorie permet de gérer à la fois le caractère incertain d'une information ainsi que son imprécision. Elle représente donc une alternative à la théorie des probabilités (qu'elle généralise) qui ne tient compte que de l'incertitude de l'information. Par ailleurs, d'autres approches ont été développées pour tenir compte de l'imprécision de l'information. Dans les domaines de la reconnaissance des formes et de la classification, les plus connues sont certainement la théorie des sous-ensembles flous et la théorie des possibilités qui lui est sous-jacente. Une autre méthode, plus récente, a été proposée pour gérer le caractère imprécis d'une information ; il s'agit des $Okm$ (Overlapping K-Means) qui utilisent des recouvrements. On entend par recouvrement le fait qu'un élément puisse appartenir à plusieurs classes simultanément. Dans cet article nous proposons d'utiliser les Okm pour fournir une initialisation des masses d'évidence pour la théorie des fonctions de croyance. Il n'existe pas en effet de méthode générique pour l'initialisation des masses d'évidence. L'utilisation des Okm pour le calcul des masses d'évidence permet de pleinement tirer partie de la notion d'imprécision de la théorie des fonctions de croyance puisque les Okm fournissent des informations sur l'appartenance simultanée à plusieurs classes.
Dans la section 2 les bases de la théorie des fonctions de croyance sont présentées succinctement, ainsi que quelques-unes des méthodes courantes utilisées pour l'initialisation des masses d'évidence. Dans la section 3, nous présentons la méthode des Okm et dans la section 4, l'adaptation que nous en faisons pour l'initialisation des masses d'évidence. Nous donnons ensuite dans la section 5 un tout premier exemple d'utilisation sur des données réelles issues des bioprocédés. Enfin nous entamons une discussion ouverte sur cette approche dans la section 6 avant de conclure.

## 2 La théorie des fonctions de croyance

### 2.1 Rappel sur les fonctions de croyance

La théorie des fonctions de croyance est une généralisation de la théorie bayésienne qui tient compte des notions d'incertitude et d'imprécision de l'information. Elle a été introduite par Dempster (Dempster (1968)) puis a été formalisée mathématiquement par Shafer (Shafer (1976)). Elle a été utilisée récemment avec succès dans le domaine des bioprocédés (Lardon et al. (2004) Lardon (2004)).
Considérons l'ensemble de tous les évènements possibles (on parle d'ensemble de toutes les hypothèses) ; cet ensemble

est appelé *ensemble de discernement* et est noté $\Theta$. Toutes ces hypothèses sont mutuellement exclusives et sont nommées *singletons*. La théorie des fonctions de croyance porte sur l'ensemble des sous-ensembles $E$ de $\Theta$. Cet ensemble de sous-ensembles de $\Theta$ est noté $2^{\Theta}$. $E$ peut être composé d'un singleton ou d'une union de plusieurs singletons. Une fonction de masse $m(.)$ peut être alors définie de $2^{\Theta}$ vers [0,1] avec les propriétés suivantes :

$$\sum_{E \subseteq \Theta} m(E) = 1 \qquad (1)$$
$$m(\oslash) = 0$$

$m(E)$ est la masse d'évidence associée à $E$.
Les fonctions de *plausibilité* ($Pl(.)$) et de *croyance* ($Bel(.)$) sont définies de $2^{\Theta}$ vers [0,1] comme suit :

$$pl(E) = \sum_{F \cap E \neq \emptyset} m(F) \qquad (2)$$
$$bel(E) = \sum_{F \subseteq E} m(F)$$

Pour obtenir une fusion de l'information de deux sources différentes 1 et 2, il existe une combinaison de leur masses d'évidence appelée règle de Dempster :

$$(m_1 \oplus m_2)(E) = m_{1,2}(E) = \frac{1}{1-K} \sum_{F \cap G = E} m_1(F).m_2(G) \quad E, F, G \subseteq 2^{\Theta} \qquad (3)$$

où $K$ est défini comme suit :

$$K = \sum_{F \cap G = \emptyset} m_1(F).m_2(G) \qquad (4)$$

Le dénominateur $1 - K$ est un facteur de normalisation. Plus précisément K représente la mesure du conflit entre les sources 1 et 2. Plus K est important, plus les sources sont en conflit et moins la fusion a de sens. Si $K = 1$ alors le conflit est total et la fusion n'a pas de sens. On peut généraliser la règle de Dempster à $n$ sources :

$$(\oplus m_i)_{i=1,\ldots,n}(E) = \frac{1}{1-K} \sum_{X_1 \cap \ldots \cap X_n = E} \left(\Pi_{i=1}^{n} m_i(X_i)\right) \qquad (5)$$
$$E, X_i \subseteq 2^{\Theta}, \qquad K = \textstyle\sum_{X_1 \cap \ldots \cap X_n = \emptyset} \left(\Pi_{i=1}^{n} m_i(X_i)\right)$$

Si les sources sont en conflit fort (K est grand) alors la règle de Dempster peut conduire à des résultats erronés (Zadeh (1984)). La raison de ce comportement de la règle de Dempster provient du fait que la masse d'évidence affectée à l'ensemble vide est nul. Cette contrainte $m(\emptyset) = 0$ implique que l'intersection entre deux hypothses est vide. Partant de cette contrainte, deux points de vue sont alors possibles :

- soit l'on travaille dans un *monde fermé*. On considère que les hypothèses décrivent totalement le problème à résoudre. Dans ce cas la solution au problème de classification se trouve forcément parmi les hypothèses données. C'est le point de vue classique de la théorie des fonctions de croyance.
- soit l'on travaille dans un *monde ouvert*, et dans ce cas les hypothèses modélisent partiellement le problème à résoudre (Smets (1990)). Soit la solution au problème de classification se trouve parmi les hypothses données, soit il s'agit d'une nouvelle hypothèse omise ou du moins inconnue. Ce point de vue semble *en général* plus réaliste par rapport aux applications pratiques.

  Il faut noter que d'autres combinaisons ont été proposées comme alternative à la combinaison de Dempster (voir par exemple Lefevre et al. (2002)).

## 2.2 Initialisation des masses d'évidence

Comme nous l'avons signalé, il n'existe pas de méthode générique pour l'initialisation des masses d'évidence. En l'absence de données réelles sur ces masses d'évidence, on pourrait utiliser à l'instar de la théorie des probabilités, une

hypothèse simplificatrice comme l'équiprobabilité pour ces masses, mais cela réduirait considérablement l'intérêt de l'utilisation de la théorie des fonctions de croyance.
Plusieurs approches ont donc été proposées pour le calcul des masses d'évidence (voir notamment Vannoorenberghe (2003) et Martin (2005)). On peut citer entre autre, les fonctions à support simple (Shafer (1976)) , la méthodes basée sur les densités de probabilités (Appriou (1991), Smets (1993)) , la méthode utilisant des arbres de décisions (Vannoorenberghe et Denoeux (2002)) et celle utilisant les k-plus proches voisins (Denoeux (1995)). Les résultats et les performances de ces méthodes dépendent du domaine d'application, mais il faut noter que la plupart d'entre elles (sauf celle utilisant des arbres de décisions) fournissent des masses d'évidences uniquement sur des singletons et sur l'ensemble $\Theta$, ou bien sur des singletons, sur leur complémentaires, et sur $\Theta$. Aussi, bien qu'elles fournissent des informations sur l'imprécision des informations, on constate que cette imprécision reste quand même restreinte puisqu'elle est fortement liée à chaque singleton pris séparément et à $\Theta$ : la plupart de ces méthodes ne fournissent pas d'informations sur les unions de singletons. L'approche utilisant les Okm permet de pallier ce problème.

# 3 Les Okm

La méthode des Okm (Cleuziou (2007)) est une méthode de classification permettant de créer des recouvrements au niveau de la classification. On rappelle qu'un recouvrement désigne le fait qu'un élément puisse appartenir à plusieurs classes simultanéments. Cette notion de recouvrement permet un certain enrichissement de la classiffication.
Dans le cas d'un document par exemple, choisir une classe thématique et une seule pour ce document peut réduire considérablement la représentation que l'on conservera de ce document dans la classification. En revanche, autoriser ce document à s'afficher selon plusieurs thèmes rendra une image certainement plus juste de son contenu. La qualité d'un recouvrement pourra alors être mesurée relativement à l'écart entre le contenu réel des objets et l'"image" que la classification (ici le recouvrement) établie renvoie d'eux. Nous formalisons cette intuition dans le critère suivant :

$$W(\mathcal{R}) = \sum_{x_i \in X} d^2(x_i, \overline{x_i})$$

L'image d'un objet $x_i$ dans un recouvrement $\mathcal{R}$ est notée $\overline{x_i}$ dans ce critère et correspond à un compromis entre les différentes classes auxquelles cet objet appartient. Ainsi pour un recouvrement $\mathcal{R}$ en $k$ classes $\{R_1, \ldots, R_k\}$ de centres respectifs $\{c_1, \ldots, c_k\}$, $\overline{x_i}$ est défini par le centre de gravité de l'ensemble $\{c_j | x_i \in R_j\}$.

