Manipulation interactive d'ensemble de motifs : application aux parcours hospitaliers

Yves Mercadier*, Jessica Pinaire*,**,***
Jérôme Azé*,Sandra Bringay*,**** Maguelonne Teisseire*,♯

* LIRMM, UMR 5506, Université Montpellier, France
yves.mercadier@ac-montpellier.fr,
** CHU, Département d'information médicale, BESPIM, Nîmes, France
jessica.pinaire@chu-nimes.fr
*** équipe d'accueil 2415, Institut Universitaire de Recherche Clinique,
Université Montpellier, Montpellier, France
paul.landais@umontpellier.fr
**** AMIS, Université Paul Valéry, Montpellier, France
Sandra.Bringay@univ-montp3.fr
♯ TETIS, IRSTEA, Montpellier, France
maguelonne.teisseire@teledetection.fr

Résumé. Dans cette démonstration, nous proposons une application de visualisation des résultats de la fouille de données séquentielles. Pour illustrer le fonctionnement de cette application, nous avons utilisé des données PMSI hospitalières, plus précisément dans le cas de l'infarctus du myocarde (IM). Les résultats obtenus ont été soumis à un spécialiste pour discussion et validation.

1 Introduction et motivations

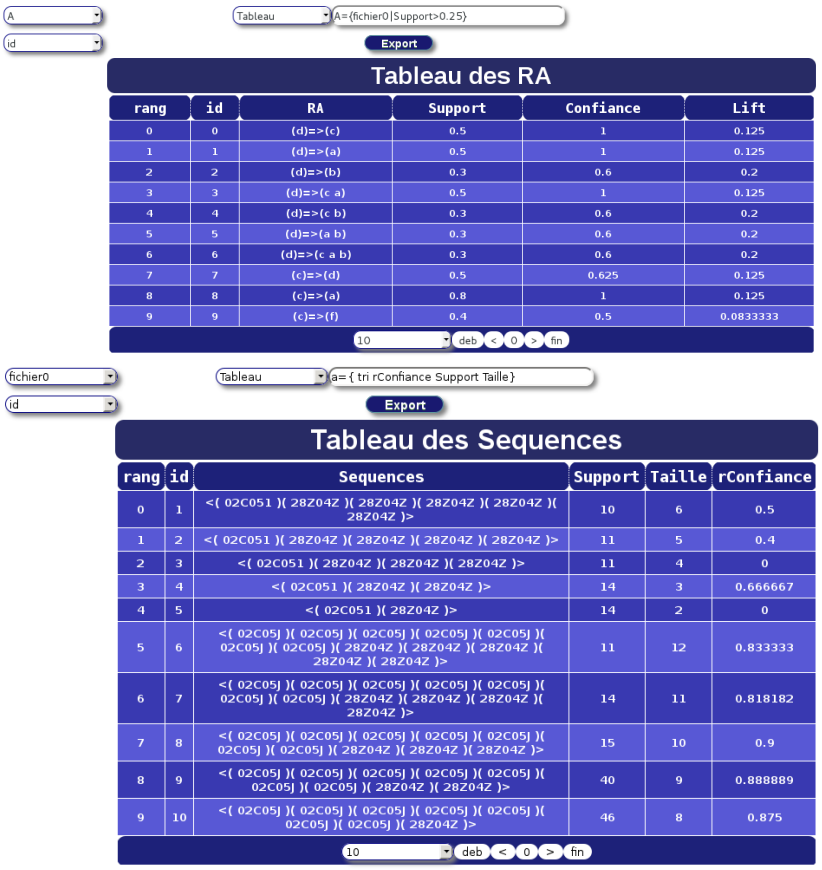
Parmi les méthodes d'extraction de connaissances, nous nous intéressons aux méthodes d'extraction de motifs, comme les motifs séquentiels définis dans Agrawal et Srikant (1995). Il en existe un très grand nombre permettant d'identifier des régularités dans un jeu de données. L'ingénieur de la connaissance utilise alors son expérience pour sélectionner des motifs répondant aux besoins des experts métier sur le jeu de données. Pour cela, il peut utiliser plusieurs mesures d'intérêt pour filtrer les motifs et les résultats obtenus sont parfois difficiles à comparer. De nombreuses mesures d'intérêt existent et ont été décrites dans Lenca et al. (2003). Pour accompagner cette tâche, nous proposons une application interactive, nommée "HIMIKO", permettant la sélection d'ensembles de motifs.

