Actes des Ateliers d'EGC 2012

Atelier aIde à la Décision à tous les Etages (AIDE)

Site de l'atelier AIDE@EGC2012

Descriptif de l'atelier

Le terme « outils d'aide à la décision » est utilisé par différentes communautés qui ne communiquent pas toujours ensemble et qui y associent des acceptions différentes :

- En informatique décisionnelle, l'aide à la décision consiste à analyser des indicateurs en fonction d'axes d'analyses le plus souvent représentés dans un espace multidimensionnel. Ces données décisionnelles sont extraites généralement des sources de production pour être intégrées et synthétisées au sein d'un entrepôt de données (« datawarehouses »).
- En fouille de données (« datamining »), l'aide à la décision consiste en l'extraction de connaissances dans de grands volumes de données pour en extraire des informations pertinentes. La fouille de données repose sur des algorithmes permettant de déterminer des corrélations ou des prédictions et peuvent facilement se coupler avec un entrepôt de données.
- En simulation, l'aide à la décision peut recourir à la construction (parfois en collaboration avec le décideur et/ou en incluant les utilisateurs de manière participative) de modèles concernant le système en jeu pour en prédire ou en anticiper les évolutions sous différents scénarios de gestion. En fonction du domaine considéré (physique, météorologique, social, ...), ces simulations peuvent générer de grandes masses de données qui peuvent faire appel aux communautés précédemment citées pour en analyser le contenu.
- En systèmes interactifs d'aide à la décision, l'aide à la décision est associée à l'analyse multicritères. Une telle analyse vise à répondre à différents objectifs le plus souvent contradictoires afin d'aider à prendre une décision ou à évaluer plusieurs options dans des situations où aucune solution ne s'impose aux autres. Certaines méthodes reposent sur l'optimisation d'indicateurs ou sur une agrégation partielle, totale, locale ou itérative.

Cet atelier vise à créer un espace de rencontre, d'échange, de réflexion pour des chercheurs se positionnant selon les points de vue précédemment cités, et ce avec l'objectif double de pouvoir :

1. Prendre connaissance des différentes définitions partagées ou non, des concepts associés et faire le point sur les avancées dans ces différents domaines ;
2. Faire émerger des synergies entre ces différents domaines afin de permettre des collaborations futures pour proposer des solutions intégrées et innovantes dans le cadre de l'aide à la décision.

Comité de relecture

- Marie-Aude AUFAURE, Ecole Centrale Paris
- Olivier BARRETEAU, Cemagref, Montpellier
- Fadila BENTAYEB, Université Lumière Lyon 2
- Stéphane BONNEVAY, Université Claude Bernard Lyon 1
- Bernard ESPINASSE, Université Paul Cézanne Aix-Marseille 3
- Frédérick GARCIA, INRA, Toulouse
- Nouria HARBI, Université Lumière Lyon 2
- Christine LARGERON, Université Jean Monnet Saint Etienne
- Christophe LE PAGE, CIRAD-Green, Montpellier
- Sabine LOUDCHER, Université Lumière Lyon 2

- Patrick MARCEL, Université François Rabelais de Tours
- Elsa NEGRE, Université Paris Dauphine
- François PINET, Cemagref, Clermont-Ferrand
- Olivier TESTE, Université Toulouse 3 Paul Sabatier
- Ronan TOURNIER, Université Toulouse 1 Capitole

Organisateurs d'AIDE 2012

- Frédéric AMBLARD (IRIT, Université Toulouse 1 Capitole)
- Cécile FAVRE (ERIC, Université Lumière Lyon 2)
- Franck RAVAT (IRIT, Université Toulouse 1 Capitole)

Articles acceptés

- L'art de la fouille pour l'aide à la décision
  - Hanen Brahmi, Nadia Araour, Sadok Ben Yahia, "La fouille intelligente des règles d'association à partir des données décisionnelles".
  - Antoine Rolland, "Aide à la décision multicritère et apprentissage automatique pour la classification".
- Retours d'expériences sur l'aide à la décision
  - Igor Crévits, Alexis Tsoukiàs, "Opérationnalisation du Processus d'Aide à la Décision".
  - Lassaâd Mejri, Ahmed Maalel, Henda Hajjami Ben Ghezela, "Simulation de cas d'accidents pour l'aide à la décision. Application à la sécurité des systèmes de transport ferroviaires automatisés".
  - Marcel Guenoun, Joris Peignot, Adrien Peneranda, "L'utilisation des systèmes d'information décisionnels dans les collectivités territoriales: premier état des lieux et recherches en cours".
- De la fouille pour l'aide à la décision
  - Sofia Benbelkacem, Baghdad Atmani, Abdelhak Mansoul, "Planification guidée par Raisonnement à base de cas et Datamining : Remémoration des cas par Arbre de décision".
  - Bilal Idiri, Aldo Napoli, "Découverte de règles d'associations pour l'aide à la prévision des accidents maritimes".
- Les décideurs dans l'aide à la décision
  - Aziza Sabri, Laila Kjiri, "Une démarche d'analyse à base de patrons pour la découverte des besoins métier d'un Système d'Information Décisionnel".
  - Rym Khemiri, Fadila Bentayeb, Omar Boussaid, "Recommandation interactive de requêtes décisionnelles".

La fouille intelligente des règles d'association à partir des données décisionnelles

Hanen Brahmi, Nadia Araour, Sadok Ben Yahia

Faculté des Sciences de Tunis.
Département des Sciences de l'Informatique.
Campus Universitaire 1060.
{hanenbrahmi, nadia.araour}@gmail.com
sadok.benyahia@fst.rnu.tn

Résumé. Les systèmes décisionnels sont dédiés au management de l'entreprise pour l'aider à la prise de décision. Dans ce contexte, la fouille des données est une tâche complexe et difficile qui exige de gros efforts à déployer pour garantir une analyse pertinente. Parallèlement, les ontologies sont à l'heure actuelle au coeur des travaux menés en ingénierie des connaissances. Elles ont fait preuve de leur capacité à résoudre des ambiguïtés sémantiques et syntaxiques. Par conséquent, plusieurs travaux de recherche ont recours aux ontologies dans le processus global d'exploration des données décisionnelles. L'objectif de cet article est de présenter et d'analyser l'état de l'art des travaux menés dans ce cadre. En particulier, il présente les ontologies ainsi que leur utilisation pour la fouille des règles d'association.
Mots clés. Fouille de données, Entrepôt de données, Ontologies, Règles d'association.

1 Introduction

Les systèmes d'information décisionnels constituent un domaine d'expertise de plus en plus répandue dans les entreprises soucieuses d'extraire l'information pertinente et cachée dans leurs systèmes opérationnels (Kimball, 2005). Ces systèmes sont dédiés au pilotage des activités de l'entreprise. De plus, ils constituent une synthèse des informations opérationnelles, internes ou externes, choisies pour leur pertinence et leur transversalité fonctionnelles. Les systèmes décisionnels sont basés sur des structures particulières de stockage volumineux appelées *entrepôt de données*.

Les entrepôts de données sont construits par extraction de données à partir de sources d'information et par intégration de ces données dans un espace de stockage commun à tous les utilisateurs décisionnels (Kimball, 2005)L'entrepôt de données est important pour plusieurs applications, spécialement pour celles évoluant dans des environnements à grande échelle et faisant appel à des sources d'information distribuées. En plus de sa vocation de stockage, un entrepôt vise aussi l'exploitation des données dans un processus d'analyse en ligne OLAP (*On Line Analysis Processing*) et d'aide à la décision.

La fouille de données vise à développer des méthodes et des techniques aidant à découvrir des tendances particulières, éventuellement non soupçonnées, à partir de l'analyse de données

témoignant de l'activité d'une organisation ou d'un système. Cet maillon ne peut être dissocié de la chaine de l'Extraction de Connaissances à partir des Données (ECD). L'avantage apporté aux décideurs repose sur la confrontation des informations extraites avec la connaissance tacite ou explicite des experts du domaine. La réussite de la démarche de fouille de données repose en particulier sur la première phase du processus dans laquelle l'analyste a pour objectif d'identifier et d'expliciter à travers le dialogue avec les experts du domaine, d'une part le problème à résoudre et d'autre part les connaissances expertes sur le domaine.

Parallèlement, les ontologies sont à l'heure actuelle au coeur des travaux menés en ingénierie des connaissances. Ces travaux visent l'établissement de représentations à travers lesquelles les machines puissent manipuler la sémantique des informations. Les ontologies apparaissent comme des composants logiciels s'insérant dans les systèmes d'information en leur apportant une dimension sémantique qui leur faisait défaut jusqu'ici. Le champ d'application des ontologies ne cesse de s'élargir et couvre les systèmes d'aide à la décision, les systèmes de résolution de problèmes ou les systèmes de gestion de connaissances. Plus précisément, les ontologies ont donné une grande satisfaction pour résoudre les conflits sémantiques et structurels.

Plus généralement, l'intégration des ontologies dans l'ECD ainsi que dans les entrepôts de données constitue un problème essentiel et un défi pour la recherche actuelle dans ces domaines. C'est ce que l'on appelle souvent le problème de "la fouille intelligente". Il s'agit à la fois de l'expression, de la structuration, de la formalisation des connaissances expertes appelées également connaissances à priori et de leur confrontation aux connaissances extraites de l'analyse des données. Par exemple, l'intégration des ontologies dans la fouille des entrepôts de données permet une analyse en ligne plus élaborée dépassant la simple exploration des données. La robustesse de la fouille de données, la maniabilité de la structuration multidimensionnelle et la richesse des ontologies peuvent apporter des amélioration aux capacités de l'OLAP afin d'extraire des connaissances syntaxiquement correct et sémantiquement raisonnable.

Le reste de l'article est organisé comme suit. Dans la section 2, nous présentons les ontologies et ses spécificités. Dans la section 3 nous classons les approches intégrant les ontologies dans le processus d'ECD. Dans la section 4, nous discutons les propositions qui couplent entre les entrepôts de données et les ontologies. La conclusion et les travaux futurs font l'objet de la section 5.

2 Les ontologies

2.1 Définitions

Définie initialement par Gruber (1993), une ontologie est *une spécification explicite d'une conceptualisation*. La conceptualisation réfère à un modèle abstrait qui identifie les concepts appropriés de certains phénomènes dans le monde. Explicite signifie que le type de concepts utilisés et les contraintes sur leur utilisation doivent être explicitement définis. Cette définition a été étendue par Guarino (1995). L'auteur a souligné le caractère formel qui correspond au fait qu'une ontologie doit être compréhensible par la machine. Cette dernière doit être capable d'interpréter la sémantique de l'information fournie. Ainsi, une ontologie est *une représentation explicite de la sémantique d'un domaine*. Dans le même contexte, Fankam et al. (2009) caractérisent une ontologie comme *une représentation formelle, référençable et consensuelle de l'ensemble des concepts partagés d'un domaine*. Plus précisément, une ontologie est *for-*

melle puisqu'elle permet des raisonnements automatiques ayant pour objet d'inférer de nouveaux faits et d'effectuer des vérifications de consistance. En outre, elle est *consensuelle* c'est à dire admise par l'ensemble des membres et des systèmes d'une communauté. De plus, chaque entité ou relation décrite dans l'ontologie peut être directement *référencée* par un symbole, à partir de n'importe quel contexte.

2.2 Composantes d'une ontologie

Gomez Perez (1999) affirme que les connaissances traduites par une ontologie sont à véhiculer à l'aide de ses composantes : concepts, relations, fonctions, axiomes et instances.

1. Les *concepts ou les classes* correspondent aux abstractions pertinentes du domaine, retenues en fonction des objectifs qu'on se donne et de l'application envisagée pour l'ontologie. Ces concepts peuvent être classés selon plusieurs dimensions telles que le niveau d'abstraction (concret ou abstrait), l'atomicité (élémentaire ou composée) et le niveau de réalité (réel ou fictif).
2. Les *relations* traduisent les associations pertinentes qui existent entre les concepts. Ces relations incluent les associations de généralisation/spécialisation (sous-classe-de), d'agrégation ou de composition (partie-de), etc. En effet, elles permettent d'apercevoir la structuration et l'interrelation des concepts, les uns par rapport aux autres.
3. Les *fonctions* sont les cas particuliers des relations. Dans lesquelles, un élément de la relation est définit en fonction des éléments précédents.
4. Les *axiomes* sont des expressions qui sont toujours vraies. Leur inclusion dans une ontologie peut avoir plusieurs objectifs : (i) définir la signification des composants ; (ii) définir des restrictions sur la valeur des attributs ; (iii) définir les arguments d'une relation et ; (iv) vérifier la validité des informations spécifiées ou en déduire de nouvelles.
5. Les *instances* constituent la définition extensionnelle de l'ontologie. Ces objets véhiculent les connaissances (statiques, factuelles) à propos du domaine du problème.

2.3 Typologie des ontologies

Les types d'ontologies mis au point sont très divers. Cette sous-section n'a pas l'ambition de fournir une typologie exhaustive des ontologies telle que celles de (Guarino, 1995; Tapucu et al., 2009). Elle présente néanmoins les types d'ontologies les plus couramment utilisés. D'une façon générale, on identifie les catégories suivantes : les ontologies de haut niveaux et les ontologies spécialisées.

D'une part, qu'elles soient appelées "modèles d'ontologies", "ontologie générique" ou "ontologie de représentation", *les ontologies de haut niveaux* décrivent des concepts très généraux ou des connaissances de sens commun telles que l'espace, le temps, l'événement, l'action, etc. Ces concepts sont indépendants d'un problème ou d'un domaine particulier. Ce type fournit des notions générales auxquelles tous les termes des ontologies existantes doivent être reliés. Par conséquent, elles sont réutilisables d'un domaine à l'autre.

D'autre part, *les ontologies spécifiques* sont des ontologies qui spécialisent un sous-ensemble d'ontologies génériques en un domaine ou un sous-domaine. Selon Tapucu et al. (2009), elles peuvent être de domaine, d'application, de méthodes ou de tâches partagées mais cela dépend de leur usage :

- Les *ontologies floues* : la connaissance floue joue un rôle important dans beaucoup de domaines qui font face à une énorme quantité de connaissances imprécises et vagues. Pour manipuler l'incertitude des connaissances, une solution possible est d'incorporer la théorie floue dans l'ontologie. Les ontologies floues contiennent des concepts flous et des appartenances floues.
- Les *ontologies d'application* définissent la structure des connaissances nécessaire à la réalisation d'une tâche particulière.
- Les *ontologies de domaine* spécifie le point de vue sur un domaine cible. Les concepts d'une ontologie sont liés à un domaine de connaissances générique comme l'informatique, le commerce, la loi, etc. Une ontologie de domaine est constituée de : (i) un vocabulaire contrôlé qui représente un ensemble de termes décrits par des énoncés logiques et dont les relations sont spécifiées au moyen de règles ; (ii) une hiérarchie de concepts afin de classer les entités d'un domaine ; (iii) une théorie du domaine qui permet de vérifier la consistance du contenu de l'ontologie ; et (iv) un moyen pour partager et réutiliser la connaissance.
- Les *ontologies de méthodes* décrivent le processus de raisonnement d'une façon indépendante d'un domaine. Elles nous fournissent les concepts et les relations pour spécifier un processus de raisonnement.
- Les *ontologies de tâches* décrivent un vocabulaire en relation avec une tâche ou une activité générique comme le diagnostic ou la vente. En effet, ces ontologies identifient un lexique systématisé de termes utilisés pour résoudre les problèmes associés à des tâches particulières.

Pour terminer, les ontologies de haut niveaux, de domaine, d'application et floues saisissent les connaissances statiques indépendamment de la façon dont on résout les problèmes alors que les ontologies de tâches et de méthode sont axées sur les connaissances visant à résoudre des problèmes.

3 Les ontologies dans le processus d'ECD

Une panoplie de travaux de recherche ont proposé d'améliorer le processus d'ECD en employant les ontologies. Dans ce contexte, Euler et Scholz (2004) utilisent une ontologie pour définir une représentation conceptuelle pour la phase de compréhension des données. Ils proposent un niveau d'abstraction conceptuel au dessus du niveau logique d'une base de données afin de faciliter le processus d'ECD. Pour ce faire, les auteurs utilisent un utilitaire graphique permettant la modélisation de la connaissance experte du domaine. Ainsi, une séquence d'opérations de pré-traitement a été définie sur la base de la description conceptuelle des données. Seulement, Brisson et Collard (2008) définissent un outil, appelé *Intelligent Discovery Assistant*, basé sur une ontologie de fouille de données pour assister l'utilisateur dans le processus d'ECD. Cet outil a comme rôle d'énumérer les différents modèles valides de fouille de données et d'aider dans l'exploration et le choix de l'espace des modèles possibles et valides.

3.1 Les ontologies et la fouille des règles d'association

Nous présentons les propositions qui se sont focalisées sur le problème de l'intégration des ontologies dans le processus d'extraction des règles d'association en vue d'obtenir des règles informatives et sémantiques. Nous soulignons que l'extraction de règles associatives est l'un des principaux problèmes de la fouille de données. Il a été introduit par Agrawal et al. (1993) dans le but d'analyser les bases de données de transactions de ventes. Il a pour but de découvrir

des relations significatives entre les données de la base de données. Une règle associative est une relation d'implication X $\Rightarrow$ Y entre deux ensemble d'articles X et Y. X est appelé prémisse, alors que Y est la conclusion de la règle.

Dans ce contexte, Zeman et al. (2009) utilisent une ontologie de domaine dans la préparation des données pour une méthode de récupération des connaissances appelées *GUHA*. Cette méthode repose sur des fondements théoriques basés sur des calculs d'observations et des statistiques. À cet effet, les auteurs ont implémenté l'application Ferda (ou Ferda DataMiner) [1]. Dans cette dernière, nous identifions deux moyens d'utilisation des ontologies : (i) la construction d'une catégorisation d'attributs la plus adéquate et ; (ii) l'identification et l'exploitation des attributs liés sémantiquement.

(Hou et al., 2005; Bellandi et al., 2008) utilisent une ontologie pendant la phase de prétraitement où les données sont levées aux concepts les plus généralisés. Ensuite, le processus de l'extraction des règles est exécuté en se basant sur un algorithme standard. Les auteurs considèrent que les règles obtenues peuvent être plus faciles à interpréter, du fait qu'elles contiennent des concepts de haut niveaux qui représentent des renseignements plus riches que les termes spécifiques dans la base de données.

D'autres approches, comme ExCIS (Brisson et Collard, 2009), utilisent la connaissance de domaine dans les deux phases de pré-traitement et de post-traitement des données. Dans la phase de pré-traitement, les auteurs emploient une ontologie pour guider la construction des bases de données spécifiques pour des tâches particulières de fouille. L'étape suivante est l'application d'un algorithme de fouille de données standard qui extrait des motifs de ces data sets. Finalement, dans la phase de post-traitement, les règles générées peuvent être interprétées, filtrées et/ou généralisés selon une ontologie.

Dans (Antunes, 2007), l'ontologie permet la définition d'un domaine particulier de contraintes qui peuvent être divisées en deux catégories, à savoir des *contraintes d'abstraction* et des *contraintes d'élagage*. Les premières sont utilisées pour définir la généralisation de certains items. Les dernières sont utilisées pour exclure des items de l'analyse. Les deux types de contraintes sont définis sur la base des conditions exprimées dans l'ontologie.

Dans de nombreuses applications du monde réel, la structure taxonomique est floue. Par conséquent, Escovar et al. (2006) proposent l'algorithme XSSDM qui utilise une ontologie floue pour représenter les relations de similarité sémantique entre les données. Cet algorithme considère une nouvelle mesure appelée similarité minimale (*Minsim*). Les associations floues peuvent être exprimées dans les règles extraites dans le cas où le degré de similarité est supérieur ou égal à *Minsim*. Dans un même contexte, Miani et al. (2009) proposent un algorithme, appelé NARFO, permettant l'extraction des règles généralisées en se basant sur une ontologie floue. Cette dernière sert comme une base de connaissances pour soutenir le processus de découverte des règles.

Marinica et Guillet (2010) proposent l'intégration des connaissances de l'utilisateur dans la découverte de règles d'association afin de réaliser une phase de post-traitement plus efficace. Plus précisément, l'approche ARIPSO intègre deux représentations des connaissances de l'utilisateur complémentaires : d'une part des ontologies de domaine associées aux attributs de la base de données ; et d'autre part des schémas de règles généralisant les impressions générales afin de sélectionner les règles intéressantes.

1. http ://ferda.sourceforge.net

Pour terminer, Mansingh et al. (2011) proposent une approche qui combine les connaissances représentées dans une ontologie d'application avec la mesure objective de fiabilité pour créer des partitions significatives dans l'ensemble des règles d'association extraites. Les auteurs dégagent cinq partitions qui intéressent les experts du domaine à savoir : (i) Des nouvelles règles avec une importance élevée qui peuvent conduire à la formation de nouvelles croyances et des ajouts à l'ontologie ; (ii) Des règles connues avec une importance élevée qui confirment les croyances antérieures de l'expert du domaine ; (iii) Des règles connues avec une importance faible qui représentent les aspects de la connaissance de l'expert et qui ne sont pas soutenus par les données ; (iv) Des règles manquantes qui n'ont pas été extraites par induction de règles d'association. La présence de ces règles s'explique par des valeurs de seuils erronées et ; (v) Des règles contradictoires permettant l'expression de la négation.

3.2 Discussion

TAB. 1 – *Comparaison des approches.*

Approches	**Phase d'ECD**				**Type d'ontologie**			**Type de règles**			**Mesures**		
	Pré-traitement	Traitement	Post-traitement	Evaluation	Domaine	Application	Floue	Généralisées	sous contrainte	Floue	Support	Confiance	Autres
Zeman et al. (2009)	×				×			×			×	×	
Hou et al. (2005)	×			×	×			×			×	×	
Bellandi et al. (2008)	×			×	×			×			×	×	
Ferraz et Garcia (2008)	×		×		×			×			×	×	×
Brisson et Collard (2009)	×		×	×	×			×			×	×	
Escovar et al. (2006)	×	×					×	×		×	×	×	×
Miani et al. (2009)	×	×					×	×		×	×	×	
Marinica et Guillet (2010)			×		×			×			×	×	×
Mansingh et al. (2011)			×	×		×		×			×	×	×
Antunes (2007)		×			×			×	×		×	×	

Dans le Tableau 1, nous évaluons les approches qui intègrent les ontologies dans le processus d'extraction des règles d'association. D'une manière générale, les approches se sont focalisées sur la fouille des : (i) *règles d'association généralisées* intégrant les connaissances sous la forme d'une taxonomie des attributs (Srikant et Agrawal, 1994) ; (ii) *règles d'association floues* mesurant un lien entre deux classes en s'appuyant sur les valeurs floues (entre 0 et 1) de leurs caractéristiques (Cuxac et al., 2005) et ; *règles d'association sous contraintes* permettant de définir une sémantique sur les règles recherchées en intégrant des contraintes liées au domaine (Boulicaut et Jeudy, 2010).
Nous distinguons quatre grands groupes d'approches :

1. Le premier groupe s'inscrit dans l'étape de pré-traitement des données (ou préparation des données). Au cours de cette étape, les ontologies peuvent faciliter l'intégration des données hétérogènes et guider la sélection des données pertinentes pour être exploitées. Par conséquent, l'utilisation d'une ontologie peut contribuer à l'identification des groupes d'attributs ou de valeurs en exploitant les relations sémantiques.
2. Motivé par la réduction et le ciblage de l'espace de recherche, le deuxième groupe d'approches concerne la phase du traitement. Les ontologies permettent la spécification des contraintes en vue de guider les algorithmes d'extraction des règles d'association telles

que des contraintes d'agrégats, des contraintes syntaxiques, des contraintes de combinaisons, etc. La fouille des règles sous contrainte permet de découvrir les relations de régularités, d'exceptions et des contrastes qui associent les données.

3. Le troisième groupe s'inscrit dans une phase d'évaluation. Dans cette phase, les règles découvertes sont interprétées selon les concepts de l'ontologie et confrontées aux connaissances à priori. L'intérêt de cette tendance est de déterminer l'intérêt des règles en vérifiant si elles confirment, contredisent ou révèlent de nouvelles connaissances quand elles sont comparées à la connaissance disponible dans l'ontologie. Ceci permet aux experts du domaine de visualiser et de valider les unités d'extraction.
4. D'autre approches emploient les ontologies dans la phase de post-traitement. En effet, l'usage du modèle des règles d'association en fouille de données est limité par la quantité prohibitive de règles générées. Par conséquent, il requiert la mise en place d'un post-traitement efficace et adapté à la fois aux préférences du décideur et à la structure des données étudiées. De plus, certaines approches dans ce groupe réduisent le nombre de règles à l'aide des mesures d'intérêt autres que la confiance et le support telles que *l'amélioration* et la fiabilité. De nombreux travaux de synthèse récapitulent les mesures objectives développées et comparent leurs définitions et leurs propriétés (Zhang et al., 2009).

4 Les ontologies et les entrepôts de données

Stoffel et al. (1997) pourraient être les premiers qui ont attaqué le problème du couplage entre les ontologies et les entrepôts de données. Ils ont présenté un schéma sémantique d'indexation qui intègre l'ontologie et les données relationnelles pour la construction d'un d'entrepôt de données médicale.

Priebe et Pernul (2003), ont exploité les problèmes liés à l'intégration des ontologies dans la découverte des connaissances à partir des entrepôts de données. En particulier, ils ont utilisé une approche basée sur l'ontologie pour fabriquer un portail intelligent de connaissances qui intègre le système OLAP et la recherche des informations structurées et non structurées d'une entreprise. L'ontologie fournit les informations nécessaires pour mapper entre les différents modèles de méta-données pour que les composants du système puissent communiquer intelligemment.

Nabli et al. (2009) ont proposé une architecture de construction d'ontologie décisionnelle à base de concepts multidimensionnels. Cette architecture se prête à construire l'ontologie d'une manière incrémentale et progressive et se base sur quatre phases : (i) l'extraction des concepts multidimensionnels à partir de sources de données hétérogènes ; (ii) la comparaison des concepts extraits de la phase précédente avec celle de l'ontologie afin de déduire les relations sémantiques ; (iii) l'alimentation de l'ontologie par les concepts et les relations extraits et ; (iv) l'optimisation des relations multidimensionnelles.

4.1 Les ontologies et la fouille des règles multidimensionnelles

Les travaux de recherche sur la fouille des entrepôts de données sont principalement concentrée sur l'extraction des règles d'association multidimensionnelles. Ce type de règles met en jeu plusieurs attributs provenant de multiples dimensions. Comme exemple, Plantevit et al. (2010) ont proposé une méthode originale d'extraction des motifs séquentiels multidimensionnels prenant en compte les hiérarchies des dimensions. L'originalité de l'approche réside dans

l'idée que les auteurs ne fixent pas un unique niveau de hiérarchie mais que les motifs séquentiels extraits sont automatiquement associés aux niveaux les plus adéquats. Le lecteur intéressé par un panorama des travaux consacrés à la découverte des règles d'association à partir des entrepôts de données peut se référer à (Brahmi et al., 2010). Récemment, des travaux s'intéressant à la fouille des règles d'association multidimensionnelles en présence d'ontologies sont en train d'émerger. Principalement, nous repérons quatre propositions qui s'inscrivent dans ce cadre :

- **Fouille évolutive des règles d'association** : Tseng et al. (2007) ont étudié le problème de l'intégration des ontologies dans le processus d'extraction des règles d'association multidimensionnelles à partir des entrepôts de données. Les auteurs révisent ce problème à partir d'un point de vue unifié en prenant en compte des ontologies qui couvrent des relations de *classification* et de *composition*. Dans ce contexte, les auteurs ont introduit deux algorithmes appelés AROC et AROS. Ces derniers sont dérivés, respectivement, des algorithmes CUMULATE et STRATIFY (Srikant et Agrawal, 1994). En effet, les méthodes proposées permettent de découvrir les associations qui peuvent s'étendrent entre les différents niveaux hiérarchiques des relations de classification et de composition dans l'ontologie.
- **Maintenance avec des informations ontologiques** : Au fil du temps, des nouvelles transactions peuvent continuellement être ajoutées à la base de données multidimensionnelle. Parallèlement, les composantes de l'ontologie évoluent. Par conséquent, les analystes ont besoin d'ajuster à plusieurs reprises les contraintes liées à la confiance et au support pour découvrir des règles informatives qui reflètent la réalité. Dans un tel environnement, comment découvrir efficacement les règles d'association est devenue un problème crucial. Comme solution, Tseng et Lin (2009) ont proposé un algorithme, appelé MFIO, permettant la maintenance des motifs fréquents en tenant compte de tous les facteurs en évolution : (i) mise à jour des transactions dans la base de données ; (ii) l'évolution de l'ontologie ; et (iii) le raffinement du support minimum. En résumé, l'algorithme se base sur quatre phases : (1) génération des k-itemsets candidats ; (2) différenciation entre les itemsets affectés par la mise à jour de ceux non affectés ; (3) parcours de la nouvelle base avec la nouvelle ontologie ; et (4) traitement des ensembles concernés par la maintenance afin de réduire le coût de calcul du support.
- **Fouille basée sur des requêtes intelligentes** : Wu et al. (2011) ont implémenté un prototype d'un système de fouille de règles d'association multidimensionnelles. À cet effet, les auteurs ont construit quatre catégories d'ontologies liées au domaine d'analyse du panier de la ménagère : (1) *l'ontologie du schéma* : décrit la structure du schéma de l'entrepôt en précisant les caractéristiques des dimensions (*e.g.*, dimension dominent, dimension optionnel), les relations qui existent entre les attributs ainsi que les types des mesures (*e.g.*, additives, semi-additives, non-additives) ; (2) *l'ontologie de contraintes* : décrit les contraintes entre les attributs (dépendance fonctionnelle, transition, etc) ; (3) *l'ontologie de domaine* : recueille le domaine lié à l'entrepôt et les connaissances d'experts et ; (4) *l'ontologie des préférences des utilisateurs* : intègre les modèles communs de fouille. En s'appuyant sur ces ontologies, le système offre à l'utilisateur la possibilité de guider le processus d'extraction des règles d'association afin d'éviter l'obtention de résultats inutiles.

4.2 Discussion

Dans le Tableau 2, nous évaluons les approches qui intègrent les ontologies dans l'extraction des règles à partir des entrepôts de données. D'abord nous rappelons que les premières tentatives d'extension des outils OLAP pour la fouille de données remontent à 1997 avec les

travaux de (Han, 1997). Ces travaux ont abouti à la création du système DBMiner. Ce dernier est doté d'outils d'exploration graphique et de visualisation spatiale des cubes de données. Han définit la notion de l'*OLAP Mining* comme un mécanisme qui intègre des tâches de fouille de données dans des requêtes décisionnelles. Ce mécanisme peut s'appliquer à différents niveaux de granularité des données et à différentes parties d'un entrepôt de données. L'OLAP Mining permet un processus d'analyse où les techniques de fouille sont utilisées, au même titre que les opérations OLAP, pour extraire des connaissances.
Les approches (Tseng et al., 2007; Tseng et Lin, 2009; Wu et al., 2011) proposent des solutions à quelques insuffisances liées à l'OLAP Mining. Nous identifions :

- **Le manque de représentations sémantiques des données** : Les entrepôts de données sont principalement basés sur un modèle de données relationnel organisé sous la forme d'un schéma en étoile ou d'un schéma en flocon de neige (Kimball, 2005). Malheureusement, ces modèles ne peuvent pas refléter les relations sémantiques qui existent entre les données. Par exemple, le schéma en étoile décrit les relations structurelles entre la table de faits et les tables des dimensions. Il fournit des informations essentielles pour exécuter des opérations OLAP typiques : ROLL-UP, DRILL-DOWN, SLICE, DICE, etc. De plus les contraintes sur les mesures ne sont pas exprimées. Par exemple, le schéma en étoile ne souligne pas la différence entre les mesures additives (quantité, montant) et celles semi-additives (le solde d'un compte ne peut pas être groupé avec la dimension temps).
- **Des difficultés dans la compréhension des intentions des utilisateurs** : Le processus de fouille de données est plus ou moins un processus subjectif qui dépend de l'intention de l'individu. En effet, cette intention est représentée par les paramètres fixés lors du lancement des requêtes. Ces dernières peuvent être un atout important pour partager des information riches. Par exemple, les résultats peuvent être analysés afin de répondre aux futur recommandations des utilisateurs. Cependant, les systèmes d'OLAP Mining actuels offrent peu de mécanismes permettant de capturer et d'analyser les anciennes requêtes de fouille lancées par les utilisateurs.
- **Le manque de soutien des utilisateurs** : Les utilisateurs non experts ou débutants sont confrontés à des difficultés majeurs en vue de répondre aux questions suivants : (i) Comment choisir/appliquer l'outil d'extraction approprié ? (ii) Comment sélectionner le sous-ensemble de données approprié ? (iii) Comment formuler une requête appropriée ?
- **Le manque d'une mise à jour active et efficace des connaissances** : Le monde réel n'est pas statique. Habituellement, nous sommes confrontés à un environnement dynamique. Cet aspect rend le processus de fouille interactif, itératif et évolutif. Par conséquent, comment mettre à jour efficacement les connaissances découvertes devient un défi majeur.

Pour résoudre ces problèmes, les trois approches emploient les ontologies afin de profiter de leur richesse sémantique. Wu et al. (2011) ont recours à la structure relationnelle ROLAP (Relational OLAP) pour employer le langage de requêtes SQL afin d'extraire les données nécessaires aux calculs et à la construction des règles d'association. Autrement, (Tseng et al., 2007; Tseng et Lin, 2009) ont adapté l'algorithme standard APRIORI afin d'établir l'interaction nécessaire avec la structure multidimensionnelle des données. En outre, nous remarquons que les propositions se limitent à des relations de compositions et de classifications figurant dans un même domaine d'application : analyse du panier de la ménagère. Enfin, les approches

permettent d'extraire des règles appartenant à un seul niveau de granularité (Wu et al., 2011) ou à plusieurs niveaux (Tseng et al., 2007; Tseng et Lin, 2009) qui s'étendent à des concepts plus généralisés de l'entrepôt de données. Plus précisément, le processus de fouille est guidé en fixant les contraintes en relation avec les connaissances définies dans l'ontologie.

TAB. 2 – *Comparaison des approches .*

Approches	**Relations ontologiques**			**Type de stockage**		**Niveau d'abstraction**		**Intéraction avec l'utilisateur**		**Domaine d'application**	
	Classification	Composition	Autres	ROLAP	MOLAP	Unique	Multiple	Oui	Non	Général	Spécifique
Tseng et al. (2007)	×	×		×			×		×		×
Tseng et Lin (2009)	×	×		×			×		×		×
Wu et al. (2011)	×	×		×		×		×			×

5 Conclusion et perspectives

L'analyse en ligne OLAP, les ontologies et la fouille de données sont trois champs de recherche qui ont connu depuis quelques années, des évolutions parallèles et indépendantes. L'études menée dans ce papier montre l'importance et l'intérêt de l'association entre ces trois domaines scientifiques. D'où, comme un futur travail s'inscrivant dans le cadre de la fouille intelligente des entrepôts de données, nous proposons : (i) d'améliorer l'efficacité de l'extraction en s'appuyant sur des représentations condensées des connaissances extraites (fermés, fermés non-dérivables, etc). L'utilisation de formes condensées peut permettre des élagages supplémentaires et ainsi améliorer la robustesse de l'extraction ; (ii) de gérer des ontologies qui renferment des relations autres que celles de compositions et de classifications afin de s'adapter aux besoins de l'utilisateur ; (iii) d'employer les méta-règles inter-dimensionnelles afin de piloter le processus de recherche des règles et ; (iv) d'opérer directement sur des cubes de données MOLAP (Multidimensional OLAP) afin d'utiliser plutôt le langage de requêtes multidimensionnelles MDX (Multi-Dimensional eXpression).

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data (SIGMOD 1993), Washington, USA*, pp. 207–216.

Antunes, C. (2007). ONTO4AR : A Framework for Mining Association Rules. In *Proceedings of the International Workshop on Constraint-Based Mining and Learning (CMILE Ű PKDD), Warsaw, Poland*, pp. 37–48.

Bellandi, A., B. Furletti, V. Grossi, et A. Romei (2008). Ontological Support for Association Rule Mining. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria*, pp. 110–115.

Boulicaut, J.-F. et B. Jeudy (2010). Constraint-based Data Mining. In *Data Mining and Knowledge Discovery Handbook*, pp. 339–354.

Brahmi, H., R. B. Messaoud, S. B. Yahia, et O. Boussaid (2010). Extraction des Règles Inter-dimensionnelles à partir des Cubes de Données OLAP : État de l'Art. In *Actes du 5ème Atelier des Systèmes Décisionnels, ASD'2010, Sfax, Tunis*, pp. 73–84.

Brisson, L. et M. Collard (2008). An Ontology Driven Data Mining Process. In *Proceedings of the 10th International Conference on Enterprise Information Systems, Barcelona, Spain*, pp. 54–61.

Brisson, L. et M. Collard (2009). *How to* S*emantically* E*nhance a* D*ata* M*ining* P*rocess ?*, Volume 19, Chapter Enterprise Information Systems, pp. 103–116. Springer Berlin Heidelberg.

Cuxac, P., M. Cadot, et C. François (2005). Analyse Comparative de Classifications : Apport des Règles d'association floues. In *Actes des 5èmes journées d'Extraction et Gestion des Connaissances, EGC'2005, Paris, France*, pp. 519–530.

Escovar, E. L. G., C. A. Yaguinuma, et M. Biajiz (2006). Using Fuzzy Ontologies to Extend Semantically Similar Data Mining. In *Proceedings of the 21st Brazilian Symposium of Databases, Florianópolis, Brazil*, pp. 16–30.

Euler, T. et M. Scholz (2004). Using Ontologies in a KDD Workbench. In *In Workshop on Knowledge Discovery and Ontologies at ECML/PKDD, Pisa, Italy*, pp. 103–108.