L'algorithme OKM que nous détaillons dans cette section présente un squelette (figure 1) similaire à l'algorithme des $k$-moyennes (voir Cleuziou (2007)). L'initialisation qui consiste à tirer aléatoirement $k$ centres puis à dériver un premier recouvrement est suivie par l'itération de deux étapes : (1) la mise à jour des centres de classes puis (2) l'affectation des objets à ces centres.

**Initialisation** : $t$=0
choisir aléatoirement $k$ centres $C^t = \{c_1^t, c_2^t, \ldots, c_k^t\}$ dans $X$,
Pour chaque $x_i \in X$ : Affecter($x_i$,$C^t$),
en déduire un recouvrement initial $\mathcal{R}^t = \{R_1^t, R_2^t, \ldots, R_k^t\}$.
**Faire**
$t$=$t$+1
• Mettre_à_jour($C^{(t-1)}$,$\mathcal{R}^{(t-1)}$) (*en déduire* $C^t$),
• Pour chaque $x_i \in X$ : Affecter($x_i$,$C^t$) (*en déduire* $\mathcal{R}^t$),
**Tant que** $\mathcal{R}^t \neq \mathcal{R}^{t-1}$

FIG. 1 – *Squelette de l'algorithme* OKM.

L'intérêt de l'algorithme OKM réside dans la méthode employée pour Mettre_à_jour les centres et pour Affecter chaque objet à un ou plusieurs centres. Ces deux opérations doivent d'une part assurer la cohérence des classes en regroupant ensemble des objets similaires et d'autre part permettre la convergence de la méthode par décroissance du critère $W(.)$.

Étant donné un ensemble $C = \{c_1, c_2, \ldots, c_k\}$ correspondant aux centres des $k$ classes respectives $R_1, R_2, \ldots, R_k$ d'un recouvrement $\mathcal{R}$, la méthode d'affectation d'un objet $x_i$, présentée en figure 2 consiste à parcourir l'ensemble des centres de classes du plus proche au plus éloigné (suivant une métrique $d$) et à affecter $x_i$ tant que son image est améliorée ($d(x_i, \overline{x_i})$ diminue). La nouvelle affectation de l'objet $x_i$ ne sera finalement conservée que si l'image de $x_i$ s'en trouve améliorée par rapport à l'ancienne affectation. Cette dernière précaution permet d'assurer la décroissance du critère $W(.)$ lors de l'étape d'affectation.

Affecter($x_i$,$C$) :

**Initialisation** : Soit $c^*$ le centre de $C$ le plus proche de $x_i$ ($\forall c_j \in C,\ d(x_i, c^*) \leq d(x_i, c_j)$),
$A = \{c^*\}$ (liste des affectations),
$C = C \setminus \{c^*\}$.

**Faire**
Soit $c^*$ le centre de $C$ le plus proche de $x_i$ et $\overline{x_i}^A$ le centre de gravité des éléments de $A$,
Si $d(x_i, \overline{x_i}^{A \cup \{c^*\}}) < d(x_i, \overline{x_i}^A)$ alors $A \leftarrow \{c^*\}$ et $C = C \setminus \{c^*\}$,
Sinon **STOP**
**Tant que** $C \neq \emptyset$

Soit $A'$ l'ancienne affectation de $x_i$,
Si $d(x_i, \overline{x_i}^A) < d(x_i, \overline{x_i}^{A'})$ alors affecter $x_i$ aux centres de $A$,
Sinon conserver l'ancienne affectation $A'$.

FIG. 2 – *Méthode d'affectation utilisée dans l'algorithme* OKM.

Enfin, la mise à jour du centre $c_j$ de la classe $R_j$ est définie dans l'algorithme OKM par :

$$c_{j,v} = \frac{1}{\sum_{x_i \in R_j} \frac{1}{\delta_i^2}} \times \sum_{x_i \in R_j} \frac{1}{\delta_i^2} . \hat{x_i}_v^j \tag{6}$$

Dans cette expression, $c_{j,v}$ désigne la $v^{\text{ième}}$ composante du vecteur $c_j$, $\delta_i$ correspond au nombre de classes de $\mathcal{R}$ auxquelles $x_i$ appartient et $\hat{x_i}_v^j$ symbolise la $v^{\text{ième}}$ composante du centre $c_j$ "idéal" pour l'objet $x_i$, c'est à dire le centre $c_j$ tel que $d(x_i, \overline{x_i}) = 0$. De façon plus précise on a $\hat{x_i}_v^j = \delta_i . x_{i,v} - (\delta_i - 1) . \overline{x_i}_v^{A \setminus \{c_j\}}$ où $A$ désigne l'ensemble des centres des classes auxquelles $x_i$ appartient. Il découle de ce qui précède une définition plus intuitive du nouveau centre $c_j$ qui correspond finalement au centre de gravité du nuage de points $\{(\hat{x_i}^j, p_i) | x_i \in R_j\}$ où chaque $\hat{x_i}^j$ est pondéré par $p_i = \frac{1}{\delta_i^2}$.

Il a été montré que chaque mise à jour d'un centre dans OKM permet d'assurer la décroissance du critère $W(.)$ mais également que le nouveau centre calculé est celui qui minimise ce critère (voir Cleuziou (2007)). Notons pour conclure sur la présentation de l'algorithme, que la méthode des $k$-moyennes peut être considérée comme un cas particulier de OKM. En effet si on restreint dans OKM chaque objet à n'appartenir qu'à une seule classe ($\delta_i$=1) on retrouve exactement le processus de classification utilisé dans l'algorithme $k$-moyennes. Il s'agit donc d'un algorithme non-déterministe puisque le résultat dépendra de l'initialisation ; de plus, chaque classe n'étant plus indépendante l'une de l'autre dans un recouvrement, l'algorithme OKM dépendra également de l'ordre de parcours des classes lors de l'étape de mise à jour des centres.

# 4 Utilisation des Okm pour l'initialisation des masses d'évidence

L'utilisation des Okm pour générer des masses d'évidence nécessite quelques adaptations de l'algorithme original. En premier lieu, l'algorithme présenté ci-dessus utilise des données multivariées. Pour la génération des masses d'évidence, nous utiliserons une version monovariée de cet algorithme, la prise en compte des différentes composantes des données se faisant par l'intermédiaire de la fusion d'informations de la théorie des fonctions de croyance : chaque composante est donc considérée comme une source d'informations. Par ailleurs, dans l'algorithme présenté, dans la partie Affecter, le calcul de distance $d$ entre $x_i$ et le centre de gravité des éléments de $A$ s'arrête dès que l'on a trouvé la distance minimale (voir 2). Pour le calcul des masses d'évidence, on s'arrêtera au même niveau c'est-à-dire que l'on calculera des masses d'évidence uniquement pour les recouvrements trouvés par les Okm (on notera que dans la méthode que nous proposons nous calculons systématiquement les masses d'évidences pour les singletons). Les figures 3 et 4 présentent l'algorithme modifié des Okm et la figure 5 résume l'approche utilisé pour le calcul des masses d'évidence.

La masse d'évidence d'une classe ou d'une union de classe par rapport à un élément $x_i$ (que l'on notera $x$ ultérieurement pour plus de lisibilité) sera égale à la distance entre $x$ et le centre de gravité du centre de cette classe ou de cette union de classe, pondérée par l'inverse de la somme de toutes les distances des différents éléments focaux relativement à $x$.
Ainsi par exemple pour la masse d'évidence du recouvrement $C_1 \cup C_2$ relativement à un élément $x$ (où $C_1$ et $C_2$ sont

**Initialisation** : $t$=0
choisir aléatoirement $k$ centres $C^t = \{c_1^t, c_2^t, \ldots, c_k^t\}$ dans $X$,
Pour chaque $x_i \in X$ :
calculer et stocker la distance entre $x_i$ et chaque $c_j^t$, $j = 1, \ldots, k$,
Affecter($x_i$,$C^t$),
en déduire un recouvrement initial $\mathcal{R}^t = \{R_1^t, R_2^t, \ldots, R_k^t\}$.
**Faire**
$t$=$t$+1
• Mettre_à_jour($C^{(t-1)}$,$\mathcal{R}^{(t-1)}$) (*en déduire* $C^t$),
• Pour chaque $x_i \in X$ : Affecter($x_i$,$C^t$) (*en déduire* $\mathcal{R}^t$),
**Tant que** $\mathcal{R}^t \neq \mathcal{R}^{t-1}$

FIG. 3 – *Squelette de l'algorithme* OKM *modifié.*

Affecter($x_i$,$C$) :
**Initialisation** : Soit $c^*$ le centre de $C$ le plus proche de $x_i$ ($\forall c_j \in C,\ d(x_i, c^*) \leq d(x_i, c_j)$),
$A = \{c^*\}$ (liste des affectations),
$C = C \setminus \{c^*\}$.
**Faire**
Soit $c^*$ le centre de $C$ le plus proche de $x_i$ et $\overline{x_i}^A$ le centre de gravité des éléments de $A$,
Si $d(x_i, \overline{x_i}^{A \cup \{c^*\}}) < d(x_i, \overline{x_i}^A)$ alors $A \leftarrow \{c^*\}$ et $C = C \setminus \{c^*\}$,
stockage de $d(x_i, \overline{x_i}^A)$
Sinon **STOP**
**Tant que** $C \neq \emptyset$
Soit $A'$ l'ancienne affectation de $x_i$,
Si $d(x_i, \overline{x_i}^A) < d(x_i, \overline{x_i}^{A'})$ alors affecter $x_i$ aux centres de $A$,
Sinon conserver l'ancienne affectation $A'$.

