Utilisation d'outils de Visual Data Mining pour l'exploration d'un ensemble de règles d'association

Gwenael Bothorel
DSNA DTI R&D APO/IRIT
7, Avenue Edouard Belin
31055, Toulouse, France
gwenael.bothorel@aviation-civile.gouv.fr

Mathieu Serrurier
ADRIA/IRIT
118, Route de Narbonne
31400, Toulouse, France
serrurier@irit.fr

Christophe Hurter
DSNA DTI R&D ICS/IRIT
7, Avenue Edouard Belin
31055, Toulouse, France
christophe.hurter@aviation-civile.gouv.fr

RESUME
La fouille de données a pour objectif d'extraire un maximum d'informations pertinentes à partir d'une grande masse de données. Elle est réalisée de manière automatique, ou en explorant les données à l'aide d'outils interactifs de visualisation. Les algorithmes de fouille de données automatique permettent une extraction exhaustive de tous les motifs satisfaisant un ensemble de métriques. La limite de ces algorithmes est que la quantité d'information extraite peut être plus importante que le volume de données initial. Dans cet article, nous nous focalisons sur l'extraction de règles d'association à l'aide de l'algorithme Apriori. Après avoir décrit un modèle de caractérisation d'un ensemble de règles d'association, nous proposons d'utiliser un outil interactif de visualisation pour explorer les résultats de cet algorithme. L'intérêt est double. D'une part, cela permet de les visualiser selon différents points de vue (métriques, constituants des règles…). D'autre part, cela permet d'explorer et donc de sélectionner facilement dans la masse de règles celles qui sont les plus pertinentes.

Mots clés
Fouille de données, règles d'association, sémiologie graphique, données aéronautiques.

ABSTRACT
Data Mining aims at extracting maximum of knowledge from huge databases. It is realized by an automatic process or by data visual exploration with interactive tools. Automatic data mining extracts all the patterns which match a set of metrics. The limit of such algorithms is the amount of extracted data which can be larger than the initial data volume. In this article, we focus on association rules extraction with Apriori algorithm. After the description of a characterization model of a set of association rules, we propose to explore the results of a Data Mining algorithm with an interactive visual tool. There are two advantages. First it will visualize the results of the algorithms from different points of view (metrics, rules attributes…). Then it allows us to select easily inside large set of rules the most relevant ones.

Categories and Subject Descriptors
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms : Design, Algorithms.

Keywords : Association Rules Mining, Visual Semiology, Aeronautical Data.

1. INTRODUCTION
La fouille de données, ou Data Mining (DM), est un processus d'extraction de connaissances à partir d'une très grande masse de données. Le principe consiste à rechercher des structures reliant ces données. Cette recherche peut se faire de manière automatique, en mettant en œuvre, par exemple, des algorithmes dont le but est de trouver des règles d'association (e.g. *pizza, chips -> bière*), comme Apriori [1]. Un avantage de cette approche algorithmique est son caractère exhaustif grâce auquel la totalité des règles d'association, qui satisfont des contraintes sur un ensemble de métriques, sera trouvée. Cependant, le nombre de règles construites peut parfois être plus important que la masse de données initiale. Dans ce cas-là, on se retrouve face à un autre problème de fouille de données qui consiste à identifier les sous-ensembles de règles les plus pertinentes.

Plusieurs outils visuels ont été développés afin de gérer cette masse de règles. Par exemple dans [9], les règles sont représentées sous forme de graphes ce qui permet, à l'aide d'outils interactifs, de filtrer efficacement les règles les plus pertinentes. Une autre approche consiste à utiliser une visualisation interactive qui affiche l'ensemble des règles dans un espace à trois dimensions [3] en prenant en compte différentes mesures de qualité. Dans [5] un ensemble d'outils de visualisations issus de techniques IHM est testé (Barres 2D et 3D, FishEyesView…) pour représenter des règles d'association. Toutes ces approches proposent chacune un type de visualisation préétablie, ce qui limite les choix de méthode d'interaction.