Fankam, C., L. Bellatreche, D. Hondjack, Y. A. Ameur, et G. Pierra (2009). SISRO, Conception de Bases de Données à partir d'Ontologies de Domaine. *Technique et Science Informatiques 28*, 1233–1261.

Ferraz, I. N. et A. C. B. Garcia (2008). Ontology in Association Rules Pre-Processing and post-processing. In *Proceedings of the Third International Conference on Information Technology and Applications, Washington, USA*, pp. 87–91.

Gomez Perez, A. (1999). Ontological Engineering : A State of the Art. *Expert Update 2*, 33–44.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition - Special issue : Current issues in knowledge modeling 5*, 199–220.

Guarino, N. (1995). Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human-Computer Studies 43*, 625–640.

Han, J. (1997). OLAP Mining : Integration of OLAP with Data Mining. In *Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics, DS-7, Leysin, Switzerland*, pp. 3–20.

Hou, X., J. Gu, X. Shen, et W. Yan (2005). Application of Data Mining in Fault Diagnosis Based on Ontology. In *Proceedings of the Third International Conference on Information Technology and Applications, Washington, USA*, pp. 260–263.

Kimball, R. (2005). L*e* D*ata* W*arehouse :* G*uide de* C*onduite de* P*rojet*. Eyrolles.

Mansingh, G., K.-M. Osei-Bryson, et H. Reichgelt (2011). Using Ontologies to Facilitate Post-processing of Association Rules by Domain Experts. *Information Sciences 181*, 419–434.

Marinica, C. et F. Guillet (2010). Knowledge-Based Interactive Postmining of Association Rules Using Ontologies. *IEEE Transactions on Knowledge and Data Engineering 22*, 784–797.

Miani, R. G., C. A. Yaguinuma, M. T. P. Santos, et M. Biajiz (2009). NARFO Algorithm : Mining Non-redundant and Generalized Association Rules Based on Fuzzy Ontologies. In *Proceedings of the 11th International Conference on Enterprise Information Systems, ICEIS*

2009, Milan, Italy, pp. 415–426.

Nabli, A., J. Feki, et F. Gargouri (2009). Advanced Internet Based Systems and Applications. Chapter An Ontology Based Method for Normalisation of Multidimensional Terminology, pp. 235–246.

Plantevit, M., A. Laurent, D. Laurent, M. Teisseire, et Y.W.Choong (2010). Mining Multidimensional and Multilevel Sequential Patterns. *Transactions on Knowledge Discovery from Data (TKDD) 4*(1), 1–37.

Priebe, T. et G. Pernul (2003). Ontology-based Integration of OLAP and Information Retrieval. In *Proceedings of the 14th International Workshop on Database and Expert Systems Applications, Washington, DC, USA*, DEXA '03, pp. 610–614.

Srikant, R. et R. Agrawal (1994). Mining Generalized Association Rules. In *Proceedings of the 21th International Conference on Very Large Data Bases (VLDB'95), San Francisco, USA*, pp. 407–419.

Stoffel, K., J. Saltz, J. Hendler, J. Dick, W. Merz, et R. Miller (1997). Semantic Indexing For Complex Patient Grouping. In *In Proceedings of the Annual Conference of the American Medical Informatics Association*.

Tapucu, D., G. Diallo, Y. Aït-Ameur, et M. O. Ünalir (2009). *Ontology*-B*ased* D*atabase* A*pproach for* H*andling* P*references*.

Tseng, M.-C. et W.-Y. Lin (2009). Incremental Mining of Ontological Association Rules in Evolving Environments. In *Proceedings of the 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Tainan*, pp. 142–151.

Tseng, M.-C., W.-Y. Lin, et R. Jeng (2007). Mining Association Rules with Ontological Information. In *Proceedings of the Second International Conference on Innovative Computing, Informatio and Control, ICICIC '07, Washington, DC, USA*, pp. 300–303.

Wu, C.-A., W.-Y. Lin, C.-L. Jiang, et C.-C. Wu (2011). Toward Intelligent Data Warehouse Mining : An Ontology-Integrated Approach for Multi-dimensional Association Mining. *Expert System Application. 38*, 11011–11023.

Zeman, M., M. Ralbovský, V. Svatek, et J. Rauch (2009). Ontology-Driven Data Preparation for Association Mining. In *Proceedings of the 8th Znalosti Conference, Brno.*, pp. 1–12.

Zhang, Y., L. Zhang, G. Nie, et Y. Shi (2009). A Survey of Interestingness Measures for Association Rules. *Proceedings of the 2nd International Conference on Business Intelligence and Financial Engineering, Los Alamitos, CA, USA*, 460–463.

Summary

Decisional support systems are dedicated to the management of the enterprise to help decision makers in the decision process. In this respect, data mining is a complex and difficult task that requires a considerable intellectual effort. From the other hand, ontologies are at the present time at the main of works led in knowledge engineering and gave proof in several domains. In this paper, we explain the specificities of ontologies. Based on this we describe today's approaches of ontology-integrated association rule mining by pointing out common aspects and differences.

Aide à la décision multicritère et apprentissage automatique pour la classification

Antoine Rolland*

* Laboratoire ERIC - Université Lumière Lyon II
av Pierre Mendes-France, 69676 BRON Cedex
antoine.rolland@univ-lyon2.fr

Résumé. Les méthodes d'aide multicritère à la décision pour la problématique du tri d'une part, et les méthodes d'apprentissage automatique d'autre part poursuivent le même objectif : permettre d'affecter des objets à des catégories prédéfinies, ordonnées ou non. Cependant, les différences existantes entre les deux domaines d'application font que jusqu'à aujourd'hui ce deux domaines ont eu peu d'interactions. Il nous semble cependant qu'il y a la matière à exploration. Nous proposons donc de présenter quelques méthodes et paradigmes pour la problématique du tri en analyse multicritère avant de proposer quelques pistes pour des applications croisées en apprentissage automatique et en aide à la décision.

1 Introduction

Prendre une décision, c'est trancher entre plusieurs possibilités et choisir celle qui sera effectivement mise en œuvre. Les différentes communautés qui se sont intéressées à la prise de décision (économie, sociologie, psychologie, mathématiques, statistique, informatique...) se sont toutes penchées d'une manière ou d'une autre sur cette question, pour essayer de comprendre comment l'être humain prend une décision, pour l'accompagner dans cette prise de décision ou pour pouvoir reproduire cette décision de manière automatique. Prendre une décision est un acte simple quand toutes les alternatives sont connues, qu'elles sont peu nombreuses, qu'elles peuvent être évaluées de manière unique, et qu'il n'y a qu'une seule personne qui décide. Cependant, si les alternatives ou leurs conséquences sont imparfaitement connues, si leur nombre est trop important pour qu'une approche systématique puisse être envisagée, si elles doivent être évaluées selon plusieurs critères partiellement conflictuels, ou si plusieurs personnes sont amenées à se prononcer, alors la prise de décision devient un acte complexe. Chacune de ces difficultés a donné naissance à un champ particulier de recherche : décision dans l'incertain, optimisation combinatoire, décision multicritère, décision multi-agents... A la suite de Roy (1996), on distingue souvent trois types de problématique en aide à la décision :

- la problématique du choix, qui consiste à vouloir *choisir* la ou les solutions considérées comme optimales pour le problème considéré ;
- la problématique du classement (ranking), qui consiste à vouloir classer du premier au dernier toutes les solutions connues du problème considéré ;
- la problématique du tri (sorting), qui consiste à affecter les solutions à des catégories (ordonnées ou non).

Nous nous intéresserons ici au cas où les solutions étudiées sont évaluées suivant plusieurs points de vue, ou critères, qui peuvent potentiellement être conflictuels. Nous nous focaliserons en particulier sur la problématique du tri dans des catégories ordonnées.

1.1 Paradigme de l'aide à la décision multicritère

En aide à la décision multicritère (MCDA ou MultiCriteria Decision Aiding), on suppose que les différentes alternatives se présentant au décideur peuvent être décrites sur un certain nombre de propriétés ou attributs. Si l'on est capable d'établir une échelle de préférence sur un attribut, on parle alors de critère. Les valeurs des alternatives sur ces critères représentent la prise en compte de points de vue diversifiés, en général non réductibles à un seul critère. Ces critères sont souvent conflictuels. La difficulté est d'arriver à obtenir une comparaison relative des alternatives, afin de guider le choix du décideur vers la solution qui lui paraîtra optimale. En effet, la notion d'optimisation "dans l'absolu" est vide de sens en décision multicritère, car il n'existe généralement pas d'alternative optimisant tous les critères simultanément. Il est donc nécessaire de prendre en compte de l'information supplémentaire, en particulier l'importance relative de chaque critère et les compensations possibles entre les différents critères.

Le but des méthodes en MCDA est donc moins de proposer automatiquement une unique solution considérée comme optimale, que d'accompagner le décideur dans sa démarche de décision. La phase de définition du problème, des alternatives et des critères est généralement considérée comme plus importante que la phase purement calculatoire mettant en œuvre une méthode particulière d'agrégation de préférence (voir Schärlig (1985), Vincke (1992), Roy (1996)). En particulier, l'interaction avec le décideur tout au long du processus fait que le résultat d'une méthode d'aide à la décision ne semble pas arriver de nulle part, mais permet bien au décideur de prendre sa décision de manière éclairée.

1.2 Aide à la décision et apprentissage automatique

Il existe de nombreuses méthodes d'aide multicritère à la décision. Chaque méthode met en jeu un certain nombre de paramètres. Quelle que soit la méthode d'aide à la décision retenue, la question de fixer les valeurs des paramètres de la méthode est une question cruciale en aide à la décision. Deux approches existent à ce propos :

- demander au décideur de fixer les valeurs des paramètres directement : cette voie est délicate, du fait que le décideur ne maîtrise pas l'effet que peut avoir tel ou tel paramètre sur les comparaisons d'alternatives, et que les concepts utilisés dans chacune des méthodes peuvent parfois être difficile à appréhender pour un non-expert (par exemple la différence entre les poids des critères dans une méthode de moyenne pondérée et les degrés d'importance des critères dans une méthode ELECTRE n'est pas évidente).
- retrouver les valeurs des paramètres de la méthode à partir d'un jeu de données existant. Il s'agit d'inférer les paramètres de la méthode en prenant comme corpus d'apprentissage un jeu de données sur lequel on recueille les préférences du décideur. C'est cette méthode, proche dans son esprit de l'apprentissage automatique, que nous allons étudier par la suite, sous le nom d'élicitation des paramètres.

Cependant, il existe de grandes différences d'approches entre l'apprentissage automatique et l'élicitation des préférences en MCDA. Waegeman et al. (2009) détaillent huit principaux points de divergence que nous reprenons dans la table 1. De manière générale, on peut résumer

TAB. 1 – *Principales différences entre aide multicritère à la décision et apprentissage automatique, selon Waegeman et al. (2009)*

Thème	MCDA	Apprentissage automatique
Information disponible	orientée données	orientée résultat
Interaction avec le décideur	oui	non
Décideur	peut changer d'avis	préférences fixes
Cohérence	données cohérentes	bruit possible
Taille de l'échantillon	petit	grand
Modèle	souvent additif	très divers
Régularisation	non	oui
Adaptabilité	mauvaise	bonne
Données	vectorielles	de toute sorte

cela en disant que l'apprentissage automatique travaille avec des données très diverses, sur lesquelles on ne peut agir, disponibles dans des corpus de grande taille mais pouvant contenir des erreurs. Au contraire, les méthodes d'élicitation en MCDA utilisent généralement des données vectorielles provenant de corpus de petite taille, sans incohérence, et modifiables par dialogue avec le décideur. Malgré ces différences, nous pensons que les méthodes en MCDA peuvent être intéressantes également dans le cadre de l'apprentissage automatique. Nous allons donc succinctement présenter ces méthodes et les techniques d'élicitation usuelles (section 2) puis nous indiquerons des pistes de recherche pour rapprocher les deux domaines (section 4)

2 Méthodes multicritères pour le tri

Les méthodes de tri en MCDA s'intéressent à des objets appelés alternatives, décrites par plusieurs attributs, aussi appelés critères s'ils sont dotés d'une relation binaire de préférence. Le but de ces méthodes est d'affecter chaque alternative à l'un des catégories prédéfinies, qu'elles soient ordonnées ou non (Roy (1996)). Formellement, nous allons considérer ici un problème de tri en aide à la décision multicritère où les alternatives de l'ensemble $A = \{a_1, a_2, ..., a_j, ..., a_m\}$ sont évaluées sur n critères $g_1, g_2, ..., g_i, ..., g_n$, avec $i \in N = \{1, ..., i, ..., n\}$. Le critère g_i prend ses valeurs sur un ensemble X_i, i.e., $g_i : A \mapsto X_i$, doté d'une relation de préférence $\succsim_i$. X est alors le produit cartésien des différentes échelles de critère ($X = \prod_{i \in N} X_i$). Toute alternative de A peut donc être décrite par un vecteur de X. Les catégories ordonnées prédéfinies forment une partition de A et sont notées C_1, C_2, ..., C_p où C_i est préférée à C_j si $i > j$. On définit $C_t^{\geq}$, $t = 1, \ldots, p$, par $C_t^{\geq} = \cup_{s \geq t} C_s$.

Il existe aujourd'hui un large panel de méthodes d'aide à la décision multicritère. On se réfèrera à Bouyssou et al. (2006) pour une vue complète des méthodes les plus connues, et à Zopounidis et Doumpos (2002) pour les méthodes de tri. On distingue généralement trois grands types de méthodes pour l'aide à la décision multicritère, en reprenant la terminologie de Roy (1996) pour les deux premières :

– les méthodes à score unique de synthèse : il s'agit, pour chaque alternative, d'agréger toutes les valeurs prises sur chacun des critères pour obtenir un "score" unique pour cha-

cune des alternatives, un critère unique de synthèse . Il suffit ensuite pour comparer deux alternatives de comparer leurs scores respectifs. Ces méthodes ont été historiquement développées sous le nom de MultiAttribute Utility Theory (MAUT) depuis les années 70 principalement aux Etats-Unis (voir Keeney et Raiffa (1976)). L'introduction d'utilité non-additive a permis de généraliser ces méthodes. On consultera Marichal (2009) pour une vue d'ensemble de ces méthodes formelles. Dans le cadre particulier des méthodes de tri, on peut mentionner la méthode UTADIS proposée par Jacquet-Lagrèze (1995).
- les méthodes à relation de synthèse : il s'agit ici au contraire de comparer, critère par critère, les deux alternatives afin d'obtenir autant de relations de préférence partielles entre les deux alternatives qu'il existe de critères, puis d'essayer d'agréger ces préférences partielles en une préférence globale. Ces méthodes, basées sur une représentation des préférences par des graphes, a été développée depuis 1975, en particulier en Europe et dans la communauté francophone.
- plus récemment, les méthodes utilisant les règles de décision et le principe de dominance ont été proposées par Greco et al. (2001b).

Afin d'illustrer ces trois approches, et devant la multiplicité des méthodes multicritères existantes, nous choisissons de présenter ci-dessous trois méthodes "emblématiques" pour les problèmes de tri en MCDA. La première méthode, utilisant une fonction d'utilité non-additive (intégrale de Choquet). La deuxième, utilise une comparaison par paire (ELECTRE TRI), et la troisième les règles de décision.

2.1 Fonction d'utilité non-additive utilisant l'intégrale de Choquet

L'intégrale de Choquet est une méthode de scoring donnant une valeur pour chaque alternative à partir d'une mesure non additive sur les ensembles de critères. Une mesure non-additive, aussi appelée capacité (voir Grabisch (1996), Grabisch et Roubens (2000),Marichal (2009)), est une fonction $v : 2^N \rightarrow \mathbb{R}^+$ telle que $v(\emptyset) = 0$, et pour deux sous-ensembles A et B de N, $A \subseteq B \subseteq N$ implique $v(A) \leq v(B)$ (principe de monotonie par inclusion). On normalise généralement la capacité par $v(\emptyset) = 0$ et $v(N) = 1$. Une mesure non-additive permet par exemple que $v(A \cup B) \neq v(A) + v(B)$. L'intégrale de Choquet $u(a)$ est alors définie par :

$$u(a) = \sum_{i=1}^{n} g_{\tau(i)}(a)(v(\{\tau(i), \ldots, \tau(n)\}) - v(\{\tau(i+1), \ldots, \tau(n)\}))$$

où τ est une permutation sur N telle que $g_{\tau(1)}(a) \leq g_{\tau(2)}(a) \leq \ldots \leq g_{\tau(n)}(a)$. L'intégrale de Choquet est un opérateur d'agrégation très général, pouvant modéliser les opérateurs additifs ou non-additif tels que le min, le max, la somme pondérée, OWA, etc (voir Grabisch et al. (2009) et Grabisch et Labreuche (2008) pour les utilisations de l'intégrale de Choquet). L'utilisation d'une intégrale de Choquet nécessite que les ensembles X_i soient commensurables, par exemple $[0, 1]$ $\forall i \in N$. Formellement, l'intégrale de Choquet est une méthode de classement (ranking method), mais elle peut être utilisée comme méthode de tri (sorting method) en définissant un niveau de référence λ_t par catégorie C_t tel que :

$$a \in C_t^{\geq} \iff u(a) \geq \lambda_t$$

Comme toutes les méthodes à base d'utilités, la méthode de tri utilisant l'intégrale de Choquet repose sur plusieurs hypothèses fortes. En particulier, on suppose :

– la commensurabilité des différentes échelles de valeurs sur les critères entre elles
– la commensurabilité des échelles de valeurs sur les critères et la mesure v choisie
– la connaissance complète de la mesure v, et en particulier de la compensation intercritères.

Ces hypothèses permettent un calcul facile de la relation de préférence entre les différentes alternatives, et en particulier de l'affectation des alternatives aux différentes catégories. Mais cette modélisation est très couteuse en information sur les paramètres de la méthode. De plus, la facilité des calculs est obtenue au prix d'un éloignement certain de la manière dont raisonne un décideur.

2.2 ELECTRE-Tri

Une deuxième approche utilise une méthode à base de surclassement appelée ELECTRE-Tri (Roy (1991), Figueira et al. (2005)). Cette méthode utilise des profils prédéterminés, éléments de X correspondant à des " alternatives fictives", pour séparer les différentes catégories. Une alternative est alors affectée à la catégorie C_i (ou supérieure) si elle est préférée au profil marquant la limite inférieure de la catégorie. La relation de préférence utilisée dans ELECTRE-Tri est une relation de surclassement utilisant la concordance et la non-discordance comme dans les méthodes ELECTRE en général (Roy (1996)). De manière formelle, nous avons $p-1$ profils $q^2, \ldots, q^p$ prenant leurs valeurs sur X, n indices d'importance des critères $\omega_1, \ldots, \omega_n$ et λ un niveau seuil tel que : a est affecté à la catégorie C_t ou supérieure si

$$\sum_{i|g_i(a)\geq p_i^t} \omega_i \geq \lambda$$

Cette approche est moins gourmande en informations initiales, supporte des relations purement ordinales sur les valeurs des critères, voire des données imprécises ou manquantes. A contrario, les résultats obtenus sont généralement moins discriminants que les méthodes à base d'utilités. En particulier, des phénomènes d'incomparabilité entre alternatives et profils peuvent survenir, aboutissant à une incertitude possible sur les affectations à des catégories (voir Roy (1968, 1996)).

2.3 Règles de décision

Une troisième approche, proposée par Greco et al. (2001b) et Greco et al. (2002), utilise les règles de décision à travers les ensembles approximatifs et le principe de dominance (Dominance-based Rough Set Approach). On dit qu'une alternative a domine une alternative b si $\forall i \in N$, $g_i(a) \succsim_i g_i(b)$. Les alternatives sont affectées aux différentes catégories en fonction de niveaux de référence sur chaque critère. Les fondements axiomatiques de ces méthodes ont été étudiés, en particulier dans le cadre de la problématique du tri, par Greco et al. (2001a) pour les relations de surclassement. Formellement, un modèle à base de règles de décision est capable d'affecter l'alternative a à une catégorie grâce à des règles du type "si $g_1(a) \succsim_1 \alpha_1$ et $g_2(a) \succsim_2 \alpha_2$ et ... alors a est affecté à la catégorie C_i".

Le grand intérêt de ce type de méthode réside dans l'interprétation immédiate des résultats et la représentation sous forme de règles de la logique du décideur. Cela se fait au détriment d'un modèle synthétique : en effet, le nombre de règles de décision peut rapidement être trop important pour être appréhendé rapidement.

3 Elicitation des paramètres des méthodes de tri

Récemment, plusieurs auteurs se sont particulièrement intéressés aux processus d'élicitation pour les méthodes multicritères d'aide à la décision, tant en étudiant les fondements théoriques de ces processus qu'en proposant des outils concrets pour l'élicitation. L'idée fondatrice est de partir d'exemples d'affectations données par le décideur pour calculer les paramètres de la méthode choisie redonnant ces affectations. On a donc en entrée de la procédure d'élicitation l'ensemble des alternatives et de leurs valeurs sur chacun des critères, généralement présenté sous forme d'un tableau appelé table de performance, ainsi que l'affectation de chacune des alternatives à une catégorie.

Dans le cadre de l'intégrale de Choquet, des procédures d'estimation des coefficients d'une capacité ont été proposées par Grabisch et al. (2008). Le point de départ de ces procédures est le suivant : connaissant la table de performance des alternatives sur tous les critères et le score global de chaque alternative, le but est de trouver les paramètres de la capacité qui fait en sorte que les résultats du calcul de l'intégrale de Choquet soient les plus proches possibles de ces valeurs globales. Les méthodes retenues utilisent la programmation linéaire avec plusieurs fonctions objectifs possibles, comme les moindres carrés ou la variance minimale. Le package `kappalab` permettant l'élicitation des capacités est disponible sous 'R' et dans le projet Decision Deck (Meyer et Bigaret (2011)) proposé par Meyer (2011).

L'avantage de ces méthodes est de toujours aboutir à une solution ; l'inconvénient est que la solution proposée peut être loin de retrouver la classification initiale si certaines incohérences existent chez le décideur. En outre, la complexité des calculs fait que le programme ne fonctionne que dans les cas où ni les critères ni les alternatives ne sont trop nombreuses.

Les procédures d'élicitation des paramètres de la méthode ELECTRE-Tri ont été proposées par Mousseau et Slowinski (1998) (voir aussi Ngo The et Mousseau (2002), Mousseau et al. (2001) et Bouyssou et Marchant (2007a,b)). Ces procédures utilisent un programme linéaire pour minimiser le nombre de différence entre les affectations aux catégories proposées par le décideur d'une part et par la méthode ELECTRE-TRI d'autre part. Le nombre de profils est déterminé par le nombre de catégories. Une implémentation de cette procédure dans le projet Decision Deck a été proposée par Sobrie (2011).

La encore, la taille de l'ensemble d'apprentissage est un facteur limitant l'efficacité de l'approche. Cependant, la possibilité de pouvoir fixer *a priori* certains des paramètres tels que la valeur de seuil permet potentiellement d'accélérer les calculs.

La méthode d'apprentissage des règles de décision utilise une approche fondée sur les ensembles approximatifs et la relation de dominance à partir d'un ensemble d'apprentissage, comme proposé par Greco et al. (2000). On dit que l'alternative a domine l'alternative b si $g_i(a) \succsim_i g_i(b)$ pour tout critère g_i, $i \in N$. Pour un ensemble de critères $\{g_i,\ i \in N\}$ donné, l'inclusion d'une alternative $a \in A$ dans l'union des classes $C_t^{\geq}$, $t = 2, \ldots, p$ aboutit à une incohérence au sens du principe de dominance si une des deux conditions suivantes tient :

- a appartient à la catégorie $C_t^{\geq}$ ou une catégorie supérieure, mais est dominée par une alternative b appartenant à une catégorie moins élevée que C_t
- a appartient à une catégorie inférieure à $C_t^{\geq}$ mais domine une alternative b appartenant à la catégorie C_t ou une catégorie supérieure

Si l'inclusion de $a \in A$ dans $C_t^{\geq}$, $t = 2, \ldots, p$, amène une incohérence au sens du principe de dominance, alors a appartient à $C_t^{\geq}$ avec ambiguïté. Et donc a appartient à $C_t^{\geq}$ sans ambiguïté

si $a \in C_t^{\geq}$ et s'il n'y a aucune incohérence au sens du principe de dominance. Cela signifie que toutes les alternatives dominant a appartiennent à $C_t^{\geq}$. Les affectations des alternatives de l'ensemble d'apprentissage peuvent donc être séparées entre affectations certaines et affectations possibles. Ces affectations entrainent ensuite des règles de décision certaines et des règles de décision possibles. L'affectation d'une nouvelle alternative se fait alors à l'aide des règles de décision certaines. le logiciel jMAF proposé par Stefanowski (1998) permet la génération des règles de décision à partir d'un corpus d'apprentissage.

Le grand avantage de cette méthode est de pouvoir traiter des incohérences dans les affectations proposées par le décideur. l'inconvénient est que cela peut aussi parfois aboutir à un grand nombre de règles de décision, chacune traitant un cas spécifique.

4 Perspectives de recherche

Les différences évoquées entre apprentissage automatique et MCDA nous conduisent à proposer quatre pistes de recherche où les deux champs pourraient interagir. L'objectif poursuivi est l'intégration de méthodes provenant de l'aide à la décision dans le domaine de l'apprentissage automatique. En particulier, un travail d'investigation doit être mené pour l'identification de champs d'application pour lesquelles l'utilisation des méthodes d'aide à la décision pourraient apporter une réelle plus-value.

1. un première piste de recherche concerne la **complexité des calculs** pour l'élicitation des paramètres. Les algorithmes utilisés reposent souvent sur des algorithmes de programmation linéaire, dont le temps de calcul s'accroit rapidement avec le nombre de variables en entrée (critère et/ou instances). A titre d'exemple, le calcul des capacités par `Kappalab` est limité à moins de 10 critères, ce qui est amplement suffisant en aide à la décision (il est courant de considérer qu'un problème à plus de 6 ou 7 critère est un problème mal défini) mais pas forcément en apprentissage automatique. Une prise en compte de la structure particulière des problèmes d'apprentissage automatique permettrait peut-être de réduire le temps de calcul, en particulier pour les méthodes du type "utilité" ou "ELECTRE Tri".
2. une deuxième piste a trait au **traitement des erreurs et incohérences** : la différence de point de vue entre l'apprentissage automatique, où les erreurs sont acceptées comme telles, et l'aide à la décision, où l'on suppose que les préférences indiquées sont cohérentes, fait que les méthodes de MCDA visent à intégrer toutes les données disponibles pour éliciter les paramètres, ce qui est inefficace dans le cas de grands corpus d'apprentissage. Il nous semble qu'une piste de recherche pourrait s'intéresser à la robustesse des algorithmes d'élicitation vis-à-vis de la présence ou de l'absence de telle ou telle alternative dans le corpus d'apprentissage, afin de pouvoir éliminer les "valeurs aberrantes".
3. une troisième piste consiste à pouvoir **interagir avec le décideur** dans le cadre de l'apprentissage automatique. cela semble paradoxal avec le terme "automatique" mais une interaction avec le décideur pourrait, en complément du point précédent, permettre aux méthodes d'élicitation de ne pas prendre en compte des données pouvant être considérées comme erronées. Une autre idée consiste à permettre au décideur de réduire l'ensemble d'apprentissage à quelques cas emblématiques pour que la taille du corpus soit compatible avec les capacités de calcul des méthodes d'élicitation. La question est alors

celle de l'analyse du corpus de données en entrée pour ôter l'information redondante et ne garder que les alternatives vraiment intéressantes.

4. enfin, une quatrième piste concerne l'intégration d'outil d'aide à la décision pour **l'agrégation de classifieurs**, dans le cas où l'apprentissage automatique se fait à travers plusieurs algorithmes potentiellement conflictuels (cf Kuncheva (2004), Stefanowski et Nowaczyk (2007)). L'idée est d'agir non pas directement sur le corpus des alternatives, mais sur les résultats obtenus à l'aide de différents classifieurs, soit faisant partie d'une même famille de classifieurs avec des paramètres différents, soit provenant de plusieurs approches différentes de classifieurs.

Références

Bouyssou, D., D. Dubois, M. Pirlot, et H. Prade (2006). *Concepts et méthodes pour l'aide à la décision, volume 3, analyse multicritère*. Hermes.

Bouyssou, D. et T. Marchant (2007a). An axiomatic approach to noncompensatory sorting methods in mcdm, I : The case of two categories. *European Journal of Operational Research 178*, 217–245.

Bouyssou, D. et T. Marchant (2007b). An axiomatic approach to noncompensatory sorting methods in mcdm, II : More than two categories. *European Journal of Operational Research 178*, 246–276.

Figueira, J., V. Mousseau, et B. Roy (2005). ELECTRE methods. In J. Figueira, S. Greco, et M. Ehrgott (Eds.), *Multiple Criteria Decision Analysis : State of the Art Surveys*, pp. 133–162. Boston, Dordrecht, London : Springer Verlag.

Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research 89*, 445–456.

Grabisch, M., I. Kojadinovic, et P. Meyer (2008). A review of methods for capacity identification in choquet integral based multi-attribute utility theory : Applications of the kappalab r package. *European Journal of Operational Research 186*(2), 766 – 785.

Grabisch, M. et C. Labreuche (2008). A decade of application of the choquet and sugeno integrals in multi-criteria decision aid. *4OR : A Quarterly Journal of Operations Research 6*, 1–44.

Grabisch, M., J.-L. Marichal, R. Mesiar, et E. Pap (2009). *Aggregation functions*, Volume 127 of *Encyclopedia of Mathematics and its Applications*. Cambridge, UK : Cambridge University Press.

Grabisch, M. et M. Roubens (2000). Application of the Choquet integral in multicriteria decision making. In M. Grabisch, T. Murofushi, et M. Sugeno (Eds.), *Fuzzy Measures and Integrals - Theory and Applications*, pp. 348–374. Physica Verlag.

Greco, S., B. Matarazzo, et R. Slowinski (2000). Multicriteria classification by dominance-based rough set approach,methodological basis of the 4emka system. *[Online], Available :http ://www-idss.cs.put.poznan.pl/4emka/*.

Greco, S., B. Matarazzo, et R. Slowinski (2001a). Conjoint measurement and rough set approach for multicriteria sorting problems in presence of ordinal criteria. In A. Colorni,

M. Paruccini, et B. Roy (Eds.), *A-MCD-A, Aide Multicritère à la Décision/Multiple Criteria Decision Aid*, pp. 117–144. Ispra : European Commission, Joint Research Centre, EUR 19808 EN.

Greco, S., B. Matarazzo, et R. Slowinski (2001b). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research 129*, 1–47.

Greco, S., B. Matarazzo, et R. Slowinski (2002). Rough sets methodology for sorting problems in presence of multiple attributes and criteria. *European Journal of Operational Research 138*, 247–259.

Jacquet-Lagrèze, E. (1995). An application of the UTA discriminant model for the evaluation of sc R&D projects. In P. Pardalos, Y. Siskos, et C. Zopounidis (Eds.), *Advances in Multicriteria Analysis*, Nonconvex Optimization and its Applications, pp. 203–211. Dordrecht : Kluwer Academic.

Keeney, R. et H. Raiffa (1976). *Decisions with multiple objectives : Preferences and value tradeoffs*. J. Wiley, New York.

Kuncheva, L. (2004). *Combining Pattern Classifiers. Methods and Algorithms*. Wiley.

Marichal, J.-L. (2009). *Aggregation functions for decision making, Decision-Making Process - Concepts and Methods*. ISTE/John Wiley.

Meyer, P. (2011). http ://www.decision-deck.org/ws/wsd-choquetintegral-kappalab.html.

Meyer, P. et S. Bigaret (2011). diviz : a software for modeling, processing and sharing algorithmic workflows in mcda. *Intelligent DecisionTechnologies : an International Journal*, to appear.

Mousseau, V., J. Figueira, et J. Naux (2001). Using assignment examples to infer weights for ELECTRE TRI method : Some experimental results. *European Journal of Operational Research 130*(2), 263–275.

Mousseau, V. et R. Slowinski (1998). Inferring an electre tri model from assignment examples. *Journal of Global Optimization 12*, 157–174.

Ngo The, A. et V. Mousseau (2002). Using assignment examples to infer category limits for the ELECTRE TRI method. *JMCDA 11*(1), 29–43.

Roy, B. (1968). Classement et choix en présence de point de vue multiples (la méthode electre). *Les cahiers du CERO* (8), 57–75.

Roy, B. (1991). The outranking approach and the foundations of ELECTRE methods. *Theory and Decision 31*, 49–73.

Roy, B. (1996). *Multicriteria Methodology for Decision Aiding*. Dordrecht : Kluwer Academic.

Schärlig, A. (1985). *Décider sur plusieurs critères*. Presses Polytechniques Universitaires Romandes.

Sobrie, O. (2011). http ://www.decision-deck.org/ws/wsd-electretribminference-pyxmcda.html.

Stefanowski, J. (1998). On rough set based approaches to induction of decision rules. In S. A. Polkowski L. (Ed.), *Rough Sets in Data Mining and Knowledge Discovery*, Volume 1, pp. 500–529. Physica-Verlag.

Stefanowski, J. et S. Nowaczyk (2007). An experimental study of using rule induction algo-

rithm in combiner multiple classifier. *International Journal of Computational Intelligence Research 3*, 335–342.

Vincke, P. (1992). *Multicriteria Decision-Aid*. J. Wiley, New York.

Waegeman, W., B. D. Baets, et L. Boullart (2009). Kernel-based learning methods for preference aggregation. *4OR* (7), 169–189.

Zopounidis, C. et M. Doumpos (2002). Multicriteria classification and sorting methods : A literature review. *European Journal of Operational Research* (138), 229–246.

Summary

Multicriteria decision aiding methods for sorting problems and machine learning algorithms are sharing a similar goal : they aim at assigning objects to one of the predefined ordered or not ordered categories. However, the differences between these two fields of application avoid seminal interactions. It seems to us that it could be interesting to study how they can interact. In this paper, we present some methods and paradigms for sorting multricriteria decision aiding methods, and then we suggest some tracks for further research.

Opérationnalisation du Processus d'Aide à la Décision

Igor Crévits[*], Alexis Tsoukiàs[**]

[*] LAMIH, Université de Valenciennes de du Hainaut-Cambrésis, Le Mont Houy,
59313 Valenciennes Cedex 9
Igor.Crevits@univ-valenciennes.fr
[**] LAMSADE, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny,
75775 PARIS Cedex 16
tsoukias@lamsade.dauphine.fr

Résumé. Les objectifs de l'atelier mettent en exergue les traits caractéristiques de l'aide à la décision qui s'appuie sur une représentation de problèmes types et recourt à un objet extérieur, le modèle, afin d'apporter une solution au problème représenté. On retrouve également dans les communautés évoquées les deux orientations actuelles de l'aide à la décision, à savoir la théorie de la décision, avec l'ensemble de ses approches mathématiques, et les systèmes d'information utilisant des représentations informatiques.
Nous proposons une troisième orientation méthodologique permettant de guider la construction d'une aide à la décision depuis le problème jusque la recommandation finale en passant par le modèle.

1. Introduction

L'essentiel du sujet de l'aide à la décision porte sur la représentation permettant de produire une solution, la question de la construction de cette représentation dans le cadre de la réalité de problèmes concrets mérite réflexion. Il se dégage alors la nécessité de se pencher sur une orientation méthodologique de l'aide à la décision.

Dans cette orientation, le facteur important n'est pas tant la décision comme résultat, mais le processus décisionnel avec lequel cette solution doit être cohérente. Pour cela, une analyse du problème concret, dont le processus décisionnel est une composante, est à mettre en œuvre. Dans ce travail d'analyse, deux acteurs particuliers apparaissent : le client, représentant du problème, et l'analyste, représentant de l'aide. La littérature présente plusieurs méthodes mais elles montrent deux limites essentielles (voir (Bouyssou, 2006) pour une présentation détaillée). Soit les hypothèses sur lesquelles elles se basent ne permettent pas de concevoir le modèle, les méthodes considérées sont alors trop proches du client. Soit elles sont attachées trop tôt à un cadre de modélisation, elles sont alors trop proches de l'analyste. Pour être réellement efficace, le cadre méthodologique doit permettre d'établir une cohérence entre plusieurs questions décisives : identification et représentation des facteurs clés de la réalité du problème, analyse du problème, au moyen de cette représentation, afin d'en identifier les verrous, construction de solutions cohérentes, évaluation fine de solutions au moyen de modèles et conclusion des évaluations en vue de l'intégration dans la réalité. De ce fait, la construction d'une aide à la décision constitue un processus décisionnel conduit par l'analyste et impliquant le client. Le Processus d'Aide à la Décision (PAD) (Tsoukiàs, 2007) offre un cadre méthodologique guidant l'interaction entre

le client et l'analyste autour de quatre artéfacts cognitifs partagés représentant le problème, les solutions, l'évaluation des solutions et la recommandation.

Dans les vingt dernières années, le trafic aérien a doublé et il est prévu la même augmentation dans les vingt prochaines années. La gestion du trafic aérien se trouve en difficulté pour absorber cette augmentation. Ainsi, plusieurs projets d'automatisation totale du contrôle aérien ont été conduits aux USA, en Europe et en France, mais ils n'ont pas été mis en œuvre. Il en va de même pour des projets plus ciblés d'aide aux contrôleurs aériens. Les changements organisationnels s'avèrent délicats car ils doivent s'accompagner d'une évolution de la formation des contrôleurs aériens difficilement compatible avec les objectifs de gestion du trafic. Enfin, cette activité fait apparaître de multiples protagonistes dont les enjeux sont imbriqués, le plus important d'entre eux étant la sécurité.

La réorganisation de la gestion du trafic aérien est enclenchée en Europe. Nous proposons ici une analyse rétrospective indépendante des différentes études d'automatisation. Puis, nous en tirons les enseignements pour proposer l'aide à la décision comme possibilité d'accroissement des capacités de contrôle, par le déploiement de PAD. Enfin, nous dressons quelques points de réflexion sur l'aide à la décision qui apparaissent à la lumière du contrôle aérien.