FIG. 4 – *Méthode d'affectation utilisée dans l'algorithme* OKM.

deux classes distinctes) sera égale à :

$$m(C_1 \cup C_2) = \frac{exp(-d(x, \overline{x}^{C_1 \cup C_2}))}{\sum exp(-d(x, E))} \tag{7}$$

où $E$ désigne tous les éléments focaux possibles et $exp$ est la fonction exponentielle. D'autres équations sont possibles pour le calcul des masses d'évidence et pourront être testées ultérieurement. En particulier, d'autres fonctions de type exponentielle pourront être utilisées dans l'équation 7.
On notera cependant que compte tenu de la structure fournie par la méthode des Okm, tous les éléments focaux-exception faite des singletons (i.e. des classes distinctes les unes des autres)- sont imbriqués les uns dans les autres.

# 5 Résultats expérimentaux

## 5.1 Les bioprocédés

Les bioprocédés et les technologies de l'environnement utilisent de plus en plus des outils issus de l'informatique et des mathématiques : traitement du signal, fouille de données etc. Ainsi, l'utilisation des méthodes de classification représente à la fois une alternative et un complément aux méthodes de modélisation mathématique utilisées dans les bioprocédés. Ces outils de classification fournissent un modèle comportemental utile pour la compréhension et la reconnaissance de la physiologie des micro-organismes. Plus précisément, les méthodes de classification sont appliquées sur les variables biochimiques mesurées en ligne pendant le bioprocédé : les variables biochimiques représentent donc les sources d'information pour la classification. Le regroupement des variables en "paquets" permet de caractériser les états physiologiques des micro-organismes par une ou plusieurs classes. Ces méthodes de classification ont fourni des résultats intéressants (Waissman-Vilanova (2000) Goma et al. (2004) Régis (2004)) et le recours à ces approches dans le domaine des bioprocédés se multiplie. Si l'on se réfère uniquement au niveau macroscopique qui nous intéresse, parmi les outils informatique et mathématique utilisés, on peut citer :

**Calcul des distances** :
Lancement de la méthode des OKM

**Initialisation des masses d'évidence** :
Pour chaque $x_i \in X$ :
Récupération de la distance entre $x_i$ et chaque centre de recouvrement (centre de classe ou d'union de classes),
Calcul de la masse d'évidence du recouvrement (classe ou union de classes) relativement à $x_i$ ,

FIG. 5 – *Méthode de calcul des masses d'évidence.*

- les modèles mathématiques. Il s'agit de trouver un modèle susceptible de reconstituer les phases de croissance des micro-organismes (Roels (1983)). La difficulté d'utilisation de ces modèles vient de la complexité du vivant qui nécessite de multiplier les variables dans l'équation qui définit le modèle. Par ailleurs, les modèles varient en fonction des micro-organismes utilisés (nous avons dénombré plus de 130 modèles).
- les techniques issues de l'intelligence artificielle. Elles cherchent à modéliser de façon explicite les connaissances des experts (voir par exemple Steyer (1991)). Cependant le nombre de règles d'experts peut augmenter de façon quasi exponentielle et la modélisation de ces connaissances sous forme de règles pour un système expert est un travail long et fastidieux.
- les techniques issues de la classification. On cherche à regrouper les mesures effectuées en ligne dans des classes de telle sorte que ces classes soient bien différentes les unes des autres tout en ayant pour chacune d'elles, la plus grande homogènéité possible. Ces techniques peuvent faire appel ou non aux connaissances des experts.

Nous nous intéressons surtout aux méthodes de classification. Dans la pratique, on cherche à classer de façon automatique les variables biochimiques mesurées durant l'expérience. Il s'agit de trouver des classes qui correspondent aux états physiologiques des micro-organismes du bioprocédé. La reconnaissance de ces états permet de contrôler et d'optimiser le bioprocédé. Les variables biochimiques se présentent sous la forme de séries temporelles. La classification consiste donc à segmenter les séries temporelles de telle sorte qu'une classe ou un groupe de classes consécutives correspondent à un état physiologique donné. En fait la classification peut être réalisée "manuellement" par un expert en microbiologie (voir figure 6). Celui-ci s'appuie sur une analyse a posteriori, qui nécessite une analyse biologique longue et disponible uniquement plusieurs jours après la fin du bioprocédé. On cherche donc à se rapprocher le plus possible du travail réalisé manuellement par l'expert en utilisant des méthodes automatiques. Le bioprocédés que nous étudions sont des bioprocédés

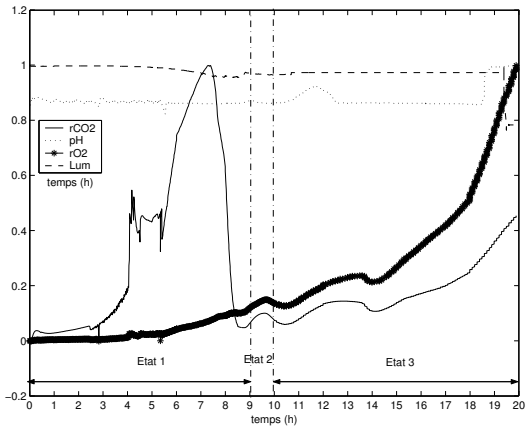



FIG. 6 – *Etats physiologiques fournis par les experts sur un bioprocédé de type batch. L'axe des abscisses représente le temps exprimé en heures, l'axe des ordonnées représente l'amplitude des variables biochimiques (les valeurs ont été normalisées). 4 variables ont été utilisées : le pH, la vitesse de consommation de l'oxygène (rO2), la vitesse de production de dioxyde de carbone (rCO2) et la luminance (Lum) qui traduit la production de biomasse. On distingue 3 états : la fermentation (état 1), la diauxie (état 2), l'oxydation (état 3).*

de fermentation utilisant les micro-organismes appelés *Saccharomyces Cerevisiae*. Le bioprocédé étudié est une fermentation batch. Dans ce type de fermentation, le substrat est placé au début du procédé puis aucune intervention extérieure n'est faite jusqu'à la fin de l'expérience. Pour ce bioprocédé, l'expert en microbiologie connaît parfaitement les différents états physiologiques (voir figure 6). Il est donc possible d'effectuer une classification supervisée en utilisant des échantil-

lons étiquetés. On pourra également évaluer le pourcentage de classification correcte.
Pour ce bioprocédé batch, l'expérience dure environ 20 heures et correspond à 1012 points de mesures des variables biochimiques. On considère le début du bioprocédé comme étant t=0 heure (0h). On rappelle que l'on cherche à détecter trois états physiologiques principaux :

- l'état 1 : la fermentation (production d'éthanol). Elle va de 0h jusqu'à environ 9h ce qui représente un total de 590 points mesurés.
- l'état 2 : la diauxie. Cet état commence à environ 9h et se termine à 9h46 ce qui représente environ 33 points de mesure. C'est le plus petit état physiologique (en temps et en quantité de données) parmi les 3 et le plus difficile à caractériser
- l'état 3 : l'oxydation (production de biomasse). Elle commence à 9h46 et se termine en même temps que la fin de l'expérience à 20h ce qui représente 389 points de mesure.

Il y a 21 variables biochimiques et chacune d'elles a donc 1012 éléments. On considère donc 3 classes qui correspondent aux 3 états physiologiques

## 5.2 Résultats de classification

Pour effectuer le calcul des masses d'évidence pour les 3 classes et les recouvrements (union de classes), nous avons fourni 3 centres de classes correspondant aux moyennes d'échantillons labellisés fournis par les microbiologistes. Il aurait été possible de choisir aléatoirement les 3 centres de classes mais nous avons cherché à optimiser les résultats. Après utilisation des Okm pour le calcul des masses d'évidence, chaque variable biochimique avait une ensemble de masses d'évidence définies pour chaque élément de mesure $x$. On rappelle que pour chaque variable biochimique prise séparément, on obtient des éléments focaux imbriqués les uns dans les autres. La classification est basée sur la recherche du maximum parmi les masses d'évidence (les notions de plausibilité et de crédibilité ne sont pas utiles ici pour prendre une décision car, du fait de la nature imbriquée des élément focaux, ces deux valeurs sont concentrées sur $\Theta$). Les résultats obtenues sont présentés sur la figure 7. Il est difficile d'évaluer cette classification car on ne peut pas vraiment la comparer avec celle fournie par les experts en microbiologie en raison de la présence des recouvrements. Quoiqu'il en soit, la classification semblent présenter beaucoup d'oscillations entre les classes et les recouvrements, et ne semblent pas fournir d'information pertinente pour la caractérisation des états physiologiques. Cependant en regardant de plus près certains intervalles temporels, les résultats sont moins chaotiques qu'il n'y paraît. Ainsi, si l'on s'intéresse à la transition entre l'état 1 ($C_1$) et l'état 2 ($C_2$) aux alentours de $t = 9h30$ (voir figure 8), on constate que la plupart des éléments appartiennent soit à $C_2$, soit à $C_3$, soit à $C_2 \cup C_3$ (seuls quelques éléments appartiennent à $C_1$). Ainsi la plupart des éléments appartiennent à la classe $C_2$ aux alentours de $t = 9h30$ avec une précision plus ou moins grande, ce qui est en adéquation avec l'intervalle temporel correspondant à l'état 2 (voir sous paragraphe 5.1).