Dans cet article, nous proposons d'utiliser l'outil de Visual Data Mining FromDady [7] pour rendre efficace l'exploration de l'ensemble des règles d'association extraites à l'aide de l'algorithme Apriori. Le Visual Data Mining (VDM) [8] est une approche manuelle de la fouille de données où l'extraction de connaissances se fait sous la forme d'une exploration visuelle des données à l'aide d'un outil interactif. Le principal avantage du VDM est sa capacité à naviguer et à interagir dans un ensemble de visualisation défini par l'utilisateur.

Notre approche permet donc d'avoir une infinité de visualisations d'un ensemble de règles d'association et non une visualisation préalablement fixée. Cela permet de visualiser l'ensemble des règles selon la façon dont elles sont construites, mais aussi en utilisant un ensemble de métriques qui vont permettre d'évaluer de manière quantitative différentes propriétés de ces règles. De plus, le Visual Data Mining est particulièrement efficace lorsqu'on dispose d'une expertise sur les données. Ainsi, en exploitant les connaissances sur les métriques des règles d'association, un outil de VDM permet simultanément de trouver une bonne visualisation pour un ensemble de règles et d'isoler rapidement de manière plus efficace les sous-ensembles de règles qui présentent un intérêt.

La première partie de cet article traite de la fouille de données visuelle en s'appuyant sur les travaux relatifs à l'analyse des visualisations et à la perception visuelle. Elle présente également l'outil d'exploration FromDaDy utilisé dans cette étude et la fouille de données automatique par l'extraction de règles d'association. Ensuite, nous détaillons comment utiliser notre outil de VDM pour produire des visualisations des règles d'association. Finalement nous montrons l'utilisation de ces outils pour exploration et filtrer notre ensemble de règles.

2. RAPPELS ET CONTEXTE

2.1 Langage de description pour le Visual Data Mining

En s'appuyant sur la sémiologie graphique de Bertin [2], Card et Mackinlay ont créé un modèle pour caractériser les visualisations [4]. Celui-ci se présente sous la forme d'un tableau avec trois groupes de colonnes. Tout d'abord les données, puis les variables de perception automatique, basées sur les propriétés visuelles telles que la position et la couleur, et enfin les propriétés relatives à la perception contrôlée, telles les données textuelles.

Données				Perception automatique						Perception contrôlée	
Variable	D	F	D'	X	Y	Z	T	R	-	[ ]	CP

Tableau 1 : modèle de Card et Mackinlay

Chaque ligne du tableau correspond à une donnée en entrée. La colonne D indique le type de données (Nominale (qualitative), Ordonnée ou Quantitative). Celle-ci est éventuellement recodée en D' via un filtre ou une fonction d'encodage F qui peut provoquer un changement de type. Les données de perception automatiques sont la position spatio-temporelle (XYZT). La colonne R indique la propriété rétinienne, caractérisée par exemple par la couleur et la taille. Les colonnes – et [ ] indiquent les propriétés de connexion et d'encapsulation. L'indication de la perception contrôlée sera un indicateur de forte charge cognitive (lecture de texte) ce qui n'est pas le cas de la perception automatique.

Les outils de VDM doivent utiliser un langage de description visuelle des données. Le modèle de Card et Mackinlay est exploité dans l'outil FromDaDy (From Data to Display) [7] qui a pour vocation première l'exploration visuelle de grandes quantités de données. Il utilise des représentations de type scatterplot (nuage de points représenté en deux dimensions) configurables avec des informations rétiniennes telles que la taille du point ou l'épaisseur de la ligne, la couleur et la transparence.

L'utilisateur définit sa visualisation en reliant les variables de la base de données aux variables visuelles (voir exemple Figure 1). La visualisation se met à jour en temps réel.