2. Gestion du trafic aérien

La gestion du trafic aérien a pour objectif de faire face à la demande d'utilisation de l'espace aérien tout en garantissant la sécurité des vols. La demande est caractérisée par des itinéraires et créneaux horaires émanant des compagnies aériennes. Ils servent de base aux autorités de gestion de l'aviation civile qui procèdent à des ajustements, uniquement si cela s'avère nécessaire, traduisant ainsi un objectif de régularité du trafic. La sécurité concerne l'utilisation simultanée de l'espace aérien. Elle relève de la responsabilité des autorités de gestion lesquelles sont organisées ou déléguées par les états. Ces deux objectifs se déclinent de plusieurs façons selon que le vol n'est pas encore opéré ou est en cours de déroulement. La sécurité constitue la priorité de toutes les parties prenantes. Plusieurs moyens organisationnels et opérationnels permettent d'assurer cette gestion.

D'un point de vue organisationnel, les itinéraires sont structurés en routes standards qui sont des successions de points appelés balises. L'espace aérien est divisé en secteurs géographiques regroupant des niveaux de vol séparés de 1000 pieds. Un itinéraire passe par plusieurs secteurs. La régularité du trafic consiste à respecter l'itinéraire et les horaires des vols. La sécurité vise à éviter que des vols suivent le même itinéraire au même moment. La gestion, Des flux, a pour objectif de prévenir la saturation de l'espace aérien et des aéroports en régulant le trafic. Les vols doivent donc être annoncés à l'organisme de régulation qui fixera leur heure de décollage qui pourra être retardée si le besoin s'en fait sentir.

Au niveau opérationnel, le contrôle de trafic aérien, a en charge de permettre aux compagnies aériennes une exécution des vols sûre, rapide et efficace. Trois types de services sont fournis. Le service d'information est chargé de fournir aux avions toutes les informations utiles au bon déroulement des vols. Le service d'alerte est chargé de la recherche des avions en difficulté et du sauvetage de leurs occupants. Les services de contrôle assurent la prise en charge lors des trois phases de l'exécution des vols : en survol, en approche des aéroports et sur les aéroports.

A chaque secteur est associée une position de contrôle qui dispose de plusieurs moyens d'information et d'action. Un système d'information délivre sur chaque position la partie de plan de vol qui concerne son secteur quelques minutes avant que le vol se présente. Cette information est capitalisée sur une bande papier appelée strip. A chaque balise est associé un horaire estimé de passage, concrétisant ainsi l'objectif de régularité du trafic. Les contrôleurs disposent également d'une vue radar représentant la situation instantanée du trafic dans le secteur.

Chaque secteur est pris en charge par deux contrôleurs. Le contrôleur organique supervise la transition des vols avec les secteurs adjacents. Le contrôleur supervise des vols à l'intérieur du secteur. Il gère les situations de rapprochement où deux vols empruntent deux routes qui se croisent sur une balise de telle sorte qu'ils se trouveront dans un avenir proche sous une distance minimale de séparation de 5 NM. Une telle situation, où les vols sont dits non séparés, est appelée « conflit ». Elle constitue la caractérisation concrète de l'objectif de sécurité. La situation avérée est appelée « airmiss ». La gestion d'un conflit a pour objectif de rétablir la séparation des vols par une modification de trajectoire, de la façon la moins pénalisante possible pour les vols.

L'augmentation du trafic aérien ne peut être satisfaite que dans la limite où les capacités de gestion permettent de continuer de garantir la sécurité actuelle des vols. Le dénominateur commun entre les deux niveaux de gestion réside dans la capacité de l'espace aérien à être utilisé par plusieurs vols simultanément. La capacité du niveau opérationnel, le contrôle de trafic, est donc déterminante, en particulier en survol.

3 Projets d'automatisation du contrôle

Plusieurs projets d'automatisation totale ou partielle du contrôle ont été conduits aux USA (AERA), en Europe (ARC2000) et en France (SAINTEX, techniques réactives, optimisation des mouvements). Ces projets abordent l'enjeu fondamental de sécurité de la navigation aérienne de différentes façons : résolution de conflits simples (AERA, SAINTEX), résolution globale de tous les conflits d'un secteur (techniques réactives), planification des vols par asservissement des horaires de passage sur les balises afin d'éliminer les conflits à la base (ARC2000) et résolution de conflits complexes et optimisation des mouvements des vols. On trouve dans (Durand, 2004) une description de ces travaux.

Ces différentes études apportent un éclairage sur le problème.Elles montrent que la seule issue pour faire face à l'augmentation de trafic réside dans l'accroissement des capacités de gestion opérationnelle durant les phases de survol qui relève du contrôle en-route. En effet, les dispositions organisationnelles de régulation des vols ne permettent de faire face à des fluctuations de trafic que dans la limite de la structuration du trafic mise en place. Une structuration et une régulation occupant plus largement l'espace et le temps permettrait d'accroitre considérablement la quantité de trafic gérable tout en garantissant la sécurité. Cependant, une telle structuration conduirait à un allongement des routes et une programmation des vols qui ne serait pas cohérentes avec les besoins des passagers. L'augmentation du trafic ne serait donc supportée que par les compagnies aériennes qui ne rendraient plus le service attendu par les passagers.

Par ailleurs, un problème fondamental de la résolution automatique de conflits concerne la prévision des trajectoires des avions. A l'origine de la résolution d'un conflit se trouve des vols dont la position doit être estimée de façon précise pour produire des déviations pertinentes. Cette prévision est difficile compte tenu des facteurs d'incertitude. Les capacités de prévision sont donc liées aux moyens technologiques au sol, comme les moyens radars, ou embarqués à bord des avions, ainsi qu'aux liaisons de données sol-bord. Ces projets montrent la nécessité d'un contexte technologique pour devenir réalisable qui ce pèserait fortement sur les compagnies aériennes niant ainsi l'objectif de service public (directe ou délégué) des autorités de gestion.

Enfin, bien qu'affichant des objectifs d'automatisation totale, le contrôleur était conservé, sans que sa place ne soit clairement définie. Les contrôleurs constituent en fait les garants ultimes de la qualité de service car ils assurent le maintien de la cohérence entre les enjeux de sécurité et de régularité du trafic lorsqu'ils deviennent contradictoires, c'est-à-dire lorsque les vols sont mis en œuvre.

Aucun des projets conduits n'a abouti en une vingtaine d'années. Le problème a été abordé de façon stricte par une seule vision extérieure des objectifs de gestion, selon une logique de résultat. La trop grande complexité des conflits envisagés et la présence nécessaire des contrôleurs a conduit à une série d'études dans une orientation de coopération Homme-Machine (SPECTRA et AMANDA) (Annebicque, 2012), (Debernard, 2009), (Pacaux-Lemoine, 2002), (Pacaux-Lemoine, 1996), (Vanderhaegen, 1994). La gestion du trafic est vue comme un problème de supervision où un système de gestion automatique de conflits assiste le contrôleur selon ses besoins. Les conflits gérés par le système sont d'une complexité identique à ceux gérés par les contrôleurs. Contrairement aux approches d'automatisation, le contrôleur conserve toute sa place sans que la capacité de gestion de conflits supplémentaire n'introduise d'incohérences d'objectifs. Cependant, même si les objectifs sont identiques pour le contrôleur et le système automatique de gestion des conflits, il demeure quelques incohérences qui ont été constatées de façon expérimentale. Quelques rares cas de déviations trop importantes ont été soulevés dans des situations de trafic chargé. Des fortes déviations font partie de la pratique des contrôleurs. Néanmoins, elles leur apparaissent inacceptables de la part d'un système automatique.

Les investigations conduites montrent un réel intérêt qui peut être amélioré en dotant le système de capacités de gestion cohérentes avec les processus décisionnels des contrôleurs. De ce fait, la question nécessite d'être abordée sous l'angle de l'aide à la décision. Dans la suite, nous présentons une telle étude et, au préalable, le cadre méthodologique utilisé.

4 Processus d'Aide à la Décision

PAD structure les décisions relatives à la conception d'une aide à la décision selon les étapes de rationalité procédurale (intelligence, design, choice, review) de Simon (Simon, 1976). PAD considère deux intervenants : le client (lequel peut être décideur), demandeur de l'aide et l'analyste, qui disposent de connaissances méthodologiques. Tous deux cherchent à construire, sous une contrainte de temps, une représentation partagée structurée en quatre artefacts : la situation problématique, la formulation du problème, le modèle d'évaluation et la recommandation finale. Chacun amène sa connaissance afin de produire une recommandation utile à apporter une réponse au problème du client.

La *situation problématique* a pour objet de représenter l'origine du problème et l'implication du client, d'identifier les conséquences d'une décision et la façon la plus judicieuse d'apporter une solution. Ces précisions sont également utiles à l'analyste pour qu'il clarifie lui-même sur quel point l'aide peut porter.

Formellement, la situation problématique $\mathcal{P}$ est un triplet ($\mathcal{A}$, $\mathcal{O}$, $\mathcal{S}$) où :

- $\mathcal{A}$ est l'ensemble des participants au processus de décision,
- $\mathcal{O}$ est l'ensemble des enjeux que les participants amènent dans le processus de décision,
- $\mathcal{S}$ est l'ensemble des engagements pris par chaque participant sur ses enjeux et ceux des autres.

L'objectif est bien de clarifier, pas de figer. Les précisions apportées permettent de mieux faire évoluer la demande d'aide et la réponse associée.

La *formulation du problème* a pour objet la réponse à apporter au problème clarifié dans le cadre de la situation problématique. La construction de la formulation du problème s'appuie sur des choix autour d'une idée que se fait le client d'une certaine rationalité dans la réponse à apporter au problème. Ces choix font l'objet d'une explicitation par l'analyste en vue de leur capitalisation en une représentation formelle, étape préalable essentielle à l'application d'une méthode d'aide à la décision.

Du point de vue formel, la formulation du problème Γ est un triplet ($\mathbb{A}$, $\mathbb{V}$, Π) où :

- $\mathbb{A}$ est l'ensemble des actions potentielles dans le cadre de la situation problématique $\mathcal{P}$,
- $\mathbb{V}$ est l'ensemble des points de vue sous lesquels il est envisagé d'observer, d'analyser, d'évaluer et de comparer les actions potentielles,
- Π est la problématique décisionnelle, la typologie d'application envisageable sur $\mathbb{A}$, une anticipation de ce que le client attend.

La formulation du problème est une étape importante dans la construction de la représentation partagée entre le client et l'analyste. Elle constitue la charnière entre le problème de décision vu d'une façon implicite par le client, et une représentation manipulée par l'aide destinée à produire un résultat. La formulation du problème en explicitant le problème de décision permet à l'analyste de préparer son travail de modélisation d'une façon intelligible pour le client, favorisant ainsi l'expression de son point de vue et son intervention.

Le *modèle d'évaluation* a pour objet d'estimer finement, sur la base d'une représentation numérique dotée de propriétés formelles, l'impact de la solution envisagée dans la formulation du problème.

Le modèle d'évaluation $\mathcal{M}$ est un n-uple (A, D, E, H, $\mathcal{U}$, $\mathcal{R}$) où :

- A est l'ensemble des alternatives sur lesquelles le modèle d'évaluation s'applique,
- D est l'ensemble des dimensions, éventuellement muni de propriétés structurelles, par lesquelles les actions potentielles de A sont manipulées par le modèle,
- E est l'ensemble des échelles associées à chaque élément de D,
- H est l'ensemble des critères sous lesquels les éléments de A sont évalués afin de prendre en compte les préférences du client, restreintes à chaque critère,
- $\mathcal{U}$ est l'ensemble des distributions d'incertitudes associées à D et/ou H,
- $\mathcal{R}$ est l'ensemble des opérateurs de synthèse d'information des éléments de A ou de $A \times A$, notamment les opérateurs d'agrégation.

Il s'agit d'une représentation conforme en grande partie aux modèles d'aide à la décision classiquement utilisés. Le modèle d'évaluation peut être soumis à plusieurs validations conformément à son cadre théorique ou à la réalité du problème.

La *recommandation finale* Φ s'attache à deux questions :
- la traduction du résultat du modèle d'évaluation en une représentation cohérente avec le processus de décision et intelligible pour le client,
- l'insertion de la traduction précédente dans le cadre (organisation, participants) du processus de décision.

PAD a été déployé dans différentes situations comme l'organisation administrative (Ostanello, 1993), la conception de logiciels (Stamelos, 2003) ou la maintenance d'infrastructures routières (Tsoukiàs, 2012). Il s'agit de problèmes moins complexes que la gestion des conflits aériens. L'aide à la décision y est vue comme une résolution de problèmes. L'aide aux décisions de gestion des conflits aériens relève d'un accroissement d'une rationalité en place. Les décisions considérées sont fortement constituées par un haut niveau de formation en amélioration permanente et de qualification professionnelle exigeante. Ce domaine constitue un réel challenge pour l'aide à la décision et une épreuve déterminante pour un cadre méthodologique.

5 Processus d'Aide à la Décision pour la représentation et l'aide à la gestion des conflits

En nous appuyant sur PAD, nous réalisons une analyse du problème pour proposer plusieurs aides à la décision. Les études relèvent de l'initiative des autorités de gestion (DGAC - Direction Générale de l'Aviation Civile) qui en sont donc les clients. L'approche décisionnelle de la résolution de conflits nécessite de considérer les contrôleurs comme décideurs différent du client. De ce fait, en préalable à la définition d'une aide à la décision pour la gestion des conflits, une analyse du comportement décisionnel des contrôleurs a été réalisée (Annebicque, 2012). Nous en présentons les grandes lignes.

La gestion des conflits passe par quatre étapes distinctes séquentielles qui s'affinent de façon fluide :
- la *détection* vise à identifier les situations conflictuelles à l'arrivée d'un nouveau vol dans le secteur,
- la *résolution* a pour but de rechercher le paramètre de vol à modifier pour dévier un avion, sans le retarder trop fortement afin d'accroitre l'écart entre les avions en conflit, sans retarder trop fortement le vol et sans générer d'autres conflits (une résolution opérée en ne considérant que l'un des deux vols en conflit est qualifiée de naturelle par les contrôleurs ; la déviation du deuxième avion sur la balise du conflit vers le premier est cinématiquement la moins pénalisante),
- l'*ordre* concerne la construction et la mise en œuvre de la modification,
- la *remise sur route* vise à restaurer les paramètres de vols afin de replacer l'avion sur sa trajectoire initiale.

La gestion d'un conflit est donc enclenchée dès qu'un nouvel avion se présente. Elle peut donc se terminer immédiatement si aucun conflit impliquant cet avion n'est détecté. La détection concerne donc des couples d'avions dissociés et non pas l'ensemble du trafic.

Les quatre étapes s'appuient sur la structuration du trafic pour intervenir de façon cohérente et maîtrisée. Le contrôleur radar constitue le seul décideur assurant la cohérence de deux enjeux. La sécurité représente l'enjeu qui le concerne directement. Mais, il porte

également l'enjeu de régularité des compagnies aériennes. Ces deux enjeux se précisent avec les principes d'élégance et d'énergie évoqués par les contrôleurs.

Pour l'ensemble des étapes de gestion de conflits, le tableau de strips et la vue radar constituent les seules représentations du trafic. Une part de l'activité concerne la gestion, appelée tenue, du tableau de strips ce qui permet de représenter la dynamique du trafic. Les strips sont également annotés pour suivre l'évolution particulière de chacun des vols. Aucune représentation n'est disponible pour les couples d'avions ou des regroupements plus importants. Il en va de même pour l'impact des résolutions sur le reste du trafic ou l'effet des ordres sur la séparation des vols. Les contrôleurs s'appuient sur des procédures et des règles de gestion basées sur les données disponibles pour construire des représentations personnelles des conflits et des étapes de résolution qu'aucunes données concrètes ne supportent. Pour garantir l'enjeu de sécurité, les contrôleurs procèdent à des ajustements des normes (7 NM pour les conflits au lieu de 5) comme référence dans leurs modes de gestion. L'ensemble de ces caractéristiques ne permet donc pas aux contrôleurs de procéder à une gestion fine des conflits.

Les contrôleurs procèdent à des séparations du trafic par des structures ordonnées comme des ordres de passage d'avions sur des balises, des sens relatifs de déviation, des temps. Des procédures de détermination de la distance de séparation de deux vols sont utilisées pour affiner l'écart d'estimées, mais elles sont très lourdes. La gestion s'opère dans un contexte particulier. Le temps de transit des vols est court (de l'ordre d'une vingtaine de minutes). Les contrôleurs disposent de peu de temps. Les décisions sont donc prises de façon réactive et très peu spéculative. Les contrôleurs ont à gérer plusieurs conflits simultanément à des états d'avancement différents, ainsi que d'autres tâches comme le transit simple de vols, les demandes des autres secteurs et la surveillance globale. Bien que différenciées, les étapes de gestion s'enchaînent rapidement. On notera également qu'il existe de nombreuses possibilités de résoudre un conflit. Des choix sont opérés à différents niveaux dans les décisions. Cependant, l'extrême rapidité des situations conduit à ce que ces possibilités et choix ne soit pas explicités.

Compte tenu de l'ensemble de ces caractéristiques, la gestion de conflits s'apparente à une situation d'aide à la décision que PAD peut représenter de la façon suivante :

- pour la situation problématique :
  - $\mathcal{A}$: contrôleur radar,
  - $\mathcal{O}$: sécurité, régularité (appelés énergie et élégance par les contrôleurs à certains moments),
  - $\mathcal{S}$: identifier le rapprochement de deux vols sous la distance de séparation, place du conflit dans le reste du trafic,
- pour la formulation du problème :
  - $\mathbb{A}$: modifications possibles d'un paramètre de vol d'un des deux vols en conflit,
  - $\mathbb{V}$: plans de vols, estimées balise, distance de séparation,
  - Π : description (ajustement de la situation),
- pour le modèle d'évaluation :
  - A : instant d'application de l'ordre, modification de cap, modification de niveau,
  - D : temps avant impact, sorti du secteur ou arrivée du vol, valeurs de cap multiples de 5° entre 5 et 60°, niveaux adjacents,
  - E : intervalle de temps jusque l'invalidité de l'ordre,
  - H : D,

$\mathcal{U}$: norme de pratique (7 NM), ordre de maintien de trajectoire ou d'écartement sur le trafic environnant,

$\mathcal{R}$: résolution au plus tôt,

- pour la recommandation finale : remise sur route (restauration du paramètre de vol modifié pour replacer l'avion sur sa route).

Compte-tenu de la structuration du trafic et de la façon donc les contrôleurs résolvent les conflits, seule la remise sur route relève d'une réaction et non d'une décision. Des aides spécifiques peuvent donc être envisagées pour les trois décisions de gestion. Nous nous focalisons ici sur les situations problématiques $\mathcal{P}_d$, $\mathcal{P}_r$ et $\mathcal{P}_o$ et les formulations du problème Γ_d, Γ_r et Γ_o.

Pour la détection de conflits, $\mathcal{P}_d$ et Γ_d sont définies par :

- $\mathcal{A}_d$: contrôleur radar,
- $\mathcal{C}_d$: sécurité (avion en conflit avec un autre),
- $\mathcal{S}_d$: identification d'un écart insuffisant d'estimées sur la balise commune aux deux routes sur le tableau de strips, identification des routes et balises communes et détermination de la distance de séparation sur la vue radar,
- $\mathbb{A}_d$: représentation du conflit par le triplet avions et balise du conflit, écart d'estimées, distance de séparation,
- $\mathbb{V}_d$: distance de séparation des avions en conflit,
- Π_d : description.

Pour la résolution de conflits, $\mathcal{P}_r$ et Γ_r sont définies par :

- $\mathcal{A}_r$: contrôleur radar,
- $\mathcal{C}_r$: énergie, élégance (déclinaison des enjeux de sécurité et de régularité),
- $\mathcal{S}_r$: identification de l'ordre de passage des deux avions en conflit sur la balise commune sur le tableau de strips, identification du trafic environnant sur la vue radar, sélection de la résolution naturelle la plus adaptée à la situation,
- $\mathbb{A}_r$: représentation des avions influençant les résolutions naturelles,
- $\mathbb{V}_r$: distances de séparation inférieure à une norme secondaire des avions influençant les résolutions naturelles,
- Π_r : rangement, tri, description.

Enfin, pour l'ordre de résolution, $\mathcal{P}_o$ et Γ_o sont définies par :

- $\mathcal{A}_o$: contrôleur radar,
- $\mathcal{C}_o$: élégance,
- $\mathcal{S}_o$: écart d'estimées à la balise du conflit de l'avion dévié sur le tableau de strips, trafic environnant sur la vue radar, utilisation des règles de construction de l'ordre, norme de séparation personnelle, annotations de strips,
- $\mathbb{A}_o$: présentation du premier ordre produisant la séparation, représentation de l'ordre sur la route, mise en œuvre automatique de l'ordre,
- $\mathbb{V}_o$: écart à la balise du conflit, distance de séparation produite par l'ordre
- Π_o : rangement, description.

Les contrôleurs sont favorables à un outil de calcul de la distance de séparation qui allégerait les procédures qu'ils utilisent peu. Un tel outil relève d'une détection automatique de conflits. Comme cette activité est déterminante dans la gestion des conflits, il est donc raisonnable de supposer qu'un tel outil réduirait à terme les aptitudes des contrôleurs. Cependant, cette demande est cohérente avec la gestion actuelle. Le contrôleur organique, par le suivi des activités du contrôleur radar, constitue déjà une possibilité de sous-traitance

potentielle de résolution de conflits. Pourtant, les deux contrôleurs ne se coordonnent pas de cette manière. Le contrôleur organique représente donc une capacité de critique que le contrôleur radar peut solliciter. La résolution automatique doit donc être vue selon une logique de critique de l'analyse d'un conflit faite par le contrôleur radar. Par ailleurs, du fait des fortes et rapides variations de trafic, cette détection peut aussi être vue selon un principe de redondance complémentaire au contrôleur organique. Enfin, une utilisation en sous-traitance d'une détection automatique de conflits ne serait pas cohérente avec la logique plus réactive que spéculative de la gestion de conflits. Une telle utilisation détournée ne peut permettre une perte de compétences à long terme, dans la mesure où elle serait opérationnellement inefficace car elle ferait perdre au contrôleur la maîtrise de la situation. Au sujet des investigations expérimentales conduites dans SPECTRA et AMANDA, on notera qu'une détection était implicitement présente dès lors qu'une résolution automatique est proposée. Les conditions expérimentales ne permettaient donc pas de mettre en évidence la demande d'une détection automatique de conflits, ni d'étudier ses modalités de mise à disposition. Par extension, la mise à disposition d'un outil permettant d'estimer précisément les effets d'un ordre est cohérente avec les procédures de gestion actuelles. Un tel outil ne peut se substituer aux capacités du contrôleur mais lui permet d'affiner sa gestion.

Sous la représentation par PAD, un certain nombre de particularités apparaissent clairement. Une transition de l'enjeu de sécurité à l'enjeu de régularité apparait dans le passage de $\mathcal{O}_d$ à $\mathcal{O}_r$ puis à $\mathcal{O}_o$. L'introduction de l'enjeu de sécurité dès $\mathcal{P}_d$, puis de l'enjeu de régularité dans $\mathcal{P}_r$, montre une importance plus grande de l'enjeu de sécurité. Ces deux enjeux sont donc ordonnés dans $\mathcal{P}_r$, ce qui introduit une idée de préférences nominales dans l'aide à la décision. Un enjeu de surveillance est présent en permanence dans le contrôle aérien. Il n'est pas évoqué ici. Il se retrouve donc dans les étapes de gestion des conflits. Les conflits constituent donc des facteurs structurant de cette surveillance que l'aide à la décision peut accroître. Les actions potentielles considérées dans chacun des ensembles A_d et A_o contribuent toutes à la décision. Elles sont donc fragmentaires (Roy, 1996). Dans $\mathcal{P}_o$, la sécurité est garantie depuis $\mathcal{P}_d$. et $\mathcal{P}_r$. Seule la régularité du trafic est à considérer, sous sa déclinaison d'élégance. A cette étape de gestion, une logique d'optimisation peut alors être envisagée ultérieurement. Les engagements pris dans $\mathcal{S}_d$, $\mathcal{S}_r$ et $\mathcal{S}_o$ s'opèrent tous selon le même ordre : le tableau de strips puis la vue radar.

Le contrôleur radar étant le seul participant, les engagements sont pris vis-à-vis de lui-même. Les actions potentielles lui permettent d'accentuer ses engagements. De ce fait, les problématiques de description occupent une place importante dans l'aide à la décision. Des besoins en matière de formation et de maîtrise globale des situations sont également révélés. La mise en cohérence avec l'existant leur donne une importance particulière. La description ne peut être opérationnellement supportée que par la vue radar car elle est cohérente avec les besoins de représentations de regroupements d'avions exprimées par les actions potentielles. Les strips représentent fondamentalement des parties de plans de vols qu'il est difficile de mettre en cohérence avec des exigences opérationnelles ; l'importante activité de la tenue du tableau de strips en témoigne. Par ailleurs, la forme concrète prise par les strips et le tableau s'accommode mal d'un accroissement de la quantité de données et d'une représentation de groupes d'avions. Les problématiques de rangement s'appuient sur les éléments ordonnés implicitement manipulés par les contrôleurs (temps avant croisement, valeur de déviation, ordre de croisement). Ces ordres permettent d'envisager des possibilités supplémentaires dans les décisions que les contrôleurs ne manipulent pas directement mais par l'intermédiaire

d'ajustements sur les normes de séparation. L'accroissement des possibilités ainsi manipulées permettent au contrôleur de produire une gestion plus fine des conflits, sans changer leurs modes de décision. Il en va de même de la problématique de tri qui concerne le regroupement de vols (en conflit, interférant). Dans Π_r et Π_o, plusieurs problématiques apparaissent. Leur ordre est important. En effet, la description s'appuie sur une exploitation des problématiques précédentes. Dans Π_r, le tri s'appuie sur la situation mise en évidence par le rangement.

6 Discussion

L'opérationnalisation de PAD montre que la mission de l'analyste ne peut se restreindre à la seule connaissance de quelques méthodes ou cadres formels, tels que ceux de la théorie de l'aide à la décision ou des systèmes d'information. Analyse et modélisation relèvent de deux métiers différents pour lesquels PAD peut apporter un support de communication évitant une confusion entre aide à la décision et modèle. Par ailleurs, la différenciation des sous-artefacts de la situation problématique conduit à ce que l'analyste présente des aptitudes à faire comprendre ses recommandations selon qu'elles concernent des enjeux, des participants de plusieurs natures ou des engagements. Cette remarque met également en exergue que l'analyste doit être à même de saisir le problème pour construire une situation problématique et une formulation de problème représentatives d'une réalité donnée. Le terme même d'analyste n'apparait alors pas correspondre complètement à sa mission. Dans cet esprit, nous préférons également utiliser les termes d'opérationnalisation ou de déploiement de PAD sur un terrain, plutôt que d'application. Le terme application laisse trop sous-entendre que l'aide à la décision ne serait limitée qu'à la construction d'un modèle dans un cadre théorique, de la décision ou des systèmes d'information, particulier.

La partie fondamentale de la mission de l'analyste concerne la construction de la situation problématique et de la formulation du problème. Le cadre de la gestion de conflits montre que la construction de $\mathcal{P}_d$, $\mathcal{P}_r$ et $\mathcal{P}_o$ puis par Γ_d, Γ_r et Γ_o résulte d'une analyse de la gestion. Deux remarques s'imposent :

- La construction de la situation problématique et de la formulation du problème nécessite une validation. Le modèle d'évaluation n'apporte qu'une partie de cette validation car il est dédié à la formulation du problème. PAD offre une représentation extérieure du processus décisionnel de gestion de conflits qui constitue un modèle d'évaluation de la situation problématique démontrant sa solide constitution. Ce modèle d'évaluation est donc qualitatif. Des cadres normatifs peuvent servir à de telles évaluations. Enfin, puisque qu'il y a évaluation de la situation problématique pour une construction argumentée de la formulation du problème, des recommandations peuvent être émises avant les résultats produits par $\mathcal{M}$.

- PAD peut servir de base au déploiement d'autres PAD. Il s'agit ici d'une représentation concrète d'une particularité de l'aide à la décision qui nécessite de prendre des décisions. Cette descente récursive dans la décision constitue une piste prometteuse pour l'aide à la décision ainsi que l'évoque (Pomerol, 2010). Plus précisément, les trois situations problématiques et formulations du problème affermissent les étapes du processus décisionnel de gestion de conflits. Cet affermissement qualitatif ne modifie en rien la structure de la gestion de conflits. Il permet un accroissement de maturité ou de rationalité procédurale. L'évaluation numérique par $\mathcal{M}$ permet une construction plus fine des décisions levant des

limites d'appréciation du décideur. $\mathcal{M}$ permet donc un accroissement à la rationalité limitée (Simon, 1997) ('à' et non 'de', car il s'agit de dépasser les limites et non de les augmenter).

Vue par le seul cadre de modélisation de la solution, l'aide à la décision est fortement orientée résolution de problème. De ce fait, l'aide à la décision est pratiquée comme une décision avec un modèle ce qui conduit à minimiser ou nier les décisions en place. Ainsi, le client et le décideur sont souvent confondus et le point de vue du client est considéré comme significatif. La séparation du client et du décideur offre un regard plus précis et permet d'envisager l'aide à la décision comme la levée de verrous dans des décisions en place. Une critique du point de vue du client peut donc plus facilement s'installer. En particulier, le client peut avoir une conviction telle de son problème qu'il exprime en fait des solutions. Ainsi, l'introduction d'une détection automatique de conflits a toujours été considérée comme étant potentiellement une source de dégradation des compétences. L'analyse du processus décisionnel de gestion de conflits conduite auprès des contrôleurs montre le contraire.

On remarque la forte présence de la problématique de description. L'ensemble des remarques précédentes la met en exergue. L'accroissement offert à une rationalité en place nécessite une représentation décrivant concrètement cette rationalité. Par ailleurs, dans le cas du contrôle aérien, le contrôleur est le seul décideur à gérer les conflits. Pour qu'il puisse respecter ses engagements, il est nécessaire qu'il dispose de descriptions correctes des situations qu'il gère. L'aide à la décision peut lui permettre d'accroitre son efficacité en disposant de représentations des effets de ses décisions. Ces remarques montrent donc que le socle de l'aide à la décision réside dans les engagements. Les problématiques décisionnelles portent donc plus sur les relations au sein des points de vue, relations issues des participants et enjeux, qu'entre actions potentielles.

7 Conclusion

Dans ce papier, nous nous sommes intéressés à une orientation méthodologique de l'aide à la décision. Cette orientation porte un regard pratique sur l'aide à la décision afin de construire des modèles qui répondent de façon plus appropriée aux problèmes réels rencontrés par des décideurs. De ce fait, elle ne se veut pas concurrente mais complémentaire des orientations de la théorie de la décision et des systèmes d'information.

Le Processus d'Aide à la Décision fournit un outil opérationnel permettant de guider la construction d'une aide à la décision. Par la définition de quatre artéfacts (situation problématique, formulation du problème, modèle d'évaluation et recommandation finale), PAD favorise la construction d'aides à la décision correctement ancrées dans la réalité du problème posé par un demandeur.

Nous avons présenté une illustration du déploiement de PAD dans le domaine du contrôle de trafic aérien. Nous nous sommes focalisés sur l'analyse du problème afin de mieux définir les aides à la décision possibles sans aller jusque l'évaluation de ces solutions. Néanmoins, l'analyse proposée tient compte des différentes études d'automatisation dans le domaine et les repositionne au sein des objectifs et de l'organisation du contrôle aérien.

A la lumière de ce déploiement, nous soulevons plusieurs questions sur la nature de l'aide à la décision et les aptitudes que doit présenter l'analyste, aptitudes plus large que la maîtrise de la théorie de la décision ou les systèmes d'information, en particulier en ce qui concerne la communication avec le client et le décideur. L'élargissement de l'évaluation à des modèles

qualitatifs apparaît également être un complément intéressant à l'évaluation sur base de modèles numériques, ce qui permet d'envisager l'aide à la décision comme des accroissements face aux rationalités limitée et procédurale. Enfin, plus largement, un retour sur la pratique apparait être une composante fondamentale d'une orientation méthodologique de l'aide à la décision.

Bibliographie

Annebicque, D., I. Crévits, T. Poulain, S. Debernard, P. MIllot (2012) Decision Support Systems for Air Traffic Controllers based on the Analysis of their Decision-Making Processes. *International Journal of Advanced Operations Management*. In press.

Bouyssou, D., T. Marchant, M. Pirlot, A. Tsoukiàs, P. Vincke (2006) *Evaluation and Decision Models: stepping stones for the analyst*, Springer Verlag, Berlin.

Debernard, S., B. Guiost, T. Poulain, I. Crévits, D. Annebicque, P. MIllot (2009) Integrating Human Factors in the Design of Intelligent Systems: an example in Air Traffic Control. *International Journal of Intelligent Systems Technologies and Applications 7 (2)*, 205-226.

Pacaux-Lemoine, M.P., S. Debernard, I. Crévits, P. MIllot (1996) Cooperation between humans and machines : first results of an experimentation of a multi-level cooperative organisation in air traffic control. *Computer Supported Cooperative Work 5 (2)*, 299-321.

Pacaux-Lemoine, M.P., S. Debernard (2002) A Common Work Space to support the Air Traffic Control. *Control Engineering Practice, A Journal of IFAC 10 (5)*, 571-576.

Pomerol, J.C (2009) Human decision; Recognition plus reasoning, dans *Decision-Making Process: Concepts and Methods*. Wiley.

Roy, B (1996) *Multicriteria Methodology for Decision Aiding*. Kluwer Academic Publishers, Dordrecht.

Simon, H.A (1976) From substantial to procedural rationality. dans Latsis S.J. (dir.), *Method and Appraisal in Economics*, Cambridge University Press, Cambridge, 129-148.

Simon H.A (1997) *Administrative Behavior* (4ième édition augmentée; première édition 1947), The Free Press, N.Y.

Stamelos I., A. Tsoukiàs (2003) Software evaluation problem situations, *European Journal of Operational Research 145*, 273-286.

Tsoukiàs, A (2007) On the concept of decision aiding process. *Annals of Operations Research 154*, 3-27.

Vanderhaegen, F., I. Crévits, S. Debernard, P. MIllot (1994) Human-machine cooperation : toward an activity regulation assistance for different air traffic control levels. *International Journal on Human-Computer Interaction 6 (1)*, 65-104.

Tsoukiàs A., H. Ralijaona (2012) Rural Road Maintenance in Madagascar: the GENIS project, in R. Bisdorff, L. Dias, V. Mousseau, M. Pirlot (eds.), *Evaluation and Decision Models: real case studies*, to appear with Springer-Verlag, Berlin.

Ostanello A., A. Tsoukiàs (1993) An explicative model of 'public interorganizational interactions. *European Journal of Operational Research 70*, 67-82.

Simulation de cas d'accidents pour l'aide à la décision. Application à la sécurité des systèmes de transport ferroviaires automatisés.

Lassaâd Mejri*,**, Ahmed Maalel*
, Henda Hajjami Ben Ghezela*

** Laboratoire RIADI-GDL, Ecole Nationale des Sciences de l'Informatique, Tunis*
mejrilassad@yahoo.fr, maalel.ahmed@gmail.com, hhbg.hhbg@gmail.com
*** Institut Français des Sciences et Technologies des Transports, de L'Aménagement et des Réseaux. Marne-La-Vallée, France*

Résumé. Ce papier présente une recherche visant le développement d'un système d'aide à la décision concernant l'homologation de systèmes de transport ferroviaires automatisés. L'objectif est de mettre en œuvre une méthode d'évaluation du degré de conformité du système de transport automatisé à un ensemble de normes de sécurité par **la simulation des scénarios d'accident**. Pour atteindre cet objectif, nous avons envisagé une **approche Rex (Retour d'expérience)** qui tire les enseignements des accidents/incidents vécus et/ou imaginés par les experts de l'analyse de sécurité de l'INRETS (Institut National de Recherche sur les Transports et leur Sécurité) : actuellement IFSTAR. Notre approche consiste à proposer une aide à la décision du coté des experts de la certification en se basant sur une réutilisation des scénarios d'accidents déjà validés historiquement sur d'autres systèmes de transport homologués. Cette approche Rex est de grande utilité vu qu'elle offre aux experts une classe de scénarios d'accidents similaire au nouveau cas traité et se rapprochant au contexte du nouveau cas. Le **Raisonnement à Partir de Cas** est alors exploité comme mode de raisonnement par analogie permettant de sélectionner et remémorer un sous ensemble de cas historiques pouvant aider à la résolution du nouveau cas présenté par les experts.

1 Introduction

Un scénario d'accident ou d'incident est vu comme une évolution dynamique d'une situation habituelle du système de transport jusqu'à atteindre une situation à risque. Cette situation indésirable sur le plan de la sécurité est ensuite exploitée pour renforcer la sécurité du système en agissant sur les fonctions de sécurité et les automatismes adéquats. Les experts de la certification sont tenus à étudier le système de transport en vue de se prononcer sur sa conformité aux consignes reconnus de la sécurité ferroviaire et se doivent de décider de lui accorder un agrément provisoire ou définitif avant sa mise en site. De l'autre coté les constructeurs du système de transport mettent à contribution le rapport d'étude de la sécurité fourni par les experts en vue d'améliorer la sécurité de leurs systèmes de transport. La simu-

lation des accidents est une activité essentielle des experts qui met à contribution leur esprit d'analyse et de synthèse. Une analyse des différents composants du système de transport et de leurs modes de défaillances permet d'envisager les causes de l'accident sous forme de pannes résumées du système. **Ces pannes résumées** injectées dans le modèle de Pétri permettent aux experts de confectionner des scénarios d'accidents. L'ensemble des scénarios proposés par les experts permettent d'avoir une idée sur le degré de conformité du système de transport proposé par le constructeur aux normes et consignes de la sécurité. Actuellement une maquette de faisabilité de l'approche à partir de cas est en cours. Cette maquette utilise **une ontologie du domaine du transport** permettant d'unifier le vocabulaire employé en matière d'analyse de sécurité ferroviaire.