On constate que beaucoup d'éléments appartiennent à des unions de classes et qu'il ne se dégage pas de classe prédominante sur un intevalle de temps important (contrairement à la classification fournie par l'expert en microbiologie). Ceci peut s'expliquer, entre autre, par le fait que les classes soient numériquement très proches pour certaines variables comme le pH ($C_1$ et $C_2$) ou la luminance ($C_2$ et $C_3$) (voir tableau 1).
Par ailleurs, ces résultats peuvent aussi s'expliquer par le conflit fort existant entre plusieurs variables biochimiques. Ce conflit perturbe souvent les résultats de la classification (voir Régis et al. (2007)). Ces résultats sont encore partiels et leur analyse doit être approfondie.

| | pH | rCO2 | rO2 | Lum |
|---|---|---|---|---|
| Etat1 | m=4.990 | m=22.241 | m=1.659 | m=69.109 |
| | $\sigma$=0.045 | $\sigma$=20.641 | $\sigma$=1.648 | $\sigma$=0.996 |
| Etat 2 | m=4.992 | m=6.278 | m=6.745 | m=67.806 |
| | $\sigma$=0.024 | $\sigma$=0.736 | $\sigma$=0.457 | $\sigma$=0.098 |
| Etat 3 | m=5.164 | m=13.857 | m=19.548 | m=67.180 |
| | $\sigma$=0.295 | $\sigma$=7.687 | $\sigma$=12.394 | $\sigma$=3.447 |

TAB. 1 – *Propriétés statistiques des états physiologiques : moyenne $m$ et écart-type $\sigma$ pour quelques variables pH, rCO2 (vitesse de production de dioxyde de carbone), rO2 (vitesse de consommation du dioxygène), Lum (luminance)*

Bien qu'ils soit délicat d'évaluer numériquement les performances de cette approche, nous avons tout de même chercher à estimer et à comparer le pourcentage de classification correcte. Pour cela nous avons choisi d'utiliser deux variables biochimiques : ces variables sont le rCO2 et le rO2, et elles ont été choisies par les experts en microbiologie car

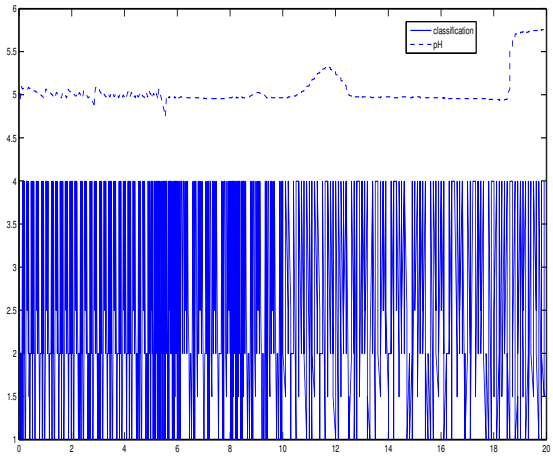



FIG. 7 – *La classification obtenue après fusion des masses d'évidence par la combinaison de Dempster. Sur le graphique le pH est également représenté. L'axe des abscisses représente le temps de mesure de l'expérience en heures. Des valeurs fictives ont été utilisée pour visualiser la classification : $C_1$ est représenté par la valeur 1, $C_2$ par 2, $C_3$ par 4, $C_1 \cup C_2$par 1.5, $C_1 \cup C_3$par 3, $C_2 \cup C_3$par 2.5, $C_1 \cup C_2 \cup C_3$ par 3.5.*

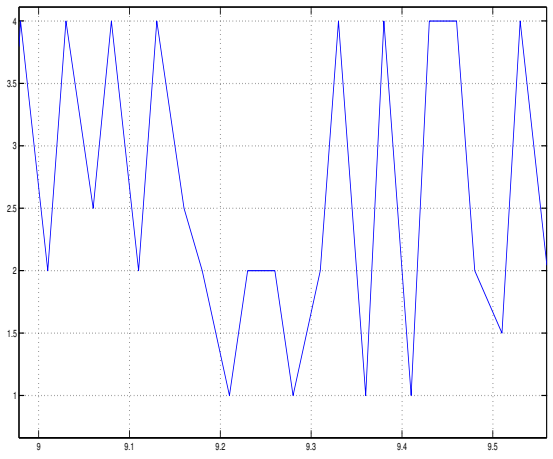


FIG. 8 – *Un zoom autour de la valeur $t = 9h30$. Des valeurs fictives ont été utilisée pour visualiser la classification : $C_1$ est représenté par la valeur 1, $C_2$ par 2, $C_3$ par 4, $C_1 \cup C_2$par 1.5, $C_1 \cup C_3$par 3, $C_2 \cup C_3$par 2.5, $C_1 \cup C_2 \cup C_3$ par 3.5.*

elles sont toutes les deux représentatives des différentes phases physiologiques du bioprocédé et ne présentent de conflit fort (voir Régis et al. (2007)). Pour comparer la classification fournie par la méthode présentée dans cet article à celle fournie par les experts, nous avons adopté la démarche suivante :

- si un élément appartient à une union de deux classes $C_i \cup C_j$, alors on considère que cet élément appartient à la classe ayant le plus grand nombre d'éléments. Par exemple si un point appartient à $C_2 \cup C_3$, on considère qu'il appartient à $C_3$ car $C_3$ (389 points) a plus d'éléments que $C_2$ (33 points).
- si un élément appartient à $\Theta$ (i.e. $C_1 \cup C_2 \cup C_3$) alors on considère qu'il n'appartient à aucune des 3 classes.

Les centres de classes ont été initialisés avec les moyennes d'échantillons labellisés fournis par les microbiologistes. Nous avons cherché à comparer cette méthode avec une autre méthode de calcul de masses d'évidence. Nous avons utilisé la méthode des k-plus proches voisins (Denoeux (1995)) qui est une méthode supervisée : cette méthode a été utilisé avec les deux variables rO2 et rCO2 (pour les détails de sa configuration voir Régis et al. (2007)). Les résultats (voir tableau 2) montrent que les deux méthodes donnent des résultats similaires alors que la méthode des k-plus proches voisins est supervisée alors que celle des OKM ne l'est pas (même si l'initialisation a été réalisée en utilisant les moyennes d'échantillons labellisés). Ces premiers résultats montrent que cette approche peut fournir des résultats comparables aux méthodes existantes tout en laissant une certaine flexibilité pour le choix des éléments focaux puisque ceux-ci ne sont pas imposés à l'initialisation.

| méthode d'initialisation des éléments focaux | % de classification correcte |
|---|---|
| k-plus proches voisins | 57.80% |
| Okm | 56.12% |

TAB. 2 – *Pourcentage de classification correcte des méthodes Okm et k-plus proches voisins utilisées pour la théorie des fonctions de croyance. Les deux classification ont été réalisé après la fusion des éléments focaux des variables rO2 et rCO2. Elles ont ensuite été comparées à une classification "idéale" fournie par les experts en microbiologie.*

# 6 Discussion

Quoiqu'il en soit on peut déjà constater que les Okm permettent un enrichissement au niveau de l'initialisation des masses d'évidence de la théorie des fonctions de croyance. En effet, les informations fournies par les Okm permettent de générer des masses d'évidence pour des éléments focaux qui sont des unions de singletons. On peut donc utiliser les concepts liés à la gestion de l'imprécision dans la théorie des fonctions de croyance, et espérer ainsi une modélisation plus fidèle de la connaissance dans des applications réelles. De plus, le calcul des masses d'évidence se fait à partir des résultats de recouvrement de la méthode Okm : le choix des éléments focaux n'est pas réalisé de façon arbitraire par un seuil et les éléments focaux peuvent varier d'une source d'information à une autre. On est ainsi plus proche de certaines applications réelles et on ne "force" pas les sources à donner systématique des valeurs "artificielles" pour des élémens focaux pour lesquels elles ne fournissent pas d'information.

Il faut cependant noter que la structure imbriquée des éléments focaux ne permet pas d'utiliser réellement les notions de plausibilité et de crédibilité car toute la masse est alors concentrée sur $\Theta$ (qui est égale à $C_1 \cup C_2 \cup C_3$ dans le cas du bioprocédé présenté ci-dessus). En revanche cette structure imbriquée permet d'envisager le passage à la théorie des possibilités (voir par exemple Bouchon-Meunier (1991)).