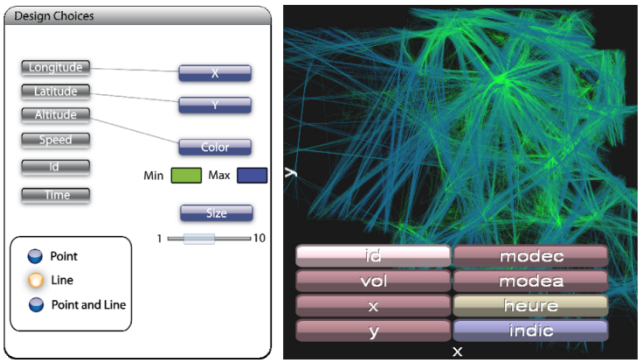

Figure 1 : exemple de configuration visuelle dans FromDaDy.

L'utilisateur interagit avec les données par manipulation directe :

- « brushing », outils de sélection de données par application d'un pinceau. Les données peintes sont alors sélectionnées.
- « pick-and-drop », outils de déplacement d'un ensemble de données préalablement sélectionnées dans une autre visualisation.
- opérations booléennes, outils de combinaison d'ensembles de sélections.

L'outil FromDaDy utilise les performances du GPU, ce qui lui permet de traiter des grandes masses de données (jusqu'à 10 millions d'enregistrements décrits selon plusieurs dizaines d'attributs).

2.2 L'extraction des règles d'association

Une des importantes applications de la fouille de données consiste à extraire des règles d'association à partir d'une grande masse de données. Dans cette étude, nous utilisons l'algorithme Apriori présenté en 1994 [1]. Une règle d'association met en lien différents attributs décrivant les données. Une règle A → B signifie : si les attributs contenus dans A sont présents dans une transaction donnée, alors les attributs contenus dans B le sont également. Apriori ne prend en compte que des attributs nominaux. Deux métriques de base sont utilisées afin d'évaluer et retenir les motifs intéressants :

- Le support $s\ (A \rightarrow B) = P\ (AB)$: probabilité de l'occurrence simultanée de A et B
- La confiance $c\ (A \rightarrow B) = P\ (B/A)$: probabilité que B soit vrai si A est vrai

La confiance mesure la probabilité que la règle soit vraie. Le support mesure la probabilité qu'a une règle d'être appliquée. Ces mesures servent de contraintes pour l'algorithme Apriori. Etant donné des seuils pour le support et la confiance, Apriori garantit l'extraction de toutes les règles d'association dont les degrés de confiance et de support sont supérieurs aux seuils. Plusieurs autres métriques sont utilisées pour évaluer les règles, comme le lift, la conviction, les mesures de Pearl, Sebag-Schoenauer, Piatetsky-Shapiro, le multiplicateur de cote, la corrélation, etc. (voir [6] pour une discussion autour de ces mesures). Elles permettent de rendre compte d'autres propriétés des règles comme la corrélation

entre les attributs ou le poids des contre-exemples. Elles sont utilisées pour filtrer les règles a posteriori.

3. VISUALISATION DES REGLES D'ASSOCIATION

Il existe des outils qui proposent de visualiser les règles d'association générées initialement de manière algorithmique. Cependant, ils ne proposent qu'un nombre limité et prédéterminé de schémas de visualisation et ne permettent pas ou peu d'interactions. En utilisant FromDaDy pour représenter les résultats d'Apriori, nous disposons d'un outil où l'utilisateur peut construire ses propres visualisations en associant à des paramètres de description des règles les caractéristiques des variables visuelles. Ces paramètres sont de deux types. Le premier type correspond aux métriques associées aux règles (support, confiance, lift…). Il s'agit de méta information qui décrit les différentes propriétés de la règle. Ces métriques sont des variables quantitatives au sens de Bertin. Il est donc théoriquement possible de représenter un nombre important de métriques en utilisant les variables visuelles adaptées au domaine d'intérêt de chacune d'elles. Par exemple, le lift est une mesure positive qui indique qu'une règle est pertinente quand sa valeur est supérieure à 1. Dans ce cas-là, un effet de seuil sur la couleur ou la taille du point est une représentation efficace de cette métrique. Le second type de paramètre correspond à la description des attributs qui constituent la règle. Du point de vue de la sémiologie graphique, ce sont des variables nominales, voire ordonnées quand elles correspondent à une discrétisation d'une variable quantitative. L'avantage de l'utilisation de FromDaDy pour représenter un ensemble de règles d'association est de pouvoir visualiser les liens entre les différentes métriques, mais aussi de mettre en évidence les liens entre ces métriques et les attributs constituant les règles. Nous avons ainsi une vision globale et complète des résultats de l'algorithme Apriori.