2 Notion de scénario d'accident

Un scénario d'accident représente une situation d'insécurité au niveau du système de transport automatisé. Cette insécurité, jugée, potentielle devrait être résolue par la proposition d'une solution à adopter pour pallier au risque qu'elle représente (exemple : collision entre deux rames de métro, déraillement de la voie ou électrocution …etc.). Un scénario est un concours de circonstances pouvant mener à un danger. Il est décrit conjointement par des attributs de situation qui donnent une idée sur l'insécurité et des attributs de solution à adopter pour anéantir ou réduire le risque d'insécurité. On recense plusieurs paramètres qui rendent compte du risque / des fonctions de sécurité / des acteurs / de la zone géographique / des pannes touchant une partie du système / des solutions adoptées pour anéantir le risque (voir figure 1).

FIG. 1 : Les paramètres de la description statique d'un scénario d'accident.

3 Travaux antérieurs

Ce problème d'analyse de sécurité a été déjà traité selon une première approche lors de nos travaux de thèse, Mejri (1995). Le travail accompli avait abouti d'abord à la réalisation d'une

maquette de faisabilité, Mejri (1995) et à la mise au point d'un système d'acquisition et de classification des scénarios d'accidents baptisé ACASYA (figure 2) Hadj mabrouk et Mejri, (1998). Dans l'objectif d'aider les experts dans leur activité d'évaluation de scénarios d'accidents, ACASYA s'articule autour de trois étapes complémentaires :

- **Classer** d'abord le scénario d'insécurité dans une classe prédéfinie à l'aide **d'un algorithme de classement**. Chaque classe de scénarios a été caractérisée par une description en termes des descripteurs les plus pertinents de la classe.
- **Evaluer** ensuite le scénario proposé par le constructeur en référence à la classe d'appartenance trouvée dans l'étape précédente afin de restreindre l'espace d'exploration. L'évaluation consiste à tester la complétude et la cohérence d'un scénario d'insécurité.
- **Générer** de nouvelles situations d'insécurité en mettant à contribution les éléments de description superflus et/ou manquants détectés dans l'évaluation par EVALSCA.

Les premiers résultats d'évaluation de nos travaux de recherche appliqués au domaine de l'analyse de sécurité, Mejri, (1995), nous ont permis de chercher ainsi à identifier un modèle d'acquisition de connaissances plus générique Mejri & Caulier, (2005) et une démarche plus générale pour la résolution de problèmes, Mejri & al (2009).

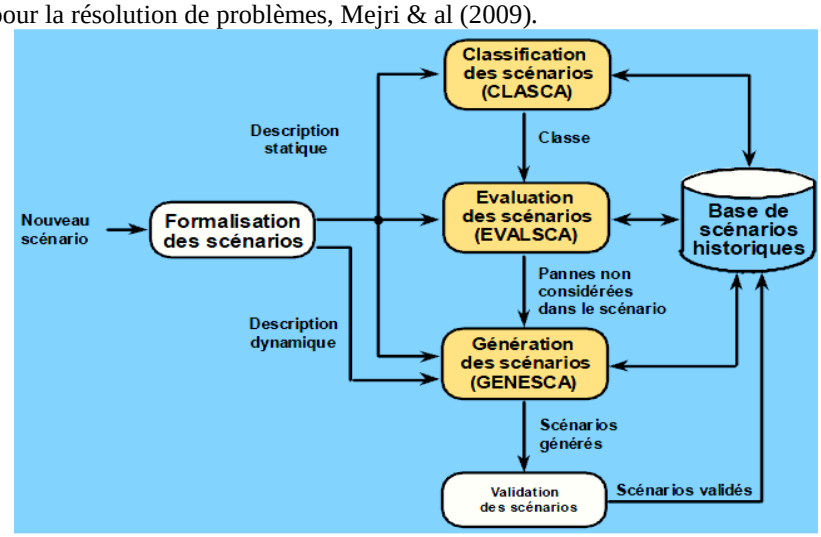


FIG. 2 - Architecture du système ACASYA, HADJ-mabrouk & al. (1994,1998), MEJRI (1995)

4 Nos orientations de recherche actuelles

Dans les travaux antérieurs plusieurs limites ont été décelées et méritent une attention particulière pour nos travaux actuels :

- ACASYA s'est limité à exploiter la description statique du scénario d'accident. Bien que cette description est apparentée aisément à un exemple d'apprentissage (fiche de paires

<attribut/valeur>), elle ne permet pas à elle seule de rendre compte de la richesse d'un accident qui fait intervenir plusieurs acteurs matériels / logiciels et humains et implique plusieurs automatismes et fonctions de sécurité, etc. **Il serait alors naturel de tirer profit de l'ensemble des autres formes de description de scénario : description dynamique, description textuelle et description graphique**.

- L'aspect statique des scénarios d'accident représente le contexte général du problème posé par l'accident ainsi que les mesures préventives ou correctives (les solutions) adoptées. Mais il ne donne aucune idée sur le déroulement du scénario dans le temps et dans l'espace. C'est seulement la description dynamique qui se charge de cela. **Cette description dynamique se prête bien à la simulation.**
- Le langage de description utilisé dans le volet statique des scénarios d'accident manque une cohérence globale et ne se réfère à aucune terminologie ou vocabulaire bien défini dans le domaine. Il est alors opportun de compléter ce langage en vue de **mettre en place une ontologie du domaine** qui dresserait tous les concepts utilisés, les relations entre les concepts, les instances de ces concepts dans le domaine. Cette ontologie serait un cadre de référence pour décrire les scénarios d'accidents et pour mesurer leur complétude et pertinence du point de vue syntaxique et sémantique.
- ACASYA opère par classification du scénario d'accident proposé par le constructeur pour rendre compte de son acceptabilité en référence à un ensemble de classes de scénarii prédéfinis par les experts. ACASYA induit par apprentissage pour chacune des classes prédéfinies **une description caractéristique de la classe** sous la forme <attribut/valeur/fréquence d'apparition dans la classe>. Le nouveau scénario (description statique) est alors rapproché selon une mesure de similarité à la description d'une classe de scénarii : celle pour laquelle il représente le meilleur score de similarité en termes d'attributs/valeurs communs. Il serait plus opportun de parcourir les différents scénarii de la classe en question en vue de mesurer la similarité inter-scénarii en vue de repérer le meilleur scénario d'accident historique (le plus similaire au nouveau scénario) stocké dans la base. Ceci s'apparente plutôt à un **Raisonnement à partir de Cas** et non à un mécanisme purement inductif. En effet, nous considérons le RàPC comme **le moyen du Retour d'expérience (Rex)** dans le domaine de l'analyse de sécurité.

Les sections suivantes vont être alors consacrées à présenter le modèle du cas tel que nous l'envisageons ainsi qu'aux travaux entrepris dans le groupe pour mettre en place l'ontologie du domaine. Ce modèle de cas envisagé intègre la description dynamique et favorise la simulation de cas d'accidents pour notamment rendre compte de l'analyse de sécurité. La section finale sera réservée à la présentation de la démarche générale à partir de cas adoptée pour le renforcement de l'analyse de sécurité et aider à la décision des experts.

5. Le Raisonnement à Partir de Cas (RàPC)

Un "**cas**" est une représentation structurée d'une histoire composée passée ou imaginée Aamodt, (1991). Selon Kolodner (1993), un cas peut être défini comme un ensemble de connaissances contextuelles enseignant une leçon. Un Processus RàPC passe par les quatre phases suivantes :

- **Phase de remémoration** : Suite à l'élaboration d'un cas cible, une recherche est effectuée dans la base de cas afin de trouver ceux susceptibles de résoudre le problème. L'élaboration d'un cas cible est la phase préliminaire d'acquisition d'un nouveau problème par la saisie des valeurs de ses attributs. Il reste alors de trouver la solution par RàPC.
- **Phase d'adaptation** : Cette phase est indispensable car le cas remémoré n'est jamais strictement identique au nouveau cas. Elle consiste à modifier la solution du cas remémoré pour prendre en compte les différences entre spécifications de problèmes. Elle demande des connaissances d'adaptation spécifiques qui ne sont pas toujours aisés à modéliser ni à mettre en œuvre.
- **Phase de révision** : Après sa génération par le système, la solution proposée est testée pour être validée lors de la phase de révision qui est généralement externe au système.
- **Phase de mémorisation ou d'apprentissage** : Cette phase du cycle RàPC tente d'améliorer les connaissances à l'origine des échecs rencontrés par la solution et d'enrichir la base de cas avec les nouveaux cas résolus.

Trois modèles RàPC de représentation des cas existent en littérature :

- Le **modèle structurel** dans lequel toutes les caractéristiques importantes pour décrire un cas sont déterminées à l'avance par le concepteur du système. La similarité entre deux cas est mesurée en fonction de la distance entre les valeurs de mêmes attributs. La similarité globale entre deux cas est habituellement évaluée par une somme pondérée de la similarité de chacun des attributs. Tous les travaux sur l'adaptation de cas sont menés dans le cadre du modèle structurel.
- Le modèle **conversationnel**, comme son nom l'indique, mise sur l'interaction entre l'usager et le système (d'où la notion de "conversation") pour définir progressivement le problème à résoudre et pour sélectionner les solutions les plus appropriées. Dans ce schéma, l'interaction entre le système et l'usager se fait comme suit :

  - ✓ L'usager fournit une brève description textuelle du problème et le système calcule la similarité entre cette description et la section "problème" des cas. Le système propose alors une série de questions.
  - ✓ L'usager choisit les questions auxquelles il souhaite répondre. Pour chaque réponse fournie par l'usager, le système réévalue la similarité de chacun des cas.

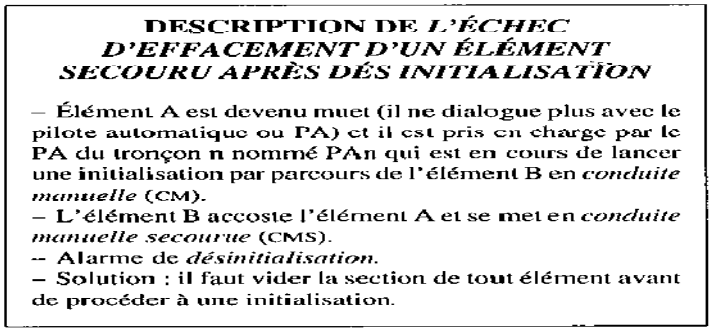
- Le modèle **Textuel**, dans lequel les cas sont non-structurés ou semi-structurés. Ils sont non-structurés si leur description est complètement "free-text". Ils sont semi-structurés lorsque le texte est découpé en portions étiquetées par des descripteurs tels que "problème", "solution", etc. Le RàPC textuel diffère de celui structurel dans lequel les textes sont tout simplement des chaînes de caractères.

6. Notion de cas d'accident

Les experts de la certification utilisent le terme scénario pour exprimer un cas d'accident. Cette notion se trouve plus riche que le concept de cas qu'on retrouve en littérature. En effet, un cas d'accident regroupe ensemble trois vues descriptives complémentaires de l'accident :

6.1. La description textuelle versus modèle textuel du cas

Il s'agit d'un texte qui explique le déroulement de l'accident. A titre d'exemple, nous donnons la description fournie par les experts pour le scénario N°34 de la base.



DESCRIPTION DE *L'ÉCHEC D'EFFACEMENT D'UN ÉLÉMENT SECOURU APRÈS DÉS INITIALISATION*

– Élément A est devenu muet (il ne dialogue plus avec le pilote automatique ou PA) et il est pris en charge par le PA du tronçon n nommé PAn qui est en cours de lancer une initialisation par parcours de l'élément B en *conduite manuelle* (CM).
– L'élément B accoste l'élément A et se met en *conduite manuelle secourue* (CMS).
– Alarme de *désinitialisation*.
– Solution : il faut vider la section de tout élément avant de procéder à une initialisation.

FIG. 3: Description textuelle d'un scénario d'accident.

La notion d'élément dans l'exemple s'apparente à un train sans conducteur. Cette description s'apparente au modèle textuel du cas dans lequel, les cas textuels sont non-structurés. Leur description est complètement en "free-text". Cependant, il est possible dans ce texte de repérer la partie problème dans les phrases 1) 2) et 3) et la partie solution dans 4).

6.2. La description statique versus modèle structurel du cas

C'est un ensemble de paramètres descriptifs du scénario sous forme d'une fiche <Attribut / Valeurs>. On recense alors à ce niveau plusieurs paramètres caractéristiques qui rendent compte de la description de l'accident. La description statique s'apparente bien **au modèle structurel du cas** et a été fortement utilisée dans le système ACASYA.

6.3. La description dynamique versus modèle conversationnel du cas

C'est une vue dynamique du déroulement séquentiel dans le temps et dans l'espace de l'accident. Cette représentation fait appel aux réseaux de Pétri comme formalisme dans lequel les places correspondent aux états, les transitions aux possibilités d'évolution d'un état à un autre. Cette description est apparentée au modèle conversationnel du cas. Dans le modèle structurel, l'usager doit avoir à priori une bonne idée de tous les facteurs pouvant influencer la résolution de son problème. Toutefois pour certains problèmes, il est difficile de déterminer à l'avance les aspects de la situation. Comme son nom l'indique, le modèle conversationnel mise sur l'interaction entre l'usager et le système pour définir progressivement

le problème et pour sélectionner les solutions les plus appropriées. La description dynamique et son exécution à partir d'une situation donnée (jetons placés dans certaines places) ne pourrait se faire que grâce à une conversation entre le système et l'expert en vu de faire évoluer le système pour parvenir à un accident. Le but de l'expert est de générer des scénarios susceptibles de parfaire l'exhaustivité du dossier de sécurité du système de transport ferroviaire.

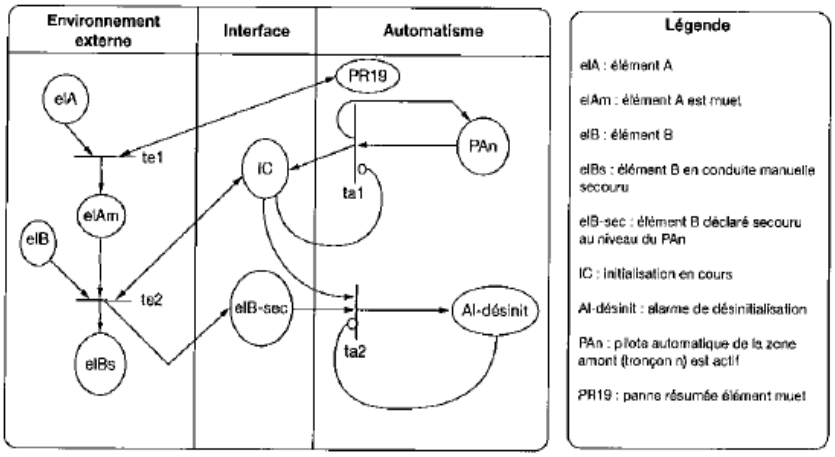


FIG. 4. Exemple de réseau de Pétri du scénario d'accident n° 34

6.4. Le modèle tridimensionnel du cas

Nous pouvons constater qu'un scénario est plus riche qu'un cas vu qu'il regroupe trois modèles de cas : Le modèle textuel, structurel et conversationnel. Dans la littérature, il est souvent question d'adopter un seul de ces trois modèles. Vu l'aspect multiforme du scénario, il est judicieux d'exploiter les trois modèles pour améliorer l'analyse de sécurité et ne pas se limiter à la description statique fortement utilisée dans ACASYA au détriment des autres.

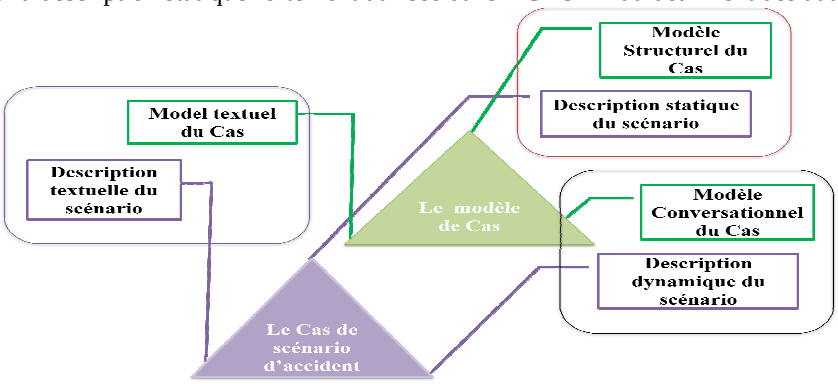


FIG. 5: Le modèle du cas de scénario d'accident.

Simulation de cas d'accidents pour l'aide à la décision.

7. Elaboration d'une Ontologie de domaine

Une ontologie peut être vue comme un treillis de concepts et de relations entre ces concepts destinés à représenter les objets du monde sous une forme compréhensible à la fois par les hommes et par les machines. Une ontologie est constituée des concepts et des relations ainsi que des propriétés et des axiomes (Uschlod et Grüninger, 1996). Nous nous sommes inspirés des travaux de (Abou Assali, 2010) pour proposer le modèle de cas. Un cas dans notre système (Maalel et al 2011) représente un accident ferroviaire.

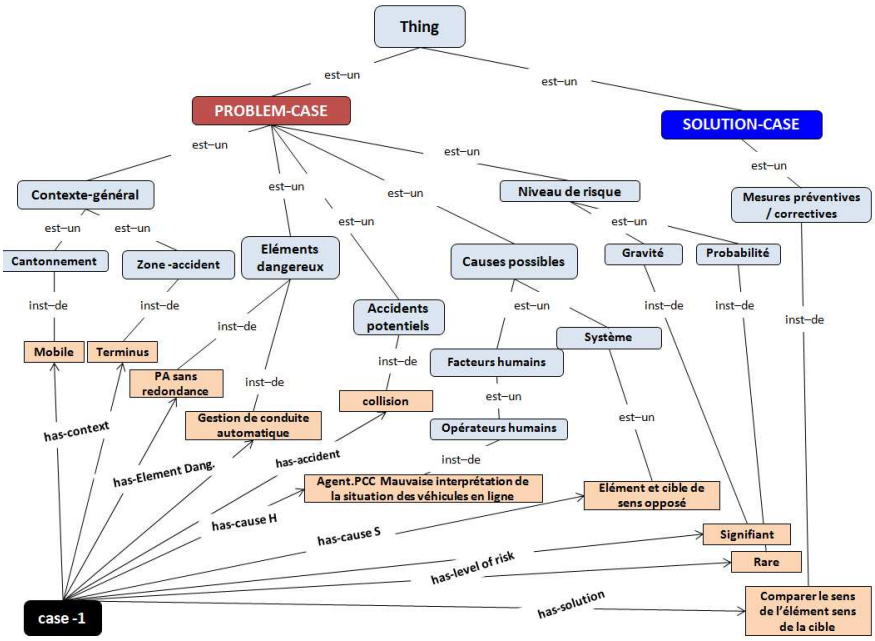


FIG. 6. *Modèle de cas dans l'ontologie du domaine.*

8. Démarche RàPC d'aide à la décision

Notre démarche pour aider les experts dans leur activité d'analyse de sécurité se base sur la réutilisation de cas et l'exploitation d'une ontologie du domaine. Elle vise à assister les experts dans le processus de prise de décision. L'analyse experte se concrétise par l'évaluation des scénarios d'accidents afin de mettre en défaut ceux proposés par les constructeurs ou aussi pour les consolider. ACASYA classifie le nouveau cas proposé par le constructeur. ACASYA opère seulement sur les cas structurels. Alors notre démarche RàPC enchaine en aval d'ACASYA les étapes suivantes :

8.1. Remémoration du cas structurel le plus proche

Il s'agit d'une étape de recherche qui parcourt tous les cas structurels d'accident stockés dans la classe repérée par ACASYA en vu d'identifier le cas structurel le plus proche du nouveau cas constructeur en référence à une mesure de similarité structurelle :

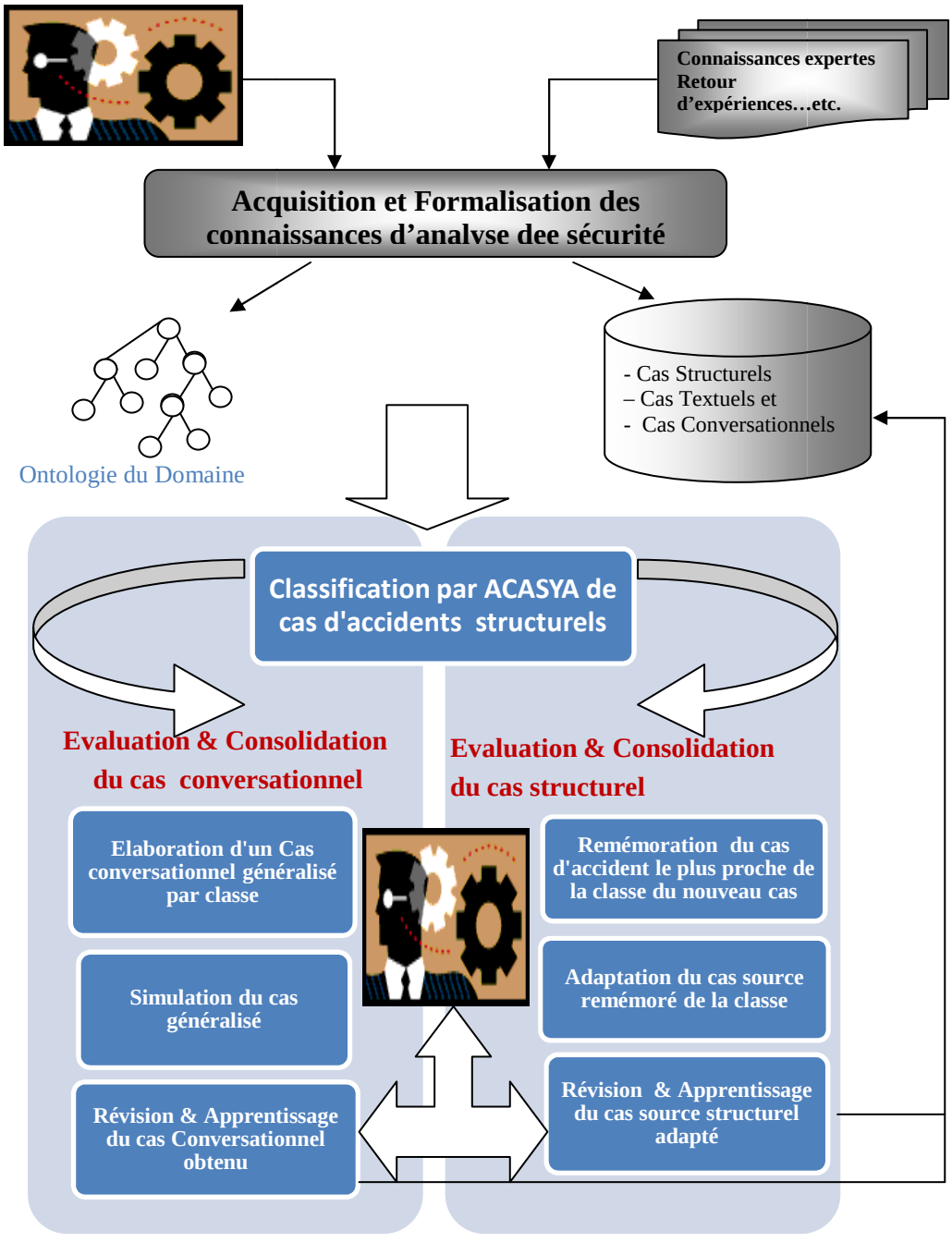


FIG. 7: Démarche de RàPC pour l'évaluation et la consolidation de cas d'accidents.

Simulation de cas d'accidents pour l'aide à la décision.

Sim_globale (Cas_cible, Cas_source) = $\sum_{i=1}^{i=n} Sim_{locale}(A_{cible,i}, A_{source,i}) * Wi$

Avec **Wi** **: Le poids de l'Attribut** $A_{source,i}$ **et** $A_{cible,i}$
Cas_ cible **: représente le cas proposé par le constructeur**
Cas_source **: un cas quelconque de la classe identifiée par ACASYA**
n **: Nombre d'attributs de la partie problème et non la solution**

Sim_locale $(A_{cible,i}, A_{source,i})$ = $\sum_{k=1}^{k=p} \partial ik$

Avec $\partial ik = 1$ *si la valeur* Vk *de* $A_{cible,i} = Vk$ *de* $A_{source,i}$
$\partial ik = 0$ *si la valeur* Vk *de* $A_{cible,i} \neq Vk$ *de* $A_{source,i}$
P : le nombre de valeurs descriptives de l'attribut

On remémore alors le cas structurel présentant le maximum de similarité globale par rapport au cas cible. L'intérêt de cette étape c'est qu'elle restreint l'espace de recherche du cas source aux seuls cas de la classe repérée par ACASYA ce qui représente en soi une avancée par rapport au RàPC classique qui repose sur un parcours exhaustif de tous les cas de la base de cas. Ce qui alourdit énormément le traitement.

8.2.Adaptation du cas remémoré

L'adaptation se réfère à des règles d'adaptation afin de permettre de tirer profit du cas source remémoré et en vu de dégager la solution au nouveau problème représenté par le nouveau cas cible. Nous avons opté à ce niveau à opérer par apprentissage automatique sur l'ensemble des cas structurels de la classe en vue d'induire des règles qui concluent sur les descripteurs de type solution et ayant la forme :

Si descripteurs de type problème Alors descripteurs de type solution

Exemple : Si (Risque = collision)et (Cantonnement=Fixe) et (Zone Accident=Ligne) et (Eléments Dangereux = PA_Sans_Redondance) Alors (Solution=SA4)

Le mécanisme d'apprentissage utilisé est CHARADE, GANASCIA (1990) dont l'intérêt est qu'il permet de structurer les règles d'un type de descripteurs vers un autre.

8.3.Révision et Apprentissage du cas adapté

Après adaptation on obtient la solution du nouveau cas cible structurel. Ce cas obtenu devrait être révisé par les experts en vue d'une validation du résultat et l'enrichissement par apprentissage de la base de cas et des indexes des cas. Dans ce contexte, l'expert pourrait soit rejeter le cas adapté soit le confirmer soit apporter lui-même des rectifications pour compléter la description du cas. L'apprentissage consiste simplement à rajouter d'abord le cas adapté en l'intégrant à une classe existante ou créer une nouvelle classe pour l'accueillir. La description caractéristique de la classe en question devrait alors être actualisée pour tenir compte du nouveau cas adapté.

8.4.Elaboration du cas généralisé

Au niveau conversationnel la démarche est un peu légèrement différente vu que la structure d'un cas conversationnel dans notre contexte d'étude diffère de celui du cas structurel. Un cas conversationnel dans notre cadre regroupe :

- Une partie problème constituée par la **situation initiale du scénario** qui définit un environnement donné susceptible d'évoluer progressivement vers une situation à risque. La partie problème inclut aussi un **ensemble d'opérateurs de transition d'un état à un autre** (ce sont les règles de transition qui correspondent aux transitions du réseau de pétri du scénario). En effet dans la formalisation, nous avons traduit les transitions de pétri en règles de production exprimées en logique de proposition. Chaque cas conversationnel est rattaché alors à un paquet de règles qui lui est spécifique.
- Une partie solution constituée par la **situation critique du scénario** c'est la situation pré finale de l'accident telle que la collision entre deux rames de métro. La partie solution inclut aussi **la trace d'évolution du scénario d'accident** c'est-à-dire l'ensemble des règles appliquées dans l'ordre ainsi que les situations intermédiaires jusqu'à la situation critique.

Un **cas généralisé** est un cas conversationnel obtenu par généralisation des cas conversationnels spécifiques à une classe de scénarios. Il regroupe des règles de transition exprimées en logique des prédicats. Ces règles générales sont obtenues en se basant sur **l'ontologie du domaine** qui illustre des relations de généralisation entre les concepts.

8.5.Simulation du cas généralisé

L'étape de simulation exploite le cas généralisé en vue de retrouver d'autres cas d'accidents non pris en compte par le constructeur. Le simulateur envisagé n'est autre qu'un système expert incluant :

- Une base de règles qui regroupe toutes les règles du cas généralisé ;
- Une base de faits alimentée par la situation initiale du cas conversationnel à tester et ce pour atteindre une séquence de situations intermédiaires jusqu'à obtenir une situation critique. C'est cette situation qui marque la fin de la simulation. L'ensemble des situations du début à la fin constitue un cas conversationnel généré par le système.
- un moteur d'inférence, qui fonctionne dans ce cas là en chaînage avant c'est-à-dire qu'il est guidé par les faits correspondants à la situation initiale. Ce moteur pourrait aussi être invoqué en chaînage arrière mais dans ce cas là il faudrait que la base de faits contienne la situation critique du cas conversationnel à tester. A ce moment, le test d'arrêt serait lorsqu'on arrive à une situation qualifiée d'initiale vu qu'elle n'est pas à risque (non présence de deux rames à proximité sur deux tronçons voisins par exemple)
- Un mécanisme d'injection de Pannes Résumées au fur et à mesure de la simulation à travers un dialogue avec l'expert via l'interface du système. Les pannes résumées sont des défaillances conséquentes qui pourraient altérer le fonctionnement normal du système de transport ferroviaire. Ces pannes on les retrouve dans la description du cas structurel, et on les retrouve aussi impliquées dans la description du cas généralisé. Cependant les insérer automatiquement lors de la simulation n'est pas une bonne solution. D'où, notre approche a consisté à impliquer l'expert du domaine dans le processus de simulation afin

de lui confier cette mission. Il incombe alors à l'expert lors de la simulation de choisir les pannes résumées à insérer dans la phase de simulation en fonction du contexte du scénario d'accident. Ceci nous paraît plus convivial et plus raisonnable et ce afin de confectionner des scénarios réalistes et non imaginaires.

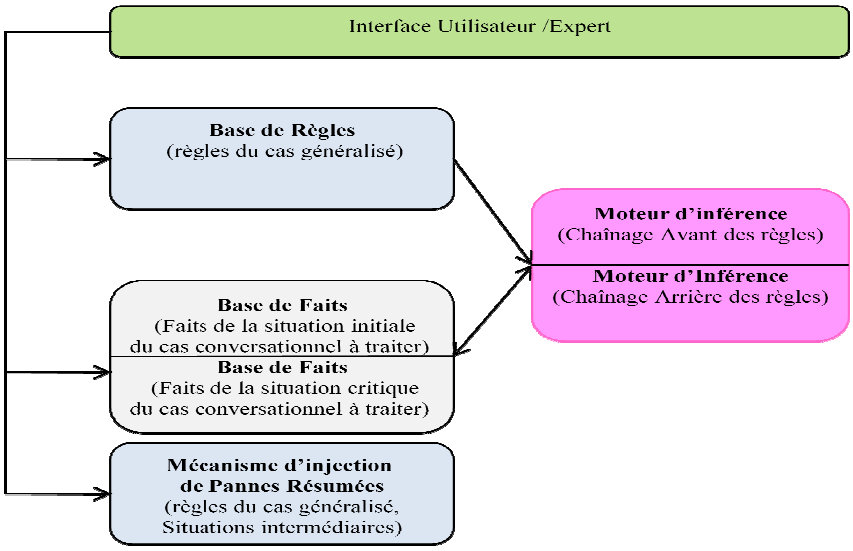


FIG. 8 : Architecture du simulateur (système expert).

7. Conclusion

Nous avons présenté dans cet article les premiers travaux concrétisés de notre approche et notamment l'architecture générale de notre système de RàPC. Cette architecture est composée de deux processus ; *hors-ligne* et *en-ligne*, qui utilisent différents types de connaissances pour la résolution de problèmes. En résultat, le travail présenté dans le cadre de cet article découle de la problématique de gestion de connaissances d'un domaine critique, celui de la sécurité et en particulier, l'accidentologie ferroviaire. Bien que le champ d'application constitue à lui-même une originalité, nous pouvons insister encore sur le caractère générique et l'ouverture du modèle de connaissances. Actuellement, nous travaillons sur l'acquisition des connaissances relatives au raisonnement et à l'amélioration du modèle de connaissances élaboré, en particulier l'ontologie, de manière à l'enrichir par les concepts relatifs à la description dynamique du scénario d'accident. Il est judicieux d'étudier les travaux récents dans le domaine du RàPC qui ont utilisé les ontologies de différentes manières : pour décrire et structurer les cas (Creek (Aamodt 1991)), pour proposer des modèles indépendants du domaine (CBROnto ,Dìaz-Agudo et Gonzàlez-Calero (2002), CCBROnto, Gómez-Gauchía et al. (2006), et Fuchs et Mille (2005) ; pour calculer la similarité sémantique (FAQFinder Burke et al. (1997)), pour traiter l'hétérogénéité des cas : Abou Assali (2010), etc.
Le but est de consolider et renforcer l'approche que nous avons présentée à travers une contribution claire et significative dans ce domaine.

Références

Abou-Assali, A. (2010). *Acquisition des connaissances d¨adaptation et Traitement de l¨hétérogénéité dans un système de ràpc basé sur une Ontologie. Application au diagnostic de la défaillance de détecteurs de gaz*. Thèse de doctorat.

Aamodt, A. (1991). *A Knowledge-Intensive, Integrated Approach to Problem Solving and Sustained Learning*. PhD thesis, University of Trondheim, Norway.

Burke, H., V.A. Kulyukin, S.L. Lytinen, N. Tomuro, et S. Schoenberg (1997). *Question Answering from Frequently Asked Question Files : Experiences with the FAQ Finder System.*. AI Magazine, 18(1) :57–66, 1997.

Dìaz-Agudo, B., et P.A. Gonzàlez-Calero (2002). *CBROnto : a task/method ontology for CBR*". Procs of the 15th International FLAIRS, 2 :101–106.

Fuchs, M. (2005). *Une modélisation au niveau connaissance du raisonnement à partir de cas*. Dans L¨Harmattan, éditeur, Ingénierie des connaissances..

Ganascia, J.G. (1990). *L'âme Machine : les enjeux de l'Intelligence Artificielle*. Le Seuil éd., janvier 1990.

Gómez-Gauchía, B. Díaz-Agudo, et P. González-Calero. (2006). *Ontology-Driven Development,of Conversational CBR Systems.* Dans Advances in Case-Based Reasoning, Proceedings of the 8th European Conference (ECCBR'06), volume 4106, pages 309–324, Turkey. Springer Berlin / Heidelberg. ISBN 978-3-540-36843-4.

Hadj-Mabrouk H. (1998) : *Acquisition et évaluation des connaissances de sécurité des systèmes industriels. Application au domaine de la certification des systèmes de transport guidés*. Thèse d¨Habilitation à Diriger des Recherches, Université de Technologie de Compiègne.

Kolodner J. (1993). *Case-Based Reasoning*. Morgan-Kaufmann Publishers, Inc., 668 pages, 1993.

Maalel A., H. Hadj Mabrouk, L. Mejri, et H. Ben Ghezela (2011). *Development of an Ontology to Assist the Modeling of an Accident Scenario: Application on Railroad Transport*. Journal of Computing, Volume 3, Issue 7, July 2011.

Mejri, L. (1994). *Dossier technique Confidentiel, 70 scénarios d¨accidents dans le transport ferroviaire*. (INRETS-CRESTA CR/A -94-16).

Mejri, L. (1995). *Une démarche basée sur l¨apprentissage automatique pour l¨aide à l¨évaluation et à la génération de scenarios d¨accidents. Application à l¨analyse de sécurité des systèmes de transport automatisés*. Université de valenciennes, 6 décembre 1995.

Mejri, L. , et P. Caulier (2005). *Formalisation of a scenario concept to dynamic problem solving*. 24 th European Annual Conference on Human Decision Making and Manual Control EAM, Athens, 17-19 October 2005.

Mejri, L., H. Hadj mabrouk, et P. Caulier. (2009). *Un modèle générique unifié de représentation et de résolution de problème pour la réutilisation de connaissances. Application à l'analyse de sécurité des systèmes de transport automatisés*. Revue Recherche, Transport Sécurité, RTS N°103.

Ushold, M., et M. Gruninger. (1996). *Ontologies: principles, methods, and applications.* Knowledge Engineering Review, 11(2):93–155.

L'utilisation des systèmes d'information décisionnels dans les collectivités territoriales: premier état des lieux et recherches en cours

Marcel GUENOUN*, Joris PEIGNOT* et Adrien PENERANDA*

*Institut de Management Public et de Gouvernance Territoriale,
21, rue Gaston de Saporta,
13625 Aix-en-Provence CEDEX 1
marcel.guenoun@univ-cezanne.fr
joris.peignot@etu.univ-cezanne.fr
adrien.peneranda@univ-cezanne.fr
http://www.managementpublic.univ-cezanne.fr/

Résumé. La présente communication trouve sa source dans l'écart entre le foisonnement des initiatives visant à l'informatisation du pilotage dans les administrations et le faible nombre d'études qui s'y intéressent. Prenant acte de la montée en puissance des collectivités dans la mise en œuvre de l'action publique, l'objectif de cette communication est de présenter un état des lieux et des enjeux du développement des systèmes d'information décisionnels dans les collectivités territoriales françaises. Nous présentons ici la première phase de ce projet de recherche, à savoir une enquête quantitative conduite auprès de 350 collectivités de grande taille (région, départements, communautés urbaines et d'agglomération et villes de 40.000 habitants) qui vise à se faire une première représentation du phénomène à analyser par la suite.

La recherche que nous souhaitons présenter trouve sa source dans l'écart entre le foisonnement des initiatives visant à l'informatisation du pilotage dans les administrations et le faible nombre d'études qui s'y intéressent.