Par ailleur, le principal inconvénient de cette approche est lié aux propriétés intrinsèques de la théorie des fonctions de croyance. En effet, la complexité et le coût en temps de calcul peuvent devenir prohibitif, compte tenu de l'augmentation de la cardinalité de l'ensemble de discernement. Ainsi en utilisant cette nouvelle méthode d'initialisation des masses par les Okm, on augmente la complexité. De plus, la méthode devra être réadaptée à chaque nouvelle application. Cette approche mérite cependant d'être appronfondie au niveau applicatif pour mieux comprendre son impact au niveau des résultats de classification.

# 7 Conclusion

Nous avons proposé une nouvelle méthode de calcul des masses d'évidence pour les éléments focaux de la théorie des fonctions de croyance. Cette méthode d'initialisation est basée sur les Okm qui permettent de créer des recouvrements au niveau de la classification. Un premier résultat expérimental a été présenté et devra être approfondi. Par ailleurs la structure imbriquée des éléments focaux permettra a priori, dans un second temps, d'utiliser la théorie des possibilités pour effectuer la classification.

## Références

Appriou, A. (1991). Probabilités et incertitude en fusion de données multi-senseurs. *Revue Scientifique et Technique de la Défense 11*, 27–40.

Bouchon-Meunier, B. (1991). La logique floue. P.U.F. (Que sais-je ?).

Cleuziou, G. (2007). Les okm : une extention des k-moyennes pour la recherche de classes recouvrantes. *EGC 07, Namur, Belgique, RNTI-E-9*, 691–702.

Dempster, A. (1968). A generalisation of bayesian inference. *Journal of the Royal Statistical Society 30*, 205–247.

Denoeux, T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE trans. on systems, man, and cybernetics 25*(5), 804–813.

Goma, G., J.-L. Uribelarrea, V. Guillouet, et C. Jouve (2004). Tackling complexity in industrial microbiology for bioprocess. In *4rth International Congress on Bioprocess in Food Industries*, Clermont-Ferrand.

Lardon, L. (2004). *Représentation et gestion des incertitudes pour le diagnostic par la théorie de Dempster-Shafer : application aux procédés biologiques*. Thèse de Doctorat, Ecole Nationale Supérieure Agronomique de Montpellier, Montpellier.

Lardon, L., A. Punal, et J.-P. Steyer (2004). On-line diagnostic and uncertainty management using evidence theory-experimental illustration to anaerobic digestion processes. *Journal of Process Control 14*, 747–763.

Lefevre, E., O. Colot, et P. Vannoorenberghe (2002). Belief function combination and conflict management. *Information Fusion 3*, 149–162.

Martin, A. (2005). La fusion d'informations. Polycopié de cours ENSIETA.

Régis, S. (2004). *Segmentation, classification, et fusion d'informations de séries temporelles multi-sources : application à des signaux dans un bioprocédé*. Thèse de Doctorat, Université des Antilles et de la Guyane.

Régis, S., A. Doncescu, et J. Desachy (2007). Théorie des fonctions de croyance pour la fusion et l'évaluation de la pertinence des sources d'informations : application à un bioprocédé fermentaire. *Traitement du signal 24*(2), 115–132.

Roels, J. (1983). *Energetics and kinetics in biotechnology*, Chapter Macroscopic theory and microbial growth and product formation, pp. 23–73. Elsevier Biomedical Press.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. New Jersey : Princeton University Press.

Smets, P. (1990). The combination of evidence in the transferable belief model. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (12), 447–458.

Smets, P. (1993). Belief functions : The disjonctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning 9*, 1–35.

Steyer, J. (1991). *Sur une approche qualitative des systèmes physiques : aide en temps réèl à la conduite des procédés fermentaires*. Thèse de Doctorat, Université Paul Sabatier, Toulouse France.

Vannoorenberghe, P. (2003). Un état de l'art sur les fonctions de croyances appliquées au traitement de l'information. *Revue I3*.

Vannoorenberghe, P. et T. Denoeux (2002). Handling uncertain labels in multiclass problems using belief decision trees. In *IPMU'02*, Annecy, France.

Waissman-Vilanova, J. (2000). *Construction d'un modèle comportemental pour la supervision de procédés : application à une station de traitement des eaux*. Thèse de Doctorat, LAAS - CNRS, Institut National Polytechnique de Toulouse.

Zadeh, L. (1984). A mathematical theory of evidence (book review). *AI magazine 5*(3), 81–83.

## Summary

We present a method for the calculation of the mass function in the belief function theory. This method uses a method of classification called the overlapping k-means (Okm) which enables to create union of classes for the mass function. The method is presented and a first experimental result is presented.

# Fusion multi-vues à partir de fonctions de croyance pour la classification d'objets

Hicham Laanaya*, Arnaud Martin*

*ENSIETA, E$^3$I$^2$-EA3876, 2 rue François Verny, 29806 Brest Cedex 9
{Hicham.Laanaya, Arnaud.Martin}@ensieta.fr,
http://www.ensieta.fr/e3i2/

**Résumé.** Nous présentons dans cet article une approche de fusion crédibiliste pour la classification d'objets à partir d'images multi-vues. Cette approche est appliquée sur des données générées à partir de trois formes de base (un cercle, un hexagone et un octogone). Ces données simulent une sortie de classifieur et représentent ainsi les positions des objets. Les résultats obtenus montrent l'intérêt de l'exploitation de l'information extraite sur plusieurs vues pour la classification d'un même objet.

## 1 Introduction

Cet article présente une approche de classification d'objets fondée sur l'utilisation des informations extraites à partir d'images de l'objet sous plusieurs vues. Sur chaque vue on extrait des informations (attributs) qu'on utilise pour la classification de cet objet, ainsi, chaque vue, après la phase de classification, donne une information sur la classe et la position de l'objet. Une fusion de ces classes parait nécessaire pour augmenter le taux de reconnaissance de l'objet (*cf.* figure 1).

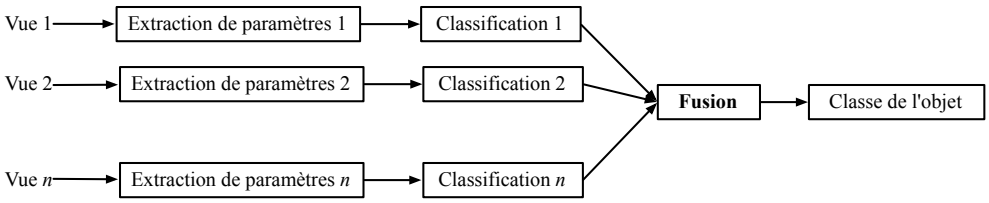



FIG. 1 – *Fusion multi-vues pour la classification d'un objet*

La fusion multi-vues pour la classification d'objets peut donc être vue comme un problème de fusion de résultats de classification de chaque vue (ou fusion de classifieurs). Les méthodes classiques issues de la théorie de l'incertain (fusion par vote, fusion par vote pondéré, fusion bayésienne, fusion par les fonctions de croyance, ...) peuvent donc être utilisées pour réaliser cette fusion (*cf.* Laanaya et al. (2008), Martin (2005a)). Dans ces théories de l'incertain deux notions bien distinctes sont essentielles afin de bien modéliser les imperfections des données : l'incertitude et l'imprécision. L'incertitude caractérise un degré de conformité à la réalité (défaut qualitatif de l'information), tandis que l'imprécision mesure un défaut quantitatif de l'information (par exemple une erreur de mesure).

Plusieurs études dans la littérature utilisent l'information extraite sur plusieurs vues pour la détection d'objets, par exemple dans la thèse de Daniel (1998), ils ont étudié l'apport de l'utilisation de plusieurs vues sonar pour améliorer les taux de classification des objets. Dans le même contexte, Quidu (2001); Aridgides et al. (2001) utilisent la classification multi-vues pour la détection des mines. En imagerie sonar, la classification multi-vues a plusieurs intérêts tels que le positionnement d'un robot autonome, et nécessite généralement une étape de recalage des images Dhibi et al. (2008). La fusion multi-vues a été aussi utilisée pour la classification des sexes en se fondant sur une séquence de marche Huang et Wang (2007). Dans le domaine de la télédétection, Milisavljevic et al. (2008) ont utilisé une approche fondée sur la théorie des fonctions de croyance et la théorie des possibilités pour la détection de mines antipersonnel.

Cet article présente une étude théorique de la fusion pour la classification multi-vues qui a pour objectif de montrer la faisabilité et l'intérêt de la combinaison des informations issues de plusieurs classifieurs à partir d'images prises sous différents angles, en particulier en imagerie sonar. Le manque de ce type de données nous a poussé à réaliser cette étude préliminaire à partir de données générées.

Notre choix s'est porté sur la théorie des fonctions de croyance qui permet, dans un même cadre théorique, de bien modéliser l'incertitude et l'imprécision et offre des avantages également pour modéliser le manque d'information selon le point de vue.