Les exemples utilisés dans l'article sont construits à partir d'une base de données aéronautique décrivant des vols commerciaux. Les règles obtenu mettent en liens les différentes caractéristiques des avions (vitesse, altitude, routes aériennes utilisées…). Si l'on choisit de présenter en X, la mesure de Piatetsky Shapiro, en Y la corrélation, la mesure de Pearl en gradient de couleur, et le support en épaisseur du point, on obtient le tableau de Card et Mackinlay ci-dessous (P pour point, C pour couleur et S pour taille) :

Données				Perception automatique							Perc. Cont.
Variable	D	F	D'	X	Y	Z	T	R	-	[]	CP
Piatetsky Shapiro	Q	f	Q	P							
Corrélation	Q	f	Q		P						
Pearl	Q	f	Q					C			
Support	Q	f	Q					S			

Tableau 2 : Structure visuelle de règles d'association

La figure 2 montre une application de ce modèle dans FromDaDy. Elle montre entre autres que les dispersions de corrélations sont plus importantes pour les valeurs plus faibles de la mesure de Piatetsky Shapiro. Le changement de courbure au début des données correspond au passage des valeurs de Piatetsky négatives (répulsion entre A et B) à des valeurs positives (attraction entre A et B), en passant par la valeur nulle correspondant à l'indépendance entre A et B.

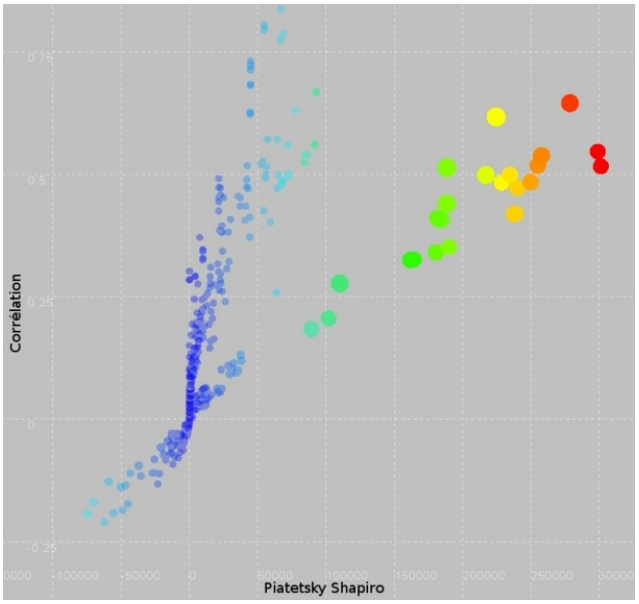

Figure 2 : visualisation de quatre règles d'association

Le tableau 3 utilise une description formelle pour une autre visualisation qui montre l'utilisation conjointe de métriques et d'attributs de règles. Ici Speed Category correspond à la vitesse moyenne en vol de l'avion. Comme Apriori ne gère pas directement les attributs numériques, cette valeur a été discrétisée. Cette variable devient donc une variable ordonnée (O). La figure 3 montre cette visualisation.