Les projets informatiques en vue d'améliorer la prise de décision sont un phénomène ancien (Rationalisation des Choix Budgétaires, ACCORD) et important dans les administrations et collectivités territoriales. Par exemple, le projet de système d'information financière de l'état, CHORUS, représente un investissement de près de 1,5 milliard d'euros sur 10 ans selon la Cour des Comptes. Autre exemple, l'Opérateur National de Paye, qui gère la modernisation de la fonction paye pour l'ensemble des administrations de l'État, a pour mission de mettre en place un Système d'Information Décisionnel devant permettre le pilotage des activités de la chaîne RH-Paye pour 2,8 millions d'agents au sein de l'État. Ce projet SID complexe et de grande envergure concerne près de 15.000 utilisateurs dans des structures ministérielles hétérogènes.

Cependant, l'essentiel des recherches en SI ayant pour objet des dispositifs développés en contexte public se focalise sur les aspects les plus visibles, comme les changements apportés par l'e-administration à la relation entre l'administration et les usagers du service public. À cet égard le lien en e-administration et décisionnel est d'ailleurs évident [Cesarini et Mezzanzanica (2007)]. Les systèmes informatisés mis en place pour le suivi de la relation avec les

usagers d'un service public génèrent, par conception, des données (statistiques de consultation, temps de traitement des demandes, nombre de dossiers suivis...) qui ont vocation à alimenter un système décisionnel. Le délai entre prestation de services publics et évaluation de la réalisation des politiques publiques pourrait donc être réduit considérablement, puisque l'analyse des données fournies en temps quasi-réel permettrait d'améliorer rapidement la qualité du service rendu aux usagers, ou d'adapter au fil de l'eau les moyens aux besoins. Dans cette optique, les objectifs d'évaluation des politiques publiques, et d'amélioration de celles-ci via une meilleure connaissance des usagers, des grandes tendances observables... serait intégrés au système dès le départ. Les données archivées par les administrations constituent un véritable « filon ». L'exploitation et l'analyse de ce capital immatériel permettrait d'avoir une connaissance fine et complète de la population d'un territoire donné [Mezzanzanica et al. (2005)], ce qui ouvre de possibilités nouvelles en termes d'ajustement des politiques publiques locales aux besoins de la population.

Prenant acte de la montée en puissance des collectivités dans la mise en œuvre de l'action publique, l'objectif de cette communication est de présenter un état des lieux et des enjeux du développement des systèmes d'information décisionnels dans les collectivités territoriales françaises.

Le choix de ce terrain est justifié tant par le poids croissant pris par les collectivités en matière d'action publique (élargissement du champ de compétences et de l'autonomie décisionnelle, accroissement de leur volume de dépenses. Les collectivités représentent aujourd'hui plus de 75% des investissements publics), que par le développement de démarches de performance en leur sein (considéré comme un prédicteur du déclenchement de démarches d'informatique décisionnelle).

Les collectivités sont confrontées à des pressions croissantes : les contraintes du cadre juridique impliquent un « droit à l'erreur » de plus en plus réduit pour les décideurs publics, ce qui les oblige à professionnaliser leur pilotage et à renforcer leurs capacités d'analyse prospective. Pour pouvoir répondre plus rapidement à des problématiques de plus en plus complexes, ils ont besoin de plus d'autonomie en matière d'analyse des données de pilotage de leur collectivité. De plus, la gestion des élus est de plus en plus surveillée et évaluée, par leurs opposants politiques, les médias, les associations, par l'Etat, par les citoyens-électeurs, par les syndicats... [Conseil Général du Gard (2011)]. La pression de la libération des données publiques (*Open Data*), qui est une obligation réglementaire depuis 2005 [1] pousse également les collectivités à plus de transparence sur leurs actions.

Les systèmes d'information des collectivités, très souvent construits au fil des besoins par empilements successifs de matériels, d'applications, de règles de gestion, de procédures... sans vision d'ensemble cohérente, montrent leurs limites. Ils ne sont pas adaptés, trop silotés, trop complexes pour répondre à ces besoins d'analyse des données dans des délais satisfaisants.

La présente communication fait état la première phase de ce projet de recherche, à savoir une enquête quantitative conduite auprès de 350 collectivités de grande taille (région, départements, communautés urbaines et d'agglomération et villes de 40.000 habitants) qui vise à se faire une première représentation du phénomène à analyser par la suite. Dans cette perspective, nous avons souhaité effectuer un large balayage sur les outils utilisés, leurs modes d'implémentation et les résistances rencontrés par les chargés de projet, avec comme mode d'entrée les démarches de pilotage. Cette enquête menée dans le cadre des travaux de l'observatoire de

1. Ordonnance 2005/650 du 6 juin 2005 et décret n. 2005/1755 du 30 décembre 2005.

la performance publique locale coordonné par l'Assocation Finances-Gestion-Évaluation des Collectivités Territoriales (AFIGESE) a été réitérée en 2010 et 2011, et se poursuit actuellement avec des études de cas encore en cours de réalisation au sein de collectivités pilotes en la matière.

1 Les collectivités territoriales, porte d'entrée du managérialisme dans le secteur public français

Les collectivités territoriales sont des divisions administratives distinctes de l'État, qui exercent les compétences qui lui sont confiées par le législateur sur la population d'un territoire donné. Pour l'essentiel ce sont les communes, département et régions. La grande diversité des compétences conduit les collectivités à exercer dans des domaines variés : ainsi un département sera amené à intervenir en matière d'insertion (RSA), d'aide sociale (allocation personnalisée d'autonomie, prestation compensatoire de handicap), d'enfance, ou encore pour la gestion matérielle des collèges, pour les routes...

La pénétration d'une logique de performance dans le secteur public depuis plus de trente ans a pour corollaire la multiplication des démarches décisionnelles. En effet, les démarches de performance visent à rationaliser l'action publique en permettant aux décideurs de fonder leur décision sur une analyse aussi précise que possible des ressources et activités consommées dans le processus de production des politiques publiques et à les confronter aux résultats obtenus.

La Loi Organique relatives aux Lois de Finances (LOLF) est emblématique de cette volonté de passer d'une culture des moyens et des procédures à une culture de performance et de résultats. La LOLF modifie le découpage du budget général de l'État en Missions, Programmes et Actions (en lieu et place d'une programmation par Ministère). Elle se veut dans la lignée de la Nouvelle Gestion Publique d'inspiration anglo-saxonne (le *New Public Management*), et serait, selon [Santo et Verrier (2007)], porteuse d'un véritable « basculement managérial » devant entraîner « la mise en place de nouveaux réflexes et de nouveaux outils de management : l'évaluation des politiques publiques, le contrôle de gestion, la comptabilité analytique, le tableau de bord prospectif, la gestion prévisionnelle des effectifs et des compétences, la mesure de la performance... ». La LOLF a inspiré des démarches de gestion dans des collectivités (Conseil Général de la Mayenne, Conseil Général de Seine Saint-Denis,...) qui ont décidé de présenter leur budget selon une logique similaire.

Selon la représentation classique du canevas de l'action publique [Santo et Verrier (2007)], les politiques publiques décidées par les élus déclenchent des réalisations par l'administration ; ces réalisations génèrent des effets sur la population ou les groupes sociaux concernés. L'adossement d'un SID à ce dispositif (FIG. 1), alimenté par les données produites à toutes les étapes de la démarche et par des sources externes, est censé produire un retour d'information permettant d'améliorer l'efficacité, l'efficience et la pertinence des politiques publiques (FIG. 2).

Les démarches de pilotage de l'action publique ont pour axiome la conviction que la pertinence de la décision est immédiatement fonction de la qualité et de l'abondance des informations disponibles. Cette logique trouve toutefois sa limite dans le fait que les jeux de pouvoir, la négociation politique et les arrangements institutionnels président plus volontiers à la plupart

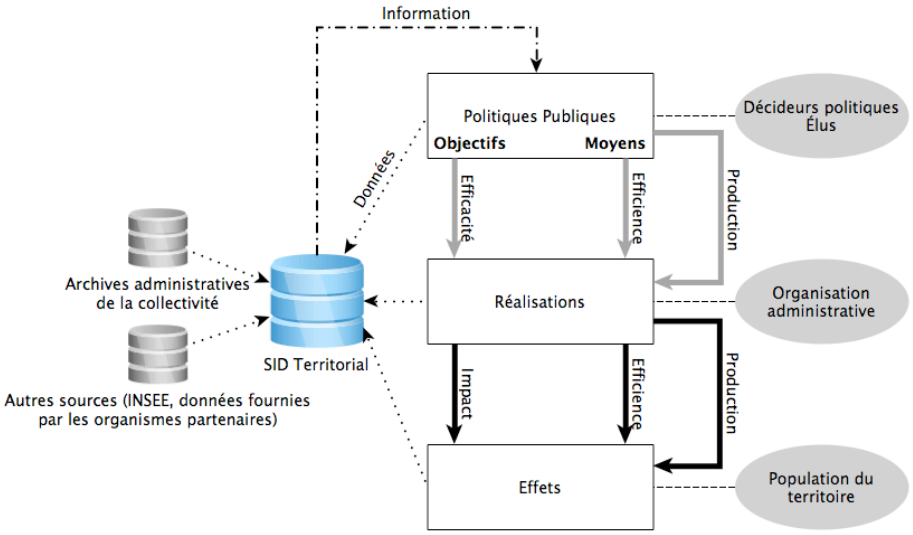


FIG. 1 – *Canevas de l'action publique, adossé à un SID territorial — adapté de : Santo et Verrier (2007).*

des décisions stratégiques dans les organisations publiques que l'analyse objective des données [Aggarwal et Mirani (1999)].

1.1 Information et décision dans les organisations publiques

La gestion des ressources informationnelles au service de la prise de décision dans les organisations publiques constitue un défi majeur [Viscusi et al. (2010)], et une opportunité d'innovation (réduction de la bureaucratie et des coûts de l'administration) [Riedl (2003); Tambouris et al. (2001)]. Les collectivités territoriales, en particulier, sont confrontées à une complexité externe et des contraintes internes tenant la croissance exponentielle des volumes de données qu'elles sont amenées à traiter. Ces facteurs amènent les collectivités à investir lourdement dans des systèmes d'information décisionnels, et attendent de ces technologies qu'elles appuient les managers publics dans leur prise de décision.

Les collectivités investissent des montants significatifs dans les technologies d'aide à la prise de décision pour faire face à leurs besoins de pilotage tout en gérant la croissance exponentielle des données que leurs systèmes informatiques collectent, mémorisent et traitent. Ces investissements sont motivés par la perspective de rendre leur administration plus performante. Cependant, en adoptant ces technologies d'aide à la décision dans leurs organisations, les collectivités locales doivent veiller aux spécificités organisationnelles de l'administration

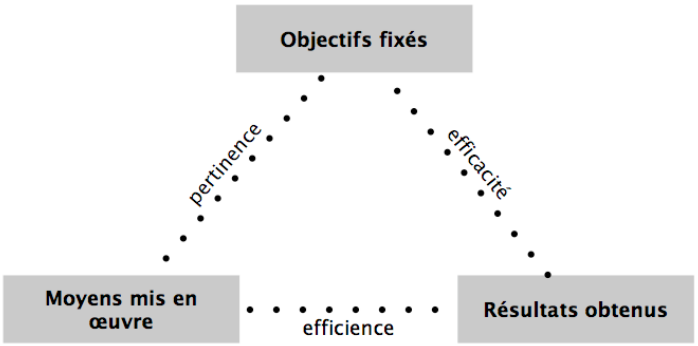


FIG. 2 – *Système de contrôle de l'action publique — source : Santo et Verrier (2007).*

publique, tels que des processus de prise de décision plus flous, ou la complexité croissante des enjeux qu'elles traitent du fait de la variété de leurs domaines d'intervention.

Les systèmes d'information décisionnels (SID) sont une catégorie de systèmes d'information destinée à « intégrer données et modèles dans le but d'améliorer le processus de prise de décision » [Arnott et Pervan (2005); Vaisman (2007)]. Comme l'ont relevé Arnott et Pervan, il existe un écart croissant entre la recherche théorique et la pratique dans la discipline des systèmes d'information en général. Cela se vérifie en particulier dans le secteur public ou la recherche ancrée sur les développement des SID est quasiment inexistante depuis 10 ans. Nous présentons d'abord quelques éléments des théories de la décision avant d'explorer le champ des systèmes d'information qui supportent ces processus de prise de décision.

L'action collective est réalisée dans les organisations par des systèmes de décision [Tabatoni et Jarniou (1975)]. Dans cette perspective, chaque décision est un choix, la sélection d'une décision pour résoudre un problème. Ces problèmes apparaissent quand le décideur perçoit un changement significatif dans le système qui détermine ses objectifs, et implique l'action [Simon (1960)]. Dans les sciences du management, décider signifie identifier et résoudre des problèmes rencontrés par les organisations [Le Moigne (1974)] en suivant la formule de Simon [Simon (1960)] : « Manager est synonyme de prendre des décisions », ce qu'il avait défini 40 ans auparavant dans un article célèbre : « Dans la société industrielle post-moderne, le problème central n'est pas comment s'organiser pour produire efficacement, bien que cela restera toujours une considération importante, mais comment s'organiser pour prendre des décisions - ce qui signifie, traiter l'information » [Simon (1973)].

La littérature sur l'information et la décision a rapidement identifiée le problème de la surcharge informationnelle des acteurs et leurs capacités cognitives limitées pour traiter l'information dans le but de prendre des décisions [March et Simon (1958)]. Ainsi le besoin de faciliter l'accès à l'aide à la prise de décision comme le développement de l'apprentissage

individuel et organisationnel est explicitement traité par la recherche sur les systèmes d'information décisionnels. Ces systèmes apparus il y a 50 ans, ont été marqués par les idiosyncrasies techniques de cette époque [Vidal et al. (2009)]. Dans le contexte technologique d'alors, ils ont été conçus pour le type d'information le plus aisément accessible, c'est à dire les données structurées générées par les systèmes transactionnels [Davenport].

Power défini les systèmes d'information décisionnels comme des « systèmes interactifs électroniques destinés à aider le décideur à utiliser technologies de la communication, données, documents, connaissance, et/ou modèles pour identifier et résoudre des problèmes, achever des processus de prise de décision, et prendre des décisions » [Power (2008)]. Ils permettent donc de fournir une vue d'ensemble des processus de l'organisation, et aident à anticiper les actions futures pour un pilotage intelligent par les managers.

Une littérature importante a été consacrée aux différences entre les organisations publiques et les entreprises privées. [Ring et Perry (1985)] ont mis en exergue 5 grandes différences :

- l'ambigüité dans la définition des politique publiques (*policy ambiguity*) ;
- l'exposition publique des actions de l'administration (*the openness of government*) ;
- le nombre de parties prenantes surveillant l'action des décideurs publics, telles que les autorités administratives de contrôle, d'inspection (*attentive publics*) ;
- les contraintes temporelles artificielles imposées par la loi ou la tenue d'élections (*time constraints*) ;
- les alliances instables à cause d'intérêts concurrents dans la définition des politiques publiques (*shaky coalitions*).

Ces différences ont des répercussions sur le pilotage des projets IT par les organisations [Rocheleau (2005)]. Un consensus existe aujourd'hui dans la recherche SI pour distinguer un management des systèmes d'information d'organisations publiques différent de celui des entreprises privées. Rocheleau établit que bien que « le management des technologies de l'information dans le secteur public présente des problèmes différents que ceux rencontrés dans le secteur privé… une attention faible a été accordée à ces différences par les chercheurs en SI.

1.2 Utilisation des SID et besoins en information

Les collectivités tendent à passer d'un pilotage de la performance par politique vers des démarches globales couvrant l'ensemble des initiatives mises en œuvre. En abandonnant une logique en silo, la qualité du système de pilotage d'une collectivité dépend de plus en plus de celle du système d'information support du pilotage. Les démarches globales de modernisation mises en œuvre par les collectivités supposent en effet la collecte, le traitement et l'utilisation de très grandes quantités d'information. Ne pas investir dans un système capable de gérer cette masse d'information au quotidien rend difficile la survie de ces démarches à court et moyen terme, et encourage la prolifération exponentielle de rapports produits à partir de données extraites ponctuellement des applications du système informatique, puis modifiées et agrégées à la main. En revanche, une démarche décisionnelle peut avoir un effet retour structurant sur le système d'information, en obligeant à réfléchir en termes de processus métier, desquels sont extraits les indicateurs de mesure, ou à améliorer la qualité des données.

L'enquête AFIGESE révèle un l'importance des investissements réalisés par les collectivités en matière de SID (TAB. 1).

En dépit de cela, de nombreux points de préoccupation sont mis en exergue : les répondants font part de leur insatisfaction vis-à-vis des systèmes existants, du fait de leur cloisonnement

Montant (euros)	Répartition (%)
+ de 500.000	12
100.000 à 500.000	47
30.000 à 100.000	24
- de 30.000	18

TAB. 1 – *Source : enquête* AFIGESE *2010*

(effet silo), des technologies décisionnelles peu diffusées et mal maîtrisées, et de la difficulté à mener de concert une démarche organisationnelle de pilotage et une démarche informatique.

71% des répondants affirment que leur organisation a mis en place un SID (sans préjuger du périmètre de l'organisation concerné). Dans les organisations n'ayant pas de SID, 78,2% déclarent vouloir en mettre un en place à très court terme.

Malgré cette adoption et adhésion très significative, 57,9% des répondants signalent que le système d'information de leur organisation ne fournit pas aux utilisateurs les informations dont ils ont besoin. Une large majorité (65%) reconnaît que l'information existe, mais celle ci est dispersée à travers des systèmes de stockage et de restitution différents (81,7%) ou silotés (80%). 42,9% estiment que le volume d'information dont ils disposent est trop important.

Ces observations sont en phase avec la littérature. Pour [Marr et Creelman (2011)], « peu d'organisations diront qu'elles manquent de données. Malheureusement, ces données sont rarement transformées en savoir utile. La plupart des organisations publiques ont construit de grandes et coûteuses baronnies, dont la seule raison d'être est de collecter et de restituer des données. Ces données servent rarement à améliorer la performance... ». Les auteurs notent que les organisations publiques sont submergées par la quantité de données qu'elles produisent, alors qu'en comparaison, les décisions stratégiques qu'elles prennent s'appuient peu sur ces données.

Le paradigme du « SIG » (Système d'Information de Gestion) reste prégnant [Le Moigne (1986)]. Celui-ci présume que toute l'information nécessaire existe, que toute décision peut se baser sur de l'information, et que le système doit fournir aux décideurs l'information dont ils ont besoin pour décider. Dans ce modèle rationnel, il est également présumé que les décideurs savent ce qu'ils doivent décider, que l'environnement est stable, et que de bonnes informations entraînent mécaniquement de bonnes décisions.

Disposer de l'information ne signifie pourtant pas que des actions adéquates seront entreprises, et ce d'autant plus que le niveau hiérarchique des décideurs s'accroît [Aggarwal et Mirani (1999)]. Au niveau le plus élevé, l'on s'appuiera davantage sur des négociations, compromis ou arrangements institutionnels que sur l'analyse des données. Des analyses très simples, faites « à la main » dans une feuille de tableur, pourront même être considérées trop sensibles dans la mesure où elles révéleront ces arrangements. Un contrôleur de gestion, ayant croisé les données du nombre d'agents par unité territoriale (les points d'accueil sur le territoire couvert par un Conseil Général) et du nombre de dossiers à traiter, mettait en lumière les disparités de traitement des usagers sur le territoire. Il témoigne ainsi que cette analyse, trop sensible politiquement, n'a pas pu sortir de sa direction.

Bien que 69,5% des répondants affirment que la mise en place d'un SID est couplée avec une démarche globale de pilotage (définition des politiques publiques de la collectivité, des

processus, activités, et indicateurs de performance associés...) les utilisateurs se plaignent de ne disposer que d'informations éparses et non agrégées, ce qui semble contradictoire. Interrogés sur les outils dont ils se servent pour gérer l'information, une majorité écrasante évoque les outils de tableur (97,8%) ou les requêtes automatisées sur les applications métier (80,4%). 38% disposent d'infocentres et 29,3% d'entrepôts de données. L'intégration des données dans un SID ne concerne que 23,9% des répondants. Ces résultats montrent que bien que les SID soient un sujet brûlant pour les collectivités, les approches intégrées et architecturées restent rares, et que le tableur, connecté ou non à une source informatisée de données, reste l'outil d'aide à la décision le plus courant. D'autres témoignages recueillis en marge de l'enquête suggèrent même qu'en dépit des investissement massifs réalisés dans l'informatique décisionnelle, une culture de « bricolage » subsiste (favorisée par la complexité d'utilisation de certaines applications décisionnelles). Des outils *ad hoc* sont développés (feuilles excel®, bases Access®) pour répondre aux besoins de *reporting* et d'analyse les plus urgents.

Notons que dans la plupart des cas, les collectivités n'ont pas la nécessité d'outils complexes, et peuvent être amenées à sur-estimer leurs besoins, sous la pression du discours des prestataires informatiques. Comme en témoigne un contrôleur de gestion d'un Conseil Général, « nous ne sommes qu'une collectivité territoriale, on ne met pas des satellites sur orbite ».

2 Motivations et freins à l'adoption des SID en collectivités

2.1 Les raisons de l'adoption des SID dans les collectivités

La plupart des répondants avancent que leurs investissements dans les SID sont liés aux besoins de pilotage de l'action publique (69%) et à la recherche de la performance (64,4%). Nous doutons cependant qu'il existe un lien empiriquement vérifiable entre l'utilisation de SID et une meilleure performance de l'organisation. Si le lancement d'un projet décisionnel précède la réflexion d'ensemble sur les politiques publiques, les processus qu'elles mobilisent, les activités qui les sous-tendent, la qualité de la donnée produite... l'outil ne sera qu'une coquille vide, découplée de l'organisation, et des sommes importantes pourront être dépensées sans qu'aucun résultat ne soit atteint. Comme dans tout projet à composante informatique, la pensée magique, selon laquelle la technologie est considérée comme une manne tombée du ciel (« dotons-nous de tel outil pour résoudre tous nos problèmes ») débouchera sur des échecs cuisants [Rochet (2012)].

En dernière analyse, un projet de SID réussi ne sera que le sous-produit d'une bonne organisation et de bons processus de prise de décision. Dans cette optique, notons que la modernisation (46%), la gestion du changement (31%) et la réduction des coûts (28,7%) sont les principales justifications avancées pour investir dans un SID, ce qui revient faire porter à la technologie (les outils d'informatique décisionnelle), abstraction faite de toute logique de management, la responsabilité de la performance de l'organisation.

2.2 Barrières à l'adoption des SID

54% des répondants pensent que la « résistance au changement » bloque l'adoption des SID. En effet, l'introduction d'un SID peut modifier l'équilibre des pouvoirs entre différentes directions. Si l'information est publiée et mise à disposition à travers l'organisation (parfois

avec des habilitations différentes selon le niveau hiérarchique de l'utilisateur), la réalité de l'organisation est exposée. Ainsi, une direction des ressources humaines peut être réticente à partager « son » information, dans la mesure où cela signifie une perte de contrôle ou de pouvoir de négociation avec d'autres directions.

33,3% des répondants évoquent le manque d'intérêts des décideurs. Cela n'est pas surprenant, la majorité des hauts fonctionnaires ne recevant pas de formation appropriée dans la discipline des systèmes d'information [Rochet (2012)]. Ils peuvent refuser de s'engager à soutenir une démarche perçue comme exclusivement technique, ou porter un intérêt limité à la démarche alors même qu'ils sont chargés de la mettre en place au sein de l'organisation. 41,4% évoquent le manque de compétences techniques en interne, 48,3% le manque de ressources humaines pour mener de tels projets.

14,9% des répondants évoquent le manque d'informations sur le retour d'investissement (ROI) des projets décisionnels. Un SID peut paraître une dépense superfétatoire pour des décideurs habitués à utiliser d'autres capteurs (intuition, expérience, réseaux personnels d'information au sein de la structure...) pour apprécier les situations, et qui perçoivent « l'informatique » de façon négative. Le ROI sera de toute façon difficile à justifier, du fait du contexte particulier des organisations publiques qui ne sont pas habituées à réfléchir en ces termes, ainsi que de la relation ambiguë entre information et décision [2] dans le contexte d'organisations publiques dont la condition de survie n'est pas de réaliser des profits. Une entreprise comme une collectivité pourront s'aider des informations fournies par le systèmes décisionnel pour redéployer leur personnel. Mais alors qu'une entreprise trouvera logique d'optimiser la répartition de ses équipes afin d'améliorer son chiffre d'affaire, un tel raisonnement est difficilement transposable dans une collectivité. Par exemple, pour un Conseil Général, il pourra être plus intéressant politiquement de maintenir une présence forte dans ses points d'accueil situés dans certains cantons, indépendamment des besoins réels du terrain. Le système d'information décisionnel, si perfectionné et fiable soit-il, reste soumis à la logique politique des élus locaux.

Le manque de ressources financières est évoqué par 21,8% des répondants seulement, ce qui confirme l'idée selon laquelle ce facteur n'est pas le plus bloquant.

Une barrière majeure reste le développement insuffisant d'une véritable « culture de gestion » dans les collectivités territoriales. En dépit du discours omniprésent sur l'indispensable évaluation des politiques publiques, il est mal perçu de devoir « rendre-compte » de ses actions de façon détaillée et transparente ; au niveau individuel (des agents du service public) la peur du « contrôle » est évoquée face à des dispositifs présentés comme au service de l'innovation organisationnelle et de la performance.

3 Tendances

Quelques tendances se dégagent des résultats de l'enquête et des retours d'expériences recueillis :

– la convergence opérationnel/décisionnel (notamment avec le suivi informatisé des demandes, qui permet en même temps le traitement, le suivi, et le reporting sur une activité), comme par exemple le suivi des demandes utilisateurs par une DSI ;

2. « Quelle que soit l'information collectée, l'on demandera toujours plus d'information ; l'information demandée n'est pas prise en compte dans la décision, ou sert à justifier a posteriori une décision déjà prise... la relation entre décision et information est donc instable et ambiguë » [March (1991)].

- la montée en maturité des utilisateurs, qui connaissent les technologies disponibles et exigent des outils simples d'emploi, ne nécessitant pas des semaines de formations pour être appréhendés ;
- l'identification d'un couplage potentiel fort entre décisionnel et systèmes d'information géographique ;

4 Conclusion, limites et recherches ultérieures

4.1 Conclusion

Le domaine de l'informatique décisionnelle reste largement à découvrir dans le monde territorial, mais sa contribution potentielle au pilotage de la performance le rend incontournable à moyen terme.

Dans les collectivités qui ont mis en place avec succès des SID, l'accent a été mis sur leur rôle structurant, dans la mesure où ils sont employés comme un levier pour développer la transparence à travers l'organisation. Ce retour positif ne peut avoir lieu que si :

- les objectifs sont clairement définis ;
- il existe un couplage explicite avec une démarche organisationnelle de pilotage ;
- les utilisateurs sont activement impliqués dans les itérations nécessaires au développement, au raffinement et à l'adaptation du système aux besoins et aux objectifs de l'organisation.

D'autre part, quelques signaux d'alerte permettent d'anticiper l'échec probable d'un projet décisionnel :

- imitation d'une mode managériale (on affirme qu'il faut faire du décisionnel, sans se demander à quelles fins) ;
- choix dicté par le discours commercial d'un prestataire informatique (on confie entièrement à un prestataire la définition de la stratégie décisionnelle au lieu de mener cette réflexion en interne) ;
- manque de réflexion en termes d'architecture du système d'information ;
- appréhension de l'outil comme solution miracle.

Il sera ainsi courant de trouver des systèmes très complexes, redondants, cloisonnés dans les silos métier, cohabitant avec des outils « manuels » (le plus souvent des feuilles de tableur) donnant lieu à une véritable double comptabilité, et générant des centaines voire des milliers de rapports qui interdisent par conception de disposer de données fiables et agrégées, et donc de contribuer de façon efficace aux processus de prise de décision.

Les projets de SID réussis sont conçus comme des projets de transformation organisationnelle. Les échecs résultent d'un manque de maturité, qui est bien souvent organisationnel avant d'être technologique (la structure lance un projet SID sans posséder la culture de gestion nécessaires). Les résistances liées aux jeux de pouvoir entre différentes directions d'une collectivité (la détention d'une information peut donner un avantage qui disparaît avec la mise en place d'un système permettant un large accès à l'information à travers la collectivité), ou le manque de clarté quant à leurs périmètres de responsabilité respectifs (qui est propriétaire et responsable de la qualité de la donnée ?), sont autant de facteurs qui peuvent aller à l'encontre du bon déroulement d'un projet décisionnel.

4.2 Limites

La principale limite de cette recherche réside dans la méthodologie adoptée pour notre phase exploratoire préliminaire. Celle-ci a consisté en la réalisation d'une enquête quantitative visant à évaluer le niveau d'adoption des SID dans les collectivités. Nous souhaitions disposer d'une vue d'ensemble, suffisamment générale pour tracer les contours d'une programme de recherche plus ciblé. à ce stade, aucune définition précise d'un SID n'avait été donnée, ce qui a limité la possibilité de construire des typologies. L'étude ne fournit pas d'indications sur les pratiques réelles des collectivités en matière d'utilisation de l'information extraite des systèmes informatiques pour informer la décision, cette compréhension étant indispensable pour élaborer des préconisations à l'intention des praticiens pour la construction de leur système décisionnel.

Du fait du public concerné par l'enquête, la communauté des contrôleurs de gestion des collectivités, il existe un biais fort en faveur des domaines concernés par cette activité (finances, ressources humaines, gestion du patrimoine immobilier). Le domaine de l'aide sociale (compétence des Départements), des transports, des subventions aux associations, sont également fortement représenté. Il est probable qu'il existe, surtout au sein de grosses collectivités, des SID « métier » silotés, dont le contrôle de gestion n'aurait pas forcément connaissance.

4.3 Recherches ultérieures

Pour cette raison, des recherches ultérieures sont envisagées afin de fournir des études qualitatives sur les liens entre systèmes d'information décisionnels et processus de décision dans les collectivités territoriales. La réalisation d'une série d'étude de cas dans des collectivités ayant déjà expérimenté des projets décisionnels, ainsi que l'exploitation des résultats de l'enquête quantitative réalisée fin 2011 s'inscrivent dans cette perspective. Ces travaux pourront déboucher par la suite sur la construction d'un référentiel ou d'une grille d'auto-évaluation destinée aux sponsors des projets décisionnels, permettant d'évaluer la maturité de leur organisation par rapport aux conditions de réussite d'un projet décisionnel.

Références

Aggarwal, A. K. et R. Mirani (1999). DSS model usage in public and private sectors : Differences and implications. *Journal of End User Computing 11*(3).

Arnott, D. et G. Pervan (2005). A critical analysis of decision support systems research. *Journal of Information Technology 20*(2), 67–87.

Cesarini, M. et M. Mezzanzanica (2007). E-Government as decision support system to improve public services provision. In *Proceedings of the European Conference on e-Government*.

Conseil Général du Gard (2011). Projet ideal — information d'évaluation et d'analyse locale : Portail décisionnel du cg 30. Présentation powerpoint - Atelier POSS LR - Montpellier.

Davenport, T. H. L'art du management de l'information.

Le Moigne, J.-L. (1974). *Les systèmes de décision dans les organisations*. Presses universitaires de France.

Le Moigne, J.-L. (1986). Vers un système d'information organisationnel. *Revue française de gestion*.

March, J. et H. A. Simon (1958). *Organizations* (Wiley ed.). New York.

March, J. G. (1991). *Décisions et organisations*. Éditions d'Organisation.

Marr, B. et J. Creelman (2011). More with less : the new performance challenges for the UK public sector (2011 and beyond).

Mezzanzanica, M., M. Cesarini, et R. Boselli (2005). Business Intelligence Exploitation for investigating territorial Systems , methodological Overviews and empirical Considerations. *Intelligence*, 1–18.

Power, D. J. (2008). Understanding Data-Driven Decision Support Systems. *Information Systems Management 25*(2), 149–154.

Riedl, R. (2003). Design principles for e-government services. In *Proceedings from eGov Day*, Austria.

Ring, P. S. et J. L. Perry (1985). Strategic management in public and private organizations : Implications of distinctive contexts and constraints. *The Academy of Management Review 10*(2), 276–286.

Rocheleau, B. (2005). *Public Management Information Systems* (illustrated ed.). Idea Group Publishing.

Rochet, C. (2012). Digitizing the public organization (à paraître). *Halduskultuur*.

Santo, M.-V. et P.-E. Verrier (2007). *Le management public* (3ème ed.). Que sais-je ? Presses Universitaires de France - PUF.

Simon, H. A. (1960). *The new science of management decision*. Harper.

Simon, H. A. (1973). Applying information technology to organization design. *Public Administration Review 33*(3), 268–278.

Tabatoni, P. et P. Jarniou (1975). *Les systèmes de gestion : politiques et structures*. Presses universitaires de France.

Tambouris, E., S. Gorilas, et G. Boukis (2001). Investigation of electronic government. In *Proceedings from 8th Panhellenic Conference on Informatics*.

Vaisman, A. (2007). Data Quality-Based Requirements Elicitation for Decision Support Systems. *Data warehouses and OLAP : concepts, architectures, and solutions*.

Vidal, P., V. Petit, F. Lacroux, M. Augier, V. Merminod, M. d. Gibon, et C. Mangholz (2009). *Systèmes d'information organisationnels*. Pearson Education.

Viscusi, G., C. Batini, et M. Mecella (2010). *Information Systems for eGovernment A Quality-of-Service Perspective*. Berlin : Springer.

Summary

This paper originated in the identification of a gap between the growing importance of decision support systems in local government organizations and the lack of attention of the research community towards this phenomenon.

Planification guidée par Raisonnement à base de cas et Datamining : Remémoration des cas par Arbre de décision

Sofia Benbelkacem, Baghdad Atmani, Abdelhak Mansoul

Equipe de recherche SIF « Simulation, Intégration et Fouille de données »
Laboratoire d'Informatique d'Oran - LIO
Département Informatique, Faculté des Sciences, Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie
sofia.benbelkacem@gmail.com
atmani.baghdad@gmail.com et mans_abdel@yahoo.fr

Résumé. Dans notre quotidien, nous faisons souvent appel à notre expérience pour résoudre certains problèmes déjà rencontrés auparavant. Le raisonnement à partir de cas copie ce comportement humain. Il permet de résoudre des problèmes en recherchant des cas analogues déjà résolus dans le passé.
Dans ce papier nous proposons une nouvelle technique de planification que nous avons baptisée DTR « Decision Tree for Retrieval ».
DTR utilise, pour le calcul de similarité, le principe de classification par arbre de décision.

1 Introduction

L'aide à la décision est l'activité de celui qui, prenant appui sur des modèles clairement explicités mais non nécessairement complètement formalisés, aide à obtenir des éléments de réponses aux questions que se pose un intervenant dans un processus de décision (Roy, 1985). De ce fait, l'aide à la décision n'a pas pour but de remplacer le décideur en lui proposant des solutions « toutes faites ». Elle cherche d'abord à le guider vers des décisions qu'il aura à prendre sous sa responsabilité. Notre contribution dans ce domaine concerne la proposition d'un système d'aide à la décision guidé par le raisonnement à base de cas.

Le raisonnement à partir de cas est un paradigme de résolution de problèmes qui se base sur des solutions connues de problèmes passés jugés similaires à ceux étudiés. Il s'applique dans divers domaines où les expériences passées représentent une part importante de l'activité (Cavallucci et al., 2010). La planification à partir de cas est un champ de recherche actif du raisonnement à base de cas. Cette technique utilise le principe du raisonnement à partir de cas pour trouver plus rapidement le meilleur plan qu'avec une technique de planification classique (Lieber, 2008).

Dans la vie courante, les individus sont souvent confrontés à des problèmes déjà vécus et qui ont probablement des solutions similaires. Par exemple pour la planification des soins, il est possible de rencontrer des patients qui auront besoin de suivre le même plan de traitement. Alors pour profiter de l'expérience passée et optimiser la durée de calcul et au lieu de synthétiser des plans à partir d'opérateurs primitifs, nous allons utiliser le principe de la planification

basée sur les cas en combinant le raisonnement à partir de cas et la planification pour mettre en place notre système de planification guidé par raisonnement à base de cas.

Dans un autre volet, en fouille de données, on utilise une variété de méthodes pour traiter de grandes quantités d'informations afin de découvrir les connaissances utiles pour la prise de décision. Nous utilisons l'une de ses méthodes : les arbres de décision, pour la phase de recherche du cas similaire.

L'article est organisé comme suit. Dans la section 2, nous présentons le principe du raisonnement à base de cas . Dans la section 3, nous introduisons les notions de plan et de planification et nous abordons la planification guidée par le raisonnement basé sur les cas. Dans la section 4, nous citons les étapes de construction de la base de cas de notre planificateur et nous donnons une brève description sur la représentation des cas. Ensuite dans la section 5, nous expliquons la méthode de recherche des cas par arbre de décision. Enfin, la section 6 est consacrée à la conclusion et aux perspectives du présent travail.

2 Le raisonnement à base de cas

Le raisonnement à base de cas est l'une des techniques de l'intelligence artificielle les plus utilisées actuellement (Lamontagne et Lapalme, 2002). Raisonner à partir de cas consiste à résoudre un nouveau problème, appelé problème cible, en utilisant un ensemble de problèmes déjà résolus. Un cas source désigne un épisode passé de résolution de problèmes et une base de cas, un ensemble de cas sources (Badra, 2009). Un cas se compose de deux parties : la partie problème et la partie solution. La partie problème est décrite par un ensemble d'indices qui déterminent dans quelle situation un cas est applicable.

Le processus du raisonnement à partir de cas opère généralement selon quatre phases séquentielles : la remémoration, l'adaptation, la révision et l'apprentissage.

2.1 Remémoration ou Recherche

Dans cette phase, il s'agit d'extraire de la base de cas, les anciens cas dont la partie problème est similaire au problème à résoudre. Des mesures de similarités sont alors à définir sur les indices constituant la partie problème du cas en question. Les cas extraits de la base de cas sont appelés cas sources. Les cas sources passent alors à la phase d'adaptation.