De nombreuses méthodes de visualisation ont été proposées pour aider l'ingénieur de la connaissance. On peut citer par exemple les travaux de Blanchard (2005) puis de Hervouet (2011) qui ont produit un système de visualisation 3D pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association. Cet outil permet notamment de comparer les règles selon différentes mesures d'intérêt. HIMIKO est dotée de deux fonctions classiques : la manipulation des ensembles de motifs et la visualisation de statistiques Lebart et al. (2006) sur

les ensembles de motifs. L'originalité de l'outil est d'intégrer une fonctionnalité permettant de comparer les ensembles de motifs selon différentes mesures d'intérêt.

Un premier objectif consiste à aider l'ingénieur de la connaissance en lui permettant de grouper des motifs et de comparer les ensembles obtenus. Un deuxième objectif consiste à proposer un outil qui puisse être appliqué à divers types de motifs suivant leur structure et à toutes les mesures d'intérêts. Nous allons dans la suite détailler ces fonctionnalités et les illustrer avec des motifs séquentiels. La figure 1 correspond à la visualisation d'un ensemble de règles d'association et de motifs séquentiels.

FIG. 1: Représentation d'un ensemble de règles d'associations puis de motfis séquentiels sous forme de tableau



2 Interface utilisateur

HIMIKO est réalisée sur le principe d'une SPA (Single Page Application) Deursen et al. (2007). Les résultats de la fouille de données, constitués des motifs associés à leurs mesures

d'intérêt, sont importés dans l'application soit au format json, soit au format tsv, soit au format csv. Tout d'abord, nous donnons la possibilité à l'expert de créer des ensembles de motifs à partir de contraintes sur les mesures d'intérêt ou sur la structure des motifs, et d'autre part à partir de plusieurs jeux de résultats d'extraction (avec des supports minimums différents par exemple). Pour cela, nous avons doté HIMIKO d'une algèbre des ensembles qui permet de comparer des ensembles de motifs. L'interface permet les opérations suivantes : *union, intersection, soustraction et soustraction symétrique*. Il est aussi possible de procéder à la caractérisation des ensembles numériques par les indicateurs statistiques suivants : *cardinal, minimum, maximum, moyenne, médiane, mode, écart-type*. La navigation, la manipulation et la comparaison des ensembles de motifs par l'expert sont alors facilitées.

La navigation entre les ensembles issus de la manipulation des données est permise par des aller-retours entre les différentes représentations. Cela ne correspond pas au terme d'hyperdata Kopecky et Pedrinaci (2011) mais serait plus proche du concept d'hyperset au sens d'une navigation inter-ensembles.

Nous présentons dans la suite un canevas des possibilités d'utilisation de l'application.

2.1 Comment créer des ensembles de motifs ?

La visualisation présentée sur la figure 1 est obtenue à partir des résultats de la fouille de données. Nous construisons les collections de motifs à partir de prédicats fondés sur des contraintes de structure de motifs ainsi que sur des contraintes de mesures d'intérêt. Pour cela, nous disposons d'une console minimale avec un jeu de commandes situé en haut du tableau.

Nous disposons du jeux de commande de recherche suivant : *debutePar<itemset>, terminePar<itemset> , <itemset>estPrecedeDe<itemset>, <itemset>estSuiviDe<itemset>, inclut<itemset>*.

On peut, pour un premier exemple, faire une recherche sur les motifs contenant l'item *05M13T* (correspondant à une hospitalisation pour douleur thoracique). Soit M l'ensemble contenant tous les motifs. Considérons l'ensemble A :

$$A = \{M_i | 05M13T \preceq M_i\}$$

Nous décrivons cette commande ainsi. L'ensemble A est constitué des motifs du premier fichier chargé tel qu'ils incluent l'item *05M13T*.

$$B = \{M_i | \textit{support}(M_i) \geqslant 10\}$$

Nous obtenons ici un ensemble B constitué des motifs issus du premier fichier chargé et respectant la contrainte de support.