L'organisation de cet article est la suivante. Dans la section 2, nous présentons les bases théoriques des fonctions de croyance. Nous donnerons ensuite une description des données générées et utilisées pour la validation de l'approche de fusion mutli-vues. Les résultats de classification sont donnés dans la section 3.2.

## 2 Fusion pour la classification multi-vues

La fusion pour la classification multi-vues exploite l'information acquise sur un objet pour plusieurs vues. Cette information, dans la plupart du temps, est entachée d'imperfections liées généralement au milieu étudié et aux capteurs utilisés. Les fonctions de croyance présentées dans la section 2.1 permettent une meilleure modélisation de ce genre d'information.

Le principe général d'une approche de fusion d'informations est décrit par la figure 2 Martin (2005b). Ainsi nous considérons les informations issues des différents classifieurs à fusionner, ainsi que des informations supplémentaires et des connaissances externes liées à l'application pour aider la combinaison.

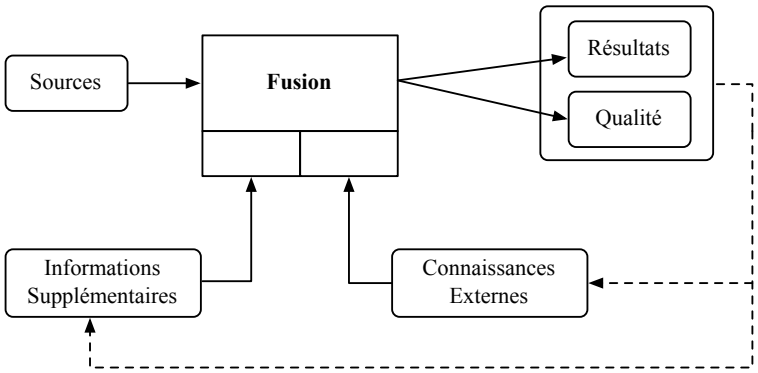



FIG. 2 – *Représentation de la fusion*

Le processus de la fusion d'information est décrit par quatre étapes : la modélisation, l'estimation, la combinaison et la décision (*cf.* figure 3). La modélisation définie le choix du formalisme, qui sera dans notre cas la théorie des fonctions de croyance. L'estimation permet de définir les fonctions choisies dans l'étape de modélisation en fonction de l'application. La combinaison est la phase de regroupement des informations. La dernière étape consiste à prendre la décision sur le résultat de la combinaison.

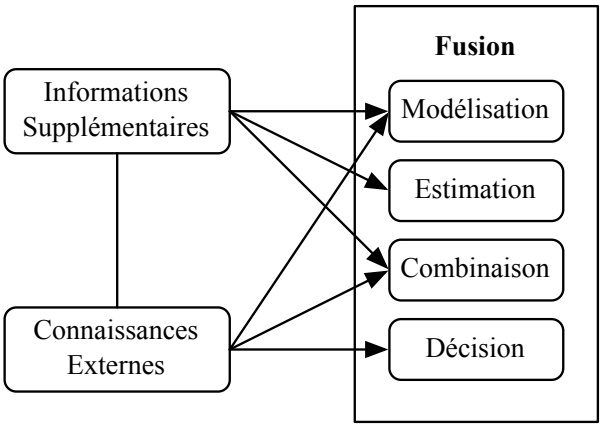



FIG. 3 – *Représentation du nœud de fusion*

### 2.1 Fonctions de croyance

Nous proposons ici l'utilisation de la théorie des fonctions de croyance pour la fusion multi-vues d'objets en vue d'une meilleure classification.

La théorie des fonctions de croyance est fondée sur la manipulation des fonctions de masse. Les fonctions de masse sont définies sur l'ensemble de toutes les disjonctions du cadre de discernement $\Theta = \{C_1, \ldots, C_N\}$ et à valeurs dans $[0, 1]$, où $C_q$ représente l'hypothèse "l'observation appartient à la classe $q$". Généralement, il est ajouté une condition de normalité, donnée par :

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

où $m(.)$ représente la fonction de masse. La première difficulté est donc de définir ces fonctions de masse selon le problème. À partir de ces fonctions de masse, d'autres fonctions de croyance peuvent être définies, telles que les fonctions de crédibilité, représentant l'intensité que toutes les sources croient en un élément, et telles que les fonctions de plausibilité représentant l'intensité avec laquelle on ne doute pas en un élément.

De façon à estimer les fonctions de masse à combiner, Appriou (2002) propose deux modèles répondant à trois axiomes qui impliquent la considération de $N$ fonctions de masse aux seuls éléments focaux possibles $\{C_q\}$, $\{C_q^c\}$ et $\Theta$. 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 :

$$\begin{cases} m_{iq}(C_q)(x) &= \dfrac{\alpha_{iq} R_i p(V_i(x)/C_q)}{1 + R_i p(V_i(x)/C_q)} \\ m_{iq}(C_q^c)(x) &= \dfrac{\alpha_{iq} R_i}{1 + R_i p(V_i(x)/C_q)} \\ m_{iq}(\Theta)(x) &= 1 - \alpha_{iq} \end{cases} \tag{2}$$

où $p$ est une probabilité, $R_i = \left(\max_{i,q} p(V_i(x)/C_q)\right)^{-1}$ est un facteur de normalisation, et $\alpha_{iq} \in [0, 1]$ est un coefficient d'affaiblissement permettant de tenir compte de la fiabilité de l'information fournie par la vue $i$ : $V_i(x)$ pour une classe $C_q$, que nous choisissons ici égale à 0.95. La difficulté de ce modèle est alors l'estimation des probabilités $p(V_i(x)/C_q)$. Dans le cas où la donnée $V_i(x)$ de la vue $i$ 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.

La combinaison des $N$ (nombre de vues dans cette étude) fonctions de masse que nous employons ici est la combinaison conjonctive non normalisée proposée par Smets (1990a) définie pour deux fonctions de masse $m_1$ et $m_2$ et pour tout $A \in 2^\Theta$ par :

$$m(A) = (m_1 \oplus m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \tag{3}$$

De nombreuses autres règles ont été proposées, un bref état de l'art ainsi que de nouvelles règles de combinaison gérant le conflit sont données par Martin et Osswald (2007).

Afin de conserver un maximum d'informations, il est préférable de rester à un niveau crédal (*i.e.* de manipuler des fonctions de masse) pendant l'étape de combinaison des informations pour prendre la décision sur les fonctions de masse issues de la combinaison. Si la décision prise par le maximum de crédibilité peut être trop pessimiste, la décision issue du maximum de plausibilité est bien souvent trop optimiste. Le maximum de la probabilité pignistique, introduite par Smets (1990b), reste le compromis le plus employé. La probabilité pignistique est donnée pour tout $X \in 2^\Theta$, avec $X \neq \emptyset$ par :

$$\mathrm{betP}(X) = \sum_{Y \in 2^\Theta, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m(Y)}{1 - m(\emptyset)}. \tag{4}$$

## 3 Expérimentations

Nous présentons dans cette partie une description des données générées utilisées pour la validation de l'approche de fusion crédibiliste pour la classification multi-vues. Nous donnons dans la section 3.2 les résultats obtenus pour différentes configurations : on utilise un nombre différent de vues et un niveau de confusion différent entre les objets considérés.

## 3.1 Données utilisées

Nous avons utilisé pour la génération de notre base de données trois formes de base : un cercle (forme invariante par rotation), un hexagone et un octogone vues à $0°$ avec $\sigma = 0$.

Ces formes ont été considérées sur $N = 36$ vues, de $0°$ à $350°$ avec un pas de $10°$, en ajoutant un bruit au niveau des bords des formes. Ce bruit est ajouté aléatoirement au niveau des bords en utilisant une gaussienne multipliée par la fonction de la forme du bruit utilisée, représentée dans la figure 4.

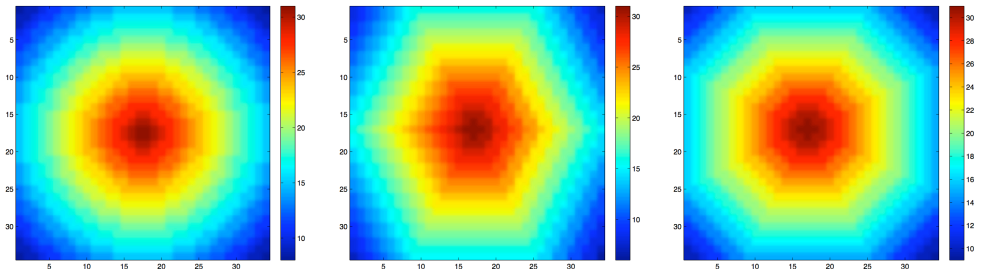


FIG. 4 – *Les fonctions utilisées pour bruiter les formes de base*

La base de données est constituée d'images binaires de taille $128 \times 128$ pixels, la forme de base se situe au centre de cette image et de taille $32 \times 32$ pixels. La figure 5 donne un exemple de quelques images de la base de données. Le tableau 1 donne quelques statistiques sur la base de données où la classe 1 désigne la classe des cercles, la classe 2 celle des hexagones et la classe 3 la classe des octogones.