Données				Perception automatique							Perc. Cont.
Variable	D	F	D'	X	Y	Z	T	R	-	[]	CP
Speed Category	N	f	O	P							
Support	Q	f	Q		P						
Confiance	Q	f	Q					C			
Corrélation	Q	f	Q					S			

Tableau 3 : Structure visuelle de règles d'association

Nous représentons en abscisse une variable nominale ordonnée qui indique que la catégorie de vitesse de l'avion apparaît dans la partie gauche de la règle. Le support est représenté en ordonnée, la confiance en gradient de couleur (du bleu pour les basses valeurs de confiance au rouge pour les hautes valeurs) et la corrélation en taille du point. Nous pouvons observer plusieurs choses sur cette visualisation. Premièrement, la corrélation fait apparaître deux groupes de règles qui correspondent aux hautes et basses valeurs de vitesse (les amas de gros points à gauche et à droite de la figure). Les valeurs de vitesse intermédiaires ne produisent pas de règles intéressantes, voire même très peu de règles. Cela montre que le pas de discrétisation utilisé pour la vitesse n'est pas forcément le plus pertinent. Nous observons aussi qu'il n'y a pas de règle avec un support haut et un degré de corrélation également haut (correspondant à de gros points avec une ordonnée élevée). Les règles avec une corrélation élevée suivent une ligne verticale par rapport au support. Les valeurs de

confiances sont distribuées de manière homogène par rapport aux autres variables. Cependant, nous identifions facilement en bas à droite de la figure un ensemble de règles qui semble intéressant vu qu'elles ont à la fois une corrélation et une confiance élevées. Nous avons montré avec ces deux figures l'intérêt de l'utilisation d'un outil de VDM pour visualiser des règles d'associations. Il permet de faire apparaître les liens entre les différentes métriques, mais aussi la façon dont les règles sont distribuées par rapport aux attributs. Certaines visualisations mettent en valeur des zones où les règles ont simultanément de bonnes propriétés par rapport à certaines métriques. En utilisant des outils d'interaction du VDM (brush, pick-and-drop…) nous pouvons ainsi parcourir efficacement l'espace des règles afin d'y trouver les plus pertinentes.