2.2 Adaptation ou Réutilisation

Cette phase consiste à proposer une solution au problème courant (cas cible) en adaptant les solutions proposées par les cas sources. L'adaptation repose souvent sur l'utilisation de connaissances dans le domaine de l'application. A l'issue de cette phase, une ou plusieurs solutions seront proposées pour le cas cible.

2.3 Révision

L'objectif de cette phase est de réviser les solutions proposées par la phase précédente en fonction de certaines règles et/ou heuristiques, qui dépendent du domaine de l'application. La

phase de révision peut être faite par des experts dans le domaine de l'application ou d'une manière automatique.

2.4 Apprentissage ou Maintenance

Cette phase a la charge d'améliorer l'expérience du système de raisonnement à partir de cas en enrichissant la base de cas par les nouveaux problèmes résolus (cas cible auquel on a apporté une solution). En effet, les cas résolus peuvent être ajoutés à la base de cas et être utilisés ultérieurement pour résoudre de nouveaux problèmes. Mais avant d'ajouter ces cas, il faut juger la pertinence de cet ajout (éviter par exemple d'ajouter des cas redondants ce qui affecte les performances du système en termes de temps et de traitement sans pour autant améliorer la qualité des solutions apportées).

3 La planification

De très nombreuses applications d'aide à la décision ont été développées (Trilling et al., 2007), (Berkoune et al., 2011). Celles-ci sont projetées sur un processus d'aide à la décision et se définissant comme une série d'étapes décisives à réaliser avec en finalité l'amélioration de la qualité des services. Dans notre cas médical, cette série d'étapes (le processus d'aide à la décision), peut être considérée comme un projet d'aide à la décision avec des tâches et des contraintes, qu'il faut réaliser selon un plan (Ruland et Bakken, 2002). De ce fait, les étapes de ce processus doivent être bien planifiées. C'est dans cette optique que nous nous sommes fixés l'implication des techniques de fouille de données dans la planification du processus de la prise de décision médicale.

La planification consiste à concevoir un futur désiré et les moyens pour y parvenir (Ackoff, 1973). C'est un processus d'aide à la décision qui vise à prévoir des ressources et des services requis pour atteindre des objectifs déterminés, selon un ordre de priorité établi, permettant ainsi le choix d'une solution préférable parmi plusieurs alternatives. Ce choix prend en considération le contexte et les contraintes internes et externes connues ou prévisibles dans le futur (Jourdain et Frossard, 1995).

3.1 Plan, planification et planificateur

En Intelligence Artificielle, le mot planification désigne un domaine de recherche qui se propose de concevoir des systèmes pouvant générer automatiquement, au travers d'une procédure formalisée, un résultat défini, sous la forme d'un système intégré de décisions appelé plan. Ce dernier se présente généralement sous la forme d'une collection organisé de descriptions d'opérations ; il est essentiellement destiné à guider l'action d'un ou plusieurs agents exécuteurs (systèmes robotiques ou humains) qui ont à agir dans un monde particulier pour atteindre un but prédéfini et qui ont donc à prendre des décisions adaptées aux situations successives qu'ils rencontrent. Par extension, le même mot désigne le processus qui élabore cet ensemble d'actions. Le planificateur ou générateur de plans est le système qui le produit (Régnier, 2005).

Classiquement un planificateur dispose en entrée d'un problème et d'un domaine de planification. Un problème de planification consiste en une description de l'état initial et du but à atteindre. Un domaine de planification est décrit par un ensemble d'actions qui vont permettre

des transitions entre les états (Baki et Bouzid, 2006). Une solution au problème de planification est un plan qui permet d'atteindre le but en partant de l'état initial. Cependant, il peut y avoir plusieurs plans possibles pour un même problème. Il faut donc trouver le meilleur plan pouvant être exécuté en respectant toutes les contraintes. Ce plan est appelé plan-solution

3.2 Planification basée sur les cas

La planification basée sur les cas est une approche qui s'inspire d'un aspect particulier du comportement humain. Généralement, l'homme ne génère pas des plans entièrement nouveaux à partir d'opérations de base, il utilise son expérience passée (succès ou échecs) pour s'aider à résoudre les nouveaux problèmes qui se posent à lui. Planifier revient alors à essayer de synthétiser un plan-solution, en réutilisant au mieux les plans déjà produits dans des situations similaires et en les modifiant pour les adapter à la nouvelle situation (Régnier, 2005).

En planification à partir de cas, un problème de planification est la spécification d'un état initial et d'un but à atteindre. Une solution est un plan permettant d'atteindre le but en partant de l'état initial (Osty et al., 2008).

La planification à partir de cas a été appliquée dans plusieurs domaines. Nous citons par exemple le système GerAmi (Corchado et al., 2008) qui a été appliqué dans le domaine médical. GerAmi est un système qui vise à aider les établissements de santé dans le traitement des personnes âgées (les patients atteint d'Alzheimer ou autres handicaps). La base de GerAmi est constituée par l'agent GerAg qui intègre un mécanisme de planification basée sur les cas dans le but d'optimiser les horaires de travail et de fournir les mises à jour des données des patients. A chaque infirmière et médecin est assigné un agent GerAg qui inclut des informations sur les patients. Les membres du personnel effectuent alors leurs fonctions en suivant le plan de leur agent. S'ils ont besoin de modifier le plan initial pour tenir compte des retards ou des alarmes, le GerAg peut replanifier en temps réel. Le système de planification à partir de cas de GerAg vise à définir les plans de travail quotidien pour chaque infirmière suivant les quatre étapes du système de raisonnement à base de cas : Remémoration, Adaptation, Révision, Apprentissage. Dans l'étape de recherche, le mécanisme de planification à partir de cas sélectionne à partir de la base de données des plans, les plans similaires au problème cible ou contenant des informations semblables (profil du patient, salles, équipements, etc.). Dans la phase de réutilisation, le système identifie tous les plans possibles qui satisfont les exigences de l'infirmière. Dans la phase de révision, l'infirmière révise le plan. Enfin, dans la phase de maintenance, le système ajoute le nouveau plan dans la base de données des plans.

4 Construction de la base de cas

Notre objectif consiste à proposer un système d'aide aux praticiens dans le traitement de la tuberculose. Pour qu'un système guidé par raisonnement à partir de cas puisse fonctionner, il doit disposer d'un certain nombre de cas constituant la base de cas. Ces cas doivent couvrir le mieux possible le domaine ciblé afin que des solutions intéressantes soient trouvées (Camelin et Compin, 2008). Pour la construction de la base des cas des différents traitements de la tuberculose, nous avons adopté les étapes suivantes :

4.1 Description du projet

Comme nous l'avons déjà mentionné dans la section 3.1, un planificateur dispose en entrée, d'un problème et d'un domaine de planification décrit par un ensemble d'actions. On appelle projet l'ensemble des actions à entreprendre afin de répondre à un besoin défini dans des délais fixés. Nous allons décrire le projet en représentant l'enchaînement des tâches (actions) sous forme d'un tableau (voir tableau 1) afin de générer le graphe ET/OU. Prenons l'exemple du traitement d'une maladie à déclaration obligatoire : la tuberculose. La tuberculose est une maladie infectieuse causée par des bactéries (Mycobacterium tuberculosis) qui s'attaquent généralement aux poumons mais parfois aussi à d'autres organes. Le traitement de la tuberculose diffère en fonction de l'age du patient et de différents autres facteurs. Le tableau 1 représente quelques tâches du traitement à suivre pour des patients atteints de tuberculose. A chaque tâche est associée sa description et les tâches qui la précèdent.

Rubriques	Tâches	Description	Tâches précédentes
	Début	Lancement du projet	-
Nouveaux nés de 0 à 1 mois	A	Examen et traitement par un spécialiste	Début
Enfants de 1 mois à 5ans	B	1^{er} test de dépistage Mantoux	Début
	C	Examen médical+radiographie	B
	D	Traitement de la tuberculose	C,F,H
	E	Début du traitement INH	C
	F	Test IGRA	E,H
	G	Stop INH	F
	H	$2^{ème}$ test de dépistage Mantoux	C
	Fin	Fin du projet	A, D, G

TAB. 1 – *Exemple de description d'un projet de traitement de la tuberculose.*

4.2 Modélisation du projet par un graphe ET/OU

Nous avons choisi ce type de modélisation car la modélisation graphique est plus expressive et facile à assimiler. Nous décrivons ci-dessous la notion de graphe ET/OU, de tâches et de contraintes.

Un graphe ET/OU est un graphe dont les nœuds représentent des tâches et les arcs représentent les relations entre les tâches. Une tâche représente l'action réalisée pendant une durée de temps et les relations entre les tâches sont les contraintes à satisfaire (Baki, 2006).

Le choix d'un plan-solution dépend de plusieurs critères : le temps, la probabilité, le coût et d'autres critères peuvent être pris en compte.

- Temps : une tâche doit s'exécuter pendant une certaine durée et dans un délai bien précis (préciser la date de début de l'exécution de la tâche et la date de fin de son exécution).
- Probabilité : la probabilité qu'une tâche s'exécute pendant une certaine durée est une probabilité conditionnelle puisqu'elle dépend, en plus de sa probabilité d'exécution pendant cette durée, des probabilités des durées des tâches précédentes.

– Coût : l'exécution de n'importe quelle activité a un coût que ce soit en argent, en matière première, en temps, etc.

Le graphe ET/OU généré à partir du projet décrit dans le tableau 1 est illustré par la figure 1. Les tâches *Début*, $A, B, C, D, E, F, G, H, Fin$ correspondent aux nœuds du graphe ET/OU. Pour les relations entre les tâches, on a pris en compte seulement les contraintes de précédence dans cet exemple (la tâche *Début* précède les tâches A et B, la tâche B précède la tâche C, la tâche C précède les tâches D, E, H, etc.).

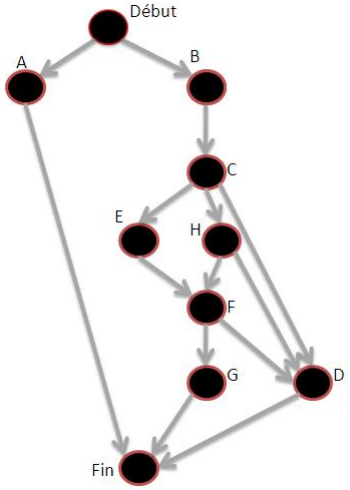


FIG. 1 – *Graphe ET/OU du projet.*

4.3 Génération des plans

Après avoir construit le graphe ET/OU, nous allons appliquer des algorithmes de planification pour déterminer les plans possibles. Par exemple, pour un enfant de 2 ans, il doit d'abord passer par le 1er test de dépistage Mantoux ; ensuite, il subit un examen médical et une radiographie ; enfin, le traitement de la tuberculose. Alors, on obtient le plan $P1$ suivant : $Début \rightarrow B \rightarrow C \rightarrow D \rightarrow Fin$

4.4 Construction et représentation des cas

Nous avons choisi de simplifier la description des cas afin de focaliser notre travail sur la remémoration par arbre de décision. Mais par la suite, nous ajouterons d'autres descripteurs liés à un problème de planification par exemple les contraintes.

Pour construire les cas, nous allons associer la durée du traitement, la probabilité de guérison et le coût du traitement à chaque plan obtenu dans l'étape précédente en fonction de ses tâches. Donc les cas seront représentés par des descripteurs (durée, probabilité, coût) qui décrivent la partie problème et le plan correspondant qui représente la partie solution. Par

exemple, si on associe au plan $P1$ une durée 07, une probabilité 0.4 et un coût 10, on obtient le cas représenté dans le tableau 2. La base de cas va être construite à partir de ces cas.

Durée	Probabilité	Coût	Plan
07	0.4	10	P1

TAB. 2 – *Représentation d'un cas.*

5 Recherche des cas par arbre de décision

Une fois la base de cas construite, nous pourrons entamer le processus de raisonnement à partir de cas. Nous nous intéressons à la première étape, la remémoration qui consiste à rechercher dans la base de cas tous les plans candidats à être un plan-solution. Cette étape nécessite d'utiliser une mesure de similarité entre cas (Rifqi, 2010). Des mesures de similarités sont alors à définir sur les indices constituant la partie problème du cas. Cette notion de similarité entre cas a fait l'objet de nombreux travaux (Cram et al., 2007). La mesure de similarité la plus couramment utilisée dans le raisonnement à partir de cas est la méthode des k plus proches voisins. C'est une méthode dédiée à la classification qui permet de prendre des décisions en recherchant des cas similaires déjà résolus (Hastie et al., 2001).

Le choix de la mesure de similarité dépend essentiellement de la méthode de représentation des cas (Hamza, 2008). Dans notre cas, nous avons opté pour la remémoration des cas par classification automatique. Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains traits descriptifs. Elles trouvent leur utilité dans un grand nombre d'activités humaines et en particulier dans la prise de décision automatisée. Les arbres de décisions sont une des nombreuses méthodes de classification.

Les travaux de raisonnement à partir de cas ont parfois recours aux arbres de décision. Nous citons par exemple le système de raisonnement à base de cas NodalCBR dans lequel la remémoration est basée sur une sélection de caractéristiques discriminantes issue d'un arbre de décision où les nœuds correspondent aux différents diagnostics et les feuilles correspondent aux différentes classes (Cunningham et Smyth, 1994). Houeland (2011) a utilisé une forêt d'arbres de décision pour déterminer la similarité entre les cas.

Parmi les méthodes de fouille de données, on trouve les arbres de décision (Breiman et al., 1984). Les arbres de décision sont des structures qui représentent des ensembles de décisions. Ces décisions génèrent des règles pour la classification d'un ensemble de données (Belacel, 1999). Le processus de la classification consiste à affecter une classe à des objets à l'aide d'un modèle entraîné sur un autre ensemble d'objets. Un arbre de décision est composé de feuilles, de nœuds et de branches. Chaque nœud de l'arbre correspond à une propriété de l'objet à classer, appelé attribut. Chaque branche de l'arbre correspond à une valeur possible de l'attribut père et chaque feuille de l'arbre correspond à une classe (Boumahdi et al., 2009).

Soit $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ l'échantillon d'apprentissage, c'est l'ensemble d'objets ou de cas qui va être utilisé pour la construction de l'arbre de décision. Chaque cas ω_i est décrit par une série de variables $X_1, X_2, ..., X_P$ dites variables descriptives. Nous désignons par l_j le nombre des différentes modalités affectées à la variable X_j . A chaque cas ω_i est associé

un attribut cible ou une classe notée Y qui prend ses valeurs dans l'ensemble des classes $C = \{c_1, c_2, ..., c_m\}$ (Atmani et Beldjilali, 2007).
Pour illustrer cette forme de notation, considérons un problème de planification du traitement de la tuberculose où les cas sont décrits par trois variables descriptives X_1, X_2, X_3 et auxquels est associée une classe $Plan1$ ou $Plan2$.
X_1 : *Durée*, qui peut prendre deux valeurs x_1^1=*Courte*, x_1^2=*Longue*
X_2 : *Probabilité*, qui peut prendre trois valeurs x_2^1=*Improbable*, x_2^2=*Peu probable*, x_2^3 =*Très probable*
X_3 : *Coût*, qui peut prendre trois valeurs x_3^1=*Faible*, x_3^2=*Raisonnable*, x_3^3=*Elevé*

ω	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$Y(\omega)$
ω_1	*Courte*	*Improbable*	*Faible*	*Plan1*
ω_2	*Longue*	*Peu probable*	*Raisonnable*	*Plan2*
ω_3	*Longue*	*Très probable*	*Elevé*	*Plan2*
ω_4	*Courte*	*Très probable*	*Faible*	*Plan2*
ω_5	*Courte*	*Peu probable*	*Elevé*	*Plan1*

TAB. 3 – *Echantillon d'apprentissage* Ω_a.

A partir de cet exemple, on construit un arbre dit de décision tel que :
– Chaque nœud correspond à un test sur la valeur $X_j(\omega)$ d'un attribut X_j qui possède l_j valeurs possibles $(x_j^1, ..., x_j^{l_j})$;
– Chaque branche partant d'un nœud correspond à une valeur x_j^v du test sur X_j avec $v = 1, ..., l_j$;
– A chaque feuille est associée une valeur c_k de l'attribut cible Y.

Pour construire l'arbre de décision illustré dans la figure 2, nous appliquons l'algorithme de la méthode SIPINA (Zighed et Rakotomalala, 1996) sur l'échantillon d'apprentissage Ω_a (tableau 3). Nous pouvons déduire quatre règles de classification $R1, R2, R3$ et $R4$ qui sont de la forme : Si Condition Alors Conclusion.
$R1$: Si ($X1$ = *Longue*) alors *Plan2*
$R2$: Si ($X1$ = *Courte* et $X2$ = *Improbable*) Alors *Plan1*
$R3$: Si ($X1$ = *Courte* et $X2$ = *Peu probable*) Alors *Plan1*
$R4$: Si ($X1$ = *Courte* et $X2$ = *Très probable*) Alors *Plan2*

L'arbre de décision peut être ensuite exploité de différentes manières : pour classer de nouvelles données, estimer un attribut, extraire des règles de classification concernant l'attribut cible, interpréter la pertinence des attributs, etc. En ce qui nous concerne, c'est surtout pour la classification de nouvelles données.

Dans notre étude, nous avons décrit les cas par une partie problème constituée de descripteurs et une partie solution représentant le plan à suivre en fonction des valeurs des descripteurs. Nous allons d'abord utiliser la base cas comme échantillon d'apprentissage pour la construction de l'arbre de décision où les nœuds correspondent aux descripteurs qui décrivent la partie problème du cas et les feuilles correspondent à la partie solution du cas (classes ou plans). Ensuite, lorsqu'un nouveau cas se présente, nous commençons la phase de recherche en utilisant l'arbre de décision. Il s'agit de classer une nouvelle donnée (le nouveau cas) selon les valeurs de ses descripteurs et d'aboutir ainsi à un plan. Par exemple, pour un nouveau cas qui

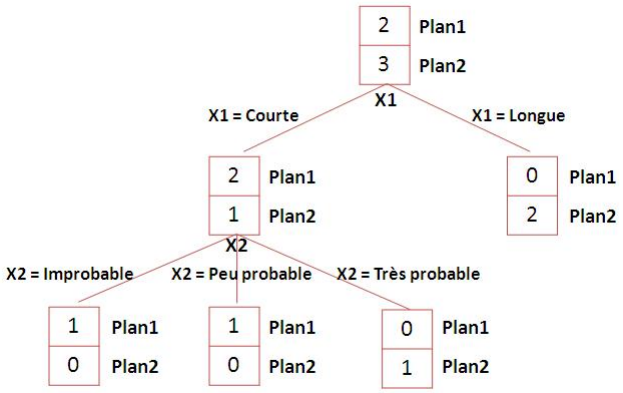


FIG. 2 – *Arbre de décision de l'échantillon d'apprentissage* Ω_a.

a comme descripteurs $X1(\omega) = Courte, X2(\omega) = Improbable, X3(\omega) = Raisonnable$, si on applique les règles de classification obtenues à partir de l'arbre de décision, on obtient la classe $c1 = Plan1$.

6 Conclusion

Notre approche consiste à combiner la planification et le raisonnement à base de cas en vue de profiter de l'expérience acquise et d'optimiser la durée de calcul. Au lieu d'utiliser des algorithmes de planification nous utilisons le raisonnement à base de cas qui va nous permettre de gagner un temps considérable en évitant de chercher à résoudre des problèmes déjà traités. Pour ce fait, nous construisons d'abord la base de cas de notre planificateur en utilisant un graphe ET/OU et des algorithmes de planification. Ensuite, nous utilisons l'arbre de décision pour la phase de remémoration du raisonnement à partir de cas. Nous avons opté pour l'arbre de décision car il représente un formalisme compréhensible par tout utilisateur (Boumahdi et al., 2009). C'est aussi une méthode que l'on peut présenter assez rapidement à un public non spécialiste du traitement des données sans se perdre dans des formulations mathématiques délicates à appréhender (Rakotomalala, 2005). Donc l'arbre de décision est un outil d'aide à la décision qui va nous permettre d'optimiser la durée de calcul.

Pour le calcul de la similarité dans la phase de remémoration, on utilise généralement les k plus proches voisins. Donc nous comptons implémenter et comparer notre approche (indexation par arbre de décision) avec les k plus proches voisins dans le domaine de tuberculose et d'un autre coté introduire la notion du traitement flou pour la description des cas.

La structure des cas que nous avons utilisée est assez simple. Nous avons décrit la partie problème des cas par une durée, un coût et une probabilité. On pourrait par la suite utiliser une représentation un peu plus complexe en ajoutant par exemple les ressources disponibles ainsi que d'autres contraintes.

Nous envisageons par la suite, l'application de notre planificateur dans un contexte médical.

Références

Ackoff, R. (1973). *Méthodes de planification de l'entreprise*. Editions d'Organisation, Paris.

Atmani, B. et B. Beldjilali (2007). Knowledge discovery in database : Induction graph and cellular automaton. *Computing and Informatics Journal 26*, 1001–1027.

Badra, F. (2009). *Extraction de connaissances d'adaptation en raisonnement à partir de cas*. Thèse de doctorat, Université Henri Poincaré Nancy 1.

Baki, B. (2006). *Planification et ordonnancement probabilistes sous contraintes temporelles*. Thèse de doctorat, Université de CAEN.

Baki, B. et M. Bouzid (2006). Planification et ordonnancement probabilistes sous contraintes temporelles. *Actes du 15e congrès francophone de Reconnaissance des Formes et Intelligence Artificielle RFIA'2006*, 99–107.

Belacel, N. (1999). *Méthodes de classification multicritère méthodologie et applications à l'aide au diagnostic médical*. Thèse de doctorat, Université Libre de Bruxelles.

Berkoune, D., M. Rekik, A. Ruiz, et J. Renau (2011). Système d'aide à la décision pour le déploiement en situation d'urgence. *9e Congrès international de génie industriel*, Québec (CANADA).

Boumahdi, M., J. Dron, S. Rechak, et O. Cousinard (2009). Utilisation de l'arbre de décision pour la détection des défauts de roulements. *Revue Internationale sur l'Ingénierie des Risques Industriels (JI-IRI) 2*.

Breiman, L., J. H. Friedman, R. Olshen, et C. Stone (1984). *Classification And Regression Trees*. New York : Chapman and Hall.

Camelin, J. et J. Compin (2008). *Modèle de Data Mining : Le type de logements en fonction des paramètres socio professionnels*. Travail d'étude et recherche.

Cavallucci, D., F. Rousselot, et J. Renaud (2010). Comparaison de la méthode de conception inventive (mci) basée sur la triz et de l'approche du raisonnement à partir de cas (ràpc). *18ème Atelier Raisonnement à Partir de Cas RàPC 2010*.

Corchado, J., J. Bajo, et A. Abraham (2008). Gerami : Improving healthcare delivery in geriatric residences. *IEEE Intelligent Systems 23*, 19–25.

Cram, D., B. Fuchs, A. Mille, et Y. Prié (2007). Raisonner à partir de l'expérience tracée : application à un environnement collaboratif. projet procogec. pp. 1–37.

Cunningham, P. et B. Smyth (1994). A comparison of model-based and incremental case-based approaches to electronic fault diagnosis. *American Association of Artificial Intelligence*.

Hamza, H. (2008). *Application du raisonnement à partir de cas à l'analyse de documents administratifs*. Thèse de doctorat, Université Nancy 2.

Hastie, T., R. Tibshirani, et J. Friedman (2001). *The Elements of Statistical Learning : Data mining, Inference and Prediction*. Springer. NewYork.

Houeland, T. (2011). An efficient random decision tree algorithm for case-based reasoning systems. *24th International Florida Artificial Intelligence Research Society Conference*.

Jourdain, A. et M. Frossard (1995). Les nouveaux outils de planification sanitaire. *Actualité et Dossier en Santé Publique 11*, I–XL.

Lamontagne, L. et G. Lapalme (2002). Raisonnement à base de cas textuel état de l'art et perspectives futures. *Revue de l'intelligence artificielle 16*, 339–366.

Lieber, J. (2008). *Contributions à la conception de systèmes de raisonnement à partir de cas*. Habilitation à diriger des recherches, Université Henri Poincaré Nancy1.

Osty, P., F. LeBer, et J. Lieber (2008). Raisonnement à partir de cas et agronomie des territoires : constructions croisées. *Revue d'Anthropologie des Connaissances 2*, 169–193.

Rakotomalala, R. (2005). Arbres de décision. *Revue MODULAD 33*, 163–187.

Régnier, P. (2005). *Algorithmes pour la planification*. Habilitation à diriger des recherches, Université Paul Sabatier.

Rifqi, M. (2010). *Mesures de similarité, raisonnement et modélisation de l'utilisateur*. Habilitation à diriger des recherches, Université Pierre et Marie Curie.

Roy, B. (1985). *Méthodologie multicritère d'aide à la décision*. Economica, Paris.

Ruland, C. et S. Bakken (2002). Developing, implementing, and evaluating decision support systems for shared decision making in patient care : a conceptual model and case illustration. *Journal of Biomedical Informatics 35*, 313–321.

Trilling, L., A. Guinet, D. LeMagny, et P. Moullier (2007). Modèle de planification des médecins anesthésistes : un problème multicritère. *7e Congrès international de génie industriel*, Trois–Rivières, Québec (CANADA).

Zighed, D. et R. Rakotomalala (1996). *Sipina-W for Windows : User's Guide*. Laboratory ERIC, University of Lyon 2.

Summary

Daily, we often use our experience to solve some problems experienced before. The case-based reasoning copy the human behavior. It solves problems by searching for similar cases already resolved in the past.
In this paper, we propose a new planning technique that we call DTR "Decision Tree for Retrieval".
DTR uses the principle of classification decision tree to calculate similarity.

Découverte de règles d'associations pour l'aide à la prévision des accidents maritimes

Bilal IDIRI*, Aldo NAPOLI*

*Mines ParisTech, CRC
Rue Claude Daunesse, 06904 Sophia Antipolis, France
Prénom.Nom@mines-paristech.fr

Résumé. Les systèmes de surveillance maritime permettent la récupération et la fusion des informations sur les navires (position, vitesse, etc.) à des fins de suivi du trafic maritime sur un dispositif d'affichage. Aujourd'hui, l'identification des risques à partir de ces systèmes est difficilement automatisable compte-tenu de l'expertise à formaliser, du nombre important de navires et de la multiplicité des risques (collision, échouement, etc). De plus, le remplacement périodique des opérateurs de surveillance complique la reconnaissance d'événements anormaux qui sont épars et parcellaires dans le temps et l'espace. Dans l'objectif de faire évoluer ces systèmes de surveillance maritime, nous proposons dans cet article, une approche originale fondée sur le data mining pour l'extraction de motifs fréquents. Cette approche se focalise sur des règles de prévision et de ciblage pour l'identification automatique des situations induisant ou constituant le cadre des accidents maritimes.

1 Introduction

L'activité maritime est un secteur important alliant intérêts publics et privés. Elle compte à elle seule 90% des échanges internationaux avec 80% du transport d'énergie (CNUCED, 2009). Pour protéger ce secteur, plusieurs dispositifs de sécurité ont été mis en place comme le développement de systèmes de surveillance maritime : SpatioNav en France, SIVE en Espagne, MEVAT en Finlande (Morel, 2009). Ces systèmes de surveillance affichent les pistes de navires additionnées à d'autres informations complémentaires (cargaison, vitesse, cap, port de départ, etc.) sur une carte numérique pour permettre aux agents d'état la surveillance du trafic maritime. Vu le nombre important de navires à surveiller, l'immensité des territoires maritimes, la multiplicité des risques et l'organisation de la criminalité en mer, les systèmes de surveillance maritime sont insuffisamment adaptés pour l'aide à l'identification des risques maritimes et doivent évoluer pour faire face à ces nouveaux défis. Parmi les travaux traitant de ces questionnements de surveillance dite "*intelligente*" ou de "*nouvelle génération*", nous pouvons citer le projet PANDA (Darpa, 2005) du ministère Américain de la défense, considéré comme le projet initiateur qui a inspiré de nombreux travaux. Nous pouvons également citer, le projet SCANMARIS (Morel et al., 2008) qui a évalué les algorithmes de détection automatique des comportements anormaux et TAMARIS (Morel et al., 2011) qui propose une couche fonc-

tionnelle pour l'authentification des comportements suspects à partir d'un ensemble d'alertes générées par SCANMARIS.

Dans le but d'améliorer la surveillance maritime, nous proposons la mise en place d'un système d'aide à l'identification des comportements anormaux de navires et la découverte *a priori* des situations à risques en utilisant une approche originale basée sur la fouille de données spatiales (Agrawal et al., 1993) (Srikant et Agrawal, 1995) (Zhenhui et al., 2010) (Lee et al., 2008) (Cao et al., 2007) (Marven et al., 2007). L'idée est d'explorer les historiques de données de déplacement de navires et d'accidentologie dans le but de découvrir les connaissances régissant la survenue des risques et identifiant les comportements anormaux des navires. Ces connaissances vont servir à l'identification automatique des risques maritimes à partir de bases de faits (flux de déplacement, météorologie, océanographie, carte de navigation).

Dans cet article, nous nous intéressons à l'extraction de motifs fréquents à partir d'une base de données d'accidents de navires dans le but de découvrir des règles [1] d'expertise pour l'aide à la détection de situations à risque (collision, échouement, avarie, etc.). Un exemple de règles d'associations est "*Les accidents de navires britanniques de type Roll-on/Roll-off* [2] *sont dans 76% des cas localisés dans les ports ou à proximité des ports*".

2 Extraction des règles d'associations pour la prévention des risques maritimes

L'extraction de règles d'associations est un problème non supervisé de data mining qui permet, à partir des itemsets [3] apparaissant fréquemment ensemble dans une base de données, d'extraire des règles de connaissance. Ce problème a été proposé pour la première fois par Agrawal (Agrawal et al., 1993) pour l'analyse du panier de la ménagère dans le but d'améliorer les ventes. La découverte de règles d'associations dans une base de données de transactions (les paniers) consiste à chercher les produits (itemsets fréquemment achetés ensemble). Nous allons voir par la suite, comment extraire ces règles dans un historique d'accidents de navires dans le but de construire une base de connaissances utile à l'identification automatique des risques maritimes.

L'extraction de règles d'associations ou le data mining en général fait partie intégrante d'un processus d'extraction de connaissance de données (ECD). Un ECD regroupe l'ensemble des méthodes et des outils qui vont nous permettre de transformer les données volumineuses et hétérogènes des accidents maritimes en connaissances utiles à la prise de décision.

Nous allons détailler par la suite chaque étape du processus d'extraction de connaissances appliqué à l'historique de données d'accidents.

2.1 Sélection des données

Nous avons à notre disposition une base de données [4] recensant les accidents qui ont affecté ou se sont produits à bord de navires entre *1991* et *2009*. Ces données concernent les navires

1. Une règle est l'unité composant la connaissance. Elle est de de la forme $A -> B$, tels que A est appelé *antécédent* de la règle et B est appelé *conséquent*. L'intersection entre A et B est vide.
2. Un navire roulier utilisé pour transporter des véhicules grâce à une ou plusieurs rampes d'accès.
3. Un itemset est un ensemble d'items, et un item est une occurence d'un objet de la base.
4. Cette base nous a été fournie par Marine Accident Investigation Branch (MAIB) à titre gracieux.

britanniques se trouvant n'importe où dans le monde et les navires d'autres nationalités se trouvant dans les eaux territoriales britanniques au moment de leur accident. La base de données contient *14 900* cas d'accidents et d'incidents qui concernent *16 230* navires.

Nous avons sélectionné dans cette base, les données qui décrivent les accidents (type d'accident, position, temps, etc.), les caractéristiques des navires (identifiant IMO, type du navire, âge du navire, longueur, etc.) et la description de l'environnement (visibilité, état de la mer, force du vent, etc.). Cette sélection de données va constituer le contexte d'exploration sur lequel va porter l'extraction de règles d'associations dans le but de trouver les relations entre les différents facteurs de situations. La sélection des attributs sur lesquels va porter notre analyse va réduire le nombre de variables à considérer, le nombre de règles générées et ainsi faciliter l'interprétation des résultats.

Dans la suite de cet article, nous allons voir comment mettre les données brutes en une forme exploitable par les algorithmes d'extraction de règles d'associations.

2.2 Préparation des données

Avant toute exploration de données par les méthodes de data mining, une étape de préparation de ces données est nécessaire pour permettre leur exploitation. La préparation des données est difficile et demande plusieurs itérations compte-tenu de son lien fort avec la qualité des résultats. En effet, la quantité et la qualité des données ont un impact direct et significatif sur la qualité des règles obtenues. Nous nous proposons d'étudier dans cette section la distribution des variables pour identifier les anomalies (données manquantes, incohérence, imprécision, etc.), les corriger et préparer le contexte d'exploration.

1. **Population d'accidents non représentative**

   Nous avons remarqué que notre base de données n'était pas représentative de la population globale d'accidents mondiaux de navires car 82% des accidents concernent des navires du Royaume-Uni. Pour avoir une population d'accidents représentative, nous avons réduit notre étude aux accidents de navires britanniques.

2. **Données manquantes**

   Dans le but d'améliorer la qualité des résultats, nous avons envisagé plusieurs approches pour nettoyer ces données : la pondération par des moyennes, des médianes ; la prédiction des valeurs manquantes (Jami et al., 2005) ; etc..

3. **Variables continues et regroupement de classes**

   Les algorithmes d'extraction de règles d'associations ne prennent pas en considération les variables (attributs) continues dans leur processus d'extraction. Pour ne pas perdre d'informations en entrée de ces algorithmes, nous avons discrétisé les variables continues en les séparant en classes (intervalles). Dans notre cas, nous avons choisi le critère d'effectifs égaux pour éviter de biaiser les résultats des algorithmes d'extraction de règles d'associations. Ces algorithmes sont basés sur la découverte d'itemsets fréquents. Une classe contenant plus d'effectifs a donc plus de chance d'apparaitre dans les règles en sortie.

   Nous avons comptabilisé des variables discrètes ayant une distribution hétérogène de leur effectif. Cette hétérogénéité révèle les classes ayant les plus grandes fréquences

d'apparition dans les règles d'associations et peuvent empêcher l'apparition des fameuses pépites d'or. Les navires de pêche, par exemple, apparaissaient presque systématiquement dans les règles d'associations car ils représentent 66% de l'effectif total. Pour faire apparaitre les autres catégories de navires, nous avons regroupé toutes les catégories de navires en trois grandes classes :

- Classe Transport : Avec un effectif de 32%, elle regroupe tous les navires de transport de personnes, d'hydrocarbures (Tanker) et de marchandises,
- Classe Plaisance : La classe des navires de plaisance représente un effectif de 1.2%,
- Classe Pêche : Les navires de pêche représentent un effectif de 66%.

4. **Données aberrantes**

   Les données aberrantes sont des données erronées. Leur identification demande d'avoir une bonne connaissance du domaine étudié. La répartition des accidents sur une carte numérique nous a permis d'identifier et d'écarter les positions aberrantes localisées sur terre, loin des zones de navigation.

   Ces erreurs de positionnement des accidents sont peut-être dues aux dysfonctionnements de GPS ou à une mauvaise saisie des coordonnées. L'analyse des valeurs extrêmes (les valeurs maximales et minimales) nous a permis aussi de détecter un ensemble de valeurs erronées comme le cas de valeurs négatives de la variable Age-of-vessel et d'autres encore.

2.3 Extraction des règles d'associations

Nous avons appliqué l'algorithme Apriori, implémenté par Christian Borgelt et implémenté dans le package Rattle 2.6.4 de R, sur les données préparées dans l'étape précédente (voir section 2.2). Nous avons fait varier les seuils du support [5] "*minsupp*", de la confiance [6] "*minconf*" et nous avons extrait plusieurs fichiers de règles d'associations. En plus des mesures support-confiance, nous utilisons aussi une autre mesure, appelée Lift, pour vérifier que les résultats obtenus ne sont pas le fruit du hasard. Si la mesure est supérieure à 1, la règle est considérée comme intéressante.

Nous avons défini trois grandes catégories à partir des règles découvertes pour les regrouper et faciliter leur exploitation :

- *Règles de prédiction* : Nous appelons règle de prédiction toute règle ayant son antécédent connu a priori, son conséquent non connu et la confiance de la règle est supérieure à 50%. Une règle de prédiction peut-être du genre "*Si nous avons un contexte Ci alors à c% il implique un accident de type Ti* ",
- *Règles de ciblage* : Ce sont les règles de connaissances générales qui identifient les relations entres les différentes dimensions (type de navire, type d'accident, zone maritime, etc.). L'antécédent et le conséquent de la règle sont connus mais pas la relation d'implication entre les deux parties. Les règles sont par exemple du genre "*Les accidents de navires de type Ti, concernent à c% les navires de type Ni*" et "*Les accidents de navires de type Ti sont localisés dans c% des cas dans la zone Zi*".,
- *Règles banales* : Ce sont les règles qui n'apportent pas d'informations nouvelles.

5. C'est un idicateur de fiabilité, il est égal au nombre d'occurences de la règle dans la base de données.

6. C'est un indicateur de précision de la règle qui est égal à la fréquence de la règle par rapport à la fréquence de l'antécédent.

2.4 Interprétation et validation des résultats

Nous avons découvert plus de 200 règles intéressantes, que nous ne pouvons pas toutes exposer. Quelques-unes de ces règles sont présentées et discutées ci-après. Ces règles ont été présentées à un sous-capitaine de la marine marchande qui nous a aidé à les interpréter.