Nous pouvons maintenant appliquer des opérations sur ces deux ensembles. Nous procédons de la façon suivante.

$$C = \{A \cup B\}$$

L'ensemble C sera le résultat de l'union de l'ensemble A avec l'ensemble B.

$$D = \{A \cap B\}$$

L'ensemble D sera le résultat de l'intersection de l'ensemble A avec l'ensemble B.

$$E = \{A \triangle B\}$$

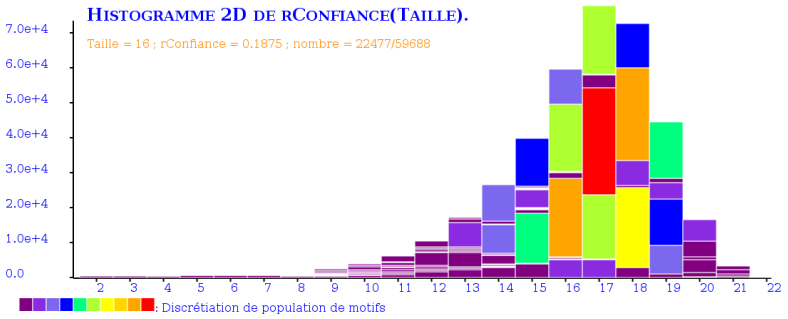
L'ensemble E sera le résultat de la soustraction symétrique de l'ensemble A avec l'ensemble B, c'est à dire l'union de A et B tout en excluant leur intersection.

2.2 Comment visualiser graphiquement les ensembles de motifs ?

Nous proposons ici deux représentations possibles des ensembles de motifs accessibles via notre interface figure 2 et figure 3.

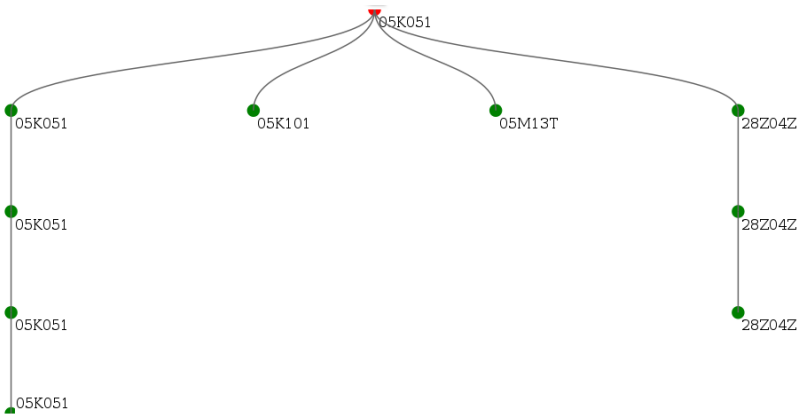
La figure 2 est l'histogramme de l'ensemble des motifs séquentiels issus de la fouille de données compacté selon deux mesures d'intérêt. Chaque barre représente un ensemble de motifs pour une valeur de la première mesure d'intérêt, ici la taille. Chaque bloc d'une barre représente la population de motifs ayant une même valeur pour la deuxième mesure d'intérêt, ici la r-confiance. Pour faciliter la compréhension du diagramme, nous discrétisons en dix intervalles la population par blocs. Lorsqu'un bloc appartient à la dernière partie (les blocs les plus importants en terme de population), il est coloré en rouge. De plus, on peut faire apparaître, par survol de la souris d'un bloc, les informations correspondantes à savoir : la valeur de la première mesure d'intérêt, la valeur de la deuxième mesure d'intérêt, le nombre de motifs constituant le bloc, le nombre de motifs constituant la barre. Finalement, il est possible de sélectionner les motifs d'un empilement par un clic de souris pour en déduire un nouvel ensemble de motifs qui pourra être nommé puis interrogé dans la console comme décrit précédemment.