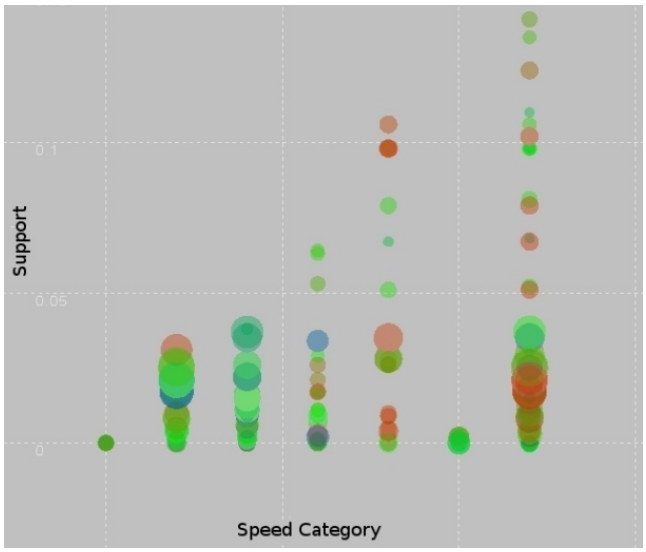

Figure 3 : visualisation de catégories de données et de règles d'association

4. EXPLORATION DE REGLES D'ASSOCIATION

Dans la section précédente, nous avons vu comment utiliser la sémiologie de Bertin pour décrire un ensemble de règles d'association selon plusieurs points de vue. Cependant les outils de Visual Data Mining permettent aussi d'interagir avec les données. Cet aspect est particulièrement intéressant quand on travaille à partir de résultats générés par un algorithme de fouille de données. En effet, comme évoqué dans l'introduction, il est courant d'obtenir un nombre important de règles avec l'algorithme Apriori. Le problème devient alors d'identifier les règles intéressantes. Pour cela, des filtres manuels sur les différentes métriques peuvent être utilisés. Cependant, les liens entre elles sont complexes et il est difficile de paramétrer efficacement ces filtres. Une des solutions consiste à renforcer les contraintes sur le support et la confiance, ou à limiter le langage utilisé, avant de lancer l'algorithme, ceci afin de limiter le nombre de règles produites. Le problème avec ce type d'approche est qu'il a été montré que des règles pertinentes par rapport à certaines métriques peuvent alors être éliminées. Au contraire, en utilisant FromDaDy, on peut relaxer les contraintes d'Apriori et utiliser des outils de visualisation et de sélection pour trouver efficacement un ensemble de règles intéressantes. En alternant les visualisations, on peut sélectionner à l'aide de brushs les règles suivant différents critères (métrique et attributs). Le retour visuel et l'interaction directe permettent d'éviter de fixer explicitement les valeurs pour les métriques comme c'est le cas pour les filtres manuels. Par sélections successives on arrive rapidement à isoler et caractériser un groupe de règles pertinentes.

5. CONCLUSIONS ET PERSPECTIVES

Nous présentons dans cet article une étude sur l'exploration visuelle de règles d'association générées par l'algorithme de Data Mining Apriori. Afin de les visualiser, leur caractérisation s'est appuyée sur le modèle de Card et Mackinlay. Ce type d'exploration permet de bénéficier simultanément de l'exhaustivité des règles calculées de manière automatique ainsi que de la richesse de l'exploration visuelle basée sur la présentation selon plusieurs points de vue et sur la sélection.

La poursuite de l'étude portera sur l'enrichissement de la présentation des règles d'association et le renforcement du lien entre la présentation des données initiales et celles des règles d'association extraites à partir de ces données. Nous pourrons par exemple visualiser de manière séparée des données d'un côté et des règles de l'autre. Les sélections sur l'une des visualisations seront automatiquement répercutées sur l'autre. Nous nous intéresserons aussi à rechercher automatiquement la visualisation des données qui supporte le mieux une règle d'association fixée.

6. REFERENCES

[1] Agrawal, R., Srikant, R., *Fast algorithms for mining association rules.* Very Large Data Bases VLDB 1994, pp. 487-499

[2] Bertin, J., Sémiologie graphique, *Sémiologie graphique - Les diagrammes – Les réseaux – Les cartes*. Paris, Editions de l'EHESS, 2005

[3] Blanchard, J., Guillet, F., Briand, H., *A User-driven and Quality-oriented Visualization for Mining Association Rules*. In Proceedings of ICDM '03. IEEE Computer Society

[4] Card, S.K., Mackinlay, J., *The structure of the information visualization design space*. In Proceedings of InfoVis 1997

[5] Chevrin, V., Couturier, O.,Mephu Nguifo, E., Rouillard, J., *Recherche anthropocentrée de regles d'association pour l'aide à la décision*. RHIM 2007 volume 8 - numéro 2, 2007

[6] Guillet, F., Hamilton, H. J., *Quality Measures in Data Mining*. Springer-Verlag New York, Inc., Secaucus, 2007

[7] Hurter, C., Tissoires, B., Conversy, S., *FromDaDy: Spreading Aircraft Trajectories Across Views to Support Iterative Queries*. IEEE Transactions on Visualization and Computer Graphics 15, 1017-1024., 2009

[8] Keim, D.A., *Information Visualization and Visual Data Mining*. IEEE Transactions on Visualization and Computer Graphics 8, 1-8, 2002

[9] Kuntz, P., Lehn, R., Guillet, F., Pinaud, B., *Découverte interactive de règles d'association via une interface visuelle.* Dans Visualisation en Extraction des Connaissances. P. Kuntz and F. Poulet (Ed.) (2006) 113-125