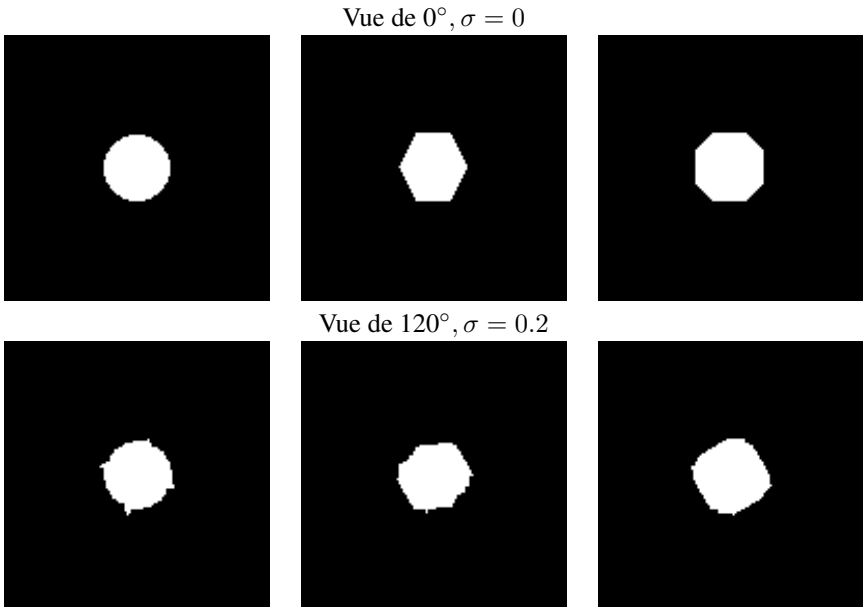



FIG. 5 – *Trois images de la base de données utilisée*

| | |
|---|---|
| Classes | {Cercle, Hexagone, Octogone} |
| Nombre de classe | 3 |
| Effectif | 1296 (432 pour chaque classe) |
| Vues | de $0°$ à $350°$ avec un pas de $10°$ |
| Certitudes | de 0.1 à 0.9 avec un pas de 0.1 |

TAB. 1 – *Quelques statistiques sur la base de données générées*

Chaque objet $o$ est caractérisé par un vecteur $V_o = (V_o^i, i = 1, 2, 3, 4)$ : chaque élément de ce vecteur représente le nombre de pixels qui entre en confusion avec les autres forme de base.

Nous avons considéré l'approche d'Appriou (2002) pour l'estimation des fonctions de masse à partir des paramètres calculés sur les objets. Ces fonctions sont estimées en utilisant une matrice de confusion sur les données d'apprentissage. Nous avons considéré deux approches pour le calcul de cette matrice de confusion : une première approche fondée sur les paramètres calculés directement sur les objets (vecteurs $V_o$) (La matrice de confusion est notée *MCnoSVM*) et une deuxième approche qui utilise une classification par SVM Vapnik (1998) sur la base d'apprentissage (matrice de confusion notée *MCSVM*). La base d'apprentissage est constituée de 648 images (216 images pour chaque classe) tirées aléatoirement en considérant toutes les certitudes.

## 3.2 Résultats

Nous donnons dans cette section les résultats obtenus pour la classification des objets en utilisant la fusion multi-vues. Pour cela, nous avons utilisé différentes valeurs pour le paramètre ($\sigma$) qui contrôle le bruit au bord des formes de base. Nous avons utilisé des valeurs entre 0 et 1 pour $\sigma$. Les résultats sont obtenus en utilisant les matrices de confusion normalisées présentées dans le tableau 2. Ces matrices ont été calculées en utilisant des bases d'apprentissage (de 648 images) tirées aléatoirement. Nous avons utilisé les SVM pour la classification en utilisant le logiciel *libSVM* développé par Chang et Lin (2001).

| $\sigma$ | *MCnoSVM* (%) | *MCSVM* (%) |
|---|---|---|
| 0.4 | $\begin{pmatrix} 81.81 & 12.50 & 5.68 \\ 13.19 & 79.27 & 7.54 \\ 12.10 & 11.05 & 76.85 \end{pmatrix}$ | $\begin{pmatrix} 71.76 & 25.93 & 2.31 \\ 25.00 & 72.22 & 2.78 \\ 0.93 & 0.46 & 98.61 \end{pmatrix}$ |
| 0.5 | $\begin{pmatrix} 77.35 & 16.32 & 6.34 \\ 21.86 & 69.01 & 9.12 \\ 9.98 & 17.11 & 72.91 \end{pmatrix}$ | $\begin{pmatrix} 65.74 & 31.02 & 3.24 \\ 32.87 & 63.43 & 3.70 \\ 3.24 & 1.85 & 94.91 \end{pmatrix}$ |
| 0.6 | $\begin{pmatrix} 72.61 & 17.79 & 9.60 \\ 11.26 & 75.11 & 13.62 \\ 16.03 & 15.37 & 68.60 \end{pmatrix}$ | $\begin{pmatrix} 61.11 & 31.02 & 7.87 \\ 34.72 & 60.19 & 5.09 \\ 3.70 & 4.17 & 92.13 \end{pmatrix}$ |
| 0.7 | $\begin{pmatrix} 84.32 & 10.65 & 5.04 \\ 5.75 & 90.72 & 3.53 \\ 7.60 & 16.33 & 76.06 \end{pmatrix}$ | $\begin{pmatrix} 59.26 & 33.33 & 7.41 \\ 34.72 & 56.02 & 9.26 \\ 6.48 & 4.63 & 88.89 \end{pmatrix}$ |
| 0.8 | $\begin{pmatrix} 91.00 & 2.47 & 6.52 \\ 6.17 & 78.76 & 15.07 \\ 10.24 & 13.11 & 76.65 \end{pmatrix}$ | $\begin{pmatrix} 59.72 & 31.02 & 9.26 \\ 30.56 & 55.56 & 13.89 \\ 12.50 & 8.80 & 78.70 \end{pmatrix}$ |
| 0.9 | $\begin{pmatrix} 88.12 & 2.26 & 9.62 \\ 10.49 & 83.60 & 5.91 \\ 12.91 & 30.60 & 56.49 \end{pmatrix}$ | $\begin{pmatrix} 60.19 & 28.70 & 11.11 \\ 34.26 & 55.09 & 10.65 \\ 11.57 & 9.72 & 78.70 \end{pmatrix}$ |

TAB. 2 – *Matrices de confusion calculées sur les données d'apprentissage et pour chaque valeur de* $\sigma$

Les matrices de confusion normalisées pour le cas de la classification avant l'étape de la fusion sont données dans le tableau 3. Nous remarquons que le taux de classification décroît en fonction de $\sigma$ : plus on a une superposition entre les classes, exprimée par l'utilisation de $\sigma$, plus le taux de classification est faible.

La figure 6 donne les résultats pour la fusion en utilisant une matrice de confusion calculée directement sur les données d'apprentissage (*MCnoSVM*) et une matrice de confusion calculée avec un classifieur de base (*MCSVM*) (SVM avec un noyau linéaire dans cette étude). Nous avons comparé les résultats par une approche classique de fusion par vote et par l'approche crédibiliste que nous avons décrit dans la section 2.1. Chaque courbe de la figure 6 représente la variation du taux de classification en fonction du nombre de vues. Notons que les vues sont tirées aléatoirement et que nous avons effectué plusieurs $10 \log_e(A_N^p)$ tirages aléatoires pour un nombre de vues égale à $p$. Nous n'avons pas effectué toutes les possibilités pour chaque nombre de vues, puisque le nombre du choix de vues possibles est égale à $A_N^p$ (où $N = 36$ est le nombre de vues maximal et $p$ le nombre de vues à choisir. Exemple : $A_{36}^{10} = 922.393.263.052.800$). Nous avons divisé la base de données en deux bases, une pour l'apprentissage du classifieur (et aussi pour le calcul des matrices de confusion utilisées pour l'estimation des fonctions de masse) et une base pour le test. Ces deux bases de données contiennent le même nombre de données avec le même nombre d'images pour chacune des trois classes.

Nous remarquons que pour les différentes valeurs de $\sigma$, on arrive à dépasser le taux de classification obtenu sans fusion (*cf.* tableau 3) à partir de l'utilisation de 2, 3 ou 4 vues. Nous remarquons aussi qu'on atteint les 100% de taux de bonne classification pour $\sigma = 0.4, 0.5, 0.6$ et $0.8$. Les différents résultats de classification pour $\sigma = 0.4, 0.5, 0.6, 0.7, 0.8$ et $0.9$ ne sont pas significativement différents.