- *Règle 1 (Règle de prédiction)* : Vessel-Category=Fish catching –> Incident-Type= Machinery Failure (supp=0.39 ; conf= 0.60 ; lift=1.23)
  La première règle nous informe que les incidents de navires de pêche sont causés dans 60% des cas par une panne mécanique.
- *Règle 2 (Règle de ciblage)* : Vessel-Category=Fish catching –> Vessel-Type= Trawler (supp= 0.14 ; conf= 0.43 ; Lift= 3)
  La deuxième règle, nous informe que si un accident concerne un navire de pêche alors dans 43% des cas c'est un chalutier. Selon le sous-officier, les chalutiers sont les plus exposés au risque de naufrage car ils tirent un chalut qui peut s'accrocher et entrainer vers le fond le chalutier sans que les pêcheurs n'aient le temps de l'abandonner.
- *Règle 3 (Règle Banale)* : Vessel-Category=Passenger –> Pollution-Caused=No (supp= 0.15 ; conf= 0.73 ; lift= 1.2)
  Enfin, la dernière règle, présente une règle banale (inutile). La règle signifie que les accidents de navires transportant des voyageurs ne causent pas de pollution ce qui est normal car ils transportent des passagers et non des substances polluantes.

3 Conclusion et perspectives

Nous avons présenté dans cet article, une approche originale de découverte de règles d'associations appliquée à des données spatiales statiques d'accidentologie de navires. Le résultat obtenu est un ensemble de règles de connaissance de prévision et de ciblage des situations à risque. Nous nous sommes focalisés dans cet article sur les risques liés à la sécurité (collision, échouement, etc.) et nous projetons de travailler par la suite sur les données spatiales dynamiques de déplacement de navires (données AIS). L'exploration automatique de ces données peut révéler des modèles intéressants (les outliers de trajectoire, les périodiques, etc.) pour l'identification automatique des comportements anormaux de navires.

Dans la perspective d'améliorer les systèmes de surveillance maritime, les connaissances obtenues vont être intégrées dans un moteur de règles pour les exécuter d'une manière continue sur des flux de données de déplacements de navires, météorologique et de contexte. Le déploiement de ces règles va permettre d'identifier automatiquement dans ces flux de données, les situations vérifiant une ou plusieurs règles de connaissance, identifiant ainsi en quasi temps réel les situations à risque et les comportements anormaux de navires.

Avertissement

Cet article a été accepté à la conférence EGC'2012.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining Association Rules between Sets of Items in Large Databases. In *the 1993 ACM SIGMOD International Conference on Management of Data, Washington*, Number May, Washington, D.C., pp. 207–216.

Cao, H., N. Mamoulis, et D. Cheung (2007). Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE Transactions on Knowledge and Data Engineering 19*(4), 453–467.

CNUCED (2009). Etude sur les transports maritimes. Technical report, ConférenCe des nations Unies sUr le CommerCe et le développement, Geneva.

Darpa (2005). Predictive Analysis For Naval Deployment Activities (PANDA).

Jami, S., T.-Y. Jen, D. Laurent, G. Loizou, et O. Sy (2005). Extraction de règles d'association pour la prédiction de valeurs manquantes. *Revue africaine de la recherche en informatique et mathématiques appliquées (AREMA) 3*(numéro spécial CARI'04), 103–124.

Lee, J.-G., J. Han, et X. Li (2008). Trajectory Outlier Detection : A Partition-and-Detect Framework. In *Data Engineering, IEEE International Conference on Data Engineering (ICDE'2008)*, Cancun, Mexique, pp. 140–149.

Marven, C., R. Canessa, et P. Keller (2007). *Exploratory Spatial Data Analysis to Support Maritime Search and Rescue Planning*, pp. 271–288. New York : Springer Berlin Heidelberg.

Morel, M. (2009). SisMaris : Système d'Information et de Surveillance MARitime pour l'Identification des comportements Suspects de navire. In *première Conférence Méditerranéenne Côtière et Maritime CM 2*, Hammamet - Tunisie, pp. 261–264.

Morel, M., V. Flori, O. Poirel, A. Napoli, P. Salom, et G. Proutiere Maulion (2011). Trairtement et Authentification des MenAces et RISques en mer. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG08)*, Troyes, France.

Morel, M., A. Napoli, A. Littaye, J.-P. Georgé, et F. Jangal (2008). Surveillance et contrôle des activités des navires en mer. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG08)*, Troyes, France.

Srikant, R. et R. Agrawal (1995). Mining Sequential Patterns : Generalizations and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology : Advances in Database Technology*, Volume 1057, pp. 3–14. Springer-Verlag.

Zhenhui, L., M. Ji, J.-G. Lee, L. Tang, et J. Han (2010). MoveMine : Mining Moving Object Databases. In *International Conference On Management of Data (SIGMOD)*, Indianapolis.

Summary

Maritime surveillance systems allow the recovery and the fusion of information on vessels (position, speed, etc.) for monitoring traffic on a display device. Today, the automatic identification of risks through these systems is difficult because of the complexity of formalizing the expertise, the large number of ships and the multiplicity of risks (collision, grounding, etc.). In addition, the periodic replacement of surveillance operators complicates the recognition of abnormal events which are scattered and fragmented in time and space. With the aim to up-

grade the maritime surveillance systems, we propose in this paper, a novel approach based on data mining for the extraction of frequent patterns. This approach focuses on rules for forecasting and targeting for the automatic identification of situations inducing or constituting part of maritime accidents.

Une démarche d'analyse à base de patrons pour la découverte des besoins métier d'un Système d'Information Décisionnel

Aziza Sabri *, Laila Kjiri**

ENSIAS, Université Mohammed-V-Souissi, Rue Mohammed Ben Abdellah Regragui,
B.P. 713 Agdal, Madinat Al Irfane - Rabat, Maroc
*azizasabri@yahoo.com, **kjiri@ensias.ma
http://www.ensias.ma

Résumé. Le domaine d'analyse et de conception des Systèmes d'Information Décisionnels (SID) est un secteur très demandeur en techniques et méthodes nouvelles favorisant la réussite du processus de la prise de décision et minimisant le temps de la conception. En effet, notre travail s'inscrit dans le cadre d'une démarche d'analyse guidant les analystes durant tout le processus de conception d'un SID. Nous cherchons, à travers cette démarche, à guider la découverte des besoins métier d'une organisation exprimés sous forme de buts en introduisant la notion de contexte, afin de favoriser une bonne conception d'un SID en garantissant une éventuelle réutilisation. Deux éléments essentiels ont été produits à l'issue de cet article : une démarche d'analyse des besoins métier d'un SID et le patron processus établi suivant le formalisme P-SIGMA et qui est lié à notre proposition d'analyse des besoins décisionnels.

1 Introduction

Le SID est un outil pour aider un décideur à prendre la meilleure décision. Cette décision dépend des informations qui s'y trouvent et de l'utilisateur qui a pour fonction d'interroger et aussi de comprendre ce système. Ainsi, la réussite du processus de la prise de décision exige une prise en compte d'une procédure permettant d'extraire les informations nécessaires à la prise de décision et de les structurer. En outre, la phase d'analyse des besoins de SID est reconnue actuellement comme une tâche cruciale sur laquelle peut reposer le processus de la prise de décision. Ainsi, notre travail s'inscrit dans le cadre d'une démarche d'analyse guidant les analystes durant tout le processus de conception d'un SID (Sabri et al. (2011a)). Nous cherchons, à travers cette démarche, à guider la découverte des besoins métier d'une organisation pour favoriser une bonne conception d'un SID. Également, nous adoptons une approche par réutilisation à base de patrons (Gamma et al. (1995)) pour pouvoir contribuer à l'amélioration des activités d'ingénierie des besoins et notamment dans l'étape d'élucidation des besoins décisionnels.

L'article est organisé de la façon suivante. Tout d'abord, nous présentons la démarche d'analyse des besoins décisionnels proposée. Dans la section 3, nous présentons le diagramme

d'activités d'analyse des besoins décisionnels lié à notre démarche. Dans la section 4, nous présentons brièvement les approches à base de patrons. Dans la section 5, nous définissons le patron "démarche d'analyse des besoins décisionnels". Enfin, nous terminons notre article par une conclusion.

2 Proposition d'une démarche d'analyse des besoins décisionnels

Notre présent travail se situe dans la phase de conception d'un ED qui se décompose en quatre étapes : analyse des besoins, modélisation conceptuelle, modélisation logique et modélisation physique. Nous allons nous focaliser très précisément sur la phase d'analyse des besoins en adoptant une démarche mixte (Ghozzi et al. (2005), Ravat (2007), Annoni (2007)) de manière à prendre en considération les besoins des décideurs (Sabri et al. (2011a)) et en analysant les données sources existantes.

Dans la littérature, nous distinguons trois types de méthodes d'analyse des besoins décisionnels : à base de modèles existants (Winter et al. (2003), Soussi et al. (2005)) à base de requêtes (Phipps (2002), Ghozzi et al. (2005)) et à base des modèles de buts (Mazon et al. (2005), El Golli (2008)). Cependant, ces méthodes ne traitent pas les problèmes liés à la diversité des perspectives, des points de vue, des contextes et des profils acteurs. L'incohérence et l'ambiguïté sémantique ne sont également pas traitées. Egalement, ces démarches n'automatisent pas l'extraction des faits et des dimensions à partir des besoins formulés et elles n'encapsulent pas les besoins métier selon des contextes prédéfinis pour une éventuelle réutilisation. Ainsi, nous remédions à ces problèmes par la proposition d'une démarche explicitant un ensemble d'étapes qui traite le problème lié à l'incohérence sémantique en standardisant la formulation des besoins, guide leur recensement selon des contextes métier prédéfinis et automatise l'extraction des faits et des dimensions nécessaires dans l'établissement du schéma en étoile. Cette démarche se base donc sur les activités suivantes : identification des acteurs de SID, identification du contexte métier, recensement des besoins sous forme de buts, traitement et formalisation des buts collectés. Nous détaillons, dans ce qui suit, chacune de ces activités.

2.1 Identification des acteurs de système d'information décisionnel

On distingue trois catégories d'acteurs au sein d'une organisation : tactique, stratégique et système (Annoni (2007)). Le schéma suivant illustre la classification des acteurs d'un SID :

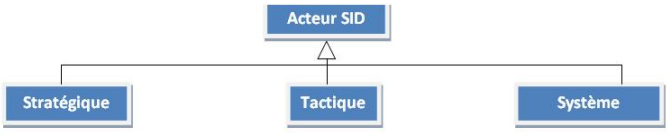


FIG. 1 – *Classification des acteurs de SID (Annoni, 2007).*

Les acteurs tactiques représentent les décideurs liés à un métier spécifique. Ils expriment principalement des besoins analytiques, fonctionnels ainsi que non fonctionnels (ergonomie d'outil). Les acteurs stratégiques sont des décideurs ayant une vision transversale des métiers liée à la direction de l'organisation. Ils expriment des besoins analytiques, fonctionnels, non

fonctionnels (sécurité, performance) mais surtout des besoins liés aux orientations stratégiques de l'organisation. Le groupe système, représenté par les concepteurs décisionnels et les administrateurs des systèmes d'information, est lié aux sources de données existantes et aux équipements décisionnels qui sont et qui seront employés.

2.2 Identification du contexte métier

Le contexte représente les informations locales concernant le point de vue d'un acteur d'une application (Rifaieh (2004)). Il permet de représenter les sémantiques reliées à la définition d'un objet (concept, composant, etc.) et ses relations avec d'autres objets. La notion de contexte peut aider à résoudre le problème de la multi-représentation des éléments et leur utilisation. De plus, le partage sémantique, en se basant sur le contexte, demande une méthode de séparation, de restructuration et d'organisation des besoins décisionnels selon leur contexte d'appartenance.

En outre, plusieurs paramètres interfèrent lors de la spécification des besoins d'un SID et peuvent influencer ce processus. Il s'agit de la multitude de contextes et de profils utilisateurs ; ceci entraîne l'expression de besoins différents, voire contradictoires en conséquence de la divergence des points de vues. Face à ces problèmes, nous allons définir les contextes métier de chaque organisation avant l'expression des besoins. Nous considérons un contexte comme étant un ensemble d'informations qui caractérise une situation. Un contexte peut détailler un autre contexte. Nous considérons également qu'un métier de l'organisation représente un ensemble d'activités définies dans des contextes différents. Le schéma suivant, que nous proposons, illustre le modèle des contextes :

FIG. 2 – *Modélisation du contexte métier.*

Nous proposons également le modèle organisationnel qui lie les besoins de SID au concept contexte. Le modèle illustre les acteurs de l'organisation qui exercent des responsabilités dans le processus d'expression de leurs besoins pour une prise de décision et la mise en place du SID adéquat.

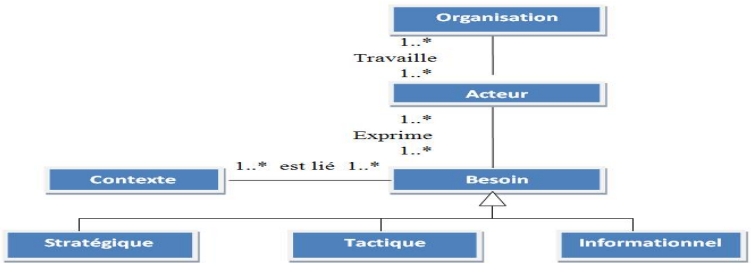


FIG. 3 – *Expression des besoins selon des contextes.*

Un besoin est associé à un ou plusieurs acteurs. Il est établi dans un contexte prédéfini. Un besoin est spécialisé en un besoin stratégique, indiquant un résultat souhaité à long terme de l'organisation. Un besoin tactique représentant une perspective particulière relative à un secteur d'activité en particulier ou un besoin informationnel formulé par des décideurs en termes d'informations recherchées.

2.3 Recensement des besoins sous forme de buts

Nous pensons que le langage le plus approprié à l'expression des besoins métier d'une organisation appartient au paradigme intentionnel (El Golli (2008)). Ce paradigme s'intéresse aux buts et aux objectifs à atteindre ainsi qu'aux résultats espérés et aux manières de les atteindre. Ainsi, cette étape consiste à collecter les besoins des acteurs de SID sous forme de buts selon des approches itératives. La collecte orientée utilisateur se base en principe sur une formulation des buts en langage naturel et cela à travers des techniques diverses (réunions, étude de la documentation existante, interviews, etc.). Dans notre démarche de recensement des besoins décisionnels, nous allons définir un diagramme de cas d'utilisation pour déterminer les acteurs de SID ainsi que leurs buts. Les acteurs de diagramme sont les acteurs de SID prédéfinis (stratégiques, tactiques et système) et les cas d'utilisation représentent les buts définis par chacun de ces acteurs. Le schéma suivant représente le diagramme UML proposé :

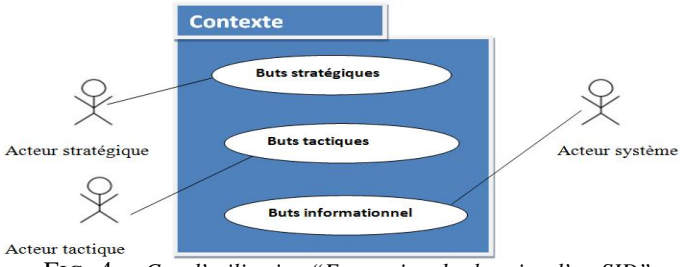


FIG. 4 – *Cas d'utilisation "Expression des besoins d'un SID".*

Nous présentons ainsi un exemple d'utilisation :

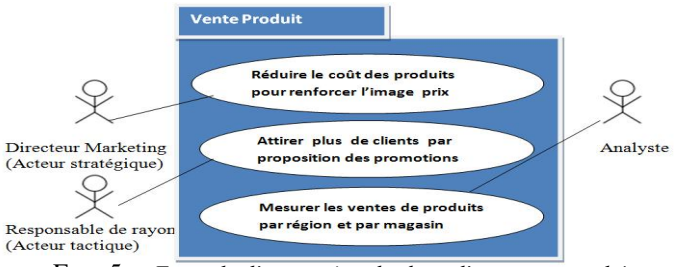


FIG. 5 – *Exemple d'expression des buts d'un supermarché.*

Chaque package de diagramme de cas d'utilisation représente le contexte défini et encapsule les buts exprimés par l'ensemble des acteurs de l'organisation suivant ce contexte.

2.4 Classification et formalisation des besoins

Après avoir recensé les besoins décisionnels sous forme des intentions (buts), nous allons les organiser suivant des règles de raffinement, de spécialisation et de généralisation. Cette classification va nous permettre de maintenir une analyse simple et systématique des buts décisionnels, d'expliciter les liens entre les buts dans un contexte donné et de faciliter la mise en place de schéma conceptuel de SID. Le modèle de classification des besoins intentionnels des décideurs est le suivant :

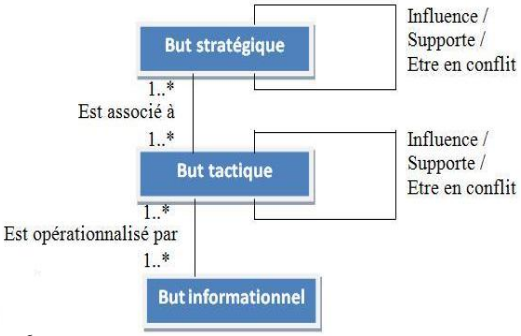


FIG. 6 – *Classification des besoins de SID (El Golli (2008)).*

Nous considérons les relations inter-buts de même type, qu'il s'agisse d'un but stratégique ou tactique : un but peut "être en conflit avec/ influencer/supporter" un autre but (El Golli (2008)). Aussi, un but stratégique est associé à un ou un ensemble de buts tactiques. Un but tactique peut être opérationnalisé par un ou un ensemble des buts informationnels. Dans cette approche, nous établissons le schéma conceptuel de SID à travers les buts informationnels. Ces buts sont opérationnels contenant des faits et des dimensions.

2.5 Modélisation des besoins décisionnels

L'étape de modélisation des besoins décisionnels consiste à établir le schéma en étoile à partir des buts informationnels prédéfinis. Cette modélisation consiste à détecter un ensemble de données spécifiques requises par des acteurs de SID pour l'exercice de leur métier (Marketing, Qualité, Finance, etc.). Afin de faciliter la tâche à l'analyste de SID pour détecter systématiquement ces données décisionnelles, nous avons défini une nouvelle version de modèle de buts informationnels (Sabri et al. (2011b)). La formulation de ces buts, suivant cette définition, va permettre de déterminer les dimensions et les faits à inclure dans le modèle multidimensionnel. En effet, le choix des faits, des tables de faits, des dimensions et des attributs d'une dimension est contextuel. Leur extraction sera donc directement faite à partir des phrases formulées sous forme de buts par les différents acteurs de SID. Ainsi, la nouvelle structure de but que nous proposons est la suivante :

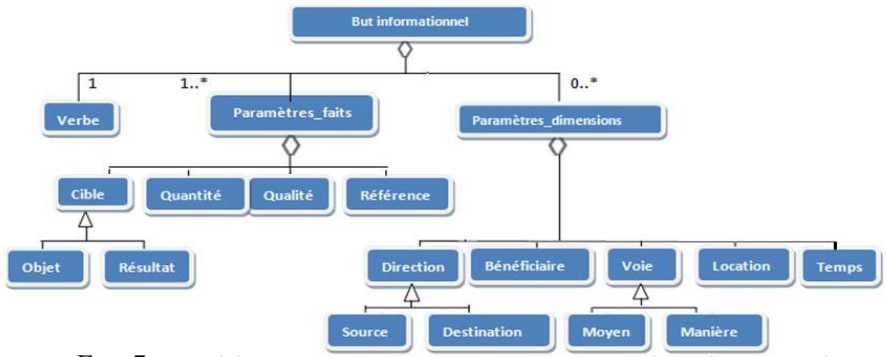


FIG. 7 – *Modèle sémantique proposé pour représenter un but informationnel.*

Une démarche d'analyse des besoins décisionnels

Chaque but informationnel sera formulé avec un "verbe", une section nommée "Paramètres_faits" et une autre nommée "Paramètres_dimensions" :

- Section "Paramètres_faits" : contient le nom et les indicateurs des tables de faits. Ainsi, la table de faits contiendra la cible, la référence, la qualité et la quantité :
  - **Cible :** La cible concerne les entités affectées par le but. Il y a deux types de cibles : l'objet et le résultat. L'objet existe avant la réalisation du but et peut éventuellement être modifié ou supprimé par celui-ci, alors que le résultat est l'entité qui résulte de la réalisation du but désigné par le verbe.
  - **Référence :** Elle est l'entité par rapport à laquelle une action est effectuée ou un état est atteint ou maintenu.
  - **Qualité :** C'est une propriété qui doit être atteinte ou préservée.
  - **Quantité :** Elle mesure la quantité qui devrait se produire.

- Section "Paramètres_dimensions" : contient les noms des tables de dimensions. Ainsi, les paramètres direction (source ou destination), location, voie (moyen ou manière), temps et bénéficiaire peuvent désigner des tables de dimensions :
  - **Direction :** Les deux types de direction sont appelés la source et la destination, et identifient respectivement l'endroit initial et l'endroit final de l'objet. La source est le point de départ du but (source d'information ou lieu physique).
  - **Voie :** spécialisée par les deux paramètres : la manière qui spécifie la façon d'atteindre le but et le moyen qui est l'outil par lequel le but est atteint.
  - **Bénéficiaire :** La personne ou le groupe en faveur de qui le but doit être atteint.
  - **Location :** Elle situe l'intention dans l'espace.
  - **Temps :** Il situe l'intention dans le temps.

Pour pouvoir établir une correspondance entre les tables de dimensions qui s'appliquent à chaque ligne de la table de faits, il faudrait savoir comment les décideurs décrivent les buts informationnels qui résultent de leurs intentions. Il importe de relier les tables de faits à des tables de dimensions représentant des descriptions susceptibles de prendre des valeurs particulières pour chaque mesure. Nous déterminons des mesures qui renseignent chaque ligne de la table des faits en répondant à la question : "Que mesure le système pour trouver l'information nécessaire aux buts informationnels du décideur ?". Ensuite, nous déterminons des tables de dimensions à partir de la section "Paramètres_dimensions".

Nous récapitulons la démarche proposée pour analyser les besoins d'un SID sous forme d'un diagramme d'activités qui explicite l'ensemble des étapes détaillées pour guider les concepteurs de SID durant tout le processus d'analyse. L'activité d'analyse peut se faire en trois principales étapes. La première consiste en la délimitation de l'environnement métier pour définir l'ensemble des contextes qui y sont attachés ainsi que la structure organisationnelle afin de localiser les acteurs à intervenir lors de l'étape suivante. La deuxième étape détaille le traitement des besoins recensés sous forme de buts. Ces buts seront classifiés, formalisés et reformulés suivant le modèle sémantique proposé dans (Sabri et al. (2011b)) afin de servir comme base d'extraction des données décisionnelles indispensables dans l'établissement du schéma conceptuel. Enfin, la troisième étape décrit la réalisation du schéma en étoile. Le schéma suivant illustre le diagramme d'activités que nous avons établi :

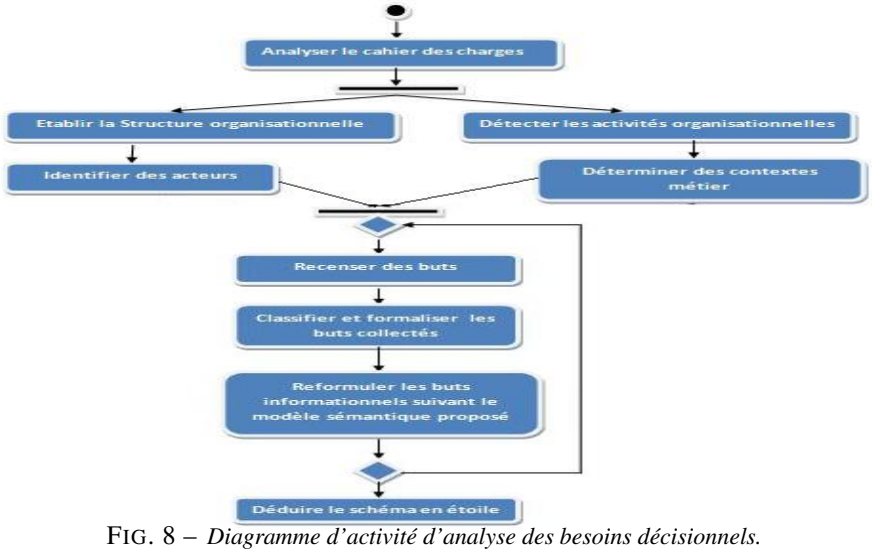


FIG. 8 – *Diagramme d'activité d'analyse des besoins décisionnels.*

L'activité "Recenser des buts" consiste à recenser les buts exprimés par chaque acteur, encapsuler les buts liés à un contexte donné dans un package et établir les diagrammes des cas d'utilisation. Quand à l'activité "classifier et formaliser les buts collectés", elle nécessite une classification des buts recensés en trois types : stratégique, tactique et informationnel. Puis, il faut associer chaque but stratégique à l'ensemble des buts tactiques qui y sont attachés et ensuite associer chaque but tactique à l'ensemble des buts informationnels qui lui sont attachés. Enfin, il faut définir les éventuelles relations (influencer, supporter, être en conflit) entre les buts de même type. Après l'activité "reformuler les buts informationnels suivant le modèle sémantique proposé", nous pouvons déduire le modèle en étoile en détectant les paramètres_faits et les paramères_dimensions et en déduisant les faits et les dimensions liés au contexte sélectionné.

3 Approches à base des patrons

Dans notre contexte, un patron est défini comme "une forme entièrement réalisée, originale ou un modèle accepté ou proposé pour une imitation ; quelque chose qui est vue comme un exemple normatif pouvant être copié, archétypé ou utilisé comme exemple" (Coad (1992)).

3.1 Classification des patrons

Dans la littérature, nous distinguons trois principaux types de patrons : les patrons d'analyse, les patrons de conception et les patrons d'implantation. Le patron d'analyse est une combinaison typique pour aider à construire des modèles pour la représentation des besoins de l'application et à transformer l'expression de ces besoins dans des modèles réutilisables (ACM (1996)). Il identifie des problèmes récurrents dans l'expression des besoins d'application de différents domaines. Les patrons d'analyse (Fowler (1997), Gzara (2000)) sont des exemples

des patrons exploités au niveau de la phase d'analyse des besoins. Quand au patron de conception (Gamma et al. (1995)), il identifie, nomme et abstrait des thèmes communs du domaine de la conception orientée objet. Il capture l'expérience et la connaissance liées à la conception en identifiant les objets, leurs collaborations et la distribution de leurs responsabilités. Concernant le patron d'implantation, il est spécifique à un langage de programmation en décrivant comment implanter certains aspects particuliers des composants ou des relations entre eux dans un langage de programmation donné (Fredj (2007)).

3.2 Le formalisme de patron P-SIGMA

Le formalisme P-SIGMA (Conte et al. (2001)) est une tentative d'uni ?cation des formalismes structurés qui ont été proposés (Coad (1992), Gamma et al. (1995), Bushmann et al. (1996)). Le Diagramme de classe suivant illustre le formalisme P-Sigma :

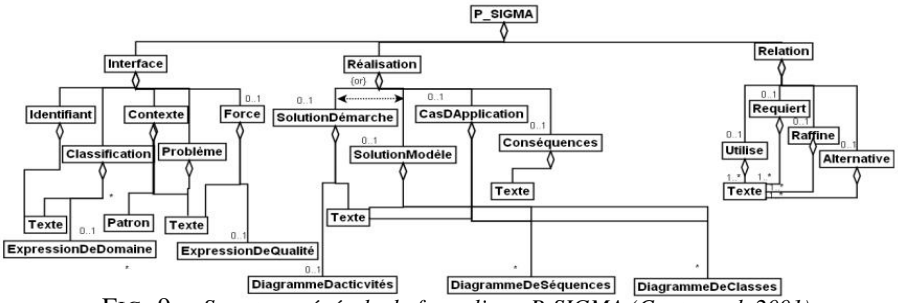


FIG. 9 – *Structure générale du formalisme P-SIGMA (Conte et al. 2001).*

Ce formalisme s'appuie sur trois principales parties : la partie Interface qui contient tous les éléments qui permettent la sélection d'un patron, la partie Réalisation qui exprime la solution d'un patron en termes de solution modèle et de solution démarche proposées par le patron et la partie Relation qui permet d'organiser les relations entre patrons. Ainsi, nous présentons les parties constituant le formalisme P-SIGMA retenu pour représenter notre démarche d'analyse des besoins décisionnels :

Partie	Rubrique	Champs
Interface	sigle	Le sigle du patron
	nom	Définit le couple problème/solution à partir duquel le patron pourra être référencé
	classification	Définit la fonction du patron par un ensemble de mots-clés du domaine
	contexte	Définit les situations (activité organisationnelle) dans lesquelles le problème est résolu et/ou le patron est appliqué
	problème	Définit le problème à résoudre
	Force	Définit les raisons de proposition de patron
Réalisation	Solution démarche	Indique la solution du problème en termes de processus à suivre : un diagramme d'activités
	Solution modèle	Décrit la solution en terme de produits attendus après l'application du patron
	Cas d'application	Décrit des exemples d'imitation de la solution modèle.
Relation	Utilise	Définit la liste des patrons utilisés par le patron concerné. Un patron P1 utilise un patron P2, si une partie des problèmes posés par P1 peuvent être résolus en partie ou complément par P2.
	Requiert	Définit la liste des patrons requis par défaut par le patron considéré. Un patron P1 requiert un patron P2, si l'application de P2 est un pré requis à l'application de P1.
	Raffine	Définit la liste des patrons raffinés par le patron considéré. Un patron P1 raffine un patron P2, si le problème posé par P1 est une spécialisation de celui posé par P2.
	Alternative	Définit la liste des patrons qui constituent une alternative au patron considéré. Un patron P1 est une alternative d'un patron P2, si P1 a le même problème que P2 mais propose une solution différente.

FIG. 10 – *Description des parties d'un patron P-SIGMA (Conte et al., 2001).*

Afin de garantir une réutilisation systématique de notre démarche, nous optons pour le formalisme des patrons P-sigma qui est commun à l'ensemble des patrons existants. Il permet d'exprimer une sémantique commune à la majorité des formalismes proposés dans la littérature, de décrire à la fois des fragments de modèles réutilisables et d'expliciter une interface de sélection des patrons afin de faciliter leur réutilisation (Fredj (2007)). Notre objectif, à travers ce formalisme, est de faciliter la sélection, la réutilisation et l'organisation des patrons et permettre, également, de définir de nombreuses informations liées à la solution-démarche traduite par le diagramme d'activités proposé ou de la solution-modèle qui consiste en la définition du patron obtenu lors de l'application de la démarche proposée.

3.3 Patron "démarche d'analyse des besoins décisionnels"

Dans cette section, nous proposons le patron processus lié à notre proposition d'analyse des besoins décisionnels suivant le formalisme P-SIGMA adopté :

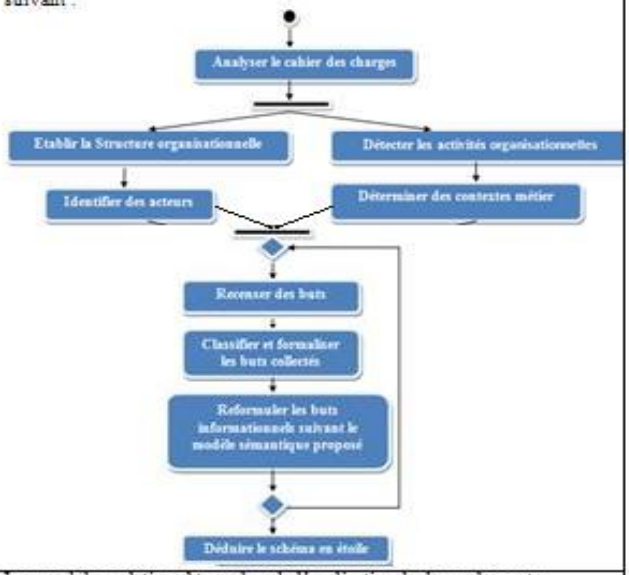
Partie	Rubrique	Champs
Interface	Sigle	ASID
	Nom	Analyser les besoins decisionnels
	Classification	SID ^ Analyse ^ Processus
	Contexte	Ce patron est reutilise lors d'une nouvelle collecte des besoins decisionnels.
	Probleme	Guider l'analyse des besoins d'un SID
	Force	Ce patron detaille les etapes a suivre pour analyser les besoins d'un SID
Realisation	Solution démarche	La solution demarche consiste en la realisation de diagramme d'activites suivant : 
	Solution modèle	Les modeles solution obtenus lors de l'application de demarche sont : • Modèle de « Classification des acteurs de SID » • Modèle de « Modélisation du contexte métier » • Modèle de « Expression des besoins selon des contextes » • Modèle de « Expression des besoins d'un SID » • Modèle de « Classification des besoins de SID » • Schéma en étoile issu de résolution des buts informationnels formulés
Relation	Utilise	Ce patron utilise d'autres patrons en cours de definition
	Requiert	
	Raffine	
	Alternative	

FIG. 11 – *Patron Processus "Analyser les besoins d'un SID".*

Le patron processus "Analyser les besoins d'un SID" capitalise l'ensemble des étapes citées dans notre démarche proposée afin de systématiser la tâche d'analyse et de conception d'un SID. Il met en avant les activités liées au problème étudié à savoir "guider l'analyse des besoins d'un SID". A travers ce patron, nous avons pu systématiser et faciliter la phase d'analyse des besoins métier décisionnels, assurer une réutilisation explicite au sein des projets décisionnels, favoriser et améliorer la communication au cours des projets, et également, nous avons pu gérer la traçabilité de la documentation entre les acteurs de SID.

4 Conclusion

Ce travail concerne notre deuxième contribution de thèse qui s'inscrit dans le cadre de la proposition d'une démarche d'analyse des besoins décisionnels, sous forme d'un patron processus, représentée par un modèle de diagramme d'activités et permettant de guider la découverte des besoins métier et la conception des SID. Dans cette perspective, la solution que nous avons proposée dans cet article est constituée de deux parties. Dans la première partie, nous avons proposé une démarche explicite pour l'analyse et la conception d'un SID. Le guidage offert par notre démarche permet de recenser les besoins métier des acteurs d'un SID sous forme de buts stratégiques, tactiques et informationnels. Ces buts sont ensuite traités et formalisés suivant un modèle sémantique en vue d'établir le modèle en étoile. Nous pensons que le terme métier d'une organisation peut être représenté par un ensemble d'activités suivant lesquelles les décideurs définissent leurs besoins dans des contextes différents. Ainsi, l'originalité de notre travail s'inscrit dans l'intégration du concept contexte pour la collecte des buts métier. Nous visons à travers ce travail, à encapsuler, sous forme de cas d'utilisation, les buts des décideurs dans un contexte prédéfini puis à les fusionner pour obtenir l'ensemble des buts concernant le métier de l'organisation. Cette proposition permet le recensement des besoins métier d'une telle organisation en vue d'une éventuelle réutilisation. Nous avons récapitulé les étapes de notre démarche dans un diagramme d'activités qui servira comme guide documenté pour analyser les besoins métier d'un SID. Dans la seconde partie, nous avons proposé un patron processus intitulé "Analyser les Besoins Décisionnels" basé sur le formalisme P-SIGMA. Ce patron capitalise le diagramme d'activités pour une éventuelle réutilisation.

Références

ACM, (1996). *Softeware Patterns.* Communications of the ACM, Vol 39, N 10.

Annoni, E. (2007). *Eléments méthodologiques pour le développement des systèmes décisionnels dans un contexte de réutilisation*. Thèse de Doctorat, Université de Toulouse 1, Toulouse, France.

Bushmann, F., Meunier, R., Rohnert, H., Sommerlad, P., Stal, M. (1996). *A System of Patterns: Pattern-Oriented Software Architecture.* John Wiley, New York, USA. 0471958697.

Coad,P. (1992). *Object-Oriented Patterns.* Communications of the ACM, Vol 35, Nř9.

Conte,A., Fredj, M., Giraudin, J.-P., Rieu, D. (2001). *P-sigma : un formalisme pour une représentation unifiée de patrons.* Inforsid'01, Mai 2001, Martigny, Suisse.

El Golli, I.G. (2008). *Ingénierie des Exigences pour les Systèmes d'Information Décisionnels : Concepts, Modèles et Processus (la méthode CADWE).* Thèse de Doctorat, Université Paris-Panthéon-Sorbonne, France.

Fredj, M. (2007). *Composants et modèles pour l'ingénierie des systèmes d'information.* Thèse de doctorat, Université Mohammed V-Agdal, Faculté des sciences, Rabat, Maroc.

Fowler,M. (1997). *Analysis Patterns- Reusable Object Models.* Addision-Wesley

Gamma, E., Helm, R., Johnson, R. E., Vlissides, J. M. (1995). *Design Patterns, Elements of reusable Object-Oriented Software.* Addison-Wesley Publishing Company. 0201633612.

Ghozzi, F., Ravat, F., Teste, O., Zurfluh, G. (2005). *Méthode de conception d'une base multidimensionnelle contrainte.* In Revue des Nouvelles Technologies de l'Information - Entrepôts de Données et l'Analyse en ligne (EDA'05), volume RNTI-B-1, pages 51-70. Cépadues éditions.

Gzara, L. (2000), *Les patterns pour l'ingénierie des systèmes d'informations Produits.* Doctorat de l'INPG, Spécialité Génie Industriel, Décembre 2000, Grenoble, France.

Mazon,J.-N.,Trujillo, J., Serrano,M.,and Piattini, M. (2005). *Designing data warehouses : from business requirement analysis to multidimensional modeling.* In 13th IEEE International Requirements Engineering Conference Workshop on Requirements Engineering for Business Needs and IT Alignment.

Moody, D., M. Kortnik (2000). *From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design.* DMDW'00, Suede.

Phipps, C. and Davis, K. C. (2002). *Automating data warehouse conceptual schema design and evaluation.* In Lakshmanan [2002], pages 23-32.

Ravat, F. (2007), *Modèles et outils pour la conception et la manipulation de systèmes d'aide à la décision.* Thèse de Doctorat, Université des Sciences Sociales, Toulouse, France.

Rifaieh, R. D. (2004). *Utilisation des ontologies contextuelles pour le partage sémantique entre les systèmes d'information dans l'entreprise.* Thèse de doctorat, Institut National des Sciences Appliquées, Villeurbanne, France.