FIG. 2: Représentation d'un ensemble de motifs séquentiels sous forme d'un histogramme empilé



La figure 3 représente un arbre correspondant à l'agrégation d'un ensemble de motifs séquentiels. Nous procédons comme suit pour réaliser cette agrégation : nous parcourons l'ensemble des motifs séquentiels étudié, nous extrayons les motifs débutant par un GHM choisi comme racine de l'arbre, nous parcourons les motifs ainsi extraits item par item, nous créons un nœud pour chaque item, nous créons un arc entre deux items successifs, nous itérons cette procédure sur l'ensemble des motifs extraits. Pour obtenir l'ensemble de la figure 3 nous recherchons les motifs contenant le GHM 05K051. Nous obtenons avec ce GHM comme racine un ensemble de neuf motifs. Nous agrégeons ces neuf motifs sur l'arbre de la figure 3.

FIG. 3: Représentation d'un ensemble de motifs séquentiels sous forme d'un arbre



2.3 Comment caractériser et comparer statistiquement les ensembles de motifs ?

Les combinaisons de mesures d'intérêt ne sont pas suffisantes pour comparer les grands ensembles de motifs. HIMIKO permet de comparer les ensembles de motifs entre eux à partir d'indicateurs statistiques.

Nous procédons ainsi : nous recherchons les motifs contenant l'item *05K051*, codant l'IM et fréquent (dont le support est supérieur ou égal à 10).

$A = \{M_i | 05K051 \preceq M_i\}$, puis

$B = \{M_i \in A | \textit{support}(M_i) \geqslant 10\}$,

et enfin affichage des caractéristiques de l'ensemble B dans la figure 4.

FIG. 4: Caractérisation statistique d'un ensemble de motifs séquentiels

Tableau du Résumé

Estimateur	Support	Taille	rConfiance
Cardinal	9	9	9
Maximun	12	6	0.666667
Minimum	10	3	0.4
Moyenne	10.44	4.56	0.5
Médiane	10	5	0.5
Mode	10	5	0.5
Ecart type	0.68	0.83	0.09

2.4 Comment mémoriser les résultats ?

À la fin de chaque étape, il est possible d'exporter de l'application les différents ensembles de motifs et leurs caractéristiques soit au format json, soit au format tsv.

3 Conclusions

Nous avons implémenté un outil interactif permettant d'explorer les ensembles de motifs en prenant en compte différentes mesures d'intérêt. Nous avons développé cet outil dans l'objectif de proposer une démarche itérative de sélections de motifs : fouille de données, manipulation des ensembles de motifs, validation par l'expert, fouille de données... HIMIKO a été utilisée pour faire émerger des connaissances médicales à partir d'une base issue des données du PMSI traitant de l'infarctus du myocarde.

L'environnement développé et son utilisation par des utilisateurs experts ont permis de soulever plusieurs limites. La première concerne les contraintes de navigation entre les types de représentation. Par exemple, il n'est pas possible de naviguer directement entre les différentes représentations d'un ensemble de motif. Une deuxième limite concerne l'affichage des ensembles de motifs sous la forme d'un arbre. Pour l'instant, l'information sur les valeurs d'une mesure d'intérêt n'est pas observable sur les arcs.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, Washington, DC, USA, pp. 3–14.

Blanchard, J. (2005). *Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association.* Thèse de doctorat, Université de Nantes Atlantique.

Deursen, A. V., A. Mesbah, et A. Mesbah (2007). Migrating multi-page web applications to single-page ajax interfaces. In *In CSMR '07 : Proceedings of the 11th European Conference on Software Maintenance and Reengineering*, pp. 181–190. IEEE Computer Society.

Hervouet, D. (2011). Visualisation des règles d'association en environnement virtuel 3D interactif. Master's thesis.

Kopecky, J. et C. Pedrinaci (2011). Restful write-oriented API for hyperdata in custom RDF knowledge bases. pp. 199 – 204. IEEE.

Lebart, L., M. Piron, et A. Morineau (2006). *Statistique exploratoire multidimensionnelle : visualisation et inférences en fouilles de données*. Sciences Sup : Cours. Dunod.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003). Critères d'évaluation des mesures de qualité des règles d'association. *Revue Nouvelles Technologies de l'Information RNTI-1*, 123–134.

Summary

In this demonstration, we propose a visualization application of the results of sequential data mining. To illustrate the operation of this application, we used hospital french DRG data, specifically in the case of myocardial infarction (MI). The results were submitted to a specialist for discussion and validation.