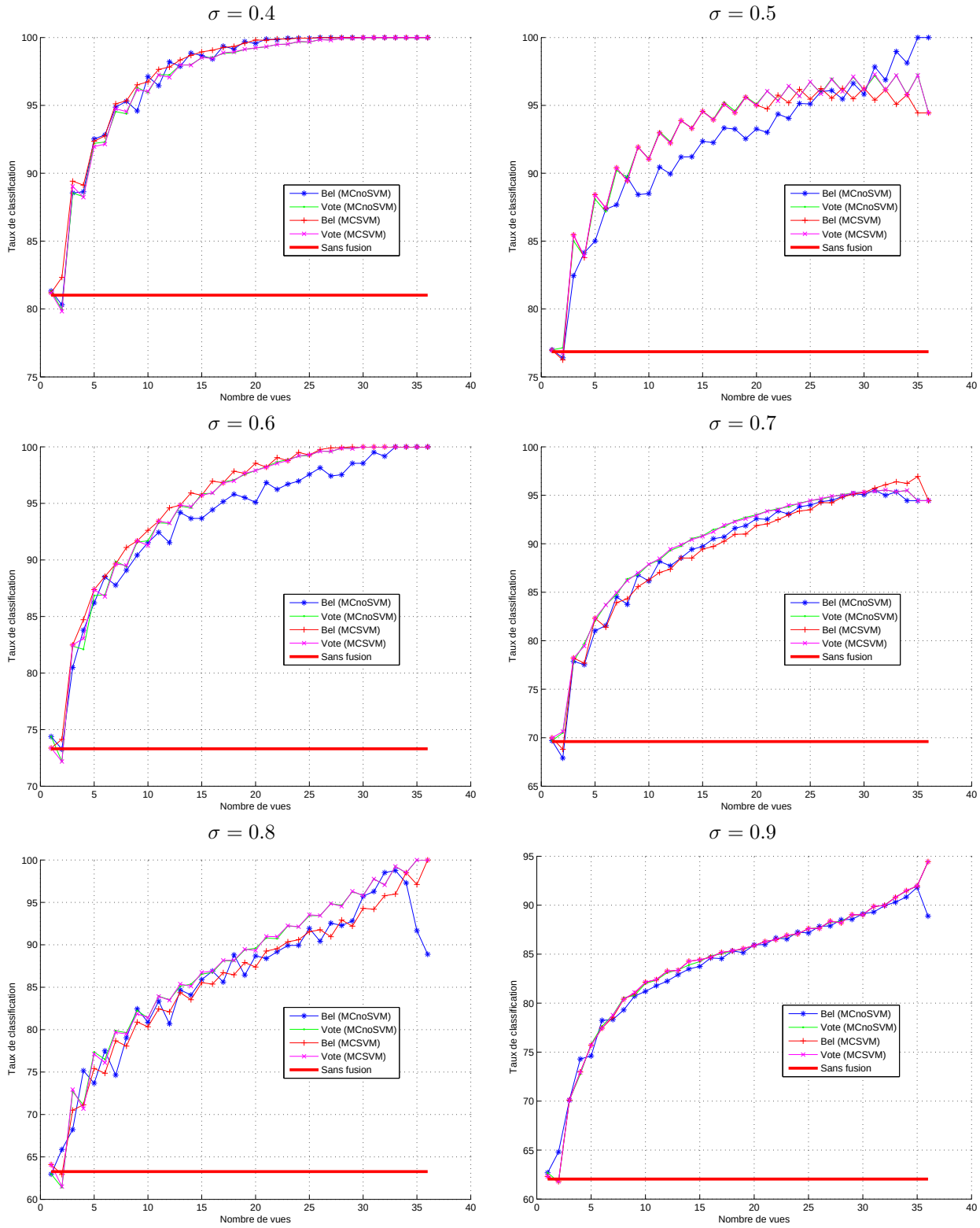



FIG. 6 – *Résultats de fusion multi-vues en utilisant deux matrices de confusion différentes pour les deux approches de fusion, vote et crédibiliste*

| $\sigma$ | Taux de classification (%) | Probabilités d'erreur (%) | Matrice de confusion (%) |
|---|---|---|---|
| 0.4 | 81.02±3.20% | (20.14 20.49 2.08) | $\begin{pmatrix} 75.00 & 25.00 & 0.00 \\ 27.31 & 71.76 & 0.93 \\ 3.24 & 0.46 & 96.30 \end{pmatrix}$ |
| 0.5 | 76.85±3.40% | (26.62 22.57 2.89) | $\begin{pmatrix} 62.50 & 31.48 & 6.02 \\ 29.17 & 70.83 & 0.00 \\ 2.31 & 0.46 & 97.22 \end{pmatrix}$ |
| 0.6 | 73.30±3.54% | (26.50 26.04 7.52) | $\begin{pmatrix} 66.20 & 27.78 & 6.02 \\ 31.94 & 63.43 & 4.63 \\ 6.48 & 3.24 & 90.28 \end{pmatrix}$ |
| 0.7 | 69.60±3.65% | (31.13 28.24 9.03) | $\begin{pmatrix} 54.63 & 37.04 & 8.33 \\ 28.24 & 64.35 & 7.41 \\ 5.56 & 4.63 & 89.81 \end{pmatrix}$ |
| 0.8 | 63.27±3.78% | (35.42 32.87 14.35) | $\begin{pmatrix} 52.78 & 31.48 & 15.74 \\ 38.43 & 54.17 & 7.41 \\ 8.80 & 8.33 & 82.87 \end{pmatrix}$ |
| 0.9 | 62.04±3.80% | (34.95 34.38 16.09) | $\begin{pmatrix} 47.22 & 42.59 & 10.19 \\ 27.78 & 59.72 & 12.50 \\ 6.48 & 14.35 & 79.17 \end{pmatrix}$ |

TAB. 3 – *Matrices de confusion normalisées avant la fusion multi-vues*

# 4 Conclusion

Nous avons employé dans cet article l'étude de la fusion pour la classification multi-vues, appliquée à des données générées modélisant la sortie d'un classifieur tentant de discriminer trois types d'objets. Le but de la fusion est ici d'améliorer le taux de bonne classification en augmentant le nombre de vues. Nous avons quantifier l'amélioration en fonction du nombre de vues ainsi que du niveau de bruit sur les résultats fournis par le classifieur. L'approche étudiée, fondée sur les fonctions de croyance a permis une amélioration significative des résultats pour la classifications multi-vues. En effet nous avons pu avoir des taux de classification qui dépassent largement le taux de classification obtenu en utilisant uniquement une seule vue. Nous avons pu voir qu'on obtient des taux de classification de 100% à partir de l'utilisation de quelques vues pour des valeurs de $\sigma$ assez petites (faible superposition entre les classes).

Les données générées ne modélisent que grossièrement les résultats possibles en sortie d'un classifieur. Nous essayerons d'appliquer cette approche sur des données réelles comme les images sonar pour la détection d'objets dans un milieu sous-marin (mines, épaves ...).

# Références


Appriou, A. (2002). *Décision et Reconnaissance des formes en signal*. Hermes Science Publication.

Aridgides, T., M. F. Fernandez, et G. J. Dobeck (2001). Side-scan sonar imagery fusion for sea mine detection and classification in very shallow water. In A. C. Dubey, J. F. Harvey, J. T. Broach, et V. George (Eds.), *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Volume 4394, pp. 1123–1134.

Chang, C. C. et C. J. Lin (2001). Libsvm : a library for support vector machines. *Software available at http ://www.csie.ntu.edu.tw/∼cjlin/libsvm/*.

Daniel, S. (1998). Fusion multisource appliquée à la reconnaissance d'objets dans le milieu sous-marin. *Thèse de l'Université de Rennes 1, Rennes, FRANCE*.

Dhibi, M., R. Courtis, et A. Martin (2008). Multi-segmentation of sonar images using belief function theory. In *Acoustics'08/ECUA08*, Paris, France.

Huang, G. et Y. Wang (2007). Gender classification based on fusion of multi-view gait sequences. In *ACCV (1)*, pp. 462–471.

Laanaya, H., A. Martin, D. Aboutajdine, et A. Khenchaf (2008). Classifier fusion for post-classification of textured images. *Information Fusion, 30 June-3 July*.

Martin, A. (2005a). Fusion de classifieurs pour la classification d'images sonar. *RNTI Extraction des connaissances : Etat et perspectives*, 259–268.

Martin, A. (Janvier 2005b). La fusion d'informations. *Polycopié de cours ENSIETA - Réf. : 1484, 117 pages*.

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

Milisavljevic, N., I. Bloch, et M. Acheroy (2008). *Multi-Sensor Data Fusion Based on Belief Functions and Possibility Theory : Close Range Antipersonnel Mine Detection and Remote Sensing Mined Area Reduction*, Chapter 4, pp. 392–418. Vienna, Austria : ARS I-Tech Education and Publishing.

Quidu, I. (2001). Classification multi-vues d'un objet immergé à partir d'images sonar et de son ombre portée sur le fond. *Thèse de l'Université de Bretagne Occidentale, Brest, FRANCE*.

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

Smets, P. (1990b). Constructing the pignistic probability function in a context of uncertainty. *Uncertainty in Artificial Intelligence 5*, 29–39.

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wesley and Sons.

## Summary

We present in this paper a belief fusion approach for pattern recognition of multi-view images. This approach is applied on generated data from three basic shapes (circle, hexagon and octagon). These data represent the object position after a classifier algorithm. The results show the interest of the use of information extracted on multiple views for the classification of the same object.