Sabri, A. Kjiri, L. (2011a). *Vers une approche à base de composants pour la conception des systèmes d'information décisionnels.* Workshop on the Evolution of the Reusability concept from Component and Service to Cloud Service (RCS2), 07/08 Juin 2011, Polytech, Campus universitaire de Lille1, Lille, France

Sabri, A. Kjiri, L. (2011b). *Vers une ontologie pour la formulation des besoins d'un Système d'Information Décisionnel.* International Workshop on Information Technologies and Communication (WOTIC'11), 13-15 Octobre 2011, ENSEM, Casablanca, Maroc.

Soussi, A., Feki, J., Gargouri, F. (2005). *Approche semi-automatisée de conception de schémas multidimensionnels valides.* In Revue des Nouvelles Technologies de l'Information - Entrepoôts de Données et l'Analyse en ligne (EDA'05), volume RNTI-B-1, pages 71-89. Cépadues éditions.

Winter,R. , Strauch, B.(2003). *A method for demand-driven information requirements analysis in data warehousing projects.* In HICSS, page 231.

Summary

The area of analysis and design of Decision Support Information System (CIS) is a highly applicant new techniques and methods for success of the process of decision making and min-

imizing the time of conception. Our work is part of an analytical approach to guide analysts through the process of designing a SID. We seek, through this process, to guide the discovery of an organization's business requirements expressed as goals by introducing the notion of context, to promote good design by ensuring a SID possible reuse. Two key elements were produced at the end of this article: an approach for analyzing business requirements of a SID and the boss following the formal process established P-SIGMA, which is related to our proposed analysis of decision-making needs.

Recommandation interactive de requêtes décisionnelles

Rym Khemiri*, Fadila Bentayeb* Omar Boussaid*

*5 av. Pierre Mendès-France - 69676 Bron Cedex
{Rym.Khemiri, Fadila.Bentayeb, Omar.Boussaid}@univ-lyon2.fr,
http://eric.univ-lyon2.fr

Résumé. Cet article présente une approche de recommandation basée sur les historiques de requêtes utilisateur. Plus précisément, nous proposons un procédé interactif d'aide à l'écriture de requêtes décisionnelles dans lequel il est proposé à l'utilisateur un choix d'attributs (axes d'analyse) et de mesures (indicateurs) fréquemment utilisés dans ses précédentes sessions. Notre intuition est que la pertinence d'une requête décisionnelle est fortement corrélée avec la fréquence d'utilisation par l'utilisateur des attributs associés à l'ensemble de ses requêtes précédentes. Notre stratégie d'aide à l'écriture de requêtes décisionnelles s'appuie sur la méthode des itemsets fréquents pour extraire les attributs fréquemment utilisés à partir d'une charge de requêtes utilisateur. Notre approche permet alors à l'utilisateur de soumettre à l'entrepôt de données des requêtes pertinentes. Pour valider notre approche, nous avons développé un prototype en Java, baptisé FIMIOQR (Frequent Itemset Mining for Interactive OLAP Query recommendation) [1].

1 Introduction

Les entrepôts de données sont des bases de données dédiées à l'analyse pour l'aide à la prise de décision. Les modèles en étoile structurent les données entreposées de manière multidimensionnelle et permettent dans un premier temps de produire des cubes de données adaptés à l'analyse. Dans un second temps, c'est à l'utilisateur de naviguer, explorer et analyser les données d'un cube afin d'en extraire des informations pertinentes.

L'exploration de données est un processus de recherche d'information destiné à détecter des corrélations cachées entre les données ou des informations nouvelles. Or, les utilisateurs doivent faire face à un volume de plus en plus important d'informations en raison de l'accroissement des capacités de stockage et de calcul. Par conséquent, il est de plus en plus difficile de savoir exactement quelles informations rechercher et où les chercher. Dans ce contexte, le recours à des techniques informatiques plus sophistiquées que la seule analyse OLAP s'avèrent nécessaires pour aider l'utilisateur dans sa tâche d'exploration des données afin de lui faciliter la recherche d'informations pertinentes. L'une de ces techniques est la recommandation d'informations et, plus particulièrement la recommandation de requêtes décisionnelles.

Ainsi, face à la nécessité d'offrir davantage de flexibilité pour répondre au mieux aux besoins des utilisateurs, plusieurs propositions sont apparues dans la littérature faisant appel aux

1. http://eric.univ-lyon2.fr/~bentayeb/logiciels.html

outils de recommandation. Les termes de requête recommandée désignent une requête existante (ou calculée) issue d'un ensemble de requêtes posées sur l'entrepôt (fichier log de requêtes par exemple) ou des préférences utilisateur. Nous pouvons citer les travaux de Giacometti et al. (2009a) et de Jerbi et al. (2009b). Dans Giacometti et al. (2009a); Negre (2011), le principe est de calculer une similarité entre la séquence courante de requêtes de l'utilisateur et les séquences de requêtes précédentes. Par ailleurs, une méthode de personnalisation qui prend en compte les préférences de l'utilisateur en combinant le contexte et le profil pour recommander des requêtes a été proposée dans Jerbi et al. (2009b).

Nous constatons que toutes les méthodes proposées dans la littérature s'appuient uniquement sur les requêtes utilisateurs déjà posées et exclut de ce fait l'utilisateur lui-même. Aussi, nous pensons qu'un système de recommandation doit pouvoir impliquer davantage l'utilisateur en l'aidant par exemple à écrire ses propres requêtes décisionnelles tout en se basant sur ses historiques de requêtes.

Nous proposons alors une approche de recommandation de requêtes différente des approches existantes dans la littérature. En effet, notre approche a pour objectif d'aider l'utilisateur à formuler de nouvelles requêtes décisionnelles de manière interactive. Il s'agit alors de guider l'utilisateur dans sa démarche de construction de sa requête décisionnelle. Les requêtes décisionnelles auxquelles nous nous intéressons sont des requêtes OLAP de la forme "SELECT ... FROM ... WHERE ... GROUP BY CUBE (ROLLU-UP)". Nous pensons que la pertinence d'une recommandation est fortement corrélée avec la fréquence de son utilisation dans l'ensemble des requêtes soumises par l'utilisateur. C'est pourquoi, nous avons eu recours à la fouille de données et plus précisément à l'algorithme Apriori qui permet d'extraire les attributs et les mesures les plus fréquemment utilisées à partir d'une charge de requêtes (fichiers logs).

Notre approche de recommandation aide l'utilisateur à construire sa requête, au fur et à mesure de son écriture, en lui suggérant des attributs fréquemment utilisés tout en respectant leur apparition aux différentes *Clauses* (SELECT, WHERE, GROUP BY, etc.) dans les requêtes précédentes.

Notre approche de recommandation de requêtes se déroule en trois phases principales. (1) Tout d'abord, à partir d'un ensemble de requêtes issues du fichier log, un prétraitement est nécessaire pour construire la matrice binaire "requêtes-attributs/mesures" nécessaire pour l'application de l'algorithme Apriori. (2) Nous procédons ensuite à l'application de l'algorithme Apriori sur la matrice "requêtes-attributs/mesures" pour obtenir l'ensemble des itemsets fréquents. (3) Nous exploitons les résultats obtenus (itemsets fréquents) pour recommander à l'utilisateur des attributs/mesures pour l'aider à anticiper dans l'écriture d'une nouvelle requête décisionnelle. Par conséquent, l'utilisateur peut construire progressivement sa requête avec chaque suggestion en choisissant ses attributs/mesures parmi les suggestions. La requête ainsi obtenue pourra être soumise par l'utilisateur à l'entrepôt de données.

Pour valider notre approche de recommandation interactive de requêtes décisionnelles, nous avons développé un prototype appelé FIMIOQR (Frequent Itemsets Mining for Interactive OLAP Query Recommendation) en Java dans l'environnement intégré Netbeans.

L'article est organisé de la manière suivante : la section 2 présente tout d'abord un aperçu de l'état de l'art des travaux de recherche relatifs à la recommandation de requêtes dans les domaines des bases de données et des entrepôts de données. Nous y mettons en évidence la position de notre contribution dans le domaine des entrepôts de données . La section 3 présente notre approche en détaillant les objectifs ainsi que la démarche de la recommandation interac-

tive des requêtes. Dans la suite de cet article, l'implémentation du prototype est présentée dans la section 4. Enfin, la section 5 résume notre proposition et en présente les perspectives.

2 Etat de l'art

La problématique de la recommandation de requêtes ne s'est pas posée que récemment dans le domaine de bases de données. Il s'agit des travaux de Stefanidis et al. (2009) proposant une approche de génération de recommandations de requêtes à l'utilisateur avec les résultats de chaque requête. Ces résultats additionnels recommandés peuvent intéresser l'utilisateur. Ils sont appelés "You May Also Like" ou résultats YMAL. L'utilité d'une requête pour un utilisateur est égale au nombre de fois que cette requête a été posée par cet utilisateur. YMAL peut exploiter l'historique de requête et aussi des sources externes (pages web pertinentes, rapports et résultats publiés, ontologies, ...).

Citons aussi les travaux de Yang et al. (2009) où les auteurs proposent une approche collaborative permettant d'aider l'utilisateur à compléter les requêtes complexes. Cette approche analyse le log de requêtes et extrait les jointures faites dans les requêtes précédentes. Ensuite ces jointures sont utilisées pour générer des recommandations à l'utilisateur courant.

Afin d'assister l'utilisateur lors de son écriture de requête, Khoussainova et al. (2010) proposent une approche collaborative de génération de recommandations. L'assistance est fournie à la demande par les suggestions pour le fragment courant de la requête en se basant sur l'analyse du log de requêtes. Cette approche dénommée "SnipSuggest" permet à l'utilisateur de composer la requête, de choisir une clause et demander des recommandations pour cette clause à n'importe quel instant. L'objectif de *SnipSuggest* est de recommander k caractéristiques qui peuvent avoir plus de chance à apparaître dans telle clause dans la requête voulue de l'utilisateur.

Dans Chatzopoulou et al. (2011), les auteurs présentent leur système *QueRIE* de recommandation de requêtes personnalisée pour une exploration interactive de la base de données après avoir présenté son modèle conceptuel dans Chatzopoulou et al. (2009). Pour générer des recommandations personnalisées à l'utilisateur courant, ce système se base sur les requêtes des autres utilisateurs et celles de l'utilisateur courant. Ces recommandations de requêtes peuvent être suggérées à tous les utilisateurs ayant des intérêts similaires.

Nous constatons que la recommandation de requêtes est peu exploitée dans les bases de données, alors même que cela allège la tâche de l'utilisateur. De la même façon, la recommandation de requêtes n'est exploitée que très récemment par peu de travaux dans les entrepôts de données et de l'analyse en ligne. Nous nous focalisons dans la suite de cet état de l'art sur les travaux de recommandation émergents.

Afin d'enrichir les possibilités d'analyse d'un entrepôt de données, la proposition Giacometti et al. (2008) permet la recommandation de requêtes pour anticiper sur une séquence de requêtes d'un utilisateur grâce à l'analyse des historiques de navigations réalisées par les autres utilisateurs. Dans ces travaux, on suppose que le log de requêtes est utilisé pour chercher les similarités entre la session courante et les sessions précédentes afin d'en extraire une seule requête comme étant une recommandation.

Pour aller au-delà de l'approche proposée dans Giacometti et al. (2009a), Giacometti et al. (2009b) traitent les fichiers log mais cette fois-ci les sessions sont associées à un intérêt, et les recommandations sont les requêtes des sessions précédentes ayant le même intérêt que

la session courante. L'importance de ces travaux réside non pas dans la recommandation des requêtes à partir des sessions qui précèdent la session courante mais dans la recommandation à partir de toutes les sessions où l'utilisateur trouve les mêmes données inattendues comme dans la session courante. Dans cette approche les préférences utilisateurs ne sont pas prises en considération.

Citons aussi la proposition de (Jerbi et al., 2009a) qui suppose qu'un profil utilisateur est fourni avec la session courante. Ils se sont inspirés des techniques de filtrage d'information en fonction de profil utilisateur pour affiner des requêtes en y ajoutant des prédicats (Koutrika et Ioannidis, 2005). L'objectif de ces travaux est de pouvoir fournir à l'utilisateur un résultat focalisé sur son centre d'intérêt, tout en exploitant des ordres (représentation qualitative des préférences). Ces derniers ne sont pas exprimés de façon absolue, mais par rapport à un contexte d'analyse donné. Ceci permet de prendre en compte le fait que les préférences peuvent varier d'un contexte d'analyse à l'autre.

Inspiré de (Kießling et Köstler, 2002), (Golfarelli et Rizzi, 2009) ; Golfarelli et al. (2011) proposent une approche où les préférences sont utilisées pour annoter la requête. Ils définissent une algèbre qui permet de formuler les préférences des attributs, des mesures et des hiérarchies. Une importante caractéristique de cette algèbre proposée est la possibilité d'exprimer les préférences des attributs de hiérarchie des ensembles de group-by ce qui peut mener par conséquent à exprimer les préférences des faits. La technique utilisée est de personnaliser une requête en traitant une sous requête de la requête courante. Dans ce cas, la requête recommandée est la sous requête qui retourne un résultat préféré non vide.

Les avantages de ces méthodes dans cette approche sont multiples (i) les utilisateurs peuvent exprimer leurs préférences pour les requêtes OLAP ; dans ce cas, le problème de perdre le temps de traitement des opérations OLAP pour trouver l'information nécessaire peut être considérablement amélioré et (ii) les recommandations sont traitées en tenant en compte le profil et les sessions utilisateur ; donc, les utilisateurs différents peuvent obtenir des recommandations différentes.

Récemment, dans Aligon et al. (2011) (basé sur les travaux de (Golfarelli et Rizzi, 2009) ; Golfarelli et al. (2011)) proposent une approche qui permet d'extraire des préférences à partir des fichiers log des requêtes MDX en utilisant des techniques de fouille de données. Ainsi, les préférences extraites (sous forme de règles d'association) permettent d'annoter les requêtes MDX.

Notre approche est différente des travaux cités précédemment dans le sens qu'elle vise à assister l'utilisateur qui interroge un entrepôt de données (requêtes décisionnelles) en lui recommandant des attributs/mesures au lieu d'une requête entière.

Nos travaux se rapprochent des travaux qui proposent des approches d'affinement de la requête utilisateur dans le domaine des bases de données Khoussainova et al. (2010). Bien que ces travaux ont quelques similarités avec notre travail actuel, les challenges et les techniques sont très différents de ce que que nous proposons. En effet, ce travail diffère de notre approche dans le sens que *SnipSuggest* recommande les additions possibles aux différentes clauses dans la requête courante de l'utilisateur, ne complète pas la requête et la recommandation est fournie à la demande. Dans notre travail, nous proposons un système pour générer des recommandations de requêtes décisionnelles à un utilisateur. Dans notre cas, chaque requête est traitée indépendamment de n'importe qu'elle requête précédente même si elles appartiennent à la même session utilisateur. Enfin, FIMIOQR ne force pas l'utilisateur à déclarer explicitement

ses préférences pour générer des recommandations puisque nous utilisons le log des requêtes au lieu du profil utilisateur.

Pour affiner la lecture des différents travaux portant sur les bases de données et les entrepôts de données, nous proposons une grille d'analyse basée sur des critères de comparaison que nous avons jugés pertinents pour étudier les différentes facettes de la recommandation. Ces critères sont inspirés en partie de l'étude de Bentayeb et al. (2009) et celle de Marcel et Negre (2011) .

Type de recommandation : ce critère représente la portée prise en compte pour la recommandation (cognitif ou collaboratif). Dans le cadre cognitif (content-based), le système de recommandation est basé sur le contenu : il s'agit de prendre en compte l'utilisateur de façon individuelle, en fonction de ses besoins. Dans le cadre social ou collaboratif, il s'agit de baser la recommandation en prenant en compte le contexte des autres utilisateurs qui auraient des préoccupations similaires (réseau social).

Objet d'exploitation : ce critère représente la source de recommandation qui peut être un profil utilisateur, l'historique des requêtes (fichier log) ou des sources externes.

Temps de recommandation : le temps de la recommandation peut être avant, pendant ou après la soumission de la requête.

Travaux	**Type de recommandation**		**Temps de recommandation**			**Objet d'exploitation**		
	Cognitif	Collaboratif	Avant	Pendant	Après	Profil	Fichier log	Source externe
Stefanidis et al. (2009)	×	×			×		×	×
Yang et al. (2009)		×		×			×	
Chatzopoulou et al. (2009), Chatzopoulou et al. (2011)		×		×			×	
Khoussainova et al. (2010)		×		×			×	
(Jerbi et al., 2009a)	×				×	×		
(Giacometti et al., 2009a), Giacometti et al. (2009b)		×	×			×		
(Golfarelli et Rizzi, 2009), Golfarelli et al. (2011)		×		×		×		
Aligon et al. (2011)		×		×		×	×	
Notre approche	×			×			×	

TAB. 1 – *Positionnement des travaux existants par rapport aux principes de recommandation définis précédemment*

3 Recommandation interactive de requêtes décisionnelles

Dans cette section, nous allons d'abord esquisser l'idée générale de notre approche. Nous allons ensuite décrire notre démarche en présentant brièvement la méthode de recherche des itemsets fréquents.

3.1 Exemple illustratif

Nous utilisons tout au long de cet article l'exemple de l'entrepôt de données FoodMart [2] d'une organisation de ventes des produits alimentaires et d'autres produits dans des magasins aux états Unis, Canada et Mexique. Nous pouvons utiliser l'entrepôt de données FoodMart pour analyser les opérations de ventes par plusieurs axes d'analyse (promotion, temps, magasin produit et client).

Nous nous intéressons au fait sales_fact_1998 qui est relié à 4 tables de dimensions *Customer*, *Product*, *Time_by_day* et *Promotion* comme illustré dans la figure 3. Il s'agit des ventes par rapport aux clients, promotion, magasin produit et temps. L'information sur les produits est collectée dans la dimension *Product*, tandis que les informations reliées au temps et dates sont sauvegardées dans la dimension time_by_day. Les informations concernant les clients (noms et adresses) sont sauvegardés dans la dimension *Customer*. Les attributs *Education*, *Gender*, *Marital_Status*, *Occupation* et *Yearly_Income* fournissent des informations additionnelles sur les clients. De plus, l'information sur les magasins d'alimentation individuels est collectée dans la dimension *Store*. Elle inclut les attributs *store_location*, *name*, *manager*, *size*, et *store_type* ainsi que les informations concernant les promotions sont sauvegardées dans la dimension *Promotion*.

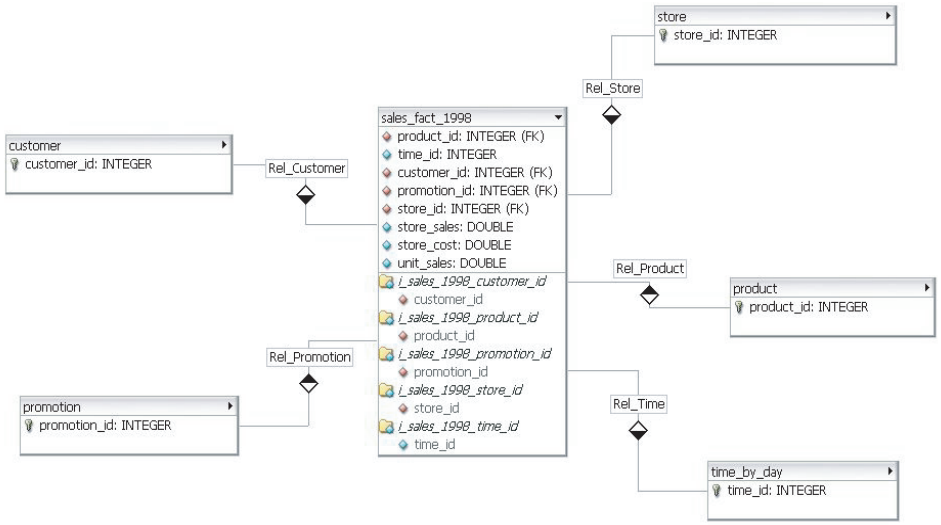


FIG. 1 – *Extrait du schéma de l'entrepôt de données FoodMart*

2. http ://www.e-tservice.com/downloads.html

3.2 Principe général de notre approche

Pour améliorer le processus d'analyse dans les entrepôts de données, nous définissons une approche d'aide à la construction de requêtes décisionnelles pertinentes. L'objectif de notre approche est de proposer à l'utilisateur un choix d'attributs/mesures susceptibles de l'intéresser pour construire sa requête. Pour cela, nous proposons d'explorer l'historique des requêtes précédentes en utilisant l'algorithme Apriori afin d'extraire les attributs/mesures les plus fréquemment utilisés. Grâce à cette exploration, l'utilisateur peut découvrir les attributs/mesures fréquemment utilisés qu'il peut ajouter à sa propre requête.

Dans les grandes lignes, notre objectif est de recommander à l'utilisateur lors de sa construction de requêtes les attributs/mesures les plus utilisés grâce à une extraction des itemsets fréquents. Dans notre cas d'étude, les objets sont des requêtes et les items sont les attributs/mesures extraits de ces requêtes.

L'approche proposée est considérée comme une aide à la construction de requête basée sur des requêtes précédentes. Elle est intégrée au début du processus habituel d'interrogation pour l'expression de la requête. Cette intégration est non intrusive et l'utilisateur peut décider ou pas d'utiliser l'aide proposée. Le scénario général que nous proposons pour notre approche est illustré par la figure 2.

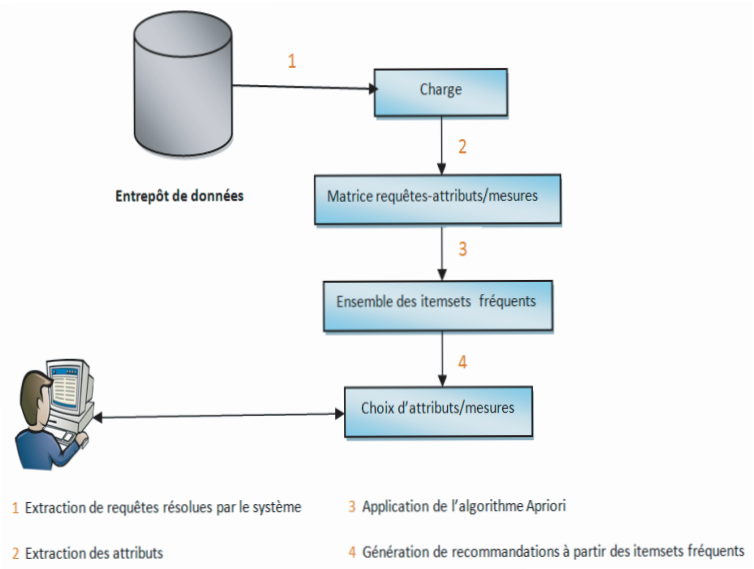


FIG. 2 – *Architecture de notre stratégie de recommandation*

3.3 Démarche

En s'inspirant de la recherche d'information (RI), généralement, les systèmes de personnalisation basés sur la fouille de l'usage du web (usage-based Web personalization systems) implique 3 phases : La préparation et la transformation des données, la découverte de motifs

(pattern discovery), et la recommandation, Mobasher et al. (2000). Dans cette section, on se concentre sur une instance spécifique de ce cadre générique.

3.3.1 Prétraitement des données

Le prétraitement de données est une phase exécutée hors ligne. Elle consiste à la préparation et la transformation des données. La préparation de données réside dans l'extraction et l'analyse de la charge de requêtes. La transformation consiste à la construction de la matrice à partir des données extraites.

Extraction et analyse de la charge La première étape de notre stratégie de recommandation de requêtes consiste à extraire du journal des transactions les requêtes adressées au SGBD (Système de gestion de base de données). Ensuite, à partir de ces requêtes, nous pouvons extraire tous les attributs et les mesures qui apparaissent dans toutes les clauses de la requête.

Dans notre approche, lors du chargement, le système prend comme entrée le log de requêtes nécessaire à l'analyse. En effet, il crée un fichier texte en utilisant une fonction permettant de mettre toute requête décisionnelle sur une ligne. La figure 3 représente un extrait de charge composé de trois requêtes. Une fois, nous avons toutes les requêtes, nous pouvons extraire les attributs/mesures (les items) de chaque clause. Nous aurons l'attribut ou la mesure préfixé(e) du nom de la clause dans laquelle il/elle apparaisse.

```
(1) select sales.time_id, sum(quantity_sold), sum (amount_sold)
    from sales, time
    where sales.time_id = time.time_id
    and time.fiscal_year = '2010'
    group by sales.time_id ;

(2) select sales.prod_id, sales.promo_id, sales.time_id, avg(amount_sold)
    from sales, product, promotion, time
    where sales.prod_id = product.prod_id
    and sales.promo_id = promotion.promo_id
    and sales.time_id = time.time_id
    and promotion.promo_category = 'cheese'
    group by Rollup (sales.prod_id, sales.promo_id, sales.time_id) ;

(3) select sales.cust_id, sales.promo_id, sales.prod_id, avg(amount_sold)
    from sales, customer, product, promotion
    where sales.cust_id = customer.cust_id
    and sales.prod_id = product.prod_id
    and sales.promo_id = promotion.promo_id
    and customer.cust_marital_status ='single'
    and product.prod_category ='Women'
    group by Cube (sales.cust_id , sales.promo_id, sales.prod_id);
```

FIG. 3 – *Extrait de charge*

Construction du contexte d'extraction À partir des attributs extraits dans l'étape précédente, nous construisons une matrice "requêtes-attributs/mesures" qui a pour lignes les requêtes de la charge et pour colonnes les attributs/mesures précédés par le nom de la clause dans laquelle ils apparaissent. Dans notre cas, chaque ligne est une requête (objet), et chaque colonne

est un attribut ou une mesure (propriété binaire d'un objet). Concrètement, la valeur 1 dans une case indique que l'utilisateur a utilisé l'attribut/la mesure dans sa requête. Il y aurait eu la valeur 0 dans le cas contraire.

Dans notre exemple, la matrice "requêtes-attributs/mesures" obtenue après l'analyse de la charge est composée de 23 colonnes et de 3 lignes (Figure 4). Pour des raisons de clarté et de lisibilité, nous avons désigné les 23 attributs respectivement par les lettres de A à W figure 4).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	1	1	1	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0
2	1	0	0	1	1	1	0	1	1	1	1	0	1	0	1	1	1	0	0	0	1	1	0
3	0	0	0	1	1	1	1	1	0	1	1	1	0	0	1	1	0	1	1	1	0	0	1

FIG. 4 – *Matrice Requêtes-attributs/mesures*

3.3.2 Recherche des itemsets fréquents

Cette phase est réalisée hors ligne où la matrice "requêtes-attributs/mesures", exploitée par l'algorithme AprioriAgrawal et al. (1996), donne lieu à un ensemble d'itemsets fréquents à partir de la matrice binaire de départ. Dans notre exemple, l'algorithme Apriori génère à chaque itération un ensemble de conjonctions de propriétés binaires susceptibles d'être fréquentes, appelées candidats. Cette génération de candidats est effectuée en tenant compte des conjonctions effectivement fréquentes de l'itération précédente. Intuitivement, un itemset est un ensemble d'items (attributs) communs à un ensemble de transactions (requêtes). Éventuellement, un itemset est dit fréquent dans son contexte lorsque son support est supérieur ou égal à un seuil appelé minsup σ (support minimal). Le support minimal est prédéterminé dans notre exemple à 50%. Ainsi, tous les itemsets ayant un support inférieur à 50% sont retirés et ceux dont le support est supérieur ou égale à 50% sont conservés. Lorsqu'il n'y a plus de candidats générés, la boucle prend fin et les conjonctions retenues sont fournies comme résultat (itemsets fréquents).

3.3.3 Génération de requêtes recommandées

Le processus de recommandation représente le composant en ligne de notre système. Dans cette étape, nous considérons la session active de l'utilisateur en conjonction avec les itemsets fréquents découverts afin de fournir des suggestions. L'ensemble de recommandations représente essentiellement une vue à court terme des attributs/mesures potentiellement utiles à l'utilisateur. Pour cela, nous utilisons directement les itemsets fréquents obtenus dans l'étape de fouille de données du log de requêtes.

Dès que l'utilisateur commence à soumettre sa requête par la saisie du premier attribut A dans la clause Select, le système recherche l'attribut A dans les différents itemsets fréquents $I_1, I_2, \ldots$ obtenus. Il propose alors à l'utilisateur la liste des attributs/mesures appartenant aux itemsets contenant A. Chaque attribut/mesure proposé est associé à la clause dans laquelle il apparaît. L'utilisateur choisit alors un attribut A_1 parmi les attributs les plus fréquemment associés à A (à partir des itemsets $I_1, I_2, \ldots$) ou introduit de nouveaux attributs/mesures. Ainsi, avec chaque saisie, une nouvelle représentation est calculée et affichée ensuite à l'utilisateur.

Enfin, lorsque l'utilisateur termine l'exploration des recommandations, l'entrepôt de données est interrogé avec les attributs/mesures choisis par l'utilisateur..

4 Implémentation

Nous avons développé un prototype appelé FIMIOQR (Frequent Itemsets Mining for Interactive OLAP Query Recommandation) en Java dans l'environnement intégré Netbeans [3].

Notre système se résume en 4 étapes : (1) l'extraction des requêtes déjà résolues par le système, (2) la visualisation des attributs/mesures présents dans le log des requêtes ainsi que les clauses dans lesquelles ils apparaissent, (3) l'affichage des itemsets fréquents issus de l'algorithme Apriori et (4) la génération des recommandations. En effet, les trois premières étapes sont juste des prétraitements et des visualisations pour l'utilisateur. C'est dans la quatrième étape que l'utilisateur va pouvoir interagir avec l'application comme illustré dans la figure 5.

Nous avons testé notre prototype sur une charge de 100 requêtes relatives à l'entrepôt de données FoodMart utilisé dans notre exemple. Nous avons enrichi, par ailleurs, la charge de requêtes de départ par les requêtes obtenues par FIMIOQR et qui sont recommandées à l'utilisateur.

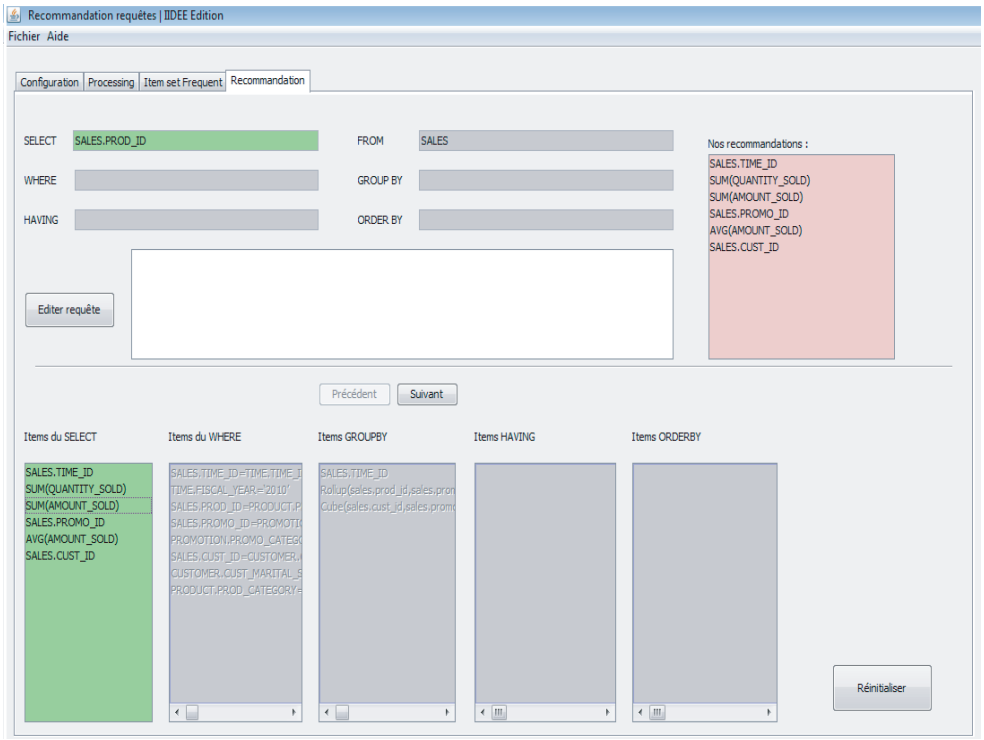


FIG. 5 – *FIMIOQR : Outil de génération de recommandations*

3. http ://eric.univ-lyon2.fr/˜bentayeb/logiciels.html

5 Conclusion et perspectives

Dans cet article, nous avons posé les bases d'un système de recommandation interactive de requêtes décisionnelles dans le cadre des entrepôts de données. Nous avons explicité les principes de notre proposition, décrit le processus centré utilisateur pour permettre l'aide à l'écriture de requêtes en se basant sur l'extraction des itemsets fréquents. Cet article se distingue des travaux existants sur la recommandation de requêtes souvent basée sur l'enrichissement de requêtes utilisateur puisque notre approche ne se base pas sur une requête existante mais aide l'utilisateur à formuler sa requête décisionnelle tout en tenant compte de ses requêtes précédentes. L'originalité de notre méthode réside dans son aspect interactif et l'utilisation de la fouille de données pour extraire les attributs/mesures fréquents utilisés par l'utilisateur dans les requêtes précédentes. Pour cela, nous avons développé un système interactif, FIMIOQR (Frequent Itemsets Mining for Interactive OLAP Query Recommendation), de recommandation basé sur les logs de requêtes de l'utilisateur sur lequel nous avons appliqué la méthode Apriori. Notre système tient alors compte (1) des requêtes précédentes soumises à l'entrepôt de données, (2) utilise les itemsets fréquents des attributs/mesures et (3) permet de proposer des attributs/mesures en adéquation avec la requête en cours d'écriture de l'utilisateur. L'approche de recommandation que nous avons proposée se base sur un processus en ligne, c'est pourquoi son efficacité et son extensibilité sont d'une importance capitale.

Une perspective directe de ce travail est son intégration dans un système de gestion de bases de données afin de tester l'efficacité de notre méthode sur des données réelles. Par la suite, nous envisageons de faire une veille sur les recommandations les plus acceptées par l'utilisateur afin de pouvoir les classer et de les recommander en priorité à l'utilisateur dans ses sessions futures.

Références

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pp. 307–328. AAAI/MIT Press.

Aligon, J., M. Golfarelli, P. Marcel, S. Rizzi, et E. Turricchia (2011). Mining preferences from olap query logs for proactive personalization. In *ADBIS*, pp. 84–97.

Bentayeb, F., O. Boussaïd, C. Favre, F. Ravat, et O. Teste (2009). Personnalisation dans les entrepôts de données : bilan et perspectives. In *5èmes journées francophones sur les Entrepôts de Données et lÁnalyse en ligne (EDA 2009), Montpellier*, Volume B-5 of *RNTI*, Toulouse, pp. 7–22. Cépaduès.

Chatzopoulou, G., M. Eirinaki, S. Koshy, S. Mittal, N. Polyzotis, et J. S. V. Varman (2011). The querie system for personalized query recommendations. *IEEE Data Eng. Bull. 34*(2), 55–60.

Chatzopoulou, G., M. Eirinaki, et N. Polyzotis (2009). Query recommendations for interactive database exploration. In *SSDBM*, pp. 3–18.

Giacometti, A., P. Marcel, et E. Negre (2008). A framework for recommending olap queries. In *DOLAP*, pp. 73–80.

Giacometti, A., P. Marcel, et E. Negre (2009a). Recommending multidimensional queries. In *DaWaK*, pp. 453–466.

Giacometti, A., P. Marcel, E. Negre, et A. Soulet (2009b). Query recommendations for olap discovery driven analysis. In *DOLAP*, pp. 81–88.

Golfarelli, M. et S. Rizzi (2009). Expressing olap preferences. In *SSDBM*, pp. 83–91.

Golfarelli, M., S. Rizzi, et P. Biondi (2011). myolap : An approach to express and evaluate olap preferences. *IEEE Trans. Knowl. Data Eng. 23*(7), 1050–1064.

Jerbi, H., F. Ravat, O. Teste, et G. Zurfluh (2009a). Applying recommendation technology in olap systems. In *ICEIS*, pp. 220–233.

Jerbi, H., F. Ravat, O. Teste, et G. Zurfluh (2009b). Preference-based recommendations for olap analysis. In *DaWaK*, pp. 467–478.

Khoussainova, N., Y. Kwon, M. Balazinska, et D. Suciu (2010). Snipsuggest : Context-aware autocompletion for sql. *PVLDB 4*(1), 22–33.

Kießling, W. et G. Köstler (2002). Preference sql - design, implementation, experiences. In *VLDB*, pp. 990–1001.

Koutrika, G. et Y. E. Ioannidis (2005). Personalized queries under a generalized preference model. In *ICDE*, pp. 841–852.
Anglais

Marcel, P. et E. Negre (2011). A survey of query recommendation techniques for datawarehouse exploration. In *Actes du congrès EDA (Entrepôts de Données et l'Analyse en ligne)*, Clermont-Ferrand, France.

Mobasher, B., R. Cooley, et J. Srivastava (2000). Automatic personalization based on web usage mining. *Commun. ACM 43*(8), 142–151.

Negre, E. (2011). Quand la recommandation rencontre la personnalisation. ou comment générer des recommandations (requêtes mdx) en adéquation avec les préférences de l'utilisateur. *Technique et Science Informatiques 30*(8), 933–952.

Stefanidis, K., M. Drosou, et E. Pitoura (2009). "you may also like" results in relational databases. In *PersDB*.

Yang, X., C. M. Procopiuc, et D. Srivastava (2009). Recommending join queries via query log analysis. In *ICDE*, pp. 964–975.

Summary

This paper introduces a new approach to help users formulating their queries. It proposes to benefit from past queries to extract frequent itemsets. The user can incrementally construct his query by recommending him frequently used attributes. An implementation illustrates the approach through a scenario